Article paper main title: Prototype of a machine learning workflow to classify land use from housing market dynamics.

**Abstract:** There is ample evidence of the role of land use and transportation interactions in determining urban spatial structure. The increased digitization of human activity produces a wealth of new data that can benefit longitudinal studies of changes in land value distributions and integrated urban microsimulation models. To facilitate that, data from various sources needs to be merged at the land parcel level to enhance datasets with additional attributes, while maintaining the ease of data storage and retrieval. This paper proposes a prototype of a workflow to augment a dataset of real estate transactions with data from multiple urban sources and use machine learning to classify land use of each record based on housing market dynamics.

**Eric J. Miller**

University of Toronto

[miller@ecf.utoronto.ca](mailto:miller@ecf.utoronto.ca)

**Stepan Oskin**

University of Toronto

[stepan.oskin@mail.utoronto.ca](mailto:stepan.oskin@mail.utoronto.ca)

# 1 Introduction

There is ample evidence of the role of land use and transportation interactions in determining urban spatial structure[cite]. Accessibility and mobility provided by transportation systems drive economic development and impact travel behaviour and location choices of households and firms. Similarly, urban sprawl and location of activities drive travel demand and the need for building transport networks[cite].

Land values in a metropolitan region are an outcome of such land use-transportation interactions[cite].

The fundamental link between transportation and urban form creates a feedback relationship between land development, travel needs, viability of alternative modes, accessibility, and other important characteristics of the urban transportation system(Kelly, 1994). Numerous "top-down" and "bottom-up" models have been designed to analyze and forecast the behaviour of urban regions and interaction of their transportation and land use systems, with the latest generation being the family of microsimulation models(Iacono, Levinson and El-Geneidy, 2008).

Among the major barriers to implementation of integrated urban models since their introduction were such aspects as data hungriness and computational requirements(Miller, Kriger and Hunt, 1998). In the past decades, continuing methodological advances in computing, such as cost-effective High-Performance Computing (HPC), detailed GIS-based datasets and machine learning methods, have turned former barriers into opportunities for model system development. Since microsimulation models are dynamic and disaggregated in their nature(Miller, 2019), their design and calibration efforts could benefit from the use of data sources that are longitudinal (time-series) and highly disaggregated spatially (to a parcel level). In our times, the amount and diversity of new data sources relating to cities grows exponentially and these new sources present new challenges and new opportunities for researchers interested in longitudinal and spatially disaggregated aspects of interactions of urban systems.

## 1.1 New data sources and their challenges

Increased digitization of human activity produces a wealth of new information that can be used to study interactions between land use and transportation at a fine spatial and temporal scale (Arribas-Bel, 2014; Chen *et al.*, 2016). An example of such digitization is the introduction of the POLARIS land registration system by the Government of Ontario in 1985, which lead to the creation of an extensive dataset of real estate transactions by Teranet Enterprises Inc.(Teranet Enterprises Inc., 2019). Teranet’s dataset captures all real estate transactions that have been recorded in the Province of Ontario since 1985, which makes it a great asset for longitudinal housing market studies and design and calibration efforts for microsimulation models.

However, despite its high spatial and temporal resolution, the version of Teranet’s dataset available to researchers at University of Toronto Transportation Research Institute (UTTRI) suffered from a severe lack of attributes describing each transaction, such as the category of the property being sold. To add new features, Teranet’s dataset can be augmented using a range of urban data sources, such as information from Census and transportation surveys and land use data. At the same time, joining these sources together can be challenging, since they all use different spatial units and are available at varying temporal spans.

The difference in units between urban data sources can be addressed by uniting all polygon-based data at the level of time indexed points, as represented by real estate transactions recorded in Teranet’s dataset. In addition, temporal spans need to be established when relating datasets to ensure proper alignment of urban data both on spatial and on temporal dimension. Finally, since none of the land use data sources available for Ontario covered the complete time interval from 1986 to 2017, a machine learning model can be trained to classify land use based on housing market dynamics. Storing this information in the form of a relational database facilitates ease of access and reduces hardware requirements, allowing a broader group of researchers to take advantage of the new powerful data source.

This paper presents an analysis of such a database for the Greater Toronto-Hamilton Area (GTHA). Specifically, it presents methods developed for dealing with two important technical issues in making use of Teranet’s dataset:

1. Merging diverse datasets spatially and temporally.

2. Using machine learning to classify land use from housing market dynamics.

Section 2 describes the data used in this study. These data come from a variety of sources, have different spatial and temporal representations, and contribute different attributes characterizing land use, buildings and market transactions within the GTHA. Section 3 describes the method developed to combine these disparate datasets into a functionally usable overall database suitable for supporting real estate market analysis and modelling. Section 4 presents a novel application of machine learning methods to classify land use based on parcel transaction data, along with other data attached to the transaction data, as described in Section 3. Section 5 evaluates the effectiveness of the machine learning model presented in Section 4. Section 6 summarizes the findings and conclusions of the study and suggests directions for further research.

# 2 Data

This paper discusses the process of development of a housing market database for the GTHA; this database was constructed by joining a detailed dataset of real estate transactions with other urban data sources and a new feature produced by a machine learning algorithm.

## 2.1 Teranet’s dataset

At the heart of the GTHA housing market database lies Teranet’s dataset—an extensive historical record of real estate transactions recorded in the Province of Ontario since the beginning of the nineteenth century; for the purposes of this paper, the period of time from 1985 to October of 2017 was considered, since records prior to 1985 appear to be less consistent.

All land owned in Canada is registered in a public land registry in the applicable province. Each province and territory in Canada has its own land registry system, whether it is a land titles system, a registry system or a combination of both, with each system having its own rules. The registry system is a public record of documents evidencing transactions affecting land. In the land titles system, the applicable provincial government determines the quality of the title, and essentially guarantees (within certain statutory limits) the title to, and interests in, the property. As of 2015, most common law provinces and territories in Canada were using the land titles system or were in the process of converting from a

registry system to a land titles system[29].

As of 2015, the Province of Ontario has largely converted from registry systems to a land titles system. In 1985, the Government of Ontario initiated the Province of Ontario Land Registration Information System (POLARIS) pilot project for the purposes of the conversion between systems and records automation. The Land Registration Reform Act (Ontario)[50] was introduced in 1990 to facilitate electronic search and registration of properties and the automation of paper-based records. POLARIS was built by the Province to house and process electronic land records, which in turn lead to the creation of an extensive dataset of land registration records managed by Teranet Enterprises Inc. Today, POLARIS

is the search/registration and property maintenance system for all automated land records in Ontario.

In 1991, the Government of Ontario established a partnership with Teranet, a Toronto-based organization, founded the same year, which provides e-services to legal, real estate, government, financial, and healthcare markets. The partnership was established to convert Ontario’s land registration system to a more modernized electronic title system. The project involved taking a 200-year-old paper-based

system and creating a database with electronic records for more than five million parcels of land. Teranet converted all qualified Registry properties in Ontario to the Land Titles system and automated existing paper Land Titles parcels. As a result, 99.9% of property in Ontario was parcelized and administered under the Land Titles system. Teranet fully automated the conversion of millions of paper-based

documents and records into the Ontario Electronic Land Registration System (ELRS)[49]. Figure 2.3 presents an example of Teranet data: records from 2014 that fall within 1,000m of an arbitrarily selected target point, colored by land use spatially joined from DMTI datasets.

Teranet’s dataset holds data on the housing market that has a very high spatial and temporal resolution; at the same time, using the version of the dataset available to UTTRI researchers in its raw form for meaningful modeling and analysis was challenging for two main reasons:

1. It lacks features describing each transaction
2. No reliable parcel-level source of land use information for GTHA covering the full period from 1985 to 2017 could be found

On the other hand, each Teranet transaction has a timestamp (date) and location

information (x and y coordinates) which presents an opportunity to join Teranet records to variety of other geocoded urban data sources, such as Census demographics, Transportation Tomorrow Survey (TTS) and parcel-level land

use information. However, joining these data sources together requires additional considerations, as they

use different spatial units and are available at different temporal spans, as will be discussed in chapter 3.

## 2.2 Select variables from the Census of Canada

# One of the major sources of demographic and statistical data in Canada are the datasets collected under the national Census program. Census datasets provide valuable insights into the latest economic, social and demographic conditions and trends in Canada and are used to plan important public services. Statistics Canada collects every five years the national Census of Canada and disseminates the information by a range of geographic units, also referred to as ”Census geography”(Map and Data Library, 2019).

## 2.3 Select variables from the Transportation Tomorrow Survey (TTS)

# Another major source of information for most transportation planning studies concerned with Southern Ontario is the Transportation Tomorrow Survey (TTS), an origin-destination travel survey(Data Management Group, 2014). The Transportation Tomorrow Survey (TTS), undertaken every five years since 1986, is a cooperative effort by local and provincial government agencies to collect information about urban travel in southern Ontario. TTS represents a retrospective survey of travel taken by every member (age 11 or over) of the household during the day previous to the telephone or web contact. The information collected and the method of collection has remained relatively consistent over the seven surveys; TTS survey data includes characteristics of the household, characteristics of each person in the household, and details of the trips taken by each member of the household, including details on any trips taken by transit(Ashby, 2018).

## 2.4 Land use from DMTI Spatial Inc. by year (2001-2014)

# DMTI Spatial Inc., a Digital Map Products company, is a major provider of location-based information in Canada. DMTI has been providing industry leading enterprise Location Intelligence solutions for more than a decade to Global 2000 companies and government agencies(DMTI Spatial Inc., 2014).

## 2.5 Detailed land use information from University of Toronto’s Department of Geography collected in 2012 and 2013

# The detailed land-use data provided by University of Toronto’s Department of Geography is a combination of parcel boundaries (from Teranet) and manually coded land-use data produced using Google maps and streetviews; it was collected by Prof. Andre Sorensen and Prof. Paul Hess’s research project.

# 3 Establishing spatial and temporal relationships among data sources

This section discusses the introduction of the new spatial and temporal relationships that were used to implement the referential integrity constraints of the new GTHA housing market database.

## 3.1 Spatial relationships between urban data sources

Most urban areas are divided into zones or planning areas on the basis of maintaining similar population sizes and following built or natural boundaries like roads or rivers. Census geography follows a certain hierarchy defined by Statistics Canada, with the largest top-level divisions being provinces and territories,

and the lowest-tier divisions to which Census data is disseminated being Dissemination Areas (DAs)(Statistics Canada, 2018). Statistics Canada defines a Dissemination Area as a small area composed of one or more neighboring

dissemination blocks, roughly uniform in population size targeted from 400 to 700 persons to avoid data suppression(Statistics Canada, 2015).



Figure 1. Spatial relationship between datasets: difference between Traffic Analysis Zones (TAZ, red) and Census Dissemination Area (DA, black).

To simulate the changes in accessibility, metropolitan regions are usually broken down into a set of small geographic zones, similar (or in many cases identical) to the set of zones used for regional travel forecasting. For TTS variables, the finest level of spatial aggregation is that of the Traffic Zone, also referred to as the Traffic Analysis Zone (TAZ). A Traffic Zone is a polygon which typically falls along the centre line of roads or the natural geographic boundaries(Data Management Group, 2019). Not as a rule, but TAZ zones roughly follow Census tract boundaries, which are slightly bigger than DA boundaries. Figure 1 presents an example of TAZ polygons overlaid with Census DA boundaries.

TTS data has been collected for changing TAZ boundaries, or in other words, different zone systems due to growing population and expanding extents of the survey in the GTHA region over the years. To make the TTS data consistent for comparing over all years from 1986 to 2016, the Data Management Group (DMG) at the UTTRI, the custodian of the dataset derived from TTS, made all surveys available in the 2001 zone system, for convenience of researchers (any zone system could have been chosen for that matter). UTTRI used the 2001 TAZ system to model travel times for the GTHA on EMME for all TTS years based on the origin-destination trip data collected in the survey. The travel time data was used to create further transportation accessibility variables.

Land use data collected by DMTI and by the Department of Geography uses the spatial unit of a parcel polygon. Teranet records have attributes representing x and y coordinates matching parcel centroids.

Below is the summary of spatial units used by the data sources that were combined into the GTHA housing market database:

• Point data

– Teranet

• Parcel-level data (polygons)

– detailed land use from the Department of Geography

– land use from DMTI

• DA-level data (polygons)

– Census variables

• TAZ-level data

– TTS variables

When joining these data sources, difference in spatial units needs to be respected, which can be more challenging when spatially joining polygons with polygons, since it might require area-weighted spatial interpolation of data to a common unit of analysis. In addition, polygon-based data can also vary with time, as is the case with DMTI’s land use information, which is available by year. To simplify relating

different polygon-based data sources with each other, all of them can be brought together to a single level of time-indexed points, such as Teranet transactions. This allows flexibility in combining data from polygon-based data sources to a common point level while maintaining the integrity of spatial and temporal relationships through polygon-to-point spatial joins.

## 3.2 Temporal relationships between data sources

In addition to using different spatial units, data sources joined with Teranet’s dataset are available at different temporal spans:

• Teranet records have individual timestamps (date) on each record

• Census and TTS variables are sampled once in 5 years

• DMTI’s land use data is available by year and covers a time span from 2001 to 2014

• Detailed land use from the Department of Geography was collected at a single point in time during the summers of 2012 and 2013

Figure 2 presents the temporal spans assigned to each data source for joining with Teranet records.

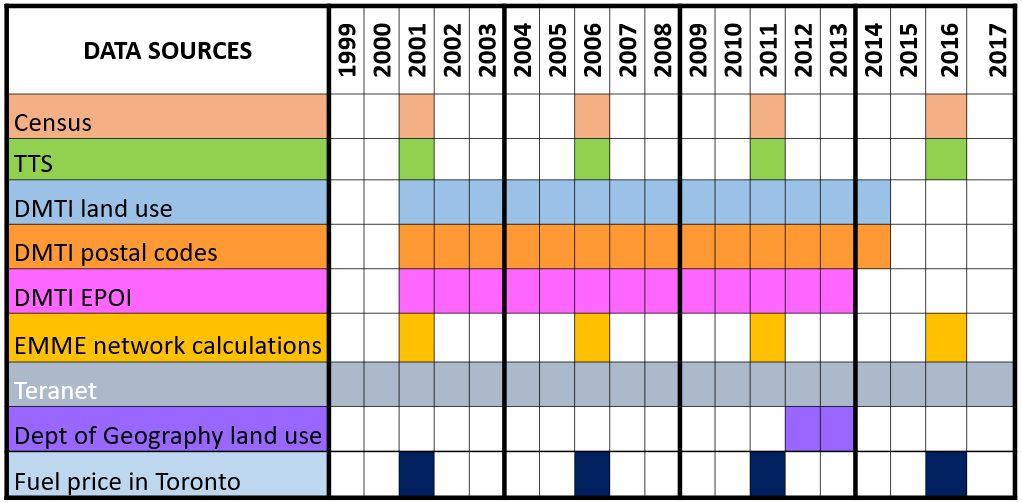


Figure 2. Temporal spans of data sources used in the GTHA housing market database.

Temporal matching between Teranet records and DMTI data could be done directly: DMTI land use for each year from 2001 to 2014 was spatially joined with a subset of Teranet records from the corresponding year; such approach ignores changes of land use that occur within a year, but recognizes land use changes between the years for which DMTI land use data was available. Since the detailed land use provided by the Department of Geography was collected at a single point in time, it was joined to all Teranet records, keeping in mind that it is most accurate around its time of collection in 2012 and 2013.

As for Teranet and Census / TTS variables, they could be matched in a number of ways:

1. By direct match with appropriate Teranet subsets (match Census / TTS variables only with Teranet records from the corresponding year).

2. Via interpolation of discrete Census / TTS variables.

3. Through assignment of temporal spans to each Census / TTS survey as new features to Teranet records (each Census / TTS survey is assigned a temporal span of 5 years; this 5-year span represents a group of Teranet records to which variables from this survey can be matched).

To avoid additional interpolation assumptions and use the actual values recorded from Census and TTS surveys, the third option has been chosen for matching Teranet records with Census / TTS variables. Each Census / TTS survey was assigned a 5-year time span centered at the survey year (i.e., 2014–2018 for

2016 survey year) and new foreign keys were introduced to Teranet records to allow matching with 5-year time spans of Census / TTS variables.

## 3.3 Data preparation

Reproducibility of the data preparation process for data sources related to the GTHA housing market database has been established via a streamlined data preparation workflow using Python via a series of jupyter notebooks. It accomplishes three

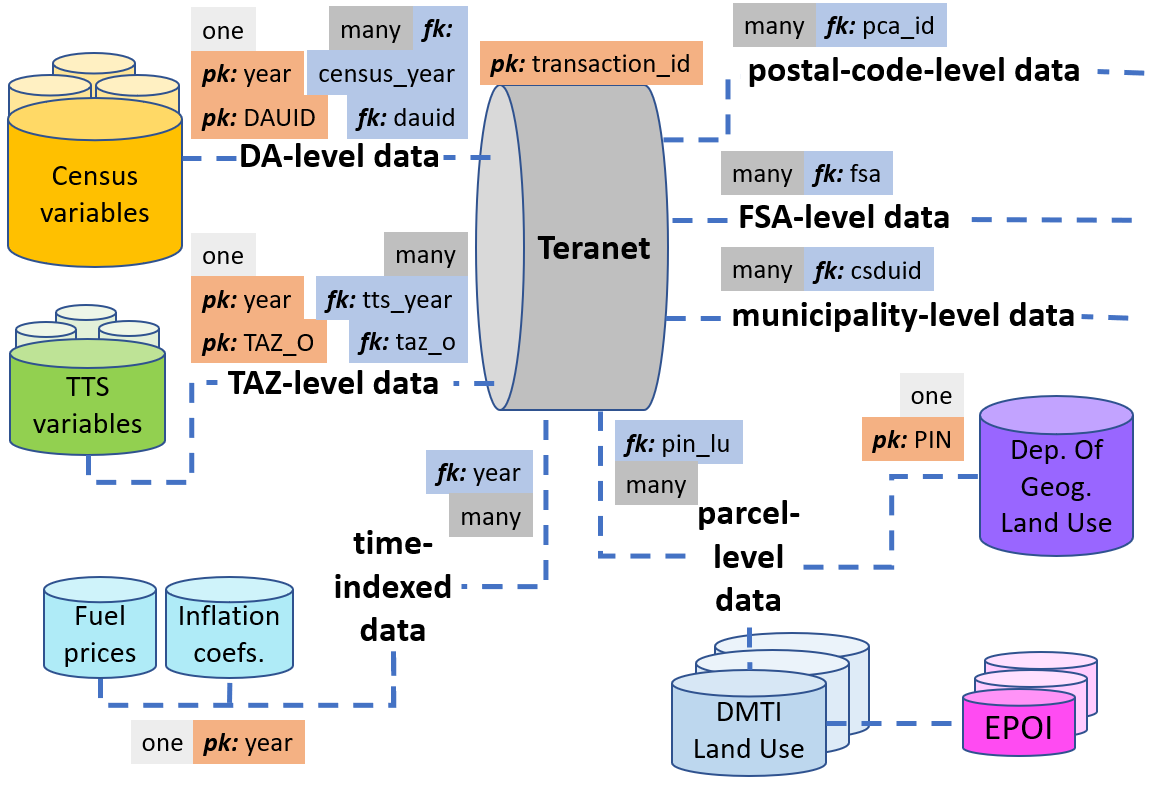
main objectives:

• Clean Teranet dataset and correct its records for consistency.

• Introduce new keys that would allow efficient joining of other data sources, such as Census and TTS variables or parcel-level land use information, while maintaining the integrity of spatial and temporal relationships.

• Engineer the new features that can be used by the machine learning algorithm along with the features from the joined datasets to classify land use.

To implement the spatial and temporal relationships between the data sources discussed in section 2, a number of new foreign keys was introduced to Teranet records via a series of spatial joins and feature engineering. Figure 3 presents the Entity Relationship (ER) diagram of the GTHA housing market database.

Figure 3. Entity Relationship (ER) diagram of the GTHA housing market database; its referential integrity constraints were set up using the new keys introduced based on the spatial and temporal relationships between data sources.

In addition to producing new keys for joining datasets, a number of the new features was engineered from original Teranet attributes. These new features were intended to give each Teranet record spatial and temporal “context” of the housing market dynamics by grouping records using different spatial (parcel-level) and temporal (yearly) criteria (e.g., rolling count of transactions coming from a particular coordinate pair, or ratio of price to median for that year, etc.).

# 4 Prototype of a machine learning workflow to classify land use

One of the major features that was missing from the available version of Teranet’s dataset was the information about the type of property being transacted, which introduced a major limitation on how Teranet’s data can be used. None of the available sources of land use information covered the full time interval from 1986 to 2017; the most detailed and accurate source was the land use data collected in 2012 and 2013 by University of Toronto’s Department of Geography (introduced in section 2.5).

To address this issue, detailed land use from the Department of Geography was used as labelled data to train a machine learning model capable of recognizing certain property types that have characteristically different behavior on the housing market. For example, this model can differentiate a detached house from a condo through such features as high / low volume of transactions from a coordinate pair, ratio of price to median price for that year, etc. This section discusses a basic prototype of a machine learning workflow to classify land use from housing market dynamics.

## 4.1 Target variable

Target variable was constructed by reducing the land use codes with the highest counts of Teranet records found in the Department of Geography’s land use dataset to the three major land use classes. Since many machine learning algorithms are subject to a frequency bias in which they place more emphasis on learning from data observations which occur more frequently, the three classes were selected to have a comparable number of Teranet records between themselves and thus produce a more balanced dataset.

In addition, the chosen groupings combine categories that have a similar distribution of price and count of sales per coordinate pair between land use categories that were grouped together to form a single class. For example, detached and semi-detached houses and townhouses would have a much smaller frequency of transaction and a higher median price per coordinate pair when compared to condos and strata townhouses.

The three target classes that were introduced were:

• Class 0: “condo”, including Apartments/Condos/Residence and Strata Townhouses.

• Class 1: “house”, including Single Detached Houses, Duplex/Semi-Detached and Townhouses.

• Class 2: “other”, including Commercial/Shopping, Mix (Commercial Residential), Industrial/Employment Lands, and everything else.

## 4.2 Dimensionality reduction and hyperparameter tuning

To reduce the dimensionality of augmented Teranet’s dataset and only include features that are the most relevant to the classification algorithm, a combination of algorithmic feature selection techniques has been utilized. Feature selection algorithms present a practical approach to feature selection at scale; such algorithms combine a search strategy for proposing new feature subsets with an objective function to evaluate these subsets; objective function plays the role of a feedback signal used by the search strategy to choose between candidate subsets. Figure 4 presents the top 11 features that were selected by at least four different feature selection methods.

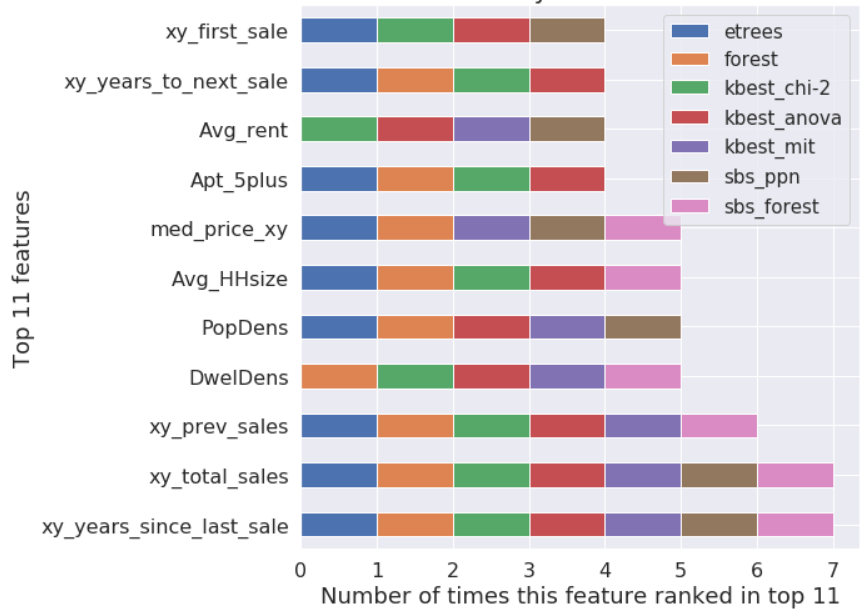


Figure 4. Feature Subset Selection (FSS) results: 11 features that were selected by at least four different feature selection methods.

There are two types of parameters in machine learning: those that are learned by parametric models from training data (i.e., weights in logistic regression), and the parameters that tune the performance of a learning algorithm, or its hyperparameters (i.e., regularization parameter in logistic regression or maximum depth of a decision tree). An acceptable bias-variance trade-off for a classifier can be found by tuning its hyperparameters, but care must be taken to ensure unbiased assessment of its generalization performance.

To facilitate unbiased performance evaluation of classifiers, all GTHA Teranet records from 2011 to 2014 have been split into two subsets using random subsampling: 70% of the data was used to train models and tune their hyperparameters, while 30% of the data has been used as a test subset for evaluation of the classifier. Train and test subsets have been stratified across the target classes: in this context, stratification means that training and test subsets will have the same proportions of class labels as the input dataset. To avoid overfitting model to test set while tuning its hyperparameters, *k*-fold cross validation was used via grid search. To improve convergence of gradient-based linear models in the presence of outliers, quantile transformation (uniform PDF) was applied to input features; to improve performance of nearest neighbors, input features were standardized.

## 4.3 Model selection

An important point to be summarized from the famous No Free Lunch Theorems (NFL)(Wolpert, 1996; Wolpert and Macready, 1997) is that no single classifier works best across all possible scenarios, as there is a lack of a priori distinctions between learning algorithms. In practice, it is essential to compare the performance of at least a handful of different classification algorithms, since each of them has its inherent biases. After tuning hyperparameters, performance of the following models have been compared: perceptron learning algorithm (η = 0.5, max\_iter=5), logistic regression (L2, L1 regularization, C=0.1), linear discriminant analysis classifier, quadratic discriminant analysis classifier, Linear Support Vector Classification (L2, L1 regularization, C = 0.1), decision tree and random forest, K-Nearest Neighbors (Manhattan and Euclidean distance, 4 neighbors), Gaussian Naive Bayes.

# 5 Evaluation of results

This section discusses the evaluation of predictive performance of machine learning algorithms used for classifying land use from the housing market dynamics.

## 5.1 Metrics for evaluating model performance

Classification accuracy (ACC) is a common metric used to compare the performance of different classifiers; it is defined as the proportion of correctly classified instances. Precision (PRE) is defined as a fraction of relevant examples (true positives) among all of the examples that were predicted to belong in a certain target class. Recall (REC) is defined as a fraction of examples which were predicted to belong to a class (true positives) with respect to all of the examples that truly belong to that class. F1 score, also known as balanced F-score or F-measure,

combines precision and recall into a single metric.

Precision, recall, and F1 score are metrics specific to binary classification systems. In case of a multi-class classification problem, these metrics can be produced using individual confusion matrices constructed separately for each class using One-versus-All technique (OvA). Micro-averaging can be useful to weigh each instance or prediction equally, while macro-averaging evaluates the overall performance of a classifier with regard to the most frequent class labels by weighting all classes equally(Raschka and Mirjalili, 2017).

## 5.2 Evaluating model performance

Four different subsets of data have been used to test the best performing models: train, test, and two additional validation subsets. The train and test subsets represent Teranet records from 2011 to 2014 randomly sampled into 70% train and 30% test subsets; these are the primary subsets that were used for training and tuning the hyperparameters and then evaluating the performance of classifiers on

unseen test data, as was described in section 4.2. The two additional validation subsets were composed of Teranet records from 2010 and 2015. Since the Department of Geography land use information (target variable) was collected in 2012 and 2013, it can be less accurate for these subsets; thus, they have not

been used for model selection, training or primary testing, but were utilized to test fitted models as an additional reference for the generalization of performance of classifiers. Figure 5 presents model performance on train, test, and two additional validation subsets.

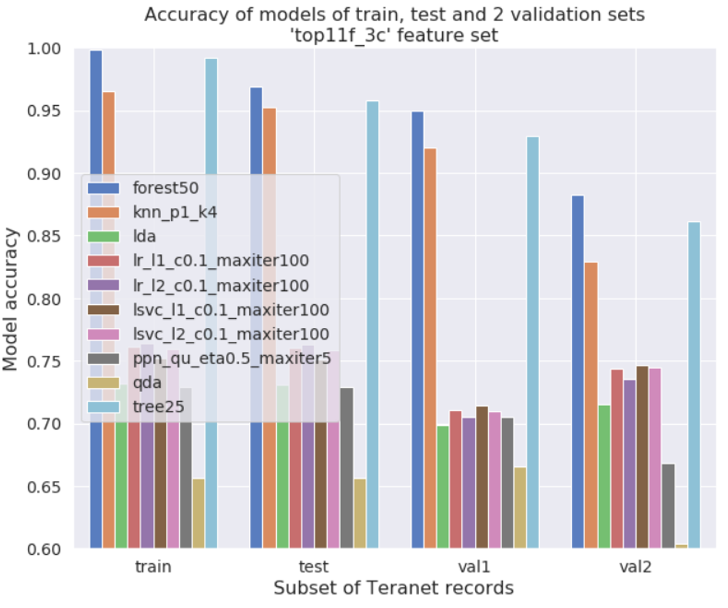
. 

Figure 5. Model performance (accuracy) on train, test, and two additional validation subsets.

As can be seen on figure 5, in terms of prediction accuracy, tree-based and nearest neighbors models dramatically outperform linear models, such as logistic regression, linear SVC, perceptron and LDA classifier. This indicates that in the current feature space target classes are not linearly separable. Best-performing linear models were able to reach classification accuracy around 75% on the test set,

with all linear models performing close to each other. This performance was consistent across the train, test, and both extra validation subsets. Linear models do not seem to overfit the data, but instead suffer from high bias. In comparison, tree-based models capable of drawing complex non-linear decision boundaries were able to achieve much higher classification accuracy, with both decision tree and random forest scoring above 95% on the test set. These models have a much lower bias on this dataset compared to linear models, but do overfit the training data to some degree under the current size of the training subset.

## 5.3 Best performing model: Random Forest

As can be seen on figure 5, random forest with 50 estimators and Gini impurity criterion showed the best results in terms of accuracy and fit times on all subsets. Figure 6 presents the classification report showing all main model performance metrics and figure 7 presents confusion matrices for the best performing model: random forest with 50 estimators using Gini impurity criterion.

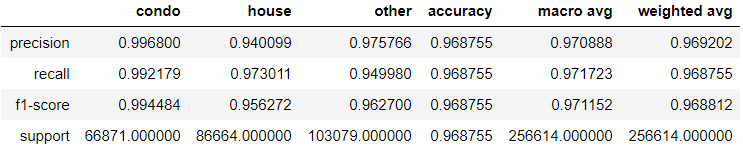


Figure 6. Best model performance on test set: classification report for random forest with 50 estimators using Gini impurity criterion.

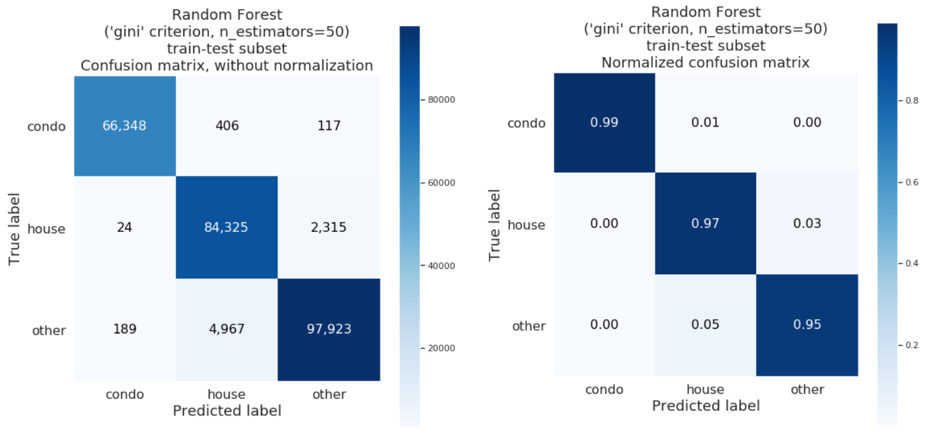


Figure 7. Confusion matrices for best performing model: random forest.

# 6 Summary, conclusions and future work

Microsimulation models present the latest generation of integrated land use and transportation models and are well suited to analyze the complex interaction of transportation and land use. New data sources that appear with the increased digitization of human activity present opportunities to look at urban processes at unprecedented spatial and temporal scale, and thus present a lot of value for design and validation of integrated urban models and for longitudinal studies concerned with evolution of urban form. Introduction of POLARIS electronic land registration system by the Province of Ontario in 1985 lead to the creation of an extensive dataset of real estate transactions by Teranet Enterprises Inc.

However, despite having very high spatial and temporal resolution, available version of Teranet’s dataset suffered from severe lack of features describing each individual transaction.

One of the major attributes missing from Teranet data is the type of property being transacted, or land use information for the parcel where a transaction is recorded. Along with selected Census and TTS variables, detailed parcel-level land use from the Department of Geography and DMTI land use data have been spatially joined to each Teranet record. However, since both of these data sources have their limitations, detailed land use data from Department of Geography has been used to train an algorithm capable of classifying land use based on the housing market dynamics; this way, land use information can be made available for each Teranet record for the full timespan covered by the Longitudinal Housing Market Research conducted by UTTRI.

To augment Teranet’s dataset, new variables were engineered from its native attributes to capture the housing market dynamics at the parcel level: for example, ‘xy total sales’ was computed as the total count of Teranet records coming from a particular coordinate pair; ‘med price xy’ represents the median price of all records coming from a coordinate pair, etc. To augment Teranet data with demographic and

transport information, the new Teranet features were spatially and temporally joined with Census and TTS variables recorded at the level of a Dissemination Area and TAZ zone, respectively. Finally, the augmented Teranet dataset has been tested with machine learning algorithms, attempting to classify land use for each Teranet record within the span of Census / TTS variables, thus recognizing land use

changes with time.

A range of preprocessing techniques has been tested with several linear, tree-based and nearest neighbors classification algorithms; tree-based models and k-Nearest Neighbors significantly outperformed linear models. The new features engineered from native Teranet attributes have shown to have strong predictive power when classifying land use. When joined with Census variables at the level of Dissemination Areas, new features engineered from Teranet’s dataset allowed the classification of land use with a high level of accuracy. A random forest model was trained using a random 70% sample of all Teranet records with new features from 2011 to 2014 stratified by target classes (“condo”, “house”, or “other”); the model achieved 97% of accuracy on the test subset composed of the remaining 30% of records from

2011 and 2014. Tree-based models did show some degree of overfitting and could benefit from further increase in the size of training data, as indicated by their learning and validation curves.

Features engineered from native Teranet attributes that capture price ratios to median and frequency of transactions from a coordinate pair have strong predictive power for land use classes, as indicated by feature selection techniques and model coefficients. This workflow could be further improved by joining more Census / TTS variables to engineer new features; target classes also could be redefined to allow

more meaningful classification. In addition, results of the classification preformed by this workflow need to be investigated. A map produced with counts of misclassified Teranet records per DA shows that errors seem to be highly concentrated and correspond to high-frequency transactions, such as condos

and mixed use properties. To facilitate further investigation of classification results, augmented Teranet dataset with new feature ‘lucr predict’ along with related Census and TTS tables has been transformed into PostgreSQL relational database to facilitate ease of access by a broader group of specialists.

# References [Section heading: Heading 1 style]

Bhat, C. R. 1998. A model of post home-arrival activity participation behavior. *Transportation Research B* 32(6):387–400.

Cho, S., D. M. Lambert, R. K. Roberts, and S. G. Kim. 2008. Moderating urban sprawl through land value taxation. Presented at the American Agricultural Economics Association Annual Meeting, July 27–29, 2008.

Giuliano, G., and S. Hanson, eds. 2004. *The geography of urban transportation*. New York: The Guilford Press.

Gifford, J. 2005b. Congestion and its discontents. In *Access to destinations*, edited by D. Levinson and K. Krizek, 39–61. London: Elsevier.

Hanson, S., and G. Pratt. 1995. *Gender, work and space*. London and New York: Routledge.

Peirce, N. 2003. Zoning: Ready to be reformed? http://www.postwritersgroup.com/peir0127.htm.

Stopher, P. R., D. T. Hartgen, and Y. J. Li. 1996. SMART: Simulation model for activities, resources, and travel. *Transportation* 23:293–312.

Whitmeyer, J. M. 2000. Power through appointment. *Social Science Research*, 29(4):535–555. doi:10.1006/ ssre.2000.0680.

Zhao, Z., K. V. Das, and C. Becker. 2008. Funding surface transportation in Minnesota: Past, present and prospects. Technical Report CTS 08-23, Center for Transportation Studies, University of Minnesota.