Nearest Neighbor Regression for Land Value Estimation in Edmonton

COMP 657 TME 3 Research Project and Report

Jamie Czerwinski

Research Problems

Evidence and economic theory suggest that a Land Value Tax (LVTs) is an efficient, fair, and effective tool for governments to raise revenue, manage the cost of real estate, incentivize development, maximize land use, and ensure that the economic value generated by civic activity flows to public -- and not private -- interests. In jurisdictions where property taxes are levied on the combined value of land and improvements, tax assessments are generally made only for the combined value of land and improvements, and not for land and improvements separately; any shift towards LVTs would require these jurisdictions to develop assessment methodologies for both the land and improvement components of real property.

Research Methods

The City of Edmonton assesses taxes on the combined value of land and improvements. It also publishes data on:

1. The assessed value of every parcel of land in the City of Edmonton
2. A set of parcels of land in the City of Edmonton that happen to be vacant parcels

The intersection of these data sets comprise a set of the land values of a selection of parcels within the City of Edmonton. If the distribution of vacant parcels provides sufficient coverage over the city, assessed property values are accurate, and vacant parcels are accurately identified, then this data can be used as a training set for machine learning algorithms that estimate the land value of every parcel of land in the city.

Research tools, environments, and data

Source Code

Source code and instructions for this project can be found at https://github.com/jczerwinski/yeg-land-value-estimation.

Data

The data for this project was sourced from the City of Edmonton's Open Data Portal. The specific data sets used were:

1. Property Assessment Data (Current Calendar Year)
2. Property Information Data (Current Calendar Year)
3. Vacant Land Inventory

The Property Assessment Data was leveraged for its geographic coordinate and assessed value attributes, the Property Information Data was leveraged for its lot size attribute, and the Vacant Land Inventory data was leveraged for its lot size and vacant indicator attributes.

The Vacant Land Inventory data was originally generated in 2014, and last updated in 2015, while the Assessment and Information Data sets were current as of April 2021. This means that the vacancy status attribute was seven years out of date at the time of this analysis and likely included some inaccuracies as a result.

Environment and Tools

This analysis was implemented in Python -- IPython to be specific. Pipenv was used to install third-party library dependencies and manage the python environment. The following third-party dependencies were used:

- Sodapy: HTTP API client library, used to access City of Edmonton Open Data assets
- Scikit-learn: Machine learning library. Used to implement the nearest neighbors regression algorithm.
- Numpy: Used for data preprocessing and manipulation
- Pandas: Used for data preprocessing and manipulation

Algorithms and programming details

This study comprises two files:

1. getData.py - Downloads, joins, and saves a local copy of the City of Edmonton data sets. This script ensures that the data need not be downloaded multiple times, saving time, bandwidth, and data quota.

2. model.py - Preprocesses the data in preparation for model training, testing, and analysis, and fits and evaluates the model.

Data was preprocessed to the following format:

- Features:

    - Latitude: the latitude of the vacant parcel
    - Longitude: the longitude of the vacant parcel
- Target:
    - Price per square metre: the unit price of a square metre of land

Only data for vacant parcels was used.

Scikit-learn's radius-based nearest-neighbor regression model was used to model the function from features to targets. It was trained with the following parameters:

- Euclidean distance based on the degrees of latitude and longitude separation between parcels of land.
- An infinite neighborhood radius -- all examples are used when making predictions
- Inverse-distance example weighting for prediction

Results

Five-fold cross-validation testing was performed. The model was scored on 13 metrics for each cross-validation fold. Results are reported in the following table:

Cross-Validation Fold	**1**	**2**	**3**	**4**	**5**
Fit Time	0.003	0.003	0.001	0.001	0.001
Score Time	0.07	0.04	0.02	0.02	0.02
Explained Variance	-0.25	0.001	-0.87	0.02	0.11
Max Error	8668	356599	14436	110327	36154
Mean Absolute Error	1067	1824	859	1256	719
Mean Squared Error	1.66e6	4.38e8	1.80e6	4.90e7	9.57e6
Root Mean Squared Error	1287	20917	1341	7000	3093
Mean Squared Log Error	3.38	2.88	2.69	1.39	0.65
Median Absolute Error	977	548	784	498	16
R-squared	-1.44	-0.0003	-1.32	0.01	0.08
Mean Poisson Deviance	1206	12234	990	4306	1953

Mean Gamma Deviance	1.60	3.34	1.38	1.66	0.86
Mean Absolute Percentage Error	20	17	12	5	2

These results are mixed. On one hand, metrics that estimate the proportion of change in the target variable explained by the model -- such as R-squared and Explained Variance -- have relatively low scores, maxing out at 0.08 and 0.11, respectively, indicating that the model predicts a relatively small proportion of change in the target. On the other hand, the model performs reasonably well on the Mean Absolute Percentage Error, with a worst-case score of 20, indicating that for most test-train splits and parcels, the estimated prediction is off by less than 20%.

These results suggest that this method has promise, but that outliers may be significantly impairing the performance of the model.

Future Opportunities

Given that it appears that outliers may be impairing the performance of this model, automated outlier exclusion or a manual audit of the assessed values of vacant lots may help remedy this impairment.

Standard geographic coordinates (latitude and longitude) were used as features, and standard euclidean distance was used to compute the distance between examples. This algorithm does not accurately calculate distances between geographic points. Utilizing the haversine formula -- which accurately calculates distances between geographic points -- could potentially improve the accuracy of this model.

Factors other than local geography may have a significant influence on land values. For example, land use zoning influences what a given parcel of land may be used for, which may influence its market value. Models that incorporate zoning and other information may improve land value assessment predictions.

Other data sources may also be useful in the development of land value models. The Multiple Listing Service maintains sale and listing data for real estate, and the Alberta Land Titles Registry maintains transaction history and sale price data for real estate. Models that incorporate this data could improve upon the ideas explored in this study.

This study uses point-in-time data sets, but tax assessment data in particular is available on an annual basis. This time series data may be useful for improving the accuracy of the model.