Data-driven analytics of New York City geospatial profiles for urban planning and building portfolio management

Zheng Yang

**Background**

Our world is rapidly urbanizing. Over 50% of world population now resides in cities with this number expected to increase to 67% by 2050. This rapid urban growth has begun to significantly increase the demands on urban systems and in turn is creating numerous challenges at the intersection of urban infrastructure, governance and the environment. Meanwhile, as the development of information technologies, massive amounts of data are being collected on an array of urban systems and provide opportunities to analyze and optimize urban planning and management. However, due to the lack of systematic studies to analyze the data, city officials struggled to understand, visualize and, most importantly, translate the massive data into policy and decision making for better urban development.

**Data and Objective**

PLUTO (Primary Land Use Tax Lot Output) data, liberated by Department of City Planning, Department of Finance, and Department of Information Technology and Telecommunications in New York City, provides detailed and fundamental information for every piece of land in the city and every single infrastructure on the land. The objective of this project is to conduct systematic data analytics to understand and make sense the geospatial profiles of 800,000+ lots and buildings, find their relationships and patterns at different scales for better urban planning and building portfolio management. For example, if the city government wants to build a recreation center, where should it be built to maximize its effects?

**Research Questions**

There are several interesting research questions that will be explored using the PLUTO dataset for the objective. More insights about the possibilities of PLUTO will be continuously added.

1. How much of land in New York is owned by public authority (e.g., city, state, or federal)? How is the land spatially connected to each other? How does the location of the land change over years?

2. What are the spatial distributions and temporal variations of commercial/residential areas in different boroughs? How does building height vary across the city? Why do some lots are far from public space?

3. How is the value of land and building related to urban context including school district, health center district, sanitation district, tax lot, etc.? How do the building characteristics (e.g., height, floor area, location, year built) impact the value of the land and building, and what are their predictive power?

4. How are the land and buildings in Manhattan different from other boroughs? How to cluster lots and buildings based on their geospatial profiles and characteristics?

5. How to measure (e.g. define walk score) the health and balance of the development and distribution of urban infrastructure in New York metropolitan area? It is of significance to support municipal urban planning and building management decision making.

6. Is there any mismatch of land use category and building class in New York City? What are representative characteristics of buildings in the special purpose and limited height districts? Can we use geospatial profile and building characteristic information to classify the land use category?

**Proposed Methods**

Different data analytics skills will be practiced in this project, such as data import, data cleaning and manipulation, data merging, exploratory analysis, statistical analysis, supervised machine learning (classification and regression), unsupervised machine learning, deep learning, network analysis, spatial analysis, parallel computing, data visualization, and web application.

**Additional Analysis**

There are more than 6000 open datasets for New York City, which can be coupled with PLUTO data for more domain-specific analysis. For example, we can merge the New York Taxi and Limousine Commission (TLC) datasets with the PLUTO dataset to further analyze how the geospatial profiles of land and buildings impact the pickup time, trip distance, and taxi fare amount of yellow taxi trip at different temporal and spatial scales in different boroughs of New York City. Another example is we can also merge the PLUTO data with New York Energy and Water (LL84) datasets, and explore what the relationships are between geospatial profiles of land/building and energy/water usage for targeted policy and energy efficiency program recommendations.