

Revealing spatiotemporal travel demand with taxi trip data: A case study of New York City

택시 이동 데이터를 활용한 시공간적 교통 수요 분석: 뉴욕시 사례 연구

지리학과 2020110218 황지연

초록

본 연구는 Xie C 외(2021)의 논문 「Revealing spatiotemporal travel demand and community structure characteristics with taxi trip data」을 바탕으로, GIS 와 공간분석 수업에서 학습한 공간통계 기법을 적용해 보는 프로젝트의 일환으로 진행되었다. 2025년 1월의 뉴욕시 택시 운행 데이터를 활용하여 시공간적 교통 수요의 분포와 패턴을 분석하는 것을 목적으로 하였으며, 이를 위해 평일과 주말의 주요 피크 시간 대를 선정하고, 각 시간대별 승하차 O/D 매트릭스를 구축하였다. 이후 택시존 단위 이동 데이터를 1 마일 격자 단위로 재할당 하기 위해 면적 가중 보간 (Areal Weighted Interpolation)을 수행하였으며, 커널 밀도 추정(KDE)을 통해 고수요 지역을 시각화 하였다. 공간적 자기상관 분석을 위해 Global Moran's I 및 Local Moran's I(LISA) 지표를 활용하였고, 이를 통해 출퇴근 및 여가 목적의 이동 흐름이 시기별로 상이하게 나타남을 확인하였다. 특히 맨해튼 중심부에서는 모든 시간대에서 높은 이동 수요가 관찰되었으며, 주말 저녁에는 문화 및 여가 공간을 중심으로 클러스터가 확산되는 양상이 나타났다. 본 연구는 공간 통계 기법을 실제 도시 교통 데이터에 적용하여 분석하고, 연구자의 배경지식을 바탕으로 분석 결과를 해석해보는 것으로서 의의가 있다.

주요어 : Global Moran's I, Local Moran's I, 면적 가중 보간(Areal Weighted Interpolation), 커널 밀도 추정 (Kernel Density Estimation, KDE), travel demand, transportation

1. 연구지역 및 데이터

1.1 연구지역

뉴욕시(New York City, NYC)는 미국 뉴욕주에 위치한 미국 내에서 가장 인구가 많은 도시로 맨해튼 (Manhattan), 퀸즈 (Queens), 브루클린 (Brooklyn), 브롱스 (Bronx), 스테튼 아일랜드 (Staten Island) 다섯 개의 자치구로 구성되어 있다(그림 3A). 현재 뉴욕시에는 세가지 유형의 택시가 운행되고 있다. 첫번째는 옐로우 택시(Yellow Taxi)로 뉴욕시 전역에서 승객을 태울 수 있으며, 두번째는 시내 중심 지역 외곽 지역에서의 택시 서비스 및 접근성을 개선하기 위해 2013년 8월에 도입된 그린 택시 (Green Taxi), 세번째는 우버 (Uber), 리프트 (Lyft)와 같은 라이드헤일링 (Ride-hailing) 서비스, 고급 차량 서비스 등과 같은 호출 차량 (FHV: For-Hire Vehicle)이다. 전체 도시 수준에서의 이동 패턴을 분석하기 위해서는 이 세가지 유형의 택시 데이터를 모두 고려해야한다. 본 연구는 이러한 세가지 유형의 택시 이동을 바탕으로 이동 수요의 시공간적 분포를 분석한다.

1.2 데이터 설명

본 연구에서 사용한 데이터는 뉴욕시 택시 및 리무진 위원회 (TLC: Taxi and Limousine Commission) 에서 공개한 택시 운행 기록 데이터로 뉴욕시 공공 데이터 플랫폼인 NYC Open Data에서 수집하였다. 해당

데이터는 옐로우 택시 (Yellow Taxi), 그린 택시 (Green Taxi), 호출 차량 (FHV: For-Hire Vehicle) 을 포함하며, 본 연구에서는 세가지 유형의 데이터를 모두 활용하였다. 단, 호출 차량의 경우 하루 1만건 이상의 운행이 기록되어 있는 HVFHV (High Volume For-Hire Vehicle) 의 데이터만 활용하였다.

데이터에서 하나의 행은 택시의 단일 이동(trip)을 나타내며 택시ID, 승하차 시각, 승하차 지점의 구역 ID, 승객 수, 이동거리, 결제 방식, 결제 금액 등의 정보를 포함하고 있다(그림 1). 본 연구에서는 2025년 1월 20일부터 2025년 1월 26일까지의 데이터를 사용하였다. 해당 기간의 원본 데이터에서 승객 수 열이 결측치인 행들은 제거하였으며, 최종적으로 5,512,100건의 택시 이동 데이터를 분석에 활용하였다.

VendorID	승객탑승시각	승객하차시각	승객 수	이동거리 (mi)	요금코드	지정전송 여부	승차지점 TLC택시존 ID	하차지점 TLC택시존 ID	결제 방식	기본요금(거리 시장기반)	추가 요금	MTA 세금	팁	유료통행료	시설개선할증료	총요금	뉴욕주운집통행료	공항요금 (JFK,LGA)	맨해튼운집구역 (CBD)운행료
1	2025-01-01 00:18:38	2025-01-01 00:26:59	1.0	1.60	1.0	N	229	237	1	10.00	3.5	0.5	3.00	0.0	1.0	18.00	2.5	0.0	0.00
1	2025-01-01 00:32:40	2025-01-01 00:35:13	1.0	0.50	1.0	N	236	237	1	5.10	3.5	0.5	2.02	0.0	1.0	12.12	2.5	0.0	0.00
1	2025-01-01 00:44:04	2025-01-01 00:46:01	1.0	0.60	1.0	N	141	141	1	5.10	3.5	0.5	2.00	0.0	1.0	12.10	2.5	0.0	0.00
2	2025-01-01 00:14:27	2025-01-01 00:20:01	3.0	0.52	1.0	N	244	244	2	7.20	1.0	0.5	0.00	0.0	1.0	9.70	0.0	0.0	0.00
2	2025-01-01 00:21:34	2025-01-01 00:25:06	3.0	0.66	1.0	N	244	116	2	5.80	1.0	0.5	0.00	0.0	1.0	8.30	0.0	0.0	0.00
...
2	2025-01-31 23:01:48	2025-01-31 23:16:29	NaN	3.35	NaN	None	79	237	0	15.85	0.0	0.5	0.00	0.0	1.0	20.60	NaN	NaN	0.75
2	2025-01-31 23:01:29	2025-02-01 00:17:27	NaN	8.73	NaN	None	161	116	0	28.14	0.0	0.5	0.00	0.0	1.0	32.89	NaN	NaN	0.75
2	2025-01-31 23:26:59	2025-01-31 23:43:01	NaN	2.64	NaN	None	144	246	0	14.91	0.0	0.5	0.00	0.0	1.0	19.66	NaN	NaN	0.75
2	2025-01-31 23:14:34	2025-01-31 23:34:52	NaN	3.16	NaN	None	142	107	0	17.55	0.0	0.5	0.00	0.0	1.0	22.30	NaN	NaN	0.75
2	2025-01-31 23:56:42	2025-02-01 00:07:27	NaN	2.29	NaN	None	237	238	0	12.09	0.0	0.5	0.00	0.0	1.0	16.09	NaN	NaN	0.00

그림 1. 택시 운행 데이터셋

1.3 분석 단위 설정

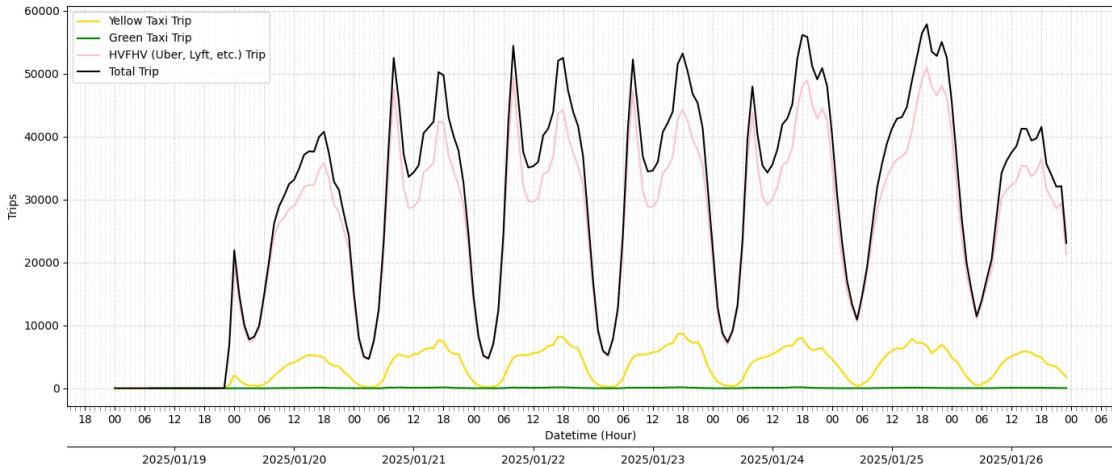


그림 2. 택시 이동 시계열 분석 결과

시간의 흐름에 따른 이동 패턴을 파악하기 위해 1시간 단위로 시계열 분석을 진행하였다(그림 2). 시계열 분석 결과, 평일 오전에는 오전 7시부터 이동량이 급격히 증가하여 오전 10시까지 지속되는 패턴을 보이며 평일 오후의 피크 시간대는 오후 4시부터 오후 7시까지이다. 이러한 특징은 공휴일인 2025년 1월 25일 월요일 (Martin Luther King Jr. Day)을 제외하고 화요일부터 금요일까지 유사한 패턴으로 나타난다. 그러나 주말에는 시간의 흐름에 따른 이동 패턴이 토요일과 일요일에 서로 다르게 나타난다. 토요일의 이동 피크 시간대는 오후 6시부터 오후 8시, 오후 9시부터 오후 11시이다. 반면에 일요일은 오후 1시부터 4시, 오후 5시부터 6시 사이에 이동 피크가 나타난다. 또한 일요일은 토요일보다 택시 이동량이 적다. 따라서 본 연구는 대표성 있는 화요일과 토요일의 데이터를 선택하였으며 시계열 분석을 통해 파악한 평일 2개의 피크 시간대와 주말 2개의 피크시간대를 분석 범위로 설정하였다(표 1).

표 1. 연구의 시간적 범위

날짜	기간	
2025년 1월 20일 화요일	피크 1	오전 7시~오전 10시
	피크 2	오후 4시~오후7시
2025년 1월 25일 토요일	피크 3	오후 6시~오후8시
	피크 4	오후 9시~오후 11시

본 연구의 공간분석 단위는 1마일 격자로 진행하였다(그림 3B). 연구에 활용한 택시 이동 데이터의 75%는 이동거리가 1마일 이상이며, 이는 1마일 격자 단위가 서로 다른 셀 간의 택시 이동에 대한 정보를 보존할 수 있음을 의미한다.

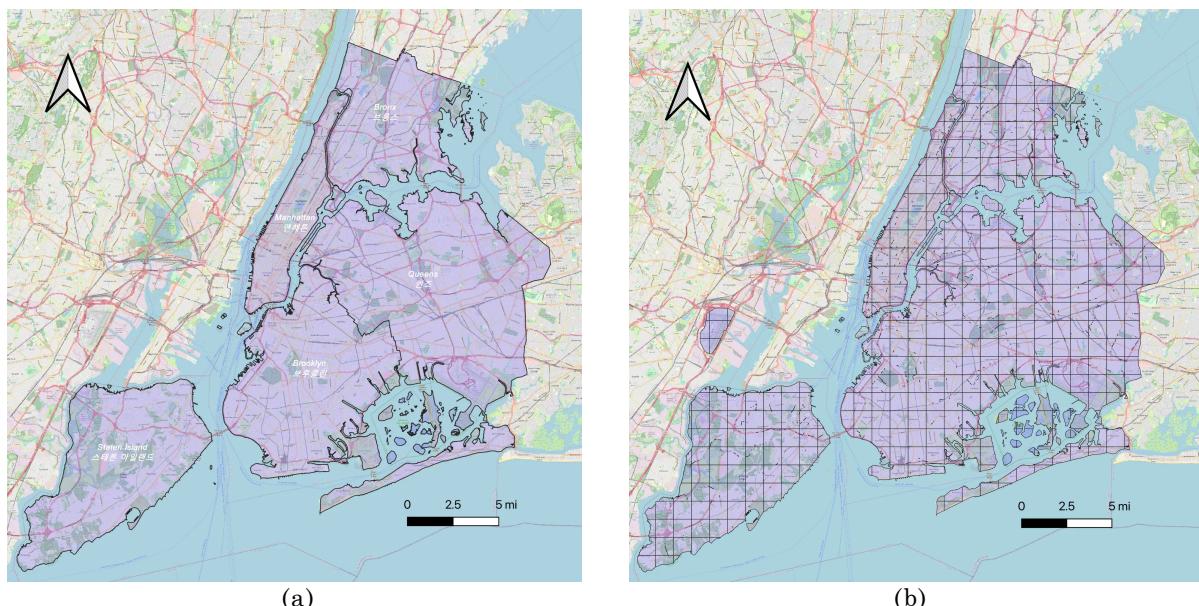


그림 3. 연구 지역 및 공간분석 단위

2. 데이터 전처리 및 방법론

2.1 데이터 전처리

2.1.1 택시 이동 O/D 매트릭스 구축

본 절에서는 뉴욕시 택시 이동 데이터를 활용하여 피크 시간대별 O/D 매트릭스를 구축하고, 이를 바탕으로 각 택시 구역별 통행량을 집계하였다. O/D 매트릭스는 승차지점(TLC 택시존 ID)과 하차지점(TLC 택시존 ID) 간의 쌍을 기준으로 집계되며, 각 쌍에 대해 탑승한 승객 수의 합을 계산하였다(그림 4A). 이를 통해 각 피크 시간대에 어떤 구간 간 이동이 활발한지를 확인할 수 있다. 이후 O/D 매트릭스에서 승차지점별 승객 수의 합을 합산하고, 하차지점별 승객 수의 합을 합산하여 출발지와 도착지별 이동량 데이터프레임을 생성하였다(그림 4B).

The figure consists of two side-by-side dataframes, labeled (a) and (b).

Dataframe (a): od_matrix_dic['od_matrix_peak1']

```
od_matrix_dic['od_matrix_peak1']

   승차지점TLC택시존ID  하차지점TLC택시존ID  count
0                      3                  3    55.0
1                      3                 18   13.0
2                      3                 20   13.0
3                      3                 31   1.0
4                      3                 32   24.0
...                   ...
18917                264                265   1.0
18918                265                42   1.0
18919                265                152   1.0
18920                265                191   1.0
18921                265                265   2.0

18922 rows × 3 columns
```

Dataframe (b): od_aggregate_dic['pickup_counts_peak1']

```
od_aggregate_dic['pickup_counts_peak1']

   zone_id  pickup_trip_count
0         3            423.0
1         4            357.0
2         5            66.0
3         6            89.0
4         7           1246.0
...       ...
253      261            297.0
254      262           1045.0
255      263           1180.0
256      264            31.0
257      265             5.0

258 rows × 2 columns
```

그림 4. Peak1 O/D 매트릭스와 출발지/도착지별 이동량 데이터프레임

2.1.2 면적 가중 보간(Areal Weighted Interpolation)

본 연구에서 활용한 2025년 택시 이동 데이터는 승하차 지점이 정확한 위경도가 아닌 263개의 뉴욕시 택시 구역으로 공개되어 있었다. 따라서 택시 구역별 이동량을 1마일 격자 단위로 할당하는 작업이 필요하였다. 해당 작업은 서로 겹치지만 경계가 일치하지 않는 폴리곤 값을 추정하는 기법인 면적 가중 보간 (Areal Weighted Interpolation)을 통해 진행하였다. 면적 가중 보간은 R의 areal 패키지를 활용하였으며, 과정은 부록의 코드1과 같다. 보간 유형은 extensive를 사용하였으며 이는 보간 대상인 pu_1, dof_1 등 이동량 변수들이 비율이 아닌 절대적인 count 값이기 때문이다. 이러한 변수들은 특정 구역 내 전체 이동량을 나타내므로 이 값을 격자 내 겹치는 면적 비율에 따라 누적 합산하여 분배하는 방식의 면적 가중 보간이 적합하다. 최종적으로 피크 시간대별 1마일 격자 단위 이동량 보간 결과는 아래 그림5와 같다.

The figure shows a single datafram titled 'areal_interpolate'.

```
areal_interpolate = pd.read_csv('~/Users/jiyeonhwang/Downloads/python/공분 프로젝트/areal_interpolate_result.csv')
areal_interpolate

   grid_id  dof_1  dof_2  dof_3  dof_4  pu_1  pu_2  pu_3  pu_4
0        0  7.120955  5.380277  5.380277  3.481356  8.703390  6.013251  6.171495  5.063790
1        1  4.382237  3.311024  3.311024  2.142427  5.356068  3.700556  3.797939  3.116258
2        2  2.866051  2.165460  2.165460  1.401180  3.502951  2.420221  2.483911  2.038080
3        3  2.478995  1.873018  1.873018  1.211953  3.029883  2.093373  2.148462  1.762841
4        4  6.373900  4.356580  4.212863  2.741188  6.522420  4.588449  4.617672  3.435832
...      ...
422     422  63.552901  44.232109  25.608393  18.041137  86.261624  57.801942  30.943474  27.886503
423     423  70.787302  46.726441  27.827085  16.941098  81.798253  74.996383  33.601128  26.917165
424     424  10.874726  7.852226  4.661351  3.104420  11.895062  11.009130  5.231547  3.877980
425     425  5.971798  3.884096  2.330458  1.359434  6.602964  6.554413  2.815970  2.184804
426     426  8.409290  5.469457  3.281674  1.914310  9.298076  9.229708  3.965356  3.076569

427 rows × 9 columns
```

그림 5. 1마일 격자 단위로 보간된 이동량 데이터프레임

2.2 방법론

2.2.1 커널 밀도 추정 (Kernel Density Estimation, KDE)

피크 시간대의 이동 수요가 공간적으로 어디에 집중되어 있는지를 파악하기 위해 커널 밀도 추정 (Kernel Density Estimation, KDE) 기법을 활용하였다. 기존의 263개 택시존 단위로 제공된 이동량 데이터를 그대로 활용할 경우 공간적 패턴이 택시존 경계에 의해 단절되어 표현될 우려가 있었다. 이에 따라, 택시존 단위의 이동량 데이터를 면적 가중 보간 (Areal Weighted Interpolation)을 통해 500m 격자 단위로 재분배하였다. 이후 각 격자의 중심점을 커널 밀도 추정의 포인트 데이터로 활용하여 연속적인 공간 분포 패턴을 보다

정교하게 시각화하고자 하였다. QGIS의 커널 밀도 추정 히트맵을 활용하였으며 픽셀 크기 100m, 대역폭(bandwidth) 750m, Quartic 커널 함수를 사용하였다.

2.2.2 전역적 공간적 자기상관 Global Moran's I

공간적으로 인접한 지역 간 이동 수요가 유사한 패턴을 보이는지 파악하기 위해서 공간적 자기상관을 측정하는 지표인 Moran's I를 활용하였다. 공간적 자기상관이란 토블러 (Tobler)의 '모든 것은 서로 관련되어 있지만, 가까운 것일수록 더 많이 관련되어 있다'는 지리학 제1법칙에 기반하며, 공간적으로 가까운 객체들이 비슷한 경향을 가지는 것을 의미한다. Global Moran's I 지표는 -1부터 1까지의 값을 가지며, -1은 강한 음의 공간적 자기상관, 1은 강한 양의 공간적 자기상관, 0은 공간적 자기상관이 존재하지 않음을 의미한다. Global Moran's I 계산은 공간 데이터 분석을 위한 오픈소스 Python 패키지인 PySAL의 libpysal, esda 서브패키지를 활용하였다. 공간적 자기상관을 계산하기 위해서 먼저 이웃을 정의하고 각 이웃에 대한 가중치를 부여해야한다. 이웃 정의에는 최소한 하나의 정점 (vertex)를 공유하는 폴리곤을 이웃으로 정의하는 Queen's Case를 사용하였으며, 가중치는 이웃 폴리곤은 1, 이웃이 아닌 폴리곤은 0으로 부여하였다. Global Moran's I 계산 과정은 부록의 코드2와 같으며, 이웃 폴리곤 시각화 결과와 인접성 행렬은 그림6과 같다.

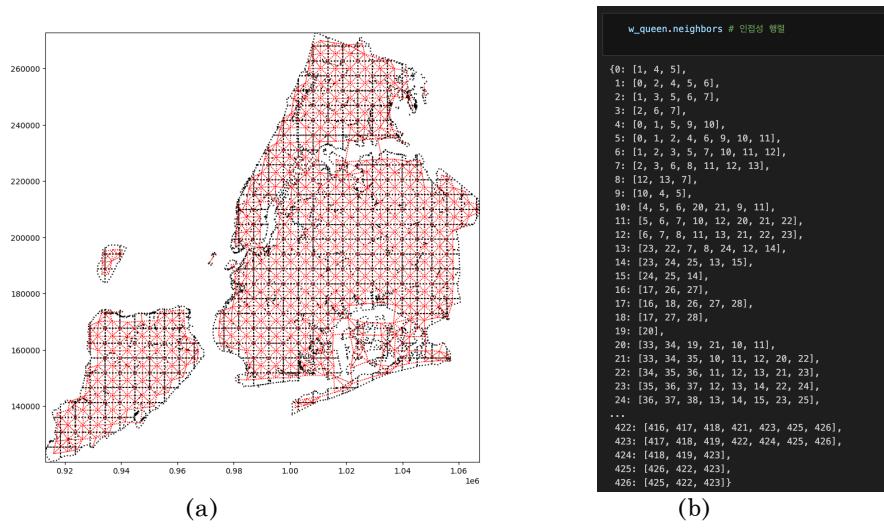


그림 6. 이웃 폴리곤 시각화 및 인접성 행렬

2.2.3 국지적 공간적 자기상관 Local Moran's I

Global Moran's I는 전역적인 공간적 자기상관의 정도를 나타내는 하나의 지표만 계산한다. 본 연구에서는 그리드 단위의 국지적인 공간적 자기상관을 파악하기 위해 Local Moran's I LISA (Local Indicator of Spatial Association) 분석을 진행하였으며, 이를 통해 공간적 자기상관의 국지적인 패턴과 클러스터를 식별하였다. LISA 분석 결과는 아래 표2와 같이 4개의 유형으로 분류된다. LISA 분석 또한 Global Moran's I와 동일한 Python 패키지를 활용하였으며 시각화는 folium 패키지를 통해 진행하였다. LISA 분석 과정은 부록의 코드3과 같다.

표 2. LISA 분석 결과의 4개 유형

HH (High-High)	높은 값이 주변의 높은 값들과 인접 (Hotspot)
LL (Low-Low)	낮은 값이 주변의 낮은 값들과 인접 (Coldspot)
HL (High-Low)	높은 값이 주변의 낮은 값들과 인접
LH (Low-High)	낮은 값이 주변의 높은 값들과 인접

3. 분석 결과

3.1 커널 밀도 추정 (KDE) 히트맵

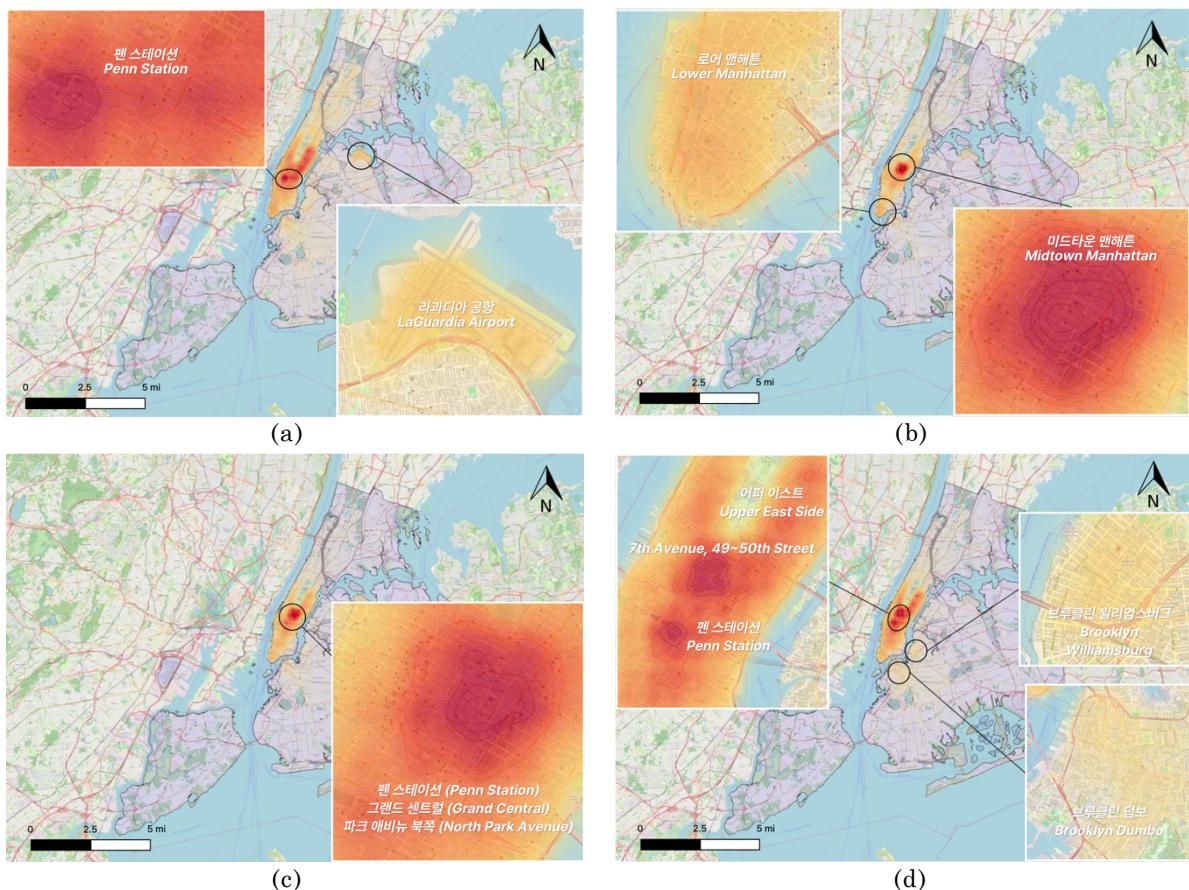


그림 7. peak1, peak2 핫스팟 (a) peak1 출발지, (b) peak1 도착지, (c) peak2 출발지, (d) peak2 도착지

그림7은 커널 밀도 추정을 통해 평일 오전 시간대인 peak1, 평일 오후 시간대인 peak2의 이동 수요 출발지와 도착지 핫스팟을 시각화한 결과이다. Peak1 평일 오전 7시부터 10시까지의 출발지는 미국 동부의 대중교통 허브인 펜스테이션과 라과디아 공항이 핫스팟으로 나타났다(그림7A). 도착지의 경우 월스트리트, 뉴욕 증권거래소, 로펌, 보험사 등 전통적인 뉴욕의 금융과 비즈니스 중심지인 로어 맨해튼과 뉴욕 최대의 상업 밀집 지구인 미드타운 맨해튼이 핫스팟으로 나타났다(그림7B). 이는 전형적인 통근 흐름을 반영하는 것으로 해석할 수 있다.

Peak2 평일 오후 4시부터 7시까지의 출발지 핫스팟 분석 결과, 펜스테이션, 그랜드 센트럴과 같은 교통 허브에서 이동 수요가 집중되었다(그림7C). 이는 퇴근 이후 주거지가 있는 외곽으로 이동하려는 통근자의 환승 흐름을 보인다고 해석할 수 있다. 또한 UN본부와 같은 국제기관, 대기업 오피스, 고급 주거지가 밀집 되어있는 파크 애비뉴 북쪽에서도 출발지 이동 수요가 높게 나타났다. 이 또한 퇴근 시간대의 이동 수요 출발지로 기능하는 것을 알 수 있다. Peak2 도착지 핫스팟 분석 결과, 쇼핑, 공연, 전시가 모여 있는 7번가 49~50번가 일대(록펠러센터 인근), 브루클린에서 주거지와 문화지구가 혼합되어 있는 윌리엄스버그와 덤보, 고급 주거지 밀집 지역인 어퍼 이스트에서 도착 수요가 집중되었다(그림7D). 이는 퇴근 후 문화활동, 모임, 귀가 등 복합적인 이동 흐름을 나타내는 것으로 해석할 수 있다.

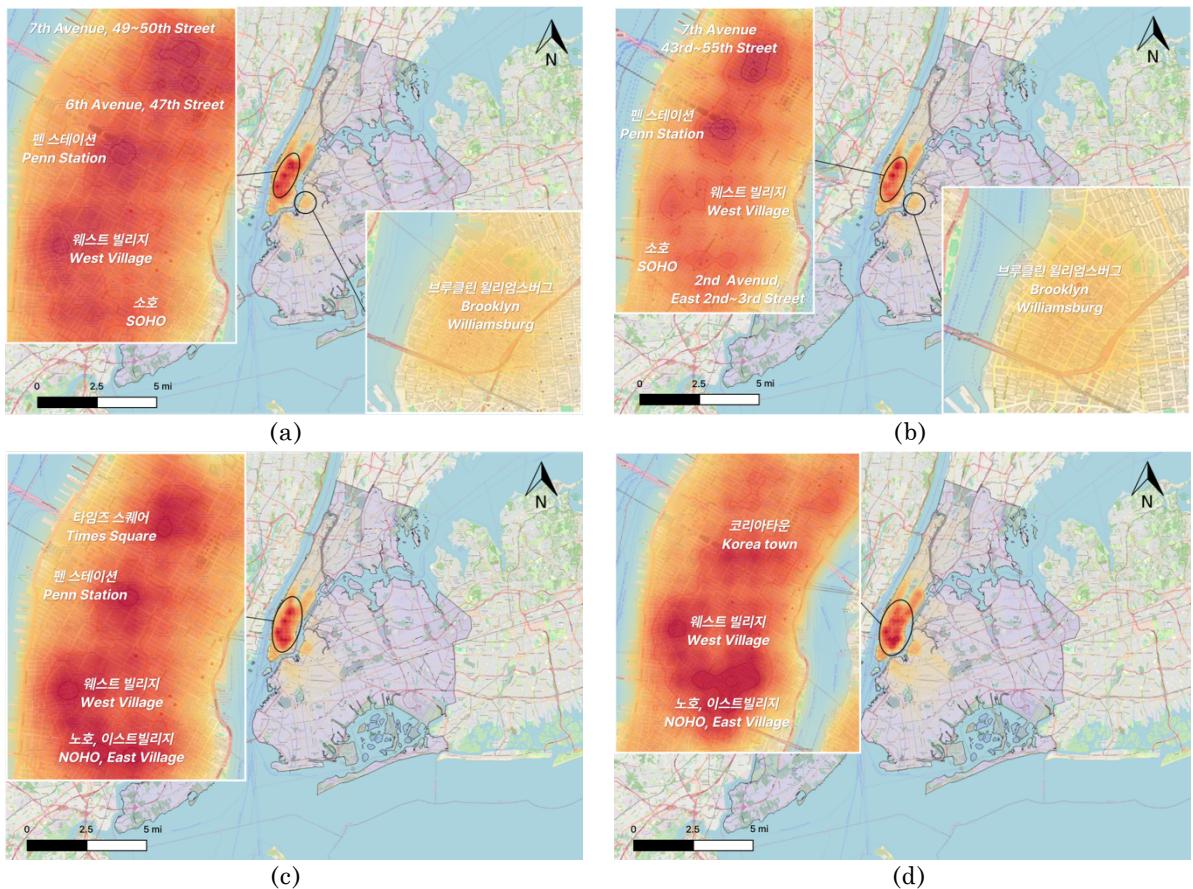


그림 8. peak3, peak4 핫스팟 (a) peak3 출발지, (b) peak3 도착지, (c) peak4 출발지, (d) peak4 도착지

Peak3 주말 오후 6시부터 8시까지의 출발지와 도착지 핫스팟은 대체로 유사하게 나타났다. 타임스퀘어와 브로드웨이, 레스토랑, 호텔이 위치한 전형적 관광지인 미드타운(6th Avenue, 47th Street), 고급 쇼핑거리로 유명한 5번가, 쇼핑·카페·갤러리가 밀집되어 있는 소호와 웨스트빌리지, 레스토랑·루프탑바와 같은 모임 장소와 주거지가 혼합되어 있는 월리엄스버그에서의 이동 수요가 높게 나타났다(그림8A, B). 이는 주말 오후 시간대의 여가 활동을 목적으로 한 이동을 반영한 결과로 해석할 수 있다.

Peak4 주말 오후 9시부터 11시까지의 출발지 핫스팟 분석 결과, 바·클럽·재즈바·루프탑 라운지와 같은 유흥 시설이 밀집 되어있는 노호와 이스트빌리지, 타임즈 스퀘어, 재즈바·공연장·레스토랑이 많은 웨스트빌리지에서 출발 수요가 집중되었다(그림8C). 이는 저녁 활동 이후 2차 장소로 이동하거나 귀가를 시작하는 출발점을 반영한다고 해석할 수 있다. 특히 타임즈 스퀘어의 경우 늦은 시간에 공연이 끝나는 브로드웨이 극장가에서의 출발 이동 수요를 반영한다고 볼 수 있다. 도착지의 경우 출발지 핫스팟과 거의 유사하나 코리아타운이 추가적으로 핫스팟으로 나타났다. 이는 늦은 밤까지 영업하는 레스토랑과 바가 밀집해 있는 특성 때문에 2차 이동 수요가 반영된 결과라고 해석할 수 있다.

3.2 공간적 자기상관 분석 결과

3.2.1 Global Moran's I

전역적 공간적 자기상관을 분석한 결과(표3), 모든 피크 시간대에서 Moran's I값이 0.6이상 0.73이하로 매우 높은 양의 공간적 자기상관을 나타내었다. 이는 택시 이동 수요의 출발지와 도착지가 모두 공간적으로 군집(Hotspot/Coldspot)되어있음을 의미한다. 또한 p-value가 모두 0.05보다 낮은 거의 0에 가까운

값으로, 계산된 Moran's I 값은 통계적으로 유의한 결과이며, 공간적 군집이 우연히 발생할 확률이 거의 없음을 의미한다.

표 3. Global Moran's I 분석 결과

O/D	시간대	Global Moran's I	p-value	z-score
출발지	Peak1	0.733	2.5e-139	25.127
	Peak2	0.664	7.6e-115	22.778
	Peak3	0.708	3.8e-130	24.273
	Peak4	0.678	1.7e-119	23.243
도착지	Peak1	0.623	2.2e-101	21.377
	Peak2	0.707	4.7e-130	24.263
	Peak3	0.696	4.2e-126	23.887
	Peak4	0.734	6.7e-140	25.179

3.2.2 LISA (Local Moran's I)

공간적 자기상관의 군집을 식별하기 위해 진행한 LISA 분석 결과는 그림9와 같이 나타났다. 맨해튼 중심부는 모든 시간대에서 출발지, 도착지 모두 H-H(High-High) 클러스터로 나타났으며 이는 해당 지역에 있는 그리드의 이동 수요도 높으면서 주변의 이동 수요도 높음을 의미한다. 도착지의 경우 출근 시간대인 peak1을 제외한 나머지 시간대에 모두 브루클린, 롱아일랜드시티와 같은 지역으로 H-H 클러스터가 분산되는 경향을 파악할 수 있었다(그림 9B, D, F, H). 해당 지역은 맨해튼 외곽의 주거 밀집 지역으로, 퇴근 후 또는 모임과 같은 여가 활동 이후의 귀가 이동으로 인한 결과라고 해석된다.

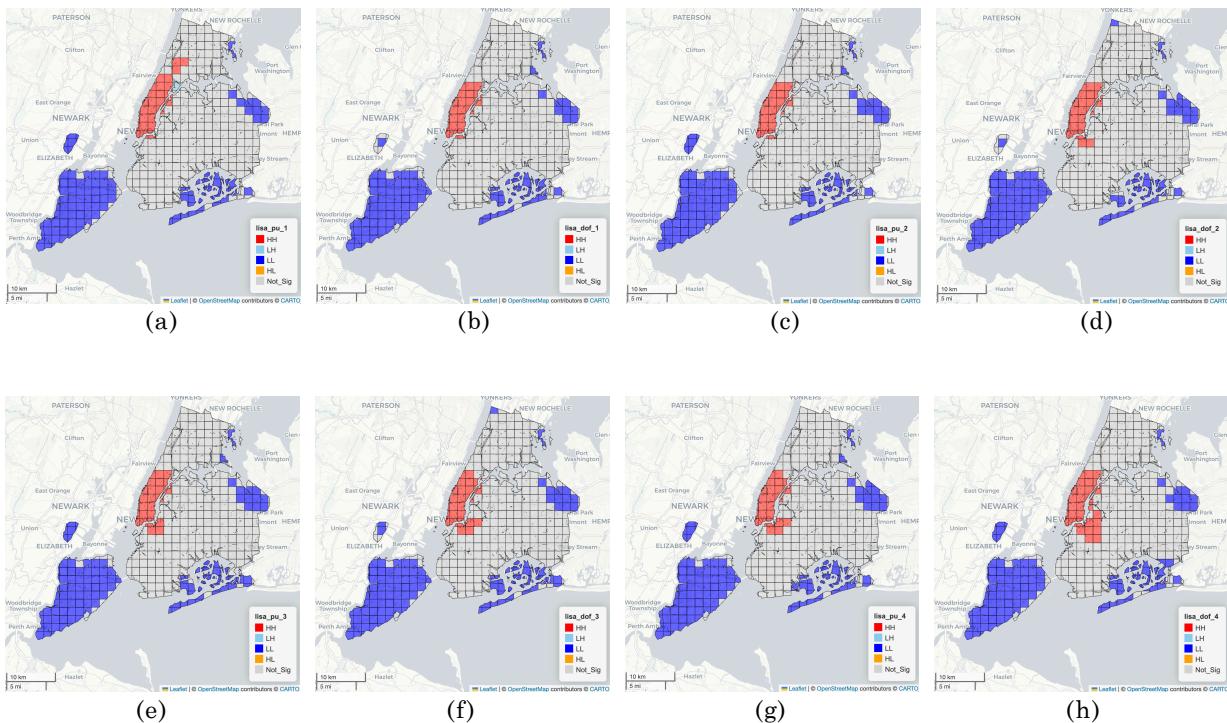


그림 9. LISA 분석 결과

(a) peak1출발지, (b) peak1도착지, (c) peak2출발지, (d) peak2도착지,

(e) peak3 출발지, (f) peak3 도착지, (g) peak4 출발지, (h) peak4 도착지

4. 논의

본 연구는 Xie C 외(2021)의 논문 「Revealing spatiotemporal travel demand and community structure characteristics with taxi trip data」을 바탕으로, GIS와 공간분석 수업에서 학습한 공간통계 기법을 적용해 보는 프로젝트의 일환으로 진행되었다. 기존 논문의 결과와 본 연구의 결과에서는 크게 2가지 차이가 존재한다. 첫째, 시계열 분석 결과를 통해 도출한 피크 시간대가 상이하게 나타났다. 기존 논문에서 시계열 분석을 통해 선정한 피크시간은 평일 오전 7시~10시, 평일 오후 5시~8시, 주말 오후 12시~3시, 주말 오후 5시~8시로 본 연구에서 선정한 피크시간과 평일 오후에 1시간, 주말에 4시간 이상 차이가 존재한다. 이는 논문과 본 연구에서 사용한 뉴욕시 택시 데이터의 시점 차이가 원인인 것으로 파악된다. 본 연구는 2025년 1월 데이터를 사용하였고, 기존 논문은 2016년 6월 데이터를 사용하였는데 6월에는 표준 시간을 1시간 앞당기는 썸머타임이 적용되어 시간의 차이가 존재한다. 썸머타임 시행 여부의 차이와 더불어, 여름과 겨울의 계절적 차이도 작용하여 일몰 시각이 1월에는 오후 4시, 6월에는 오후 8시로 다소 차이가 크다. 이에 따라 사람들의 주 활동 시간과 이동 시간에 변화가 발생하여 기존 논문과 본 연구의 피크 시간대에 차이가 존재하는 것으로 파악된다. 둘째, LISA 분석으로 파악한 클러스터에 다소 차이가 존재한다. 기존 논문에서 파악한 H-H (High-High) 클러스터는 맨해튼 지역에서만 나타났다. 하지만 본 연구에서는 출근 시간대를 제외한 모든 시간대에서 맨해튼 외곽 주거 밀집지역인 브루클린과 롱아일랜드시티가 도착지 이동 수요가 높은 H-H (High-High) 클러스터로 파악되었다. 이는 2016년과 2025년 사이에 발생한 뉴욕시의 사회적, 경제적, 공간적 변화가 원인인 것으로 유추된다. 브루클린과 롱아일랜드시티는 2016년 이후에 모두 주거지 개발이 이루어진 지역이다. 롱아일랜드시티의 경우 2016년 이후에 Court Square, Jackson Ave 주변에 수천 세대의 고층 아파트가 개발되었으며, 브루클린 또한 Pacific Park Brooklyn 프로젝트로 2018년 이후에 고층 아파트가 다수 완공되었다. 이와 같은 주거지 개발로 인한 인구 유입으로 이동 수요가 증가한 것으로 파악된다.

본 연구를 위해 참고한 논문은 뉴욕시의 택시 데이터를 활용하여 시공간적 이동 수요를 공간 통계 기법과 커뮤니티 구조 분석을 활용하여 다각도로 분석하였다는 점에서 의의가 있다. 그러나 원본 논문에는 몇 가지 문제점이 존재하여 분석 결과에 대한 해석과 방법론에 대한 신뢰도에 다소 한계가 있다고 판단된다. 다음은 주요 문제점과 향후 개선 방안에 대한 논의이다.

첫째, 논문에서는 택시 O/D 데이터의 공간 단위에 대한 구체적인 설명이 부족하였다. 특히 분석에 활용된 데이터가 정확한 위경도 좌표를 포함한 포인트 데이터인지, 아니면 TLC 택시존(Taxi Zone) 단위의 공간 데이터인지에 대한 명확한 언급이 없었다. 논문 본문에서는 택시존 기반 분석임을 암시하고 있으나, KDE 및 Moran's I 분석에 앞서 택시존 데이터를 어떤 방식으로 격자화하였는지, 택시존 단위의 값을 1마일 격자 단위로의 보간을 수행했는지 여부에 대한 설명이 전혀 제시되어 있지 않았다. 이러한 정보의 부재는 분석 절차의 재현 가능성을 저하시킬 뿐만 아니라, 결과 해석의 신뢰도를 약화시킬 수 있다. 따라서 향후 연구에서는 공간 단위의 특성과 전처리 과정을 명확히 서술하고, 격자 기반 분석을 수행한 경우 격자 생성 기준, 보간 방식 등을 구체적으로 제시할 필요가 있다. 둘째, 시각 자료의 배치에 있어 분석 절과 그림 간의 연계성이 떨어지는 문제도 확인되었다. KDE 분석 결과는 4.1절에 서술되어 있으나, 해당 결과에 대한 그림은 4.2절(공간적 자기상관 분석)에 삽입되어 있어 독자의 혼란을 유발할 수 있다. 셋째, 논문에서는 Global Moran's I 분석 시 공간 가중치를 두 격자 간의 유클리드 거리의 역수로 설정하였다고 언급하고 있으나, 모든 격자를 이웃으로 간주한 것인지, 혹은 특정 거리 이내의 격자만을 이웃으로 설정한 것인지에 대한 설명이 결여되어 있다. 공간 가중치 행렬은 공간적 자기상관 분석의 핵심 요

소로, 이웃의 정의에 따라 결과가 달라지기도 한다. 따라서 분석에 사용된 공간 가중치 행렬의 생성 방식, 이웃 설정 기준에 대한 정보가 보완되어야 할 것이다.

참고문헌 및 데이터 출처

1. Xie C, Yu D, Zheng X, Wang Z, Jiang Z (2021) Revealing spatiotemporal travel demand and community structure characteristics with taxi trip data: A case study of New York City. PLoS ONE 16(11): e0259694. <https://doi.org/10.1371/journal.pone.0259694>
2. Louail T, Lenormand M, Cantu Ros OG, Picornell M, Herranz R, Frias-Martinez E, et al. From mobile phone data to the spatial structure of cities. Sci Rep. 2015; 4: 5276. <https://doi.org/10.1038/srep05276> PMID: 24923248
3. Christopher Prener, Ph.D. Areal Weighted Interpolation, <https://cran.r-project.org/web/packages/areal/vignettes/areal-weighted-interpolation.html>
4. TLC Trip Record Data, 2025 January, NYC Open Data, <https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>
5. Taxi Zone Shapefile, NYC Open Data, <https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>
6. Borough boundaries for NYC (Clipped to Shoreline), 2025, NYC gov department of city planning, <https://www.nyc.gov/content/planning/pages/resources/datasets/borough-boundaries>

부록

코드 1. 면적 가중 보간(Areal Weighted Interpolation) 진행 코드

```
> library(sf)
> library(areal) # 면적 가중 보간을 위한 areal 패키지
> zonemap_trip <- read_sf('/Users/jiyeonhwang/Downloads/python/공분 프로젝트 /zonemap_with_trips.shp') #
python으로 전처리한 출발지/도착지별 이동량과 택시존 공간정보를 담고 있는 shp파일
> grid <- read_sf('/Users/jiyeonhwang/Downloads/python/공분 프로젝트/ grid.shp') # python으로 전처리한 1마일 격자
shp파일
> areal_interpolate <- aw_interpolate(
  grid, # 보간 대상이 되는 격자 폴리곤 레이어(target)
  tid = grid_id, # grid 객체에서 각 그리드를 식별하는 고유 ID필드
  source = zonemap_trip, # 택시존별 이동량 값이 존재하는 원본 폴리곤 레이어(source)
  sid = OBJECTID, # zonemap_trip 객체에서 각 폴리곤을 식별하는 고유 ID 필드
  weight = "sum", # 가중치 방식, sum은 source의 값을 100% target에 재분배
  output = "tibble", # 결과를 tibble 형식으로 반환
  extensive = c("pu_1", "dof_1", "pu_2", "dof_2", "pu_3", "dof_3", "pu_4", "dof_4")) # 보간할 속성 변수들
(pu_1 : pick up peak1, dof_1 : drop off peak1)
```

코드 2. Global Moran's I 계산 코드

```
importesda
import libpysal
import geopandas as gpd
grid_trip=gpd.read_file('/Users/jiyeonhwang/Downloads/python/공분프로젝트/grid_with_trips.shp') # 1 마일
격자 단위 이동량 데이터 불러오기
# Global Moran's I 계산
mi_queen = esda.moran.Moran(y, w_queen) # w_queen 은 인접 행렬, y 는 peak1 pickup
```

```
print(f"Moran's I with Queen's case contiguity: {round(mi_queen.I, 3)}, p-value: {mi_queen.p_norm},  
z-score: {round(mi_queen.z_norm, 5)}")
```

코드 3. LISA 분석 코드

```
# 색상을 지정할 컬럼 LISA 분석을 진행할 속성 컬럼 리스트  
col_list = ['dof_1', 'dof_2', 'dof_3', 'dof_4', 'pu_1', 'pu_2', 'pu_3', 'pu_4']  
# Queen's Case 인접성 행렬 계산  
w_queen = libpsal.weights.Queen.from_dataframe(grid_trip, use_index=True)  
lm_dict = {1: 'HH', 2: 'LH', 3: 'LL', 4: 'HL'} # LISA 분석 결과 매핑  
for i in col_list: # 반복문을 통한 각 피크 시간대의 출발지, 도착지 LISA 분석 진행  
    y = grid_trip[f'{i}']  
    lm_queen =esda.moran.Moran_Local(y, w_queen, seed=17)  
    lisa_queen = []  
    for idx in range(len(lm_queen.q)):  
        # p-value 가 0.05 미만인 경우 LISA 분석 결과 리스트에 저장  
        if lm_queen.p_sim[idx] < 0.05:  
            lisa_queen.append(lm_dict[lm_queen.q[idx]])  
        else:  
            lisa_queen.append('Not_Sig')  
    grid_trip[f'lisa_{i}'] = lisa_queen
```