A prototype of a machine learning workflow to classify land use from housing market dynamics. Part of a Longitudinal Analysis of housing sales in the Greater Toronto-Hamilton Area.

by

Stepan Oskin

A thesis submitted in conformity with the requirements
for the degree of Master of Applied Science
Graduate Department of Civil Engineering
University of Toronto

Abstract

A prototype of a machine learning workflow to classify land use from housing market dynamics. Part of a Longitudinal Analysis of housing sales in the Greater Toronto-Hamilton Area.

Stepan Oskin
Master of Applied Science
Graduate Department of Civil Engineering
University of Toronto
2019

Increased digitization of human activities produces a wealth of new data that can be used to model and analyze urban systems. The complex interaction of transportation and land use could be studied empirically, for which fine-scale urban data from multiple sources needs to be combined. Teranet's dataset of real estate transactions holds fine-scale information on the housing market of Ontario, but is very limited in the number of available attributes. The dataset can be augmented by joining additional attributes from various data sources, such as Census or TTS survey, based on spatial and/or temporal relationships. However, this presents a challenge since these data sources use different spatial units and are available at different temporal spans. This master's thesis proposes a machine learning workflow to augment Teranet's data and classify land use based on housing market dynamics.

Contents

Chapter 1

Introduction

1.1 Introduction

Chapter 2 presents background information on land use and transportation models, context for Teranet's dataset of land registry records, its opportunities and challenges and the proposed solution, chapter 3 discusses the nature of the spatial and temporal relationships of different data sources used in this master's thesis, chapter 4 presents the data preparation workflow designed to implement the relationships introduced in chapter 3, chapter 5 describes a prototype of a machine learning workflow to classify land use from housing market dynamics, chapter 6 presents and discusses the results and chapter 7 presents the conclusion and outlines opportunities for future work.

Chapter 2

Background information

2.1 Chapter 2: transportation and land use, land registration in Canada and Teranet

This chapter discusses the complex interaction of land use and transportation, provides a brief overview of the history of development of land use-transportation (LUT) models, presents some legal and historical background for Teranet's dataset of real estate transactions and finishes with discussing challenges of working with Teranet's data and the proposed solution.

2.2 Land Use and Transport (LUT) models

2.2.1 Complexity of urban systems and "wicked" problems

In her famous 1961 book, Jane Jacobs[13] described a city as "a problem in organized complexity"; since then, many other researchers have remarked that urban systems exhibit complex behaviour[3, 4]. Complexity of a system can be defined as a state or quality of being intricate or complicated. For a system to be complex is not necessarily the same as to be complicated; complex systems can be simple, i.e. governed by a single equation. Complexity of a system has to do with the intrinsic ability of a system to surprise us with its behaviour; that the system is hard to understand, despite the mechanics of it being relatively simple.

In 1973, a little over a decade after Jacobs, Rittel and Webber[26] presented a path-breaking conceptualization; this conceptualization characterized urban planning problems as "wicked" problems: problems which cannot be definitively described and for which it makes no sense to talk of "optimal solutions". In their paper, Rittel and Webber stated that such "wicked" problems are never "solved", and that the focus instead becomes on iteratively "re-solving" the problems over and over. More than 40 years after their original publication, Rittel and Webber's ideas remain relevant to the policy sciences today: there is an intense interest in the nature of "wicked" problems and the complex tasks of identifying their scope, viable responses, and appropriate mechanisms and pathways to improvement[9]. Interaction between land use and transportation, which is discussed in the following section, presents a prime example of urban complexities and "wicked" problems.

2.2.2 Transportation-land use cycle

Among the reasons why transportation and land use interaction is "wicked" are such aspects as pluralism of expectations among stakeholders, institutional complexity in policy making, and scientific uncertainty[25]. More importantly, there is a fundamental link between transportation and urban form: urban form has an enormous impact on the type and cost of transportation systems needed to serve residents of a metropolitan area[14]. Transportation, in turn, influences land development and location choices of people and firms, and thus completes the formation of a feedback relationship that Stover and Koepke[31] referred to as a cycle. Interconnections between points (activities) in space can be perceived through the medium of the transportation system[23]

Figure 2.1 illustrates the complex interactions between land use and transportation system as summarized by Miller, Kriger and Hunt[23].

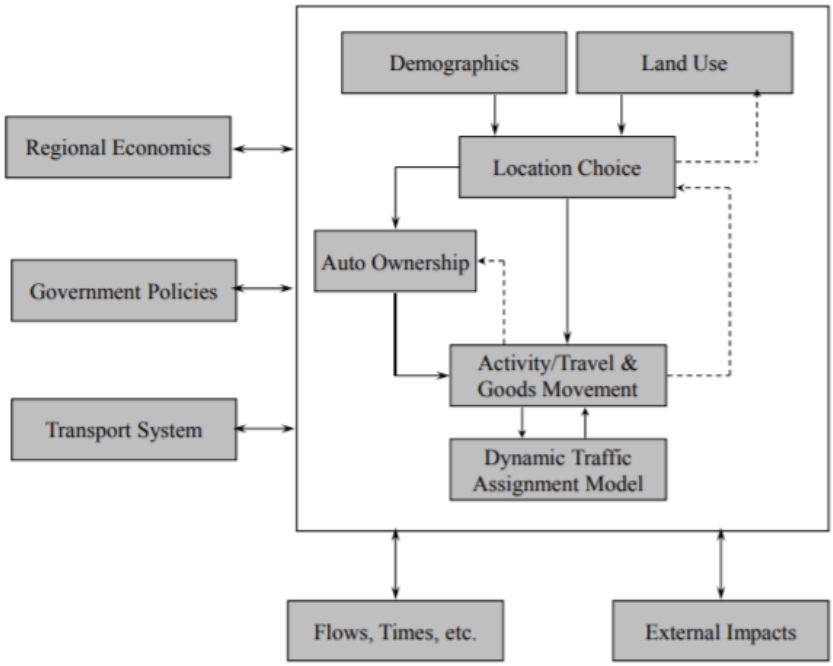

Figure 2.1: An Idealized Integrated Urban Model System, adapted from Miller, Kriger and Hunt[23].

Many different types of models are used in planning, such as demand forecasting models projecting traffic or ridership, or land use models projecting and distributing population and jobs within an area. At an earlier stage of model development, some analysts argued that there is no significant link between transportation and land use, given the near-ubiquity of the transportation (road) network[23]. However, the unprecedented urban growth of the 21st century introduced new challenges for urban systems such as extreme road congestion, equity of access to jobs and services among low-income households, energy scarcity, environmental and GHG impacts from transportation systems and public health impacts of land use patterns[19, 24].

It became apparent that these "transport problems" cannot be solved through transportation policies and investment alone, that the physical design of the city at the "macro" and "micro" scale critically interfaces with the demand for and performance of the transportation system. In addition, to accurately assess the costs and benefits of an expensive long-term transportation infrastructure investment, "feed-

back" effects of these investments on urban form, land values, property taxes, quality of life, etc. need to be quantified and included in evaluation and decision making. Thus, today there is a steadily growing recognition within the urban policy field that the interaction between transportation and land use does exist and does matter[19].

In the context of models, integrated urban models (IUMs) aim to capture the complex relationship between urban systems such as transportation and land use more accurately. Integrated land use-transportation models combine travel demand forecasting and land use forecasting functions and recognize that the distribution of population and jobs depends, in part, on transportation accessibility. The reverse is also true, and thus integrated models incorporate feedback relationship between transportation and land use, with economic decisions by households and firms acting as one of the links between the two systems[23].

2.2.3 Evolution of LUT models

The history of treating cities as systems via simulation models of transportation and land use dates back to 1950s when General System Theory and Cybernetics came to be applied in the softer social sciences[3]. The first operational simulation model that truly integrated land use and transportation is considered to be A Model of Metropolis built in 1964 by Ira S. Lowry for the Pittsburgh region based on economic base theory[15]. It was a highly aggregate model based on theories of spatial interaction, such as the gravity model that was popular in quantitative geography and transportation planning at the time[5]. Models based on spatial interaction framework continued to be developed through mid-1980s, until developments in random utility theory allowed researchers to describe choices among discrete alternatives, such as the choice of travel mode, and generate models based on the study of disaggregate behaviour[12].

Figure 2.2 provides the general overview of chronological development of LUT models summarized by Iacono[12].

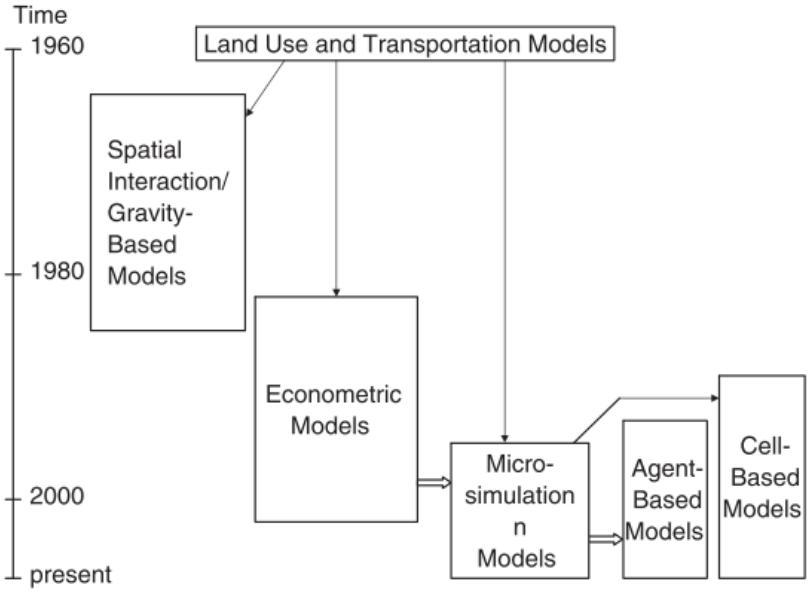

Figure 2.2: Chronological development of LUT models summarized by Iacono[12].

The modeling paradigm has changed fundamentally in the early 1990s along with the advances in computing power and efficiency of data storage. Urban systems used to be viewed as hierarchical and centrally organized equilibrium structures, or "top-down". Instead, now they were considered to be structured from the "bottom-up", dynamically retaining their integrity through interactions of numerous microelements[3]. A new broad class of LUT models that could fall under the title of "microsimulation" began to be developed. It included such classes of models as activity-based travel, cell-based models, and multi-agent models, and more recently comprehensive urban microsimulation models that fully reflect the dynamics of changes in the population and the urban environment[12].

"Micro" in the microsimulation implies that the model must be highly disaggregated spatially, socio-economically and in its representation of processes. "Simulation" implies that the model must be numerical, stochastic, have an explicit time dimension, and "evolve" into the end state rather than "solve for it"[21]. An example of such model has been developed by the University of Toronto ILUTE team; their product is an integrated urban model capable of microsimulating urban demographic evolution, housing markets and travel behaviour over extended periods of time[20]. The ILUTE system and some of the ways for its future improvement are discussed in the following section.

2.2.4 ILUTE and HoMES model systems

The Integrated Land Use, Transportation, Environment (ILUTE) model system is an agent-based microsimulation model for greater Toronto-Hamilton area; it includes such components as land use, activity/travel, urban economics, auto ownership, demographics and emissions/energy use. It uses disaggregate models of spatial socioeconomic processes to evolve the state of the greater Toronto–Hamilton area from a known base case to a predicted end state in 1-year time steps. The system state is defined in terms of the individual persons, households, dwelling units, firms, etc. that collectively define the urban region being modeled[22].

ILUTE model simulates the evolution of an urban region's spatial form, demographics, travel behavior and economic structure over time. Many markets are of interest within ILUTE, such as housing, labour, commercial, real estate, etc.) and are modeled via microsimulation. The model aims to capture the dynamics of these markets through disaggregated representations. For example, in the housing market component of ILUTE, houses are auctioned off one dwelling at a time to interested bidders in a disaggregate implementation of Martinez' Bid Choice theory[17].

The Housing Market Evolutionary System (HoMES) is the updated housing market module for the ILUTE model system. HoMES is a disaggregate, agent-based microsimulation of the owner-occupied housing market that evolves the residential location of households over time and includes the endogenous supply of housing by type and location, as well as the endogenous determination of sales prices and rents.

An overview of the framework of housing market supply, demand and clearing mechanisms utilized in HoMES provided by Rosenfield et al.[27] is presented on figure 2.3

Among the major barriers to implementation of integrated urban models since their introduction were such aspects as data hungriness and computational requirements[23]. However, continuing methodological advances, such as cost-effective High Performance Computing (HPC), detailed GIS-based datasets and machine learning methods, mean that former barriers now represent opportunities for model system development. In the case of ILUTE and HoMES, one of the possibilities for further improvement is the use of new data sources to update the housing market model. One of these new data sources, Teranet's dataset of land registry records, and main challenges of working with it are discussed in section 2.3.

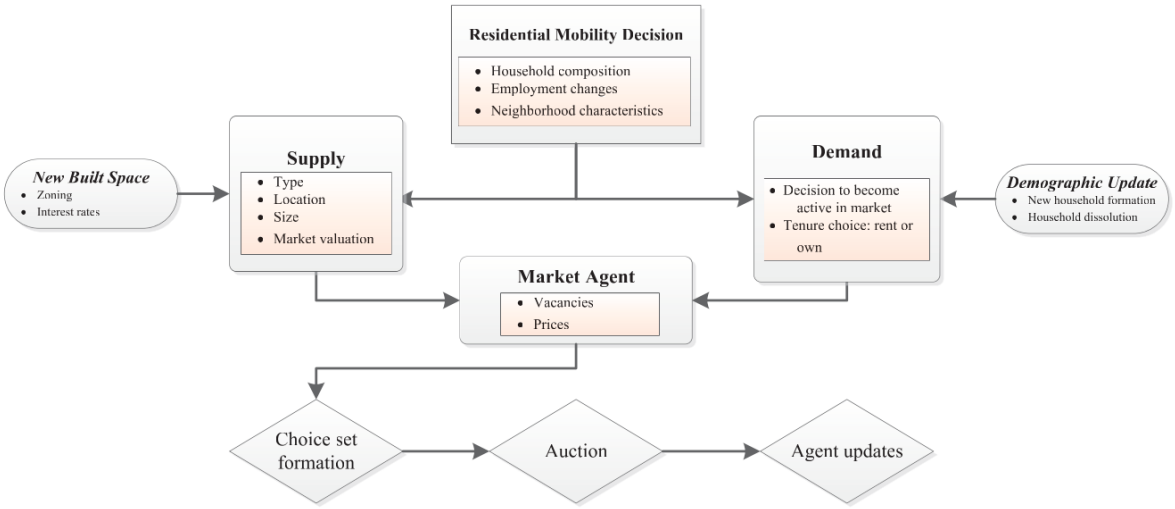

Figure 2.3: Framework of housing market supply, demand and clearing mechanisms used in HoMES module of ILUTE, as summarized by Rosenfield et al.[27].

2.3 New data sources and their challenges

2.3.1 Further improvement of integrated urban models

As an increasing amount of aspects of human life becomes digitalized, a wealth of new data is produced and can be used to model and analyze dynamics of urban systems[1, 7]. An example of such digitalization of human activity is the introduction of the Province of Ontario Land Registration Information System (POLARIS) in 1985 by the Government of Ontario[33] that will be discussed in the following section. Introduction of POLARIS lead to the creation of an extensive dataset of real estate transactions (land registration records) by the Teranet Enterprises Inc. This dataset offers a very fine resolution of housing market dynamics across both time and space, which can be beneficial for updating and testing microsimulation models, but it also presents challenges to work with that will be discussed in section 2.3.3 of this chapter; the chapter concludes with the proposed solution to address one of the main challenges of working with Teranet's data.

2.3.2 POLARIS: electronic system of land registration in Ontario

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[18].

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)[35] 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.

2.3.3 The Teranet-Ontario Partnership, Teranet's data and its challenges

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)[34].

Teranet dataset presents an extensive historical record of real estate transactions recorded in the Province of Ontario since the beginning of XIX century. However, when it comes to using new data sources in social studies, along with opportunities there are also challenges present. For example, these data sources can have issues with the quality of the data, might require a specific set of skills to take advantage of these data sources, or might not be suitable for traditional methods meant for traditional data[1], all of which are true in the case of Teranet's dataset.

One of the major attributes missing from the available version of Teranet's dataset is the information about the type of property being transacted, with records of various residential, commercial and industrial properties all being mixed together in the same dataset. At the same time, Teranet records have timestamps (dates) and location information (x and y coordinates) and thus can be joined to variety of other urban data sources, such as Census demographics, Transportation Tomorrow Survey (TTS) and parcel-level land use information. It is possible to derive this attribute from additional related sources of information, such as detailed land use or Census demographics. 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.

2.4 Chapter summary

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. 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. Since urban systems are complex in nature and require "re-solving" over and over, data science process models present a good fit for this task with their iterative structure.

Increased digitization of human activity, such as introduction of POLARIS land registration system by the Government of Ontario in 1985, produce a wealth of new information that can be used to study interaction between land use and transportation at a fine spatial and temporal scale. Teranet's dataset

of real estate transactions presents a wealth of information on the housing market of Ontario and can be used for empirical studies of transportation-land use interaction. However, along with the opportunities, the new data sources also present new challenges. Teranet's dataset has some data quality issues that need to be addressed and might require special skills to work with due to its size. But most importantly, it is very limited in the number of features available for each transaction.

One of the major attributes missing from the available version of Teranet's dataset is the information about the type of property being transacted, with records of various residential, commercial and industrial properties all being mixed together in the same dataset. At the same time, Teranet records have timestamps (dates) and location information (x and y coordinates) and thus can be joined to variety of other urban data sources, such as Census demographics, Transportation Tomorrow Survey (TTS) and parcel-level land use information. It is possible to derive this attribute from additional related sources of information, such as detailed land use or Census demographics. 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.

Chapter 3

Spatial and temporal relationships between GTHA data sources

3.1 Chapter 3: data sources used in the GTHA housing market database

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. 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. Changes to relative accessibility of a location can thus be estimated as changes in zone-to-zone travel times in a travel network[12].

The proposed GTHA housing market database combines urban data coming from a variety of sources. At the heart of it lies Teranet's dataset of real estate transactions (land registration records) recorded in the Province of Ontario since the beginning of XIX century up to October of 2017. As was discussed in section **??**, one of the main challenges of working with Teranet's data is the lack of available features. At the same time, Teranet records have timestamps (dates) and location information (x and y coordinates) and thus can be joined to variety of other urban data sources, such as Census demographics, Transportation Tomorrow Survey (TTS) and parcel-level land use information.

However, as will be discussed in this chapter, these data sources use different spatial units and are available at different temporal spans. Therefore, when joining data from these sources, special consideration needs to be taken with respect to their temporal and spatial relationships to ensure semantic interoperability. These spatial and temporal relationships are implemented via the data preparation workflow in Python and a PostgreSQL relational database, design and implementation of which is the primary focus of this master's thesis.

The nature of spatial and temporal relationships between the various data sources used in the proposed GTHA housing market database is discussed in this chapter, implementation of relationships is discussed in chapters 4 and **??** of this master's thesis.

3.2 Description of data sources

This section describes different data sources combined into the GTHA housing market database.

3.2.1 List of data sources that are currently used in the database

Currently, the following data sources are available in the GTHA housing market database:

1. Teranet's dataset
    - real estate transactions (land registration records) recorded in Ontario
2. Select variables, tables from Census of Canada
    - demographics
    - statistics
    - Dissemination Area (DA) geometry
3. Select variables, tables from Transportation Tomorrow Survey (TTS)
    - information about the transportation network
    - Transportation Analysis Zones (TAZ) geometry
4. DMTI Spatial
    - parcel-level land use by year (2001-2014)
    - enhanced points of interest (EPOI) by year (2001-2013)
    - postal geography
    - FSA geometry
5. land use information from University of Toronto's Department of Geography
    - detailed parcel-level land use information collected in 2011

Different data sources and their spatial and temporal relationships are discussed in the remainder of this chapter.

3.2.2 Teranet's dataset of land registration records

Teranet's dataset[32] of real estate transactions (land registration records) recorded in the Province of Ontario holds a wealth of information on the housing market of Ontario. Due to the introduction of POLARIS (discussed in section **??**) by the Province of Ontario in 1985, Teranet's dataset includes a complete population of real estate transactions recorded in Ontario from 1985 up to October of 2017 (records prior to 1985 appear to be incomplete, see chapter 6). Since Teranet's dataset has a high number of records, it can be used to investigate aspects relating to the housing market at a very fine spatial and temporal scale.

At the same time, Teranet's dataset is very limited on the amount of features available for each record. Since each record has a timestamp and is geocoded, Teranet's dataset can be augmented by

joining additional attributes from various data sources, such as Census or TTS survey, based on spatial and/or temporal relationships.

Characteristics of the raw Teranet dataset:

- 9,039,241 rows
- 15 columns

Characteristics of the Teranet dataset after data preparation (described in chapter 4):

- 5,188,513 rows
- 75 columns

Every record in Teranet's dataset is geocoded with a pair of coordinates (latitude and longitude). These coordinates are supposed to represent the centroid of the corresponding parcel, but might not always be that accurate.

3.2.3 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 data provide valuable insight into the latest economic, social and demographic conditions and trends in Canada and is 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"[16].

Census geography follows a certain hierarchy defined by Statistics Canada, with the largest top-level divisions being provinces and territories, lowest-tier divisions to which census data is disseminated are Dissemination Areas (DAs)[30]. Statistics Canada defines a dissemination area as a small area composed of one or more neighbouring dissemination blocks, roughly uniform in population size targeted from 400 to 700 persons to avoid data suppression[29].

3.2.4 Transportation Tomorrow Survey (TTS)

Another major source of information for most transportation planning studies concerned with Southern Ontario is the Transportation Tomorrow Survey (TTS)[10], an origin-destination travel survey. 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 and 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[2].

The finest level of spatial aggregation is that of the Traffic Zone also referred to as Traffic Analysis Zone (TAZ). The Traffic Zone is a polygon which typically falls along the centre line of roads or the natural geographic boundaries[11]. Not as a rule, but the TAZs roughly follow census tract boundaries, which are slightly bigger than 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), 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.

3.2.5 DMTI data

DMTI Spatial, a Digital Map Products company, is a major provider of location based information in Canada. DMTI's Enhanced Points of Interest (EPOI) is a vector GIS database of over 1 million business and recreational points of interest for all provinces/ territories of Canada. The attribute information contains multiple feature types and categories, information is available by year. Spatial units used by DMTI datasets used in the GTHA housing market database are points and parcel polygons, available by year.

3.2.6 Detailed land use information from geography department

The detailed land-use data collected by University of Toronto's Department of Geography is a combination of parcel boundaries (from Teranet) and manually coded land-use data using Google maps and streetviews, provided by Prof. Andre Sorensen and Prof. Paul Hess's research project. The spatial unit used by land use data provided by the Department of Geography is a parcel polygon.

3.3 Spatial relationships between datasets

This section introduces the spatial relationships between the datasets used in the GTHA housing market database.

3.3.1 Spatial units used by the data sources in the GTHA housing market database

The following spatial units are being used by different sources in the GTHA housing market database:

- Point data
  - Teranet
  - EPOI from DMTI
- 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

3.4 Temporal relationships between datasets

3.4.1 Temporal spans at which different urban data sources are available

Figure 3.1 presents temporal spans of the data sources used in the GTHA housing market database.

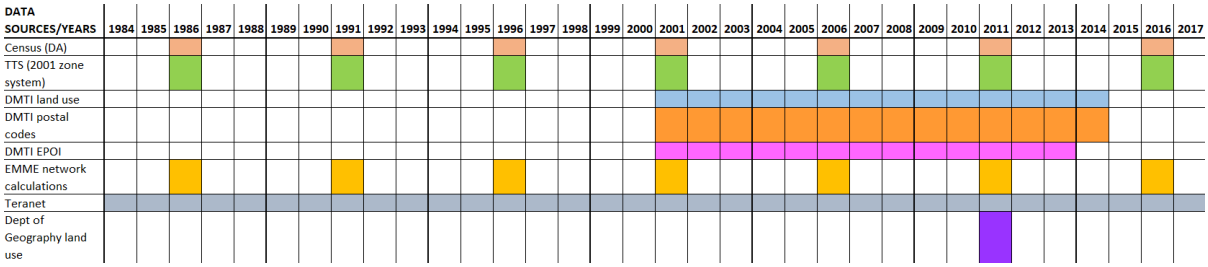

Figure 3.1: Temporal spans of the data sources used in the GTHA housing market database.

3.4.2 Possible approaches to matching temporal spans to relate data sources

Teranet and Census / TTS variables can be matched in a number of ways:

1. Direct match with appropriate Teranet subsets
   - use subsets of Teranet records from the Census / TTS years and match only with data directly recorded for that year
   - for example, take a subset of Teranet records from 2016 and match it with 2016 Census / TTS variables
   - technically, any date span can be specified when creating a Teranet subset, but in this case, appropriate Census / TTS variables must be selected manually
   - benefits:
     - precision of match: variables from Census would be composed of the actual values produced by the survey, rather than an interpolation based on assumptions
     - flexibility of use: a new Census table can be added and its variables can be match with appropriate Teranet records via simple SQL queries
   - disadvantages:
     - limited match: only Teranet data from Census years can be used, records from between Census years cannot be matched with the Census vairables
     - hard to generalize SQL queries: need custom SQL queries to match data to different tables, several queries to match Teranet data from different Census years

2. Interpolation of discrete Census / TTS variables
    - discrete Census / TTS variables can be turned into continuous via interpolation
    - Teranet records can be matched to real recorded and interpolated values by year, or finer time scale
    - benefits:
        - most Teranet records used: all Teranet records within the Census / TTS range can be used (within the interpolation region)
        - closest match: closest temporal match between Teranet records and Census / TTS variables
        - precise, if correctly assessed: in the case where correct assumptions are made while interpolating values, the most precise match
    - disadvantages:
        - more assumptions: additional assumptions need to be made about the dynamics of each Census / TTS variables between Census years
        - inaccurate, if incorrectly assessed: in case of incorrect assumptions, there is a risk of lower accuracy compared with other matching methods
        - interpolated rather then recorded: Teranet values from non-Census years will be matched to variables that are interpolated rather then recorded
        - more data pre-processing needed: each Census / TTS variable needs to be processed in order to produce interpolated values
3. Assign temporal spans for each Census / TTS survey as new features to Teranet records
    - each Census / TTS survey can be assigned a temporal span of 5 years representing a group of Teranet records to which its variables can be matched
    - Teranet records are matched by year, each year would yield an appropriate Census or TTS variable from an appropriate temporal span
    - for example, the Census of 2016 would have a temporal span of 2014–2018, and thus a Teranet record from 2015 would be matched to variables from 2016 Census.
    - Census of 1991 would have a temporal span of 1989–1993, and thus a Teranet record from 1993 would be matched with variables from 1991 Census.
    - TTS survey would be matched in a similar manner
    - benefits:
        - most Teranet records used: all Teranet / TTS records that fall within the specified temporal spans range can be matched to appropriate Census / TTS variables
        - recorded rather than interpolated: Teranet records are matched to actual recorded Census / TTS values
        - avoid interpolation assumptions: since no interpolation is performed, no additional assumptions are needed
        - no additional data pre-processing: all matching is done through an "adapter" table, original Census / TTS variables do not need to be changed

- disadvantages:
  - step-change in Census / TTS variables: when matching Teranet sources from non-Census years, instead of using interpolation, same Census / TTS variables are used for a group of 5 years centered at each Census year
  - varying accuracy: accuracy of match further away from the Census years (+/- 2 years) probably will be lower. In addition, there would be a step change from every +2 to -2 Census year (i.e., 1998 to 1999)
  - need new foreign keys: additional features specifying the temporal spans of Census / TTS variables to years for each Teranet record needs to be added to the Teranet table to facilitate temporal integrity of the joining operations
  - Census and TTS tables need to be in a "tidy" data format, with each variable being a column, and each observation being a row. Year of Census and TTS needs to be encoded as a separate variable 'year', 'year' and 'dauid' or 'taz_o' becoming the primary keys of TTS and Census tables. In a case of a large number of variables such transformation makes Census table "long" instead of "wide".

3.4.3 Matching temporal scales to facilitate linking datasets

3.5 Chapter summary

Different data sources use different spatial and temporal scales, and that's what we are going to address with data prep and the database.

Chapter 4

Data preparation

4.1 Chapter 4: streamlined data preparation workflow in Python

The "wicked" nature of transportation-land use interaction introduced in chapter 2 dictates the need to iteratively "re-solve" transportation and land use planning problems instead of focusing on finding some single "optimal solution". This approach resembles the methodologies typically employed for data science projects, where the sequence of steps is iterated over, producing a more meaningful solution on each new iteration of the cycle, as defined by such process models as CRISP-DM[28]. Similarly, data preparation can be followed in a linear manner, but is very likely to be iterative in nature[6].

This was indeed the case during the preparation of all the input data for the GTHA housing market database, especially in the case of Teranet's dataset. Some clean up steps required decisions on the criteria to filter out outliers and duplicates, which can result in removal of millions of records from the dataset; addition of new data sources requires modification of preprocessing steps to introduce new foreign keys; new features need to be engineered as the needs and understanding of the data evolve.

Since these steps are likely to be revisited in the future, a streamlined workflow has been established using Python via a series of jupyter notebooks as a part of this master's thesis to make this process efficient, modular, reproducible, easy to inspect and to modify. This section describes the concept of "Tidy Data" and database normalization and lists the steps taken in preparing all the input data before it can be entered into the housing market database.

This section describes the main requirements for structure of the data to be entered into the database, introduction of spatial and temporal relationships to allow joining different sources together, and outlines the data preparation workflow for Teranet's dataset

4.2 Tidy data and database normalization

Data preparation plays a critical role in research projects:

- it can determine the success of applications of machine learning algorithms
- it is a prerequisite for any meaningful analysis
- it is often required to allow the introduction of constraints necessary for implementation of an RDBMS .

This section describes the concept of "Tidy Data", as defined by Hadley Wickham. The concept of "Tidy Data" presents the basic ideas of normalization of a database, as defined by Edgar F. Codd, reformulated in statistical language.

4.2.1 Tidy data

Hadley Wickham in his paper "Tidy Data"[36] formalized the way how a shape of the data can be described and what goal should be pursued when formatting data. The principles of tidy data provide a standard way to organize data values within a dataset. The tidy data standard has been designed to facilitate initial exploration and analysis of the data, and to simplify the development of data analysis tools that work well together. The principles of tidy data are closely tied to those of relational databases and Codd's relational algebra[8].

4.2.2 Normalization of a database according to Codd

As an integral part of his relational model, Codd proposed a process of database normalization, or restructuring of a relational database in accordance with a series of so-called normal forms in order to reduce data redundancy and improve data integrity. Normalization entails organizing the columns (attributes) and tables (relations) of a database to ensure that their dependencies are properly enforced by database integrity constraints. As defined by Codd ([8], section 17.5.1), the basic ideas in normalization are to organize the information in a database as follows:

1. Each distinct type of object has a distinct type identifier, which becomes the name of a base relation.
2. Every distinct object of a given type must have an instance identifier that is unique within the object type; this is called its primary-key value.
3. Every fact in the database is a fact about the object identified by the primary key.
4. Each such fact contains nothing other than the single-valued immediate properties of the object.
5. Such facts are collected together in a single relation, if they are about objects of the same type. The result is a collection of facts, all of the same type.

4.2.3 Teranet's dataset in the context of a normalized database

Teranet's dataset is intended to be used as one of the tables (relations) of the proposed GTHA housing market database that would include other sources of information, such as DA-level demographics, land use information, etc. In this context, Codd's basic normalization ideas will take the following form:

1. Each distinct type of object has a distinct type identifier, which becomes the name of a base relation.
    - The Teranet dataset (filtered to include only GTHA records) presents a single type of object (relation, or table) —real estate transactions recorded in the GTHA between 1805-01-06 and 2017-10-11.
    - This condition is met.

2. Every distinct object (transaction) must have an instance identifier that is unique within the object type, or its primary-key value.

    - In case of Teranet dataset, all the native columns, including registration date, consideration amount, pin, address information, and x and y coordinates, have duplicated values present.
    - These do not necessarily represent duplicated records, as in the case with multiple transactions occurring for the same price, on the same date, at the same address, coordinates, or under the same pin.
    - Thus, no combination of Teranet columns constitutes a candidate key (unique identifier to be used in RDBMS).
    - As a matter of fact, even all the native columns together do not identify records uniquely.
    - For example, this is the case when two same price condo units are being sold on the same day in the same apartment block with no unit number specified for each transaction; the pin can be the same because transactions for some apartment blocks are recorded under a single pin with no regard to units.
    - This aspect also complicates the detection of duplicate records using native Teranet columns.
    - To address the issue of unique identifier (but not the issue of duplicate detection), a surrogate key (artificial unique identifier for RDBMS) is added to the Teranet dataset via a new attribute transaction_id.
    - It corresponds to the row number of each instance (transaction) in the Teranet dataset (filtered to include only GTHA transactions via a spatial join in Step 2.1, see section 4.3 of this document), ordered from the earliest date to the latest.

3. Every fact in the database is a fact about the object identified by the primary key.

    - This condition is mostly met, as every transaction in Teranet dataset is described by the values found in columns of a single row.
    - However, there are some records in which multiple attributes are recorded into a single column (described below).

4. Each such fact contains nothing other than the single-valued immediate properties of the object, all columns in Teranet dataset contain single-valued immediate properties of each transaction.

    - For some records, unit number, street number, street designation, street direction, or postal code are recorded as a part of street name.
    - This issue is addressed by parsing the other attributes from street name as a part of the data preparation process.

5. Such facts are collected together in a single relation, as they are all objects of the same type (a single table of real estate transactions recorded in Ontario).

Thus, Teranet dataset fits into a normalized database, with the new attribute `transaction_id` as its primary key.

4.2.4 Codd's constraints formed in statistical language by Wickham

According to Wickham[36], *tidy data* is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types.

In tidy data:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

This is Codd's 3rd normal form[8], but with the constraints framed in statistical language, and the focus put on a single dataset rather than the many connected datasets common in relational databases. *Messy data* is any other arrangement of the data.

According to Wickham, the most common problems with messy datasets are:

- Column headers are values, not variable names.
- Multiple variables are stored in one column.
- Variables are stored in both rows and columns.
- Multiple types of observational units are stored in the same table.
- A single observational unit is stored in multiple tables.

In the Teranet dataset, none of these problems are present, so it presents *tidy data*.

4.3 Step 2.1: Spatial join of Teranet points with Dissemination Area polygons

Step 2.1 of the cleaning process of Teranet data involved the spatial join of Teranet points with the polygons of Dissemination Areas (DA). Parameters that were used for the spatial join operation were `how='inner', op='within'`.

The spatial join was performed to filter out Teranet records whose coordinates fall outside of GTHA
.

In addition to that, three new attributes were produced as a results of the spatial join:

- Dissemination Area attributes `OBJECTID`, `DAUID`, and `CSDNAME` were added to each Teranet record falling within a particular DA polygon.
- The added attributes allow for extra quality control of Teranet data by comparing the column `MUNICIPALITY` of Teranet records with the column `CSDNAME` associated with DA geometry.
- New columns `OBJECTID` and `DAUID` allow Teranet records to be joined in the future with DA geometry via a regular (non-spatial) join operation (for example, to be aggregated by DAs), or to add any additional DA-level attributes, such as DA-level demographics, to Teranet records.

- These future joins can be performed via regular (non-spatial) join operations, which are much less computationally intensive than a spatial join, and thus can be performed much faster.

4.4 Introducing spatial relationships to the GTHA housing market database

4.5 Introducing temporal relationships to the GTHA housing market database

4.6 Chapter summary

All the requirements for Tidy Data and RDBMS constraints have been met by prepping the data.

Chapter 5

A prototype of a machine learning workflow to classify land use from housing market dynamics

5.1 Intro: databases

Something about Knowledge Discovery in Databases

5.2 Requirements to the information system for housing market data

On one hand, for such information-handling system to be comprehensive, it needs to combine a wide range of data sources describing these systems while maintaining semantic interoperability between these sources. In the case of land use, transportation, demographic, and real estate data, it means to take into account the varying spatial and temporal scale and resolution between these data sources when joining them together. At the same time, the system needs to be easily accessible to a wide range of researchers and students; it also needs to have powerful data processing capacity, to allow working with and performing calculations on large datasets related to real estate and land use. In addition to that, the system should have a modular structure, have a workflow that is reproducible and modifiable, so that new data sources and new relationships can be added to the system, while maintaining the existing part intact.

All of the requirements listed above present a strong case for the housing market information system to be implemented in a form of a relational database. Given the size of the datasets and the current research needs, PostgreSQL presents a good option for the database management system that fits all the discussed criteria. The focus of this master thesis is the organization and implementation of the GTHA housing market database to facilitate future research activities focused on the Longitudinal Analysis of housing sales in the Greater Toronto-Hamilton Area conducted by the University of Toronto Transportation Research Institute (UTTRI).

5.3 Philosophy of the housing market database

attribute-based database defined by spatial relationships

5.4 Entity relationship diagrams of GTHA housing market database

5.5 Attribute and referential integrity constraints

5.5.1 Primary keys used in the database

5.5.2 Referential integrity constraints used in the database

5.6 Chapter summary

All the requirements have been met via appropriate constraints.

Chapter 6

Results of the Exploratory Data Analysis (EDA)

6.1 Intro: Exploratory Data Analysis (EDA) of Teranet dataset

something about the philosophy of EDA, something about ESDA what has been done with Teranet data

6.2 Results of EDA of Teranet dataset

6.3 Results of ESDA of Teranet dataset

6.4 Chapter summary

EDA and ESDA explore some basic characteristics and establish some code to look at this data.

Chapter 7

Conculsion

a lot of different urban data has been matched across time and space into a database to facilitate future efforts in researching transportation-land use interaction

Bibliography

[1] Daniel Arribas-Bel. Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography*, 49:45–53, 2014.

[2] Bess Ashby. TTS 2016 City of Toronto Summary by Ward. Technical report, malatest, Toronto, 2018.

[3] Michael Batty. Cities as Complex Systems: Scaling, Interactions, Networks, Dynamics and Urban Morphologies. 2008.

[4] Luís M.A. Bettencourt. The origins of scaling in cities. *Science*, 340(6139):1438–1441, 2013.

[5] Richard J Bouchard and Clyde E Pyers. Use of gravity model for describing urban travel: An ayalysis and critique. *Highway Research Record*, (88):1–43, 1965.

[6] Jason Brownlee. How to Prepare Data For Machine Learning, 2013.

[7] Cynthia Chen, Jingtao Ma, Yusak Susilo, Yu Liu, and Menglin Wang. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*, 68:285–299, 2016.

[8] Edgar F Codd. *The relational model for database management : version 2*. Addison-Wesley Longman Publishing Co., Boston, MA, 1990.

[9] Kate Crowley and Brian W. Head. The enduring challenge of 'wicked problems': revisiting Rittel and Webber. *Policy Sciences*, 50(4):539–547, 2017.

[10] Data Management Group. Data Management Group at the University of Toronto Transportation Research Institute, 2014.

[11] Data Management Group. Survey Boundary Files, 2019.

[12] Michael Iacono, David Levinson, and Ahmed El-Geneidy. Models of transportation and land use change: A guide to the territory. *Journal of Planning Literature*, 2008.

[13] Jane Jacobs. The Death and Life of Great American Cities. In *New York*, volume 71, pages Alexander, C., Ishikawa, S., & Silverstein, M. (19. 1961.

[14] Eric Damian Kelly. The Transportation Land Use Link. *Journal of Planning Literature*, 9(2):p.128–145, 1994.

[15] Ira S Lowry. A Model of Metropolis. Technical report, Rand Corporation, Santa Monica CA, 1964.

[16] Map and Data Library. Canadian census geography (unit) definitions, 2019.

[17] F J Martinez. The bid-choice land-use model: an integrated economic framework. *Environment and Planning*, 24(January 1991):871–885, 1992.

[18] Heather McKean and Stella Di Cresce. The International Comparative Legal Guide to: Real Estate 2015, Chapter 6: Canada. Technical report, Global Legal Group, London, UK, 2015.

[19] Eric J Miller. Integrated urban modeling: Past, present, and future. *Journal of Transport and Land Use*, 11(1):387–399, 2018.

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

[21] Eric J Miller. Towards the Next Generation of Integrated Urban Models, 2018.

[22] Eric J Miller, Bilal Farooq, Franco Chingcuanco, and David Wang. Historical validation of integrated transport-land use model system. *Transportation Research Record*, (2255):91–99, 2011.

[23] Eric J Miller, David S Kriger, and John Douglas Hunt. *Integrated Urban Models for Simulation of Transit and Land Use Policies Guidelines for Implementation and Use*. 1998.

[24] Rolf Moeckel. Constraints in household relocation: Modeling land-use/transport interactions that respect time and monetary budgets. *Journal of Transport and Land Use*, 10(1):211–228, 2017.

[25] Guido Noto, Federico Cosenz, and Carmine Bianchi. *Urban Transportation Governance And Wicked Problems: A Systemic And Performance Oriented Approach*. Phd thesis, University of Palermo, 2015.

[26] Horst W.J. Rittel and Melvin M. Webber. Dilemmas in a General Theory of Planning. *Policy Sciences*, 4(2):155–169, 1973.

[27] Adam Rosenfield, Franco Chingcuanco, and Eric J Miller. Agent-based housing market microsimulation for integrated land use, transportation, environment model system. *Procedia Computer Science*, 19:841–846, 2013.

[28] Colin Shearer. J OURNAL Statement of Purpose E-Business and the New Demands on Data E-Commerce Places on Data Warehousing Technology WAREHOUSING. *Journal of Data Warehousing*, 5(4):13–22, 2000.

[29] Statistics Canada. Dissemination area (DA), 2015.

[30] Statistics Canada. Hierarchy of standard geographic units, 2018.

[31] Vergil G Stover and Frank J Koepke. *Transportation and Land Development*. Pearson College Div, 1988.

[32] Teranet Enterprises Inc. Product Description: Ownership Property Report. Technical report, Teranet Enterprises Inc., Toronto, 2011.

[33] Teranet Enterprises Inc. About POLARIS, Teranet, 2019.

[34] Teranet Enterprises Inc. https://www.teranet.ca, 2019.

[35] The Government of Ontario. Land Registration Reform Act, R.S.O. 1990, c. L.4, 1990.

[36] Hadley Wickham. Tidy Data. *Journal of Statistical Software*, 59(10), 2014.