Lyle Prockop

Lp2974

Master’s Thesis

February 1, 2023

Assignment 1: Data Source and Description

1. **Introduction and Research Question**

In their 2019 paper “The spatially conscious machine learning model”, Timothy Kiely and Nathaniel Bastian show that including spatial features in machine learning models generates more accurate predictions than non-spatial models for New York City real estate transactions from 2003 to 2017. My research seeks to understand whether spatial features increased or decreased in overall feature importance for more recent New York City real estate transactions (April 2020 through December 2022) given the volatile real estate market since the onset of the Covid-19 pandemic.

1. **Data Source**

The methodology will closely follow Kiely and Bastian’s (2019) and the same data sources will be used. The datasets will be:

1. *NYC Rolling Sales Data*[[1]](#footnote-1), made available for public download by the City of New York. This is a list of all real estate transactions with the date and sale price and includes records of borough, neighborhood, zip code, building type, square footage, tax class, and more. This will provide data on the occurrence of sales and price o­f sales, which are the classification and regression tasks, respectively, for the machine learning models. This data has been published annually since 2003 by the Department of Finance.
2. *Primary Land Use and Tax Lot Output (PLUTO)[[2]](#footnote-2),* made available for public download by the City of New York. This is a base file containing one record for each New York City tax lot (which is defined as a “parcel of real property”)[[3]](#footnote-3). There are over 70 fields in this data set that are categorized either as tax lot characteristics, building characteristics, or geographic/political/administrative districts. The data is updated monthly; it is aggregated by the Department of Planning and information is sourced from the Department of City Planning, Department of Finance, Department of Citywide Administrative Services, and Landmarks Preservation Committee. This data will be connected to the NYC Rolling Sales Data to provide geo-references and additional building features for the prediction models.
3. *NYC Property Address Directory (PAD)*[[4]](#footnote-4), made available for public download by the City of New York. PAD contains a full list of al­­­l BBL codes and their corresponding street addresses in the city. It will be used to map Borough-Block-Lot (BBL) codes from NYC Rolling Sales Data onto PLUTO data because NYC Rolling Sales Data only contains street addresses and not BBL codes. The data comes from the Department of City Planning and is published as part of NYC Open Data. It was made public in 2015 and its update frequency is officially listed as “As needed”. It was most recently updated on December 21, 2022.

Because these datasets are assumed to be exhaustive and accurate, selection and sampling should not be issues. Any sale registered in the city during the time period being studied is included in this dataset. The biggest potential issue is loss of data while geocoding because there are multiple steps to link the sales data to the georeferenced data.

1. **Data Description**

While data for the entire city of New York is available from the New York City government, only Manhattan data will be used for this research question. There are 56,960 recorded real estate transactions in Manhattan from January 1, 2020 through December 31, 2022. The concatenated data frames for 2020, 2021, and 2022 sales initially totaled 65,792 rows, but around 15% of them were empty.

The columns in this data table are: Borough, Neighborhood, Building class category, Tax class at present, Block, Lot, Easement, Building class at present, Address, Apartment number, Zip code, Residential units, Commercial units, Total units, Land square feet, Gross square feet, Year built, Tax class at time of sale, Building class at time of sale, Sale price, and Sale date. There will be additional features when the data is joined with PLUTO data. For the spatial models, features will be added based on different spatial weighting schemes to indicate proximity to other sales.

Fig. 1 shows descriptive statistics for the sales. The mean home sale value was $3,200,661, and the median was $915,000. This means that sale prices are skewed, as we would expect because there is a lower limit ($0) but no theoretical upper limit.

Next, Fig. 2 shows relative sale prices and sale dates, colored by zip code. While there are too many zip codes for a legend to be illustrative for this figure, it is helpful to know that places near each other have similar zip codes, and similar zip codes appear as similar colors. It seems that there are transaction clusters across space and time: there is more light green in January 2022 than any other time, and many of the higher-price sales are a similar muted dark blue. This suggests that transactions are affected by nearby transactions, which justifies exploring spatial prediction models to take nearby activity into account.

**Fig. 1: Descriptive statistics for Manhattan real estate transactions from January 1, 2020 through December 31, 2022.**

Table

Description automatically generated

**Fig. 2: Relative sale prices and sale dates.**

Chart, scatter chart

Description automatically generated

Further data cleaning and preprocessing is needed prior to fitting models, including:

1. Sub setting the data to start on April 1, 2020 rather than January 1, 2020
2. Joining this sales data with PAD data to obtain BBL codes, and then joining that set with PLUTO data to include further physical attributes of each property and georeferenced locations to build the spatial weights matrices
3. Adding spatial variables with differing spatial weighting schemes.

References

Kiely TJ, Bastian ND. The spatially conscious machine learning model. *Stat Anal Data Min: The ASA Data Sci Journal*. 2019;13:31-49. <https://doi.org/10/1002/sam.1440>

1. https://www.nyc.gov/site/finance/taxes/property-rolling-sales-data.page [↑](#footnote-ref-1)
2. https://nycplanning.github.io/db-pluto/#/ [↑](#footnote-ref-2)
3. https://nycplanning.github.io/db-pluto/#/ [↑](#footnote-ref-3)
4. https://data.cityofnewyork.us/City-Government/Property-Address-Directory/bc8t-ecyu [↑](#footnote-ref-4)