本文的主要目的是对于影响纽约市范围内Citi Bike使用率的因素进行建模，主要考察纽约市不同地区的社会经济情况对于共享单车使用的影响。在现有针对共享单车（尤其是纽约市的Citi Bike）系统的研究中，少有此类基于社会经济情况的研究，大部分研究集中于：（1）基于天气等个别影响因素对共享单车使用情况的建模，包括机器学习方法；（2）分析共享单车使用情况的空间特征，包括时间和空间滞后效应；（3）对共享单车与其他交通方式互补性的讨论。我们认为，这可能是由于在其他城市中，广泛而有效的社会经济数据难以获得；但在纽约市，我们可以利用诸多官方渠道获得各种包含地理信息的社会经济数据，使我们的研究得以进行。我们预计，除了已经广泛研究的天气因素和自行车站点密度的影响之外，包括人口密度、收入结构、公共交通可获得性、机动车保有率以及街区的通行能力等多种空间异质性的因素都会影响人们对于共享单车系统的使用倾向。通过空间统计的方法，利用这些因素进行建模，我们预计可以识别出不同因素对共享单车使用率的影响情况。同时，利用可视化手段和机器学习方法，我们可以得出一个用于预测未来纽约市Citi bike使用情况的系统，这个系统也可以基于其他城市的数据预测该地区潜在共享单车系统的使用情况。

The main objective of this paper is to model the factors that influence Citi Bike usage within the range New York City, to examine the impact of multiple socioeconomic factors on bike-sharing usage in different parts of the city. Among the existing studies on bike-sharing (especially Citi Bike in New York City) systems, there are few such socioeconomic-based studies, and most of them focus on (1) modeling bike-sharing usage based on individual influences such as weather, including using machine learning approaches; (2) analyzing spatial characteristics of bike-sharing usage, including time and space lag effects; and (3) discussion of the complementarity of bike-sharing with other transportation modes. We believe that this may be due to the fact that extensive and valid socioeconomic data are difficult to obtain in other cities; however, in New York City, we have access to a variety of socioeconomic data containing geographic information using numerous official sources, allowing our study to proceed. We expect that a variety of spatially heterogeneous factors, including population density, income structure, public transportation availability, motor vehicle ownership rates, and neighborhood accessibility, will influence people's propensity to use the bike-sharing system, in addition to the effects of weather factors and bicycle station density that have been extensively studied. Using a spatial statistical approach to model these factors, we expect to identify how these factors affect bike-sharing usage rates. Also, using visualization tools and machine learning methods, we can derive a system for predicting future Citi bike usage in New York City, which can also predict the usage of potential bike-sharing systems in other cities based on local data.

本文使用的主要数据为纽约市Citi Bike的使用情况数据，这部分数据于citi bike 的官方网站上开放下载。数据的时间范围为2013年至今，包含订单的起止时间、起止坐标、用户画像等数据。同时，我们也要获得Citi Bike的站点分布情况。此外，我们需获得纽约市的社会经济数据，包括但不限于：

人口密度

人口结构

人口收入结构

该区域内公共交通站点的数量/该区域与公共交通站点的平均距离

机动车保有率

停车指数（衡量该区域内公共停车的难度）

街区功能（商业、住宅等，可依托POI分析）

地形

这些数据通常有两个来源：（1）公共的地图提供商的公开API（例如谷歌地图）；（2）纽约市政府及下属机构的开放数据portal（例如NYC Open Data）。对于后者，一般而言获取的信息会精确到一定的行政级别，这要求我们在收集信息后对本研究地理空间上的划分做出权衡。

目前预计本文的研究将包括两个部分：数据的可视化，以及建模分析。

第一部分是对数据的初始处理以及可视化，主要是为了展示纽约市Citi Bike的使用情况特征和社会经济特征，以及为第二部分做准备。在这一部分中，我们需要首先对数据进行清洗，然后基于Citi Bike的数据选取（或生成）可以量化的指标（例如，频率、时长、距离或三者的加权平均等），以用来衡量其使用率。我们还可以利用一些可视化手段（例如聚类分析等），创造可以交互的人机界面来帮助读者更好地认识数据的地理特征。

第二部分将是对Citi Bike使用情况影响因素的建模。通过收集上文中列举的社会经济数据并对其进行清洗和标准化，我们将得以建立空间统计模型以研究不同因素的影响占比。同时，我们还必须考虑到citi Bike的使用中可能存在的空间溢出效应。因此，考虑建立一个包含空间滞后的回归模型是值得考虑的。具体模型的选取需要在对数据进行初步探索的基础上再做决定。此外，一个可以选择的选项是利用机器学习方法来进行这一部分的分析。鉴于目前我们对机器学习方法的了解仍不够深入以及相关研究有限，我们将在进一步学习并与导师进行讨论后确定时候使用此种方法。

在第二部分的最后，我们将论证我们的模型具有预测能力，包括（1）在未来纽约市某地的经济特征改变时，预测新的Citi Bike需求；以及（2）在尚未有公共自行车系统运行的美国城市，预测潜在的需求分布。我们将选取一个城市为案例进行实践。

此外，一个需要进一步考虑的因素是时间。到目前为止，我们所有模型都是基于静态的（某一时刻的）纽约市数据，这要求我们在分析自行车订单的时候也要使用静态数据（例如，某一段时间内订单量的总和，或者是以天为单位对某段时间内的订单数据进行平均）。而实际上，动态的订单也包含了诸多可以研究的内容，例如：每天中不同时间段使用情况的差别；工作日与非工作日使用情况的差别，以及COVID-19的影响等。如果要利用这一部分内容，就要求我们要进一步细化研究问题，或者是在前一个研究问题的基础上进行扩展（例如，研究通勤时间段内Citi Bike使用情况的影响因素）。这将是我们下一步（在开展初步研究的同时）需要探索的问题。

Data

The main data used in this research is the usage data of Citi Bike in New York City, which is available for download on the official Citi Bike website. The data covers the period from 2013 to the present, and includes order start and end times, start and end coordinates, and user profiles, etc. We also need to obtain the distribution of Citi Bike's stations.

In addition, we need to obtain socio-economic data for New York City, including but not limited to:

Population density

Population structure

Income structure

Number of public transportation stops in the area/average distance to public transportation stops in the area

Motor vehicle ownership rate

Parking index (measures the difficulty of street parking in the area)

Function (commercial, residential, etc., can be obtained through official data or POI analysis)

Topography

These data are usually available from two sources: (1) public APIs from public map providers (e.g., Google Maps); and (2) open data portals from the City of New York and affiliated agencies (e.g., NYC Open Data). For the latter, the information obtained will generally be generalize to a certain administrative level, which requires us to weigh the geospatial division in this study after collecting the information.

Structure

It is currently expected that the research in this paper will consist of two parts: visualization of the data, and modeling analysis.

The first part is the initial processing and visualization of the data, primarily to demonstrate the usage of Citi Bike and socioeconomic characteristics of New York City, as well as to prepare for the second part. In this part, we need to first clean the data and then select (or generate) quantifiable metrics (e.g., frequency, duration, distance, or a weighted average of the three, etc.) based on the Citi Bike data that can be used to measure its usage. We can also use some visualization tools (e.g., cluster analysis, etc.) and create human-machine interfaces that can be interacted with to help the reader better understand the geographic characteristics of the data.

The second part will be the modeling of socio-economic factors influencing Citi Bike usage. By collecting, cleaning and normalizing the socio-economic data listed above, we will be able to build a spatial statistical model to study the impact of different factors. Also, we must take into account the possible spatial spillover effects in Citi Bike usage. Therefore, it is worth considering a regression model that includes spatial lags. The selection of a specific model needs to be decided based on a preliminary exploration of the data. In addition, an optional option is to use machine learning methods for this part of the analysis. Given our current lack of understanding of machine learning methods and the limited relevant research, we will determine when to use such methods after further study and discussion with our mentor.

At the end of Part II, we will argue that our model has predictive capabilities, including (1) predicting new Citi Bike demand when socio-economic characteristics change somewhere in New York City in the future, and (2) predicting potential demand distribution in U.S. cities that do not yet have a public bike system in operation. We will select one city as a case study for practice.

In addition, a further factor to consider is time. So far, all of our models have been based on static (at a given moment in time) New York City socio-economic data, which requires that we also use static data when analyzing bike orders (e.g., the sum/average of orders over a certain period of tim). However, dynamic orders also contain numerous hidden information that can be studied, such as: the difference in usage between different times of the day; the difference in usage between workdays and weekends, and the impact of COVID-19. To utilize this component would require us to further refine the research question or expand it (e.g., examine the factors influencing Citi Bike usage during commuting time periods). This will be the next step we need to explore (while conducting the initial study) and may influence our outcome.