The Journal of Transport and Land Use http://jtlu.org
Vol. 15 No. 1 [2022] pp. 355–374

JTLU

A prototype machine learning residential land-use classifier using housing market dynamics

Shivani Raghav
PSD Citywide
sraghav@psdcitywide.com

Stepan Oskin
Prodigy
stepan.oskin@prodigygame.com

Eric J. Miller
University of Toronto
eric.miller@utoronto.ca

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 support longitudinal studies of changes in land-value distributions and integrated urban microsimulation models. To produce a comprehensive dataset, information 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 and respecting spatial and temporal relationships. This paper proposes a prototype of a workflow to augment a historical dataset of real estate transactions with data from multiple urban sources and to use machine learning to classify land use of each record based on housing market dynamics. The study finds that engineered parcel-level attributes, capturing housing market dynamics, have stronger predictive power than aggregated socio-economic variables, for classifying property land use.

Article history:
Received: October 4, 2020
Received in revised form: July 25, 2021
Accepted: September 23, 2021
Available online: July 5, 2022

1 Introduction

There is ample evidence of the role of land use and transportation interactions in determining urban spatial structure (Wegener, 1994). Accessibility and mobility provided by transportation systems drive economic development and impact travel behavior and location choices of households and firms. Similarly, urban development and location of activities drive travel demand and the need for building transport networks (Manheim, 1978). Land values in a metropolitan region are an outcome of such land use-transportation interactions (Alonso, 1964; Knight & Trygg, 1977; Martinez, 2018; Spengler, 1930).

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 behavior of urban regions and interaction of their transportation and land-use systems, with the latest generation being the family of micro-

http://dx.doi.org/10.5198/jtlu.2022.1905
ISSN: 1938-7849

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

simulation models (Iacono et al., 2008). Among the major barriers to implementation of integrated urban models since their introduction, have been such aspects as data hungriness and computational requirements (Lee, 1973; Miller et al.,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 (Miller, 2018). Since microsimulation models are dynamic and disaggregated in their nature, their design and calibration efforts could benefit from the use of data sources that are longitudinal (time-series) and highly disaggregated spatially (to parcel level) (Miller, 2019). In today's times, the amount and diversity of new data sources relating to cities are growing exponentially and these new sources present new challenges and new opportunities for researchers interested in longitudinal and spatially disaggregated 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 urban systems at a fine spatial and temporal scale (Arribas-Bel, 2014; Chen et al., 2016). Real estate pricing data sources commonly include MLS Listings, or aggregated Housing Price Indices created on a quarterly or annual basis by lending organizations or government agencies such as Statistics Canada. Teranet's sales dataset is one such fine-grained parcel-level example that captures the details all real estate transactions that have been recorded in the Province of Ontario since 1985, making it an excellent asset for longitudinal housing market studies and for design and calibration efforts of microsimulation models. However, despite its high spatial and temporal resolution, Teranet's dataset suffers from a severe lack of attributes describing the property use or structural details of the plot or building on the parcel. To derive more information, the Teranet sales dataset can be augmented with a range of urban data sources. 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 could be addressed by uniting all polygon-based data at the level of time-indexed points, as represented by 'year of sales' of property transactions recorded in Teranet's dataset. In addition, temporal spans needed to be established when relating datasets to ensure proper alignment of sales transaction data with other urban datasets, both on spatial and temporal dimensions. Finally, since none of the land-use data sources that were available for the Greater Toronto-Hamilton Area (GTHA) covered the complete time interval from 1986 to 2017, a machine learning model was trained to classify residential land use based on housing market dynamics. This information was then stored in the form of a relational database to facilitate ease of access and reduce 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 housing market database for the GTHA. Specifically, it presents methods developed for dealing with two important technical challenges in making use of Teranet's dataset:

1. Enhancing the Teranet sales transactions data with socio-economic and urban form variables, including parcel-level land-use data, by merging diverse datasets spatially and temporally.
2. Using machine learning to classify land use from historically observed housing market dynamics.

Section 2 gives a brief overview of the existing literature on the relationship between land use and housing market dynamics and machine learning techniques to forecast land use at the parcel level. Section 3 describes the data sources used in this study, that contributed different attributes characterizing the urban form, land use, households, buildings and property market transactions within the GTHA. Section 3 also describes the method developed to combine these disparate datasets into a functional lon-

gitudinal relational database, suitable for supporting real estate market analysis and modeling. Section 4 presents a novel application of machine learning methods to classify land use based on parcel-level sales transaction data augmented with other urban spatial data, as described in Section 3. Section 5 presents the results of the model and evaluates the effectiveness of the machine learning model through a validation process. Section 6 summarizes the findings and conclusions of the study and suggests areas for further research.

2 Brief literature review

There are numerous studies investigating the factors that influence housing consumers' needs and evaluation of the attractiveness of housing alternatives at any given point in time. Housing prices are often used to analyze a household's willingness to pay for housing attributes as well as neighborhood characteristics and access to different amenities from a given location. Since we do not have data on dwelling unit structural qualities in the Teranet dataset, the literature review focuses on a larger set of external price-influencing factors posited by previous research. These include the socio-economic character of the neighbourhood, urban form context in terms of location of property and the surrounding land use, transport accessibility, quality of schools, type and age of buildings, access to public amenities and macro-factors like interest rates and property taxes etc. (Case & Mayer, 1996; Dubin, 1998; Füss & Koller, 2016; Potepan, 1996; Spinney et al., 2011).

The effect of land use and transportation access on house prices has been an important area of research. Clapp et al. (2002) found spatial patterns in house price change: Access to the central business district is associated with a house price gradient; access to decentralized employment subcenters causes more localized changes in house prices; and neighborhood amenities (and disamenities) can cause house prices to change rapidly over relatively short distances. Mixed land use planning has been encouraged for its accessibility benefits to residents and indirect link to promoting sustainable transport options. Similarly, the location of jobs with respect to worker's residence is a critical influencing factor that determines the consumer's willingness to pay and decision of residential location choice. A Chicago metropolitan area study found that an increase in job accessibility leads to an increase in house prices, while mixed land use has a negative effect on housing prices (Kim & Jin, 2019).

Machine learning algorithms such as support vector machine (SVM), logistic regression and non-linear classification methods such random forest are being increasingly used for studying social and ecological processes in novel ways, in a variety of applications such as forecasting housing prices, simulating spatial patterns of urban expansion or examining land-use change dynamics (Chen et al., 2017; Fan et al., 2008; Luo et al., 2019, Wu et al., 2009). Based on changes detected in remote sensing data, land use/land cover change is mostly analyzed to comprehend the driving forces behind the spatio-temporal processes of rural-urban land-use transformation, and to predict future land-use changes. Logistic regression has been widely used to identify economic, social, and ecological causal factors such as population, proximity to roads and facilities and surrounding land-use (Shu et al., 2014; Verburg et al., 2004; Wu et al., 2009). Thus, the physical urban form or built environment and distribution of human economic activity has an influence on land-use change. Further classification of urban built-up land use into residential, commercial, industrial and other urban categories is scarcely attempted.

Nonetheless, spatial auto-correlation in property prices, irrespective of land use, is a common observation in the housing research literature (Ismail, 2006). For example, a study analyzing the spatial aggregation phenomenon using real estate transaction price records of Taitung City, Taiwan, found that property prices rise in spaces surrounded by high-priced real estate (Wang et al., 2019). Housing market dynamics can reveal important insights into the urban economics of a region and have significant

implications for economic policy, as seen in the study of price dynamics of different property types in Scotland (Katsiampa & Begiazi, 2019). The authors found evidence of i) breakpoints around the recent financial crisis in three property types (flats, terraced, semi-detached) and in the average house prices, ii) negative impact of the unemployment and interest rates on house prices irrespective of the property type and positive effect of the CPI in the prices of the detached, terraced and average houses.

Hence, if housing price dynamics of property types can reveal insights about regional economics, the reverse can be tested using housing prices and macro/ micro-economic variables to provide information about property types. The current paper uses a novel machine learning approach to explore the dynamic relationship between land use, socio-economic activity, urban form and housing prices and attempts to classify land use by capturing the inherent characteristics of property types as exhibited in the property transaction data.

3 Data sources for the longitudinal housing market relational database

This section discusses the process of developing a housing market database for the GTHA and storing it as a Relational Database Management System (RDBMS). This database was constructed by joining a detailed dataset of real estate sales transactions combined with other urban data sources and a new feature produced by a machine learning algorithm. The various data sources used to compile the housing market database are described below.

3.1 Teranet sales dataset (1800s-2017)

At the heart of the GTHA housing market database lies Teranet's sales dataset— containing historical sales data of approximately 9 million real estate transactions recorded in the Province of Ontario, since the beginning of the nineteenth century. This study considers sales transactions for parcels falling within the GTHA during the time period 1986 to 2017, since records prior to 1986 appear to be less consistent in the dataset.

The Province of Ontario Land Registration Information System (POLARIS) was built in 1985, to house and process electronic land records, which in turn led to the creation of an extensive dataset of land registration records managed by Teranet Enterprises Inc., an e-services organization. The Government of Ontario established a partnership with Teranet in 1991, to fully automate the conversion of millions of paper-based documents and records into the Ontario Electronic Land Registration System (ELRS) (Teranet Enterprises Inc., 2019).The Teranet dataset provides a very high spatial and temporal resolution, but at the same time, using data in its raw form for meaningful analysis and modeling was challenging for two main reasons:

1. Each observation contains the Parcel Identification Number (PIN), parcel address, the transaction date and consideration amount (sales price) and the X-Y coordinates of the parcel centroid, but there is no attribute to distinguish between residential, commercial and industrial transactions. In other words, parcel land use or property use information is not provided.
2. Structural data about the plot or the built structure(s) is also not part of the Teranet dataset, which means there is no information about the floor area sold or number of stories/units.
3. The quality of data is inconsistent and requires cleaning as it contains duplicate PINs, null or missing values, $1 or other low and unreasonable sales values, and extreme outliers.

At the same time, since each sales transaction record has a timestamp (date) and location coordinates, the observations could be joined to a variety of other geocoded urban data sources, such as socio-

economic data from the 5-yearly Census and the GTHA-specific household travel survey, Transportation Tomorrow Survey (TTS). As is discussed in Section 3, joining these data sources together required special consideration, as they use different spatial units and are available at different temporal spans.

3.2 Select variables from the census of Canada (1986-2016)

Census datasets provide valuable insights into the economic, social and demographic profiles and trends in Canada every five years and disseminate the information by a range of geographic units, also referred to as "Census geography" (Map and Data Library, 2019). This study uses Dissemination Area (DA) level data on dwelling characteristics and socio-demographic-economic characteristics of households.

3.3 Select variables from the Transportation Tomorrow Survey (1986-2016)

Another major source of information for most transportation data for the GTHA is the Transportation Tomorrow Survey (TTS), a household travel survey managed by the University of Toronto Transportation Research Institute's (UTTRI) DMG (Data Management Group, 2014). The TTS, conducted every five years since 1986, is a retrospective survey of travel taken by every member (age 11 or above) of each sampled household during the day previous to the telephone or web contact. TTS survey data includes dwelling, population and job densities, characteristics of the household, full-time workers, students, work-commuting mode, car ownership, and details of all the trips taken weekly by each member of the household by all modes of travel and for all trip purposes (Ashby, 2018).

3.4 Land use (2001-2014) and points of interest (2001-2016) from DMTI Spatial Inc.

DMTI Spatial Inc., a Digital Map Products company, is a major provider of location-based information in Canada. DMTI has been providing enterprise Location Intelligence solutions for more than a decade to Global 2000 companies and government agencies (DMTI Spatial Inc., 2014). This study did not directly utilize DMTI data in its analysis, but spatial joins were made for parcel-level land-use data for the period 2001-2014, as well as Enhanced Points of Interest (EPOI) from 2001-2016 from DMTI Spatial, and the augmented dataset was stored in the RDBMS for retrieval by other researchers, if needed.

3.5 Detailed land-use data from University of Toronto's Department of Geography and Planning (2012-2013) and City of Hamilton's open data portal (2010)

The detailed land-use data for the Greater Toronto Area (GTA), provided by University of Toronto's Department of Geography and Planning, consists of Teranet 2016 parcels assigned with manually coded land-use produced using Google maps and street views between 2012-2013 by Prof. Andre Sorensen and Prof. Paul Hess's student research team at the University of Toronto. Hamilton's detailed land use for 2010 was sourced from the City's open data website and combined with the GTA land use to prepare comprehensive land-use codes for the whole GTHA.

4 Creating a relational database of 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 such as 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).

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 Analysis Zone (TAZ). A TAZ is a polygon which typically falls along the centre line of roads or natural geographic boundaries (Data Management Group, 2019). TAZ boundaries often – but not always -- 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.

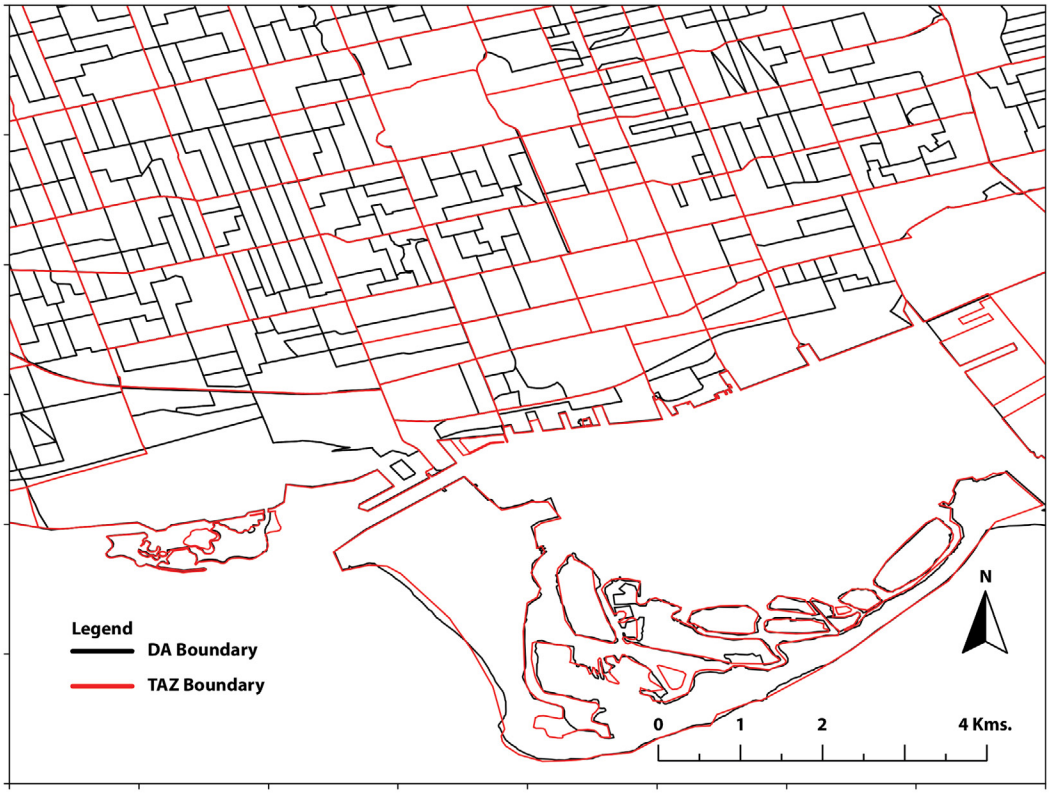

Figure 1. Spatial relationship between datasets: difference between Traffic Analysis Zones (TAZs) and Census Dissemination Areas (DAs)

To make the TTS data consistent for comparing across all years from 1986 to 2016, DMG, custodians of the TTS dataset, aggregated the data to the 2001 TAZ boundaries. The 2001 TAZ system is used to model travel times for the GTHA in the EMME network modeling system 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 for 2001 TAZs.

Table 1 below gives the summary of spatial units used by the data sources that were combined into the GTHA housing market database:

Table 1. Summary of spatial data sources and units of analysis

Spatial unit	Data source	Standardized boundary
Point data joined to Teranet parcels	• Teranet sales records (X-Y coordinates of parcel centroid) • Points of Interest from DMTI Spatial	2016 Teranet parcels
Parcel-level data (polygons)	• Detailed land use from the Department of Geography & Planning, UofT • Land use from DMTI Spatial	2016 Teranet parcels
DA-level data (polygons)	• Census variables • GIS built environment variables	2016 DAs
TAZ-level data (polygons)	• TTS variables • EMME accessibility variables	2001 TAZs

4.1 Joining different spatial datasets

When joining these data sources, differences in spatial units need to be respected, which can be more challenging when spatially joining polygons with other polygons, since it might require area-weighted spatial interpolation of data to a common unit of analysis. In addition, polygon boundaries- can also vary with time, as is the case with Census DA boundaries. Hence, all DA-level data has been interpolated to standard 2016 DA boundaries. To simplify relating different polygon-based data sources with each other, all of them have been associated to 2016 Teranet parcels via unique identifiers (primary keys). This allows flexibility in combining variables from polygon-based DA and TAZ-level data sources to common time-indexed Teranet sales transaction points (parcel centroids), while maintaining the integrity of spatial and temporal relationships through polygon-to-point spatial joins. The time frame for this study is 2001-2016, since data from DMTI Spatial is available 2001 onwards only. Figure 2 presents the temporal spans of each data source combined in the longitudinal housing market database.

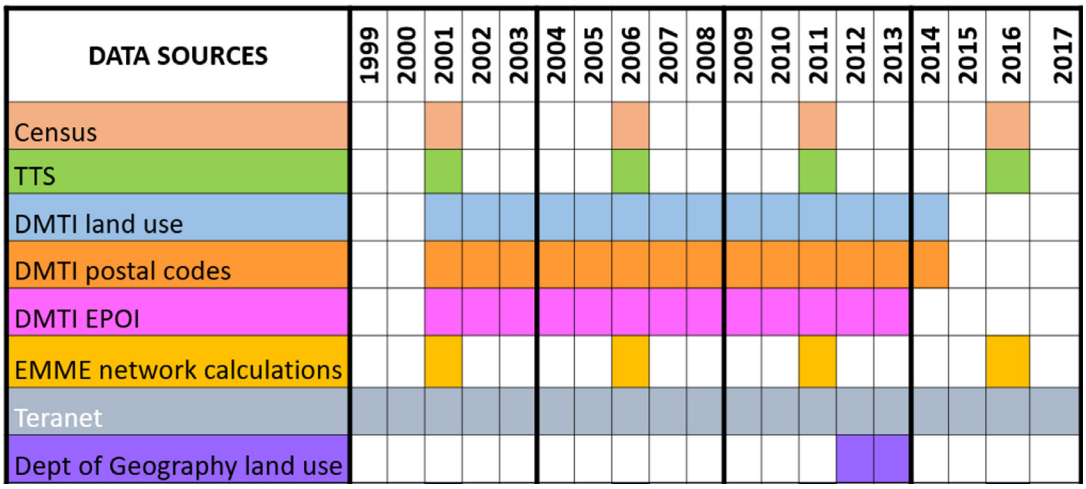

Figure 2. Temporal spans of data sources used in the GTHA housing market database (study period 2001-2016; Census, TTS variables and Teranet data are available up to 1986).

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; this approach recognizes land -se changes that happen between the years for which DMTI land-use data was available. The detailed land use provided by the Department of Geography &

Planning was joined to all Teranet records, while noting that it is most accurate for the years 2012 and 2013.

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

1. Match Census / TTS variables for a y specific year with appropriate Teranet records from the corresponding year.
2. Via interpolation of discrete Census / TTS variables.
3. Through assignment of 5-year temporal spans of Census / TTS data as new features to Teranet records.

To utilize the maximum number of Teranet records and avoid additional interpolation assumptions and use the actual values recorded from Census and TTS surveys, the third option has been chosen for joining Teranet records with Census / TTS variables. Each Census / TTS dataset was assigned a 5-year time block centered at the survey year (i.e., 2014–2018 for 2016 survey year) to generate values for non-census years and new foreign keys were added to Teranet records to correspond with the continuous dataset of Census / TTS variables. Thus, the time-series socio-economic and urban form spatial data interpolated to consistent polygon boundaries (DA/ TAZ), has been linked to Teranet parcels for the GTHA using primary keys (DA and TAZ respective unique identifiers). The housing market database contains those parcel's information which recorded residential sales transactions between 2001-2016 in the GTHA. Figure 3 shows the distribution of land-use joined parcels that have transaction price records.

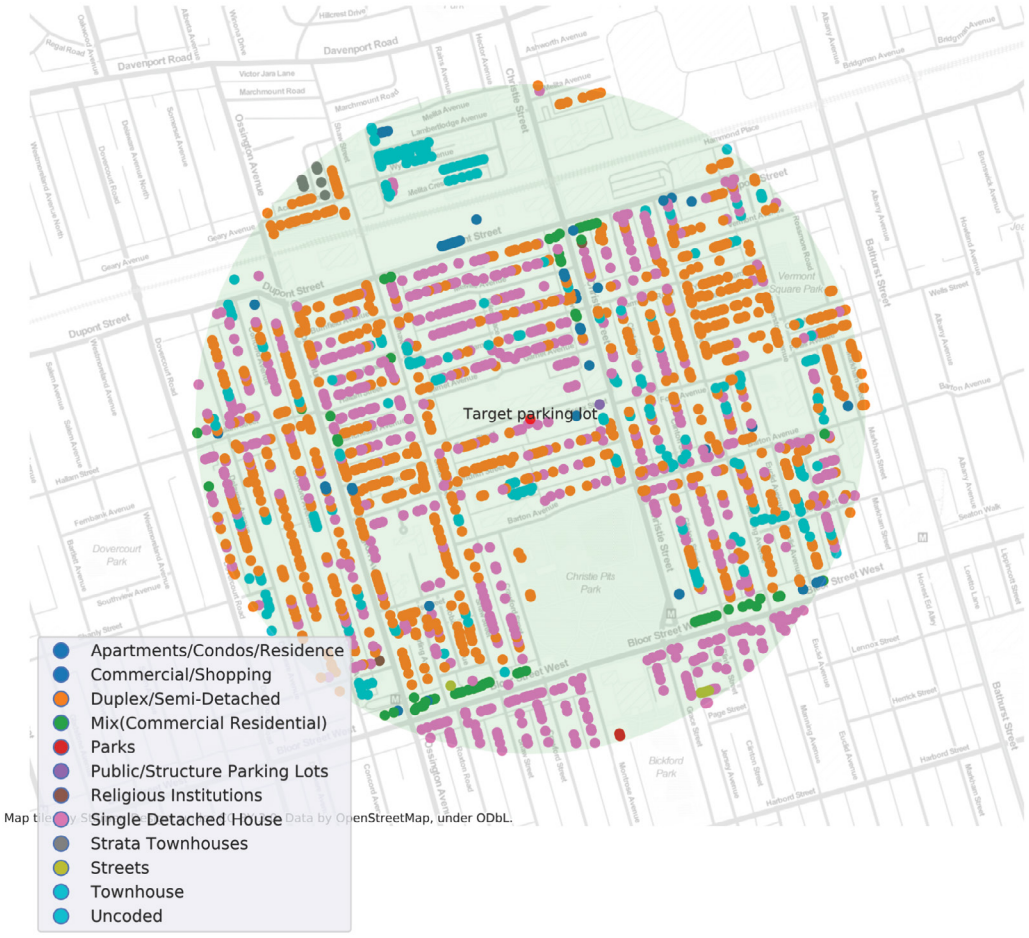

Figure 3. Detailed parcel-level land use joined to Teranet transaction records

4.2 Relational database 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:

1. Detect, clean or treat anomalies in the Teranet dataset and make the records consistent over time and space.
2. Introduce new keys that would allow efficient joining of other data sources, such as Census and TTS variables to parcel-level land-use information, while maintaining the integrity of spatial and temporal relationships.
3. Engineer new features capturing the housing market dynamics, that can be used by the machine learning algorithm along with the features from the joined datasets to identify and classify land use.

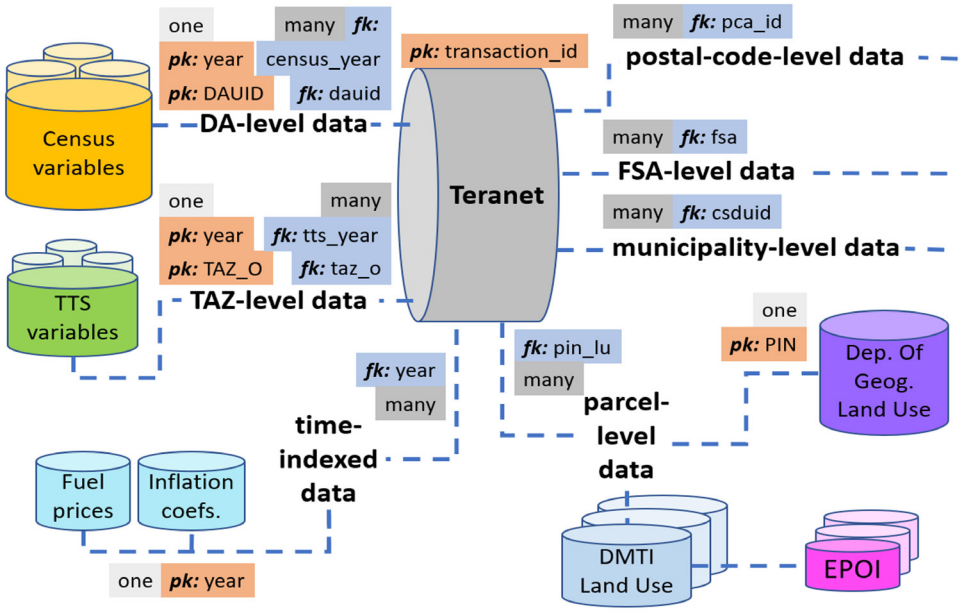

Figure 4. Entity Relationship (ER) diagram of the GTHA housing market RDBMS

To implement the spatial and temporal relationships between the data sources, and establish the referential integrity constraints of the GTHA housing market RDBMS, a number of new foreign keys were introduced to Teranet records via a series of spatial joins and feature engineering. Figure 4 presents the Entity Relationship (ER) diagram of the GTHA housing market database.

In addition to producing new keys for joining datasets, several new features were engineered from original Teranet attributes. These new features are intended to give each sales record a spatio-temporal "context" of the housing market dynamics by grouping records using different criteria (e.g., rolling count of transactions coming from a particular coordinate pair, ratio of sales price of a property to median for that year, etc.). These new features have been tested with a machine learning algorithm to classify land use from housing market dynamics.

5 Prototype of a machine learning workflow to classify land use

One of the major features that is missing in the Teranet dataset is the information about the type of property being transacted, which introduces a major limitation on differentiating the market dynamics of say, housing market from commercial or office properties. None of the available sources of land use information covered the entire time interval from 1986 to 2017. Land-use GIS shapefiles from DMTI Spatial is the only time-series land-use source available, but the polygons are coarse and there are only 7 basic land-use categories (residential, commercial, industrial & resource, open area, or water body) provided in coarse polygons, which do not follow parcel boundaries exactly and hence, introduce a margin of error when joined. The most detailed and accurate source is land-use data collected between 2012 and 2013 by University of Toronto's Department of Geography & Planning research team, which also classifies parcels by housing property type (single detached, semi-detached, townhouse, duplexes and condominium-apartments). To address this issue, the 2012-13 detailed land-use data were used as labelled data to train a machine learning model capable of recognizing certain property types that have characteristically different behavior within 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, time interval between subsequent transactions, ratio of price to median price for that year, etc. This section discusses the basic prototype machine learning workflow developed in this study to classify land use from housing market dynamics.

5.1 Target variable

The target variable was constructed from property records from 2011 to 2014, by grouping the sales transaction records by land-use codes, such that all the Teranet records are grouped into three major land-use classes, of similar transaction volume. 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 among themselves and thus, produce a more balanced dataset.

In addition, the chosen groupings of land-use types combine categories that have a similar distribution of price and count of sales per coordinate pair between categories to form a single class. For example, detached and semi-detached houses and townhouses have a much smaller frequency of transaction and a higher median price per coordinate pair when compared to condominium-apartments and strata townhouses. Strata townhouses are condominium-townhouses that are owned individually, but a fee is required to be paid to the Condominium Management Company for maintenance of common areas.

The three target classes introduced are:

- Class 0: "condo," including multi-storied Apartments/Condos/Residence
- Class 1: "house," including Single Detached Houses, Duplex/Semi-Detached and Townhouses.
- Class 2: "other," including Commercial/Office, Mix (Commercial Residential), Industrial/Employment Lands, and all other classifications.

5.2 Dimensionality reduction and hyperparameter tuning

To reduce the dimensionality of the augmented Teranet 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 5 presents the top 11 features that were selected by at least four different feature selection methods, and have been chosen to be tested with the machine learning algorithm. These include select Census variables pertaining to the DA of the property (average rent, number of apartments with 5 or more stories, average household size, population density and dwelling density) and new features engineered from the native Teranet attributes of a property (first sale transaction, number of years till the next sale, frequency of re-sale, median sale price of a property, number of previous sales, total number of sales transactions for each property and number of years since the previous sale). Table 2 provides definitions of the 11 features selected and displayed in Figure 5.

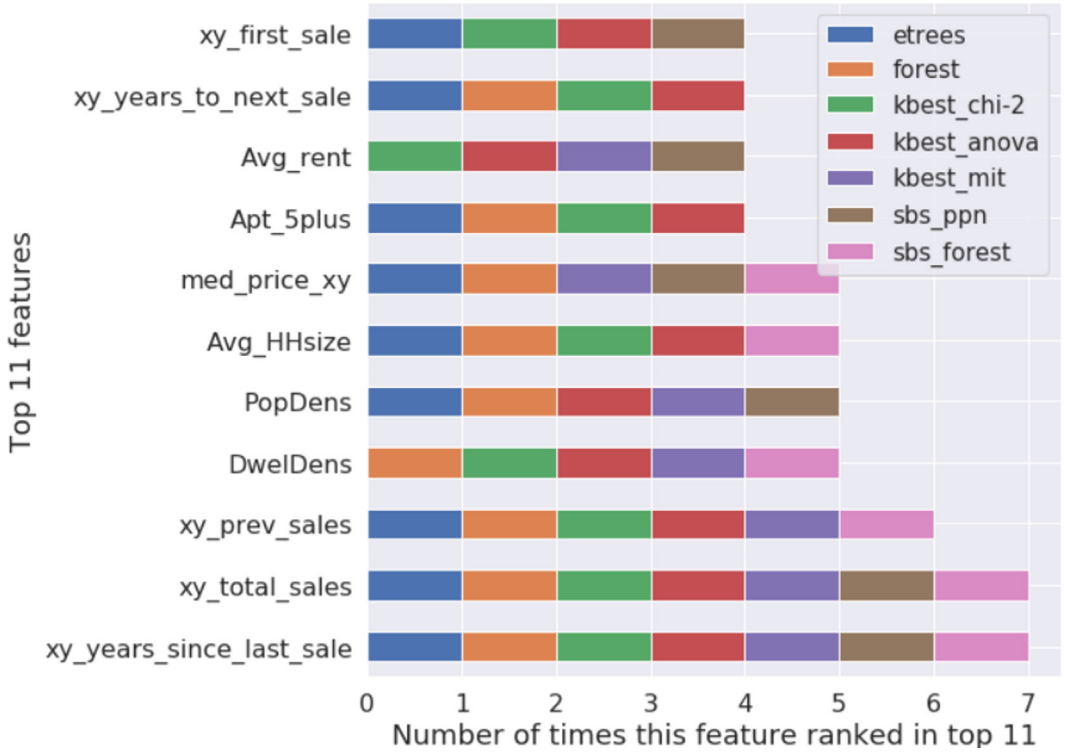

Figure 5. Feature Subset Selection (FSS) results (A number of algorithmic feature selection techniques (Select from Model, Select K-best, Sequential Backward Selection) were used to select an optimal subset of features to classify land use.)

Table 2. Variable labels and definitions for the 11 selected features.

S.No.	Variable	Label	Definition
1.	xy_first_sale	First sale transaction for a property	Set to True (1), if a record is the earliest transaction for a property, identified by an XY coordinate pair
2.	xy_years_since_last_sale	Number of years since the last sale	Number of years passed since the previous property transaction recorded for an XY coordinate pair
3.	xy_years_to_next_sale	Number of years till the next sale	Number of years until the next property transaction recorded for an XY coordinate pair
4.	xy_prev_sales	Rolling frequency of sales transactions	Rolling count of previous transactions recorded for an XY coordinate pair
5.	xy_total_sales	Total number of sales transactions	Total number of transactions recorded for an XY coordinate pair.
6.	med_price_xy	Median sale price of a property	Median price (in 2016 dollars) of all transactions recorded for an XY coordinate pair
7.	Avg_rent	Average monthly shelter cost for rented dwellings	Average monthly rent paid by a tenant in a zone (2016 CAD)
8.	Apt_5plus	Apartment dwelling type with 5 or more stories	Percentage of apartments with 5 or more stories in all dwelling types (single detached, attached, apartments with under 5 stories and apartments with 5 or more stories) in a zone
9.	Avg_HHsize	Average household size	Average number of members in a household
10.	PopDens	Population density	Total number of residents in a zone divided by the area of the zone (Persons/Sq.Km.)
11.	DwelDens	Dwelling density	Total number of residential buildings in a zone divided by the area of the zone (Dwellings/ Sq.Km.)

There are two types of parameters in machine learning: those that are learned by parametric models from training data (e.g., weights in logistic regression), and the parameters that tune the performance of a learning algorithm, i.e., its hyperparameters (e.g., 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 final evaluation of a 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 a model to the 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.

5.3 Model selection

An important point to be summarized from the famous No Free Lunch Theorems (NFL) (Wolpert 1996; Wolpert & 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 has been compared: perceptron learning algorithm (η = 0.5, maximum iterations=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), and Gaussian Naive Bayes.

6 Evaluation of model results

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

6.1 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), and micro- or macro-averaged. Micro-averaging can be useful to weight 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 & Mirjalili 2017).

6.2 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 & Planning 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 evaluation, but were utilized to test fitted models as an additional reference for the generalization of performance of classifiers. Figure 6 presents model performance on the train, test, and two additional validation subsets.

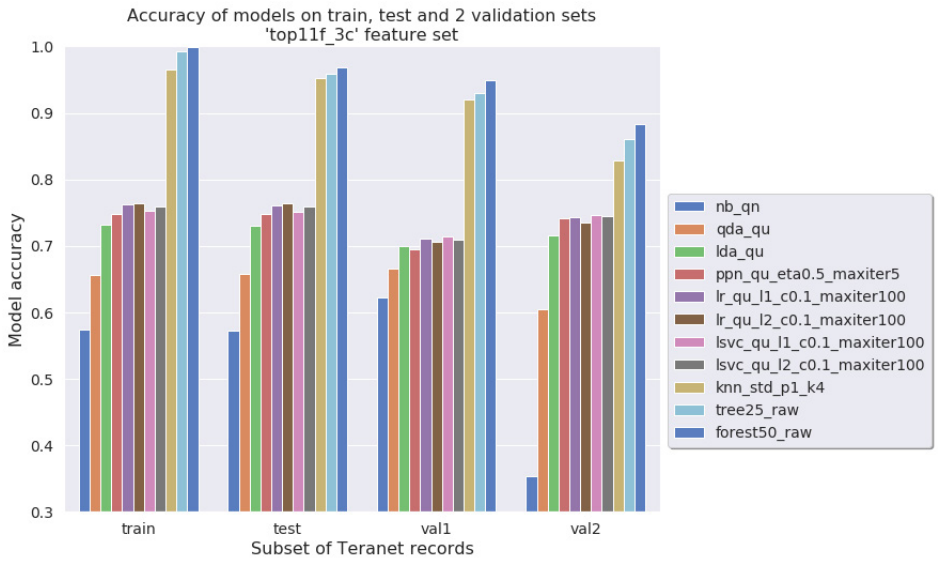

Figure 6. Model performance (accuracy) on train, test, and two additional validation subsets; it can be seen that the nearest neighbors (using standardized features) and tree-based models (using raw features) dramatically outperform linear models (using quantile-transformed (uniform PDF) features), with random forest being the best performing model at 97% accuracy on the test subset

As can be seen on Figure 6, in terms of prediction accuracy, tree-based and nearest-neighbor 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, as was validated by plotting their learning curves.

6.3 Best performing model: Random forest

As can be seen in Figure 6, random forest with 50 estimators and Gini impurity criterion showed the best results in terms of accuracy on all subsets. Figure 7 presents the classification report showing all model performance metrics discussed in Section 5.1 and Figure 8 presents confusion matrices for the best performing model: random forest with 50 estimators using Gini impurity criterion.

	condo	house	other	accuracy	macro avg	weighted avg
precision	0.996800	0.940099	0.975766	0.968755	0.970888	0.969202
recall	0.992179	0.973011	0.949980	0.968755	0.971723	0.968755
f1-score	0.994484	0.956272	0.962700	0.968755	0.971152	0.968812
support	66871.000000	86664.000000	103079.000000	0.968755	256614.000000	256614.000000

Figure 7. Best model performance on test set: classification report for random forest with 50 estimators using Gini impurity criterion; the model performs consistently well across all three target classes

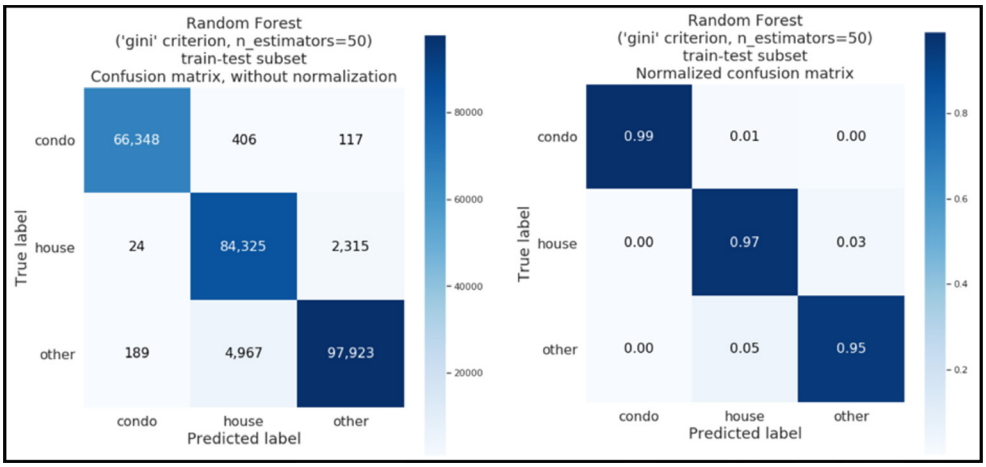

Figure 8. Confusion matrices for best performing model: random forest

It can be seen that the model is capable of recognizing all the major property classes with a high degree of accuracy. Thus, features produced from the housing market dynamics have a strong predictive power when classifying land use at a parcel level. The best performing model was used to classify land use of all Teranet records from 1986 to 2017 and save the result as a new feature in the GTHA housing market database. Figure 9 presents feature importance for random forest with 50 estimators and Gini impurity criterion.

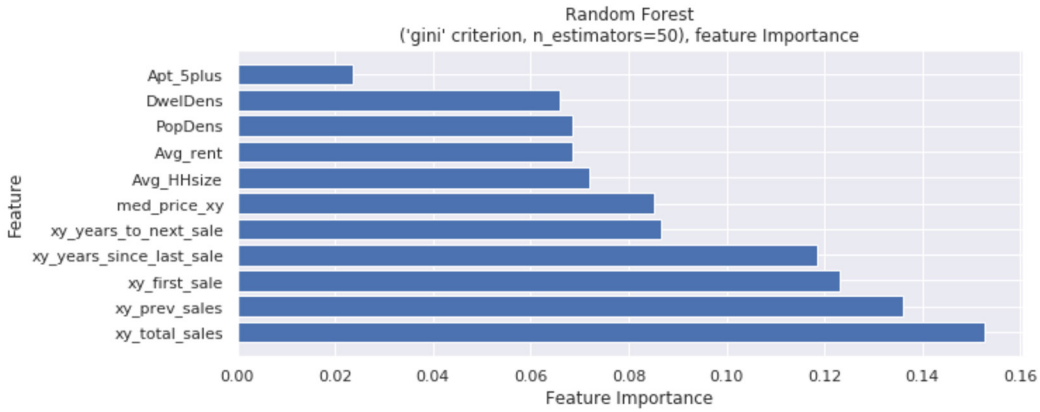

Figure 9. Feature importance for best performing model: random forest classifier with 50 estimators using Gini impurity criterion

Engineered parcel-level features capturing the housing market dynamics, have stronger predictive power than DA-level Census variables when classifying land use: total number of sales transactions for each property, number of previous sales since a reference year, first sale year, number of years since the previous sale and till the next sale from a reference year and the median price of property. This likely reflects that much greater spatial-temporal precision of the parcel-level features, relative to the Census variables, which are only collected every five years and at the DA-level of spatial aggregation. The Census variables, however, still improve model accuracy and so are retained in the analysis.

7 Discussion and conclusions

Microsimulation models represent 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 possess considerable value for design and validation of integrated urban models and for longitudinal studies concerned with evolution of urban form. Introduction of the POLARIS electronic land registration system by the Province of Ontario in 1985 led to the creation of an extensive dataset of real estate transactions by Teranet Enterprises Inc.

However, despite having very high spatial and temporal resolution, the Teranet dataset suffered from severe lack of features describing individual transactions. One of the major attributes missing from Teranet data was 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 University of Toronto's Department of Geography & Planning 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 the University of Toronto 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 till 2017 covered by the Longitudinal GTHA Housing Market database.

To augment Teranet's dataset, new variables were engineered from its native attributes to capture the housing market dynamics at the parcel level. 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 residential land use for each Teranet record within the span of Census / TTS variables, thus recognizing residential land-use changes with time.

The new features engineered from native Teranet attributes are shown to have strong predictive power when classifying residential land use. When joined with Census variables at the level of Dissemination Areas, new features engineered from Teranet's dataset allow 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 the time intervals between transactions, the total volume of transactions and their median transaction price for a single building (the "xy" variables in Figure 5 and Table 2) have strong predictive power for classifying land uses, 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 performed 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 condo-apartments and mixed-use properties. The augmented Teranet dataset with land use produced by the classification algorithm, along with related Census and TTS tables, has been transformed into a relational database to facilitate ease of access by a broader group of researchers.

While the specific lack of attributes in the Teranet database is perhaps a somewhat special case, the challenge of fusing diverse spatial-temporal datasets to create enriched databases for analysis and modelling of housing markets (and other spatial processes) is a very common one. Thus, the methods developed in this paper for dealing with inconsistent spatial analysis units and time points are of general applicability. The paper also demonstrates the power of machine learning methods for reliably imputing missing variables, again, a very generic problem. Further, the study demonstrates that parcel-level attributes engineered from property sale transaction values and volume data are strong predictors of property land use.

Finally, note that the method developed in this paper for imputing residential land uses can be extended to other land-use types (commercial, etc.), providing a suitable labelled training dataset is available.

Acknowledgements

The research reported in this paper was supported by an Ontario Research Fund Research Excellence Round 7 grant, a Natural Sciences and Engineering Research Council Discovery grant, and Teranet Enterprises Inc. The contributions of Dena Kasraian and JieLan Xu to the work are gratefully acknowledged. And we thank Profs. Andre Sorensen and Paul Hess for providing access to their parcel land-use dataset.

References

Alonso, W. (1964). *Location and land use, towards a general theory of land rent*. Cambridge: Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674730854

Arribas-Bel, D. (2014). Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography, 49,* 45–53. https://dx.doi.org/10.1016/j.apgeog.2013.09.012

Ashby, B. (2018). *TTS 2016 City of Toronto summary by Ward.* Toronto: Malatest. http://dmg.utoronto.ca/pdf/tts/2016/2016TTS_Summaries_Toronto_Wards.pdf

Case, K. E., & Mayer, C. J. (1996). Housing price dynamics within a metropolitan area. *Regional Science and Urban Economics, 26*(3–4), 387–407. https://doi.org/10.1016/0166-0462(95)02121-3

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (Aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies, 68*, 285–299.

Chen, J. H., Ong, C. F., Zheng, L., & Hsu, S. C. (2017). Forecasting spatial dynamics of the housing market using support vector machine. *International Journal of Strategic Property Management, 21*(3), 273–283. https://doi.org/10.3846/1648715X.2016.1259190

Clapp, J. M., Kim, H. J., & Gelfand, A. E. (2002). Predicting spatial patterns of house prices using LPR and Bayesian smoothing. *Real Estate Economics, 30*(4), 505–532. https://doi.org/10.1111/1540-6229.00048

Data Management Group. (2014). Data Management Group at the University of Toronto Transportation Research Institute. Retrieved from http://dmg.utoronto.ca. http://dmg.utoronto.ca

Data Management Group. (2019). Survey boundary files. Retrieved from http://dmg.utoronto.ca. http://dmg.utoronto.ca/survey-boundary-files

DMTI Spatial Inc. (2014). *CanMap ® RouteLogistics user manual V2014.2.* (2014.2). Retrieved from www.dmtispatial.com

Dubin, R. A. (1998). Predicting house prices using multiple listings data. *Journal of Real Estate Finance and Economics, 17*(1), 35–59. https://doi.org/10.1023/A:1007751112669

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research, 9,* 1971–1874. http://www.csie.ntu.edu.tw/

Füss, R., & Koller, J. A. (2016). The role of spatial and temporal structure for residential rent predictions. *International Journal of Forecasting, 32,* 1352–1368. https://doi.org/10.1016/j.ijforecast.2016.06.001

Iacono, M., Levinson, D., & El-Geneidy, A. (2008). Models of transportation and land-use change: A guide to the territory. *Journal of Planning Literature, 22*(4), 323–340.

Ismail, S. (2006). Spatial autocorrelation and real estate studies: A literature review. *Malaysian Journal of Real Estate, 1*(1), 1–13.

Katsiampa, P., & Begiazi, K. (2019). An empirical analysis of the Scottish housing market by property type. *Scottish Journal of Political Economy, 66*(4), 559–583. https://doi.org/10.1111/sjpe.12210

Kelly, E. D. (1994). The transportation land-use link. *Journal of Planning Literature, 9*(2), 128–145. https://journals-sagepub-com.myaccess.library.utoronto.ca/doi/10.1177/088541229400900202

Kim, D., & Jin, J. (2019). The effect of land use on housing price and rent: Empirical evidence of job accessibility and mixed land use. *Sustainability, 11*(3), 938. https://doi.org/10.3390/su11030938

Knight, R. L., & Trygg, L. L. (1977). *Land-use impacts of rapid transit* (DOT-TPI-10-77-29). Washington, DC: U.S. Department of Energy, Office of Scientific and Technical Information. https://www.osti.gov/servlets/purl/5952387https://www.osti.gov/servlets/purl/5952387

Lee, D. B. (1973). Requiem for large scale models. *Journal of the American Institute of Planners, 39*, 163–178.

Luo, T., Tan, R., Kong, X., & Zhou, J. (2019). Analysis of the driving forces of urban expansion based on a modified logistic regression model: A case study of Wuhan City, Central China. *Sustainability, 11*(8), 2207. https://doi.org/10.3390/su11082207

Manheim, M. L. (1978). *Fundamentals of transportation systems analysis volume 1: Basic concepts*. Cambridge, MA: MIT Press. https://mitpress.mit.edu/books/fundamentals-transportation-systems-analysis-volume-1

Map and Data Library. (2019). *Canadian census geography (unit) definitions.* Toronto: University of Toronto. https://mdl.library.utoronto.ca/canadian-census-geography-unit-definitions

Martinez, F. J. (2018). *Microeconomic modeling in urban science.* Cambridge, MA: Academic Press, Elsevier.

Miller, E. J. (2018). The case for microsimulation frameworks for integrated urban models. *Journal of Transport and Land Use, 11*(1), 1025–1037.

Miller E. J. (2019). Travel demand models, the next generation: Boldly going where no one has gone before. In K. G. Goulais & A. W. Davis (Eds.), *Mapping the travel behavior genome*. Cambridge, MA: Elsevier.

Miller, E. J., Kriger, D. S., & Hunt, J. D. (1998). *Integrated urban models for simulation of transit and land-use policies guidelines for implementation and use* (TCRP Report 48). Washington, DC: Transportation Research Board.

Potepan, M. J. (1996). Explaining intermetropolitan variation in housing prices, rents and land prices. *Real Estate Economics, 24*(2), 219–245. https://doi.org/10.1111/1540-6229.00688

Raschka, S., & Mirjalili, V. (2017). *Python machine learning (*2nd Ed). Birmingham, UK: Packt Publishing.

Shu, B., Zhang, H., Li, Y., Qu, Y., & Chen, L. (2014). Spatiotemporal variation analysis of driving forces of urban land spatial expansion using logistic regression: A case study of port towns in Taicang City, China. *Habitat International, 43,* 181–190. https://doi.org/10.1016/j.habitatint.2014.02.004

Spengler, E. H. (1930). *Land values in New York in relation to transit facilities.* New York, NY: Columbia University Press.

Spinney, J., Kanaroglou, P., & Scott, D. (2011). Exploring spatial dynamics with land price indexes. *Urban Studies, 48*(4), 719–735. https://doi.org/10.1177/0042098009360689

Statistics Canada. (2015). *Dissemination area* (DA). Census program reference materials, 2011 census dictionary. Retrieved from https://www12.statcan.gc.ca/census-recensement/2011/ref/dict/geo021-eng.cfm

Statistics Canada. (2018). *Hierarchy of standard geographic units. Illustrated glossary 92-195-X.* Retrieved from https://www150.statcan.gc.ca/n1/pub/92-195-x/2011001/other-autre/hierarch/h-eng.htm

Teranet Enterprises Inc. (2019). About POLARIS, Teranet. Retrieved from www.teranet.ca. https://www.teranet.ca/registry-solutions/about-polaris/

Verburg, P. H., Schot, P. P., Dijst, M. J., & Veldkamp, A. (2004). Land-use change modelling: Current practice and research priorities. *GeoJournal, 61*(4), 309–324. https://doi.org/10.1007/s10708-004-4946-y

Wang, W. C., Chang, Y. J., & Wang, H. C. (2019). An application of the spatial autocorrelation method on the change of real estate prices in Taitung city. *ISPRS International Journal of Geo-Information, 8*(6), 249. https://doi.org/10.3390/ijgi8060249

Wegener, M. (1994). Operational urban models state of the art. *Journal of the American Planning Association, 60*(1), 17–29. https://www.tandfonline.com/doi/abs/10.1080/01944369408975547

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation, 8*(7),1391–1420.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67–82. https://ti.arc.nasa.gov/m/profile/dhw/papers/78.pdf

Wu, B., Huang, B., & Fung, T. (2009). Projection of land-use change patterns using kernel logistic regression. *Photogrammetric Engineering and Remote Sensing, 75*(8), 971–979. https://doi.org/10.14358/PERS.75.8.971