CASA0007 Quantitative Methods Written Investigation

基于聚类方法分析伦敦Middle layer Super Output Areas (MSOAs)之间的通勤模式

Analysis of commuting patterns between Middle layer Super Output Areas (MSOAs) in London based on clustering method.

Introduction

通勤模式反映了一个城市的经济地理信息 [@Ratti2010]，研究人们的通勤模式对于理解城市的经济发展至关重要 [@Martin2018]。人口普查为此提供了大量且复杂的通勤数据，这些数据以一种Origin-destination流的形式呈现，反映了地区之间的人员流动性。此外，人群之间的差异性在分析流动数据时也不应被忽略 [@Shen2019]。因此，本项研究将以伦敦地区人口普查中的通勤数据为基础，重点探索不同Middle layer Super Output Areas(MSOAs)间不同社会等级的人们的移动模式是否表现出聚类的现象。我们首先假设这种模式是完全空间随机的，这也表明人们通常的工作地点是随机分布的。在后面的分析中将检验假设的正确性。

Literature Review

Origin-destination (OD) flow数据可以看作在时间和空间上的分辨率更粗略的轨迹数据，它保留了真实轨迹起点和终点的地理位置信息，隐含了轨迹的方向和距离 [@Guo2014]。使用聚类对OD flow进行分析的思路分为基于点的聚类和基于线的聚类 [@Guo2020]。基于点的聚类从点的相似度考虑，以欧几里得距离等度量指标为基础，使用K-means [@Heredia2022], 层次聚类 [@Zhu2014]，Density-based clustering [@Pei2015]等方法寻找点簇。而基于线的聚类从线的相似度考虑，以Dynamic Time Warping（DTW）距离或几何特征等度量指标为基础，使用TR-OPTICS [@Shuliang2018], Fast-clusiVAT [@Kumar2018], TRACLUS [@Lee2007] 等方法寻找线簇。总的来说，基于点的聚类更易实现，且计算量对比基于线的聚类较低。对于更加复杂的Origin-destination (OD) flow数据或轨迹数据，使用基于线的聚类方法可以得到更加准确的聚类结果。

Data Presentation

本项研究使用2021年英国人口普查中的出发地-目的地数据 [@zotero-272]。具体来说，Origin-destination public data一共包含30个数据集，由main flows, univariate and multivariate datasets组成。这些数据集分为四种类型，我们选取了属于Origin-destination Workplace data这种类型的按approximated social grade划分的OD流数据集。此数据集一共有3760,466行数据，通过筛选选取了伦敦MSOAs的数据，具体有565,668行。数据集中包含几个主要的关键字段，如表1所示。四种社会等级下面使用AB, C1, C2, and DE表示。通过选取流量最多的前50条数据，绘制出OD矩阵，如图1所示，清楚的呈现了始发地到目的地之间的关系。不难发现矩阵在大部分单元格是没有数值的，这表明多数情况下OD矩阵是一个稀疏矩阵。此外，有数值的单元格在特定的纵坐标（目的地）上有聚集现象。进一步结合MSOAs的位置信息，分别选取不同社会等级中流量最多的100条数据，其OD flow的空间分布如图2所示。通过图2 可以明确发现不同社会等级的人们的工作地点有明显的聚集现象，不同的社会等级之间也存在不同的聚类现象，接下来将使用定量的聚类方法进行分析。

Methodology

由于K-means是常用的，且较为简单的聚类方法，可以快速给出具有一定合理性的聚类结果，所以首先使用K-means对不同社会等级人们的OD flow数据进行聚类分析。K-means的核心是将n个数据依据数据间的相似性聚集到指定的k个簇中，数据间的相似性使用欧氏距离进行计算。K-means方法受到K值的影响很大，因此选取合理的K值将决定聚类结果的准确性。我们采用四种确定K值的方法进行综合比较，这些方法分别是Elbow Method，Silhouette Score，Calinski-Harabasz Index，Davies-Bouldin Index。选择不同的K值分别计算这四种方法的结果，如图3所示。Elbow Method需要寻找数据变化的临界点，Silhouette Score和Calinski-Harabasz Index都是值越大越好，而Davies-Bouldin Index的值是越小越好。因此通过图3我们确定对于社会等级是AB的OD flow数据，K值为5；对于社会等级是C1的OD flow数据，K值为6；对于社会等级是C2的OD flow数据，K值为2；对于社会等级是DE的OD flow数据，K值为8。选取好合理的K值后，可以应用K-means方法进行聚类分析。

考虑到实际的OD flow数据集可以存在任意形状的簇，也会存在一些异常数据，此时使用K-means可能无法获得更加合适的聚类结果，因此可以使用基于密度的聚类方法。Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) 作为一种稳定的基于密度的聚类方法，是Density-based spatial clustering of applications with noise (DBSCAN)的改进方法。HDBSCAN首先使用mutual reachability distance衡量两个点的相似度，mutual reachability distance定义为，其中d(a,b)表示点a和点b的原始距离，core\_k(x)定义为当前点x到其第k近的点的距离, mutual reachability distance使得HDBSCAN可以处理密度不均匀的数据。之后，使用最小生成树构建点与点之间的层次树模型，引入层次聚类的思想。这样使得HDBSCAN只需设置簇的最小尺寸 (minPts)，即可自动获得最优的聚类结果。对于OD flow 数据集，经过多次尝试，确定了minPts均设置为4。

Results

K-means的聚类结果如图4所示。同时，我们计算出聚类结果中工作地点出现次数最高的地区，其结果如表2所示。结合图4和表2，可以发现社会等级是AB的人们的工作地点基本位于City of London 001，而居住地点也在离市中心较近的地区；社会等级是C1的人们的工作地点大部分在Hillingdon 031，少量的工作地点分散在城市不同位置；社会等级是C2的人们的工作地点大部分也在Hillingdon 031，其余的工作地点分散在城市的东北地区；社会等级是DE的人们的工作地点大部分在Hillingdon 031和Brent 027，其余的工作地点分散在城市不同位置。

HDBSCAN的聚类结果如图5所示，其结果更加细致。社会等级是AB的聚类结果分成了5个簇，更加细致划分了不同通勤距离和方向的OD flow；社会等级是C1的聚类结果分成了更多的簇，可以发现在Hillingdon 031周围存在更多较小的簇；社会等级是C2的聚类结果也分成了更多的簇，遍布在城市的西部、南部和东北部；社会等级是DE的聚类结果大致和K-means相同。

Discussion

将K-means和HDNSCAN的聚类结果与London house prices per square metre map [@Plumplot] 、London net household income map [@Plumplota] 进行对比，我们发现社会等级是AB的人们收入更高，也居住在房价更高的地方；社会等级是C1、C2、DE的人们也有部分收入较高，但大部分居住在房价更便宜的地方。其他研究表明，社会经济地位较高的人比社会经济地位较低的人可以承受更多的通勤成本，因此可以适应更长的旅行 [@Dargay2012]，但从聚类结果中也能发现，社会经济地位较低的人中仍有部分人可能处于房价的原因选择较长的通勤距离。

此外，根据伦敦地区人口统计地图 [@Ibbetson2020]，发现Hillingdon 031周围地区人口数量很多，同时东北部地区人口数量也较多，这和我们的聚类结果相一致。此外聚类结果也和工作机会高度相关，从伦敦MSOA就业人数地图 [@Hill2018] 也可以明确看出多数工作机会分布在Hillingdon 031周围，City of London 001还有东部地区。这些都可以说明伦敦逐渐涌现多个城市中心，在未来的城市规划中需要注意区域结构发生的变化。

Conclusions

本项研究将K-means和HDBSCAN聚类方法应用于伦敦Origin-destination (OD) flow数据，分析了不同社会等级人们在MSOAs之间的通勤模式，并根据房价和收入因素解释了产生聚类现象的原因，引出了城市结构正在发生改变。本项研究也存在一些不足，只考虑了社会等级这一个内在因素，而人们的通勤模式往往与多个因素都有联系，未来的工作可以利用更多影响通勤模式的因素，使用更加复杂的聚类方法深入探索通勤模式。

References