택시와 대중교통 이용패턴 차이를 유발하는 공간적, 사회적 특성 연구: 지하철역을 중심으로

김경옥
서울과학기술대학교 글로벌융합산업공학과



목차

1 서론

2 연구방법

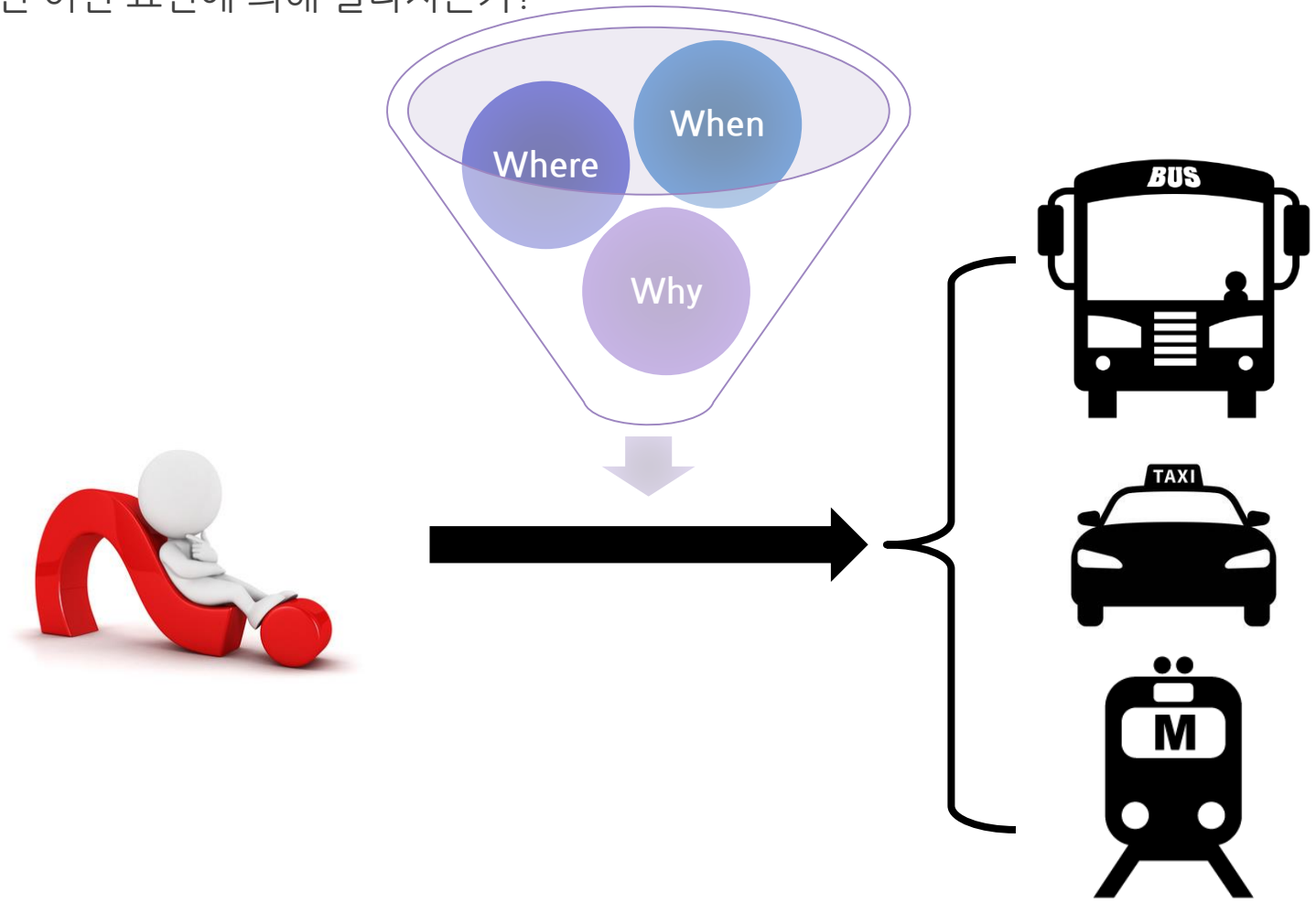
3 분석결과
- 이용패턴 군집 분석
- 이용패턴 & 이용빈도 예측

4 결론

참고문헌, 사용데이터 & 부록

1. 서론

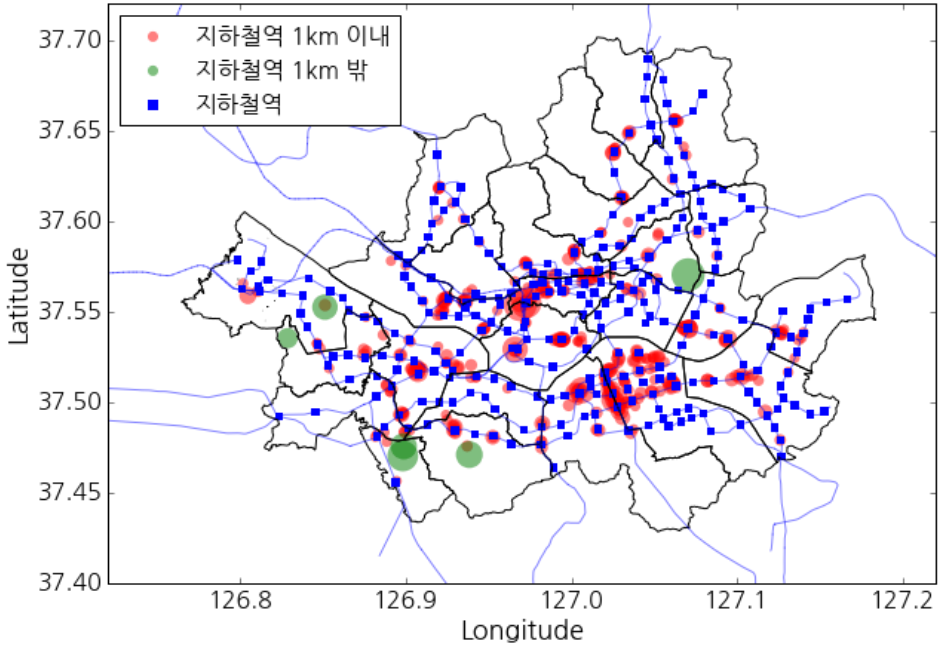
- 핵심 질문
  - 장소, 시간, 이동 목적에 따라 이동수단 선택이 달라지는가?
  - 다르다면 어떤 요인에 의해 달라지는가?



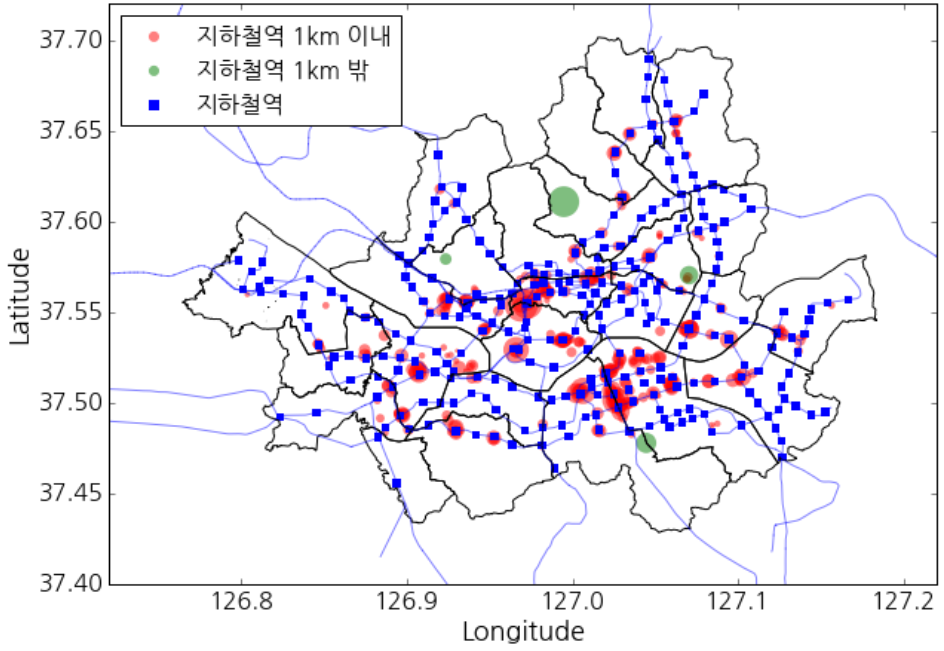
2. 연구방법 - 연구의 범위 설정

- 서울시에서 대중교통 정류장 중 이용빈도가 높은 지점은 지하철역
  - 2016년 4월 1일부터 7일까지 일주일간의 '대중교통 환승데이터' 기준
- 택시 이용빈도가 높은 지점과 지하철역과의 관계를 파악
  - 택시 이용빈도 상위 10% 지점(365개) 중 승차 기준 359개, 하차 기준 361개 지점이 지하철 반경 1km 이내 위치하며 상위 10% 지점들은 전체 승차량의 약 53%, 하차량의 46% 차지

[승차]



[하차]



지하철역을 중심으로 분석 진행

2. 연구방법 - 분석 데이터

- 지하철: 2015년 '서울시 지하철 호선별 역별 시간대별 승하차 인원 정보' 활용
  - 1시간 간격으로 역별, 호선별, 역별 승하차 인원을 월별로 제공
  - 신분당선 데이터는 제공되고 있지 않음

- 택시: 2015년 '서울시 택시운행 분석데이터 정보' 활용
  - 서울시 도로를 150m 간격으로 분할 → Link
    - 150m의 시작점, 중간점들, 끝점의 (위도, 경도) 정보가 제공됨 → 중앙점을 Link의 대표 위치로 이용
  - 30분 간격으로 Link별, 요일별, 날씨별 승하차 횟수를 1년간 합산한 수치를 제공



지하철, 택시 모두 같은 포맷의 데이터로 변환

지하철	• 서울 시외에 위치한 지하철역 제거 → 총 276개 역에 대한 분석 진행 • 호선에 상관없이 지하철역을 기준으로 합산 • 월별 기록을 합산
택시	• 30분 간격을 1시간 간격으로 합산 (0시, 0시 30분 데이터를 0시로 통합) • 요일과 날씨에 관계없이 Link별로 합산 • 서울 시내에 위치한 지하철역 반경 670m(도보 10분 기준) 내의 모든 Link의 승하차 횟수를 합산

2. 연구방법 - 분석 데이터

- 지하철, 택시 데이터를 이용패턴을 대표할 수 있는 형태로 변환

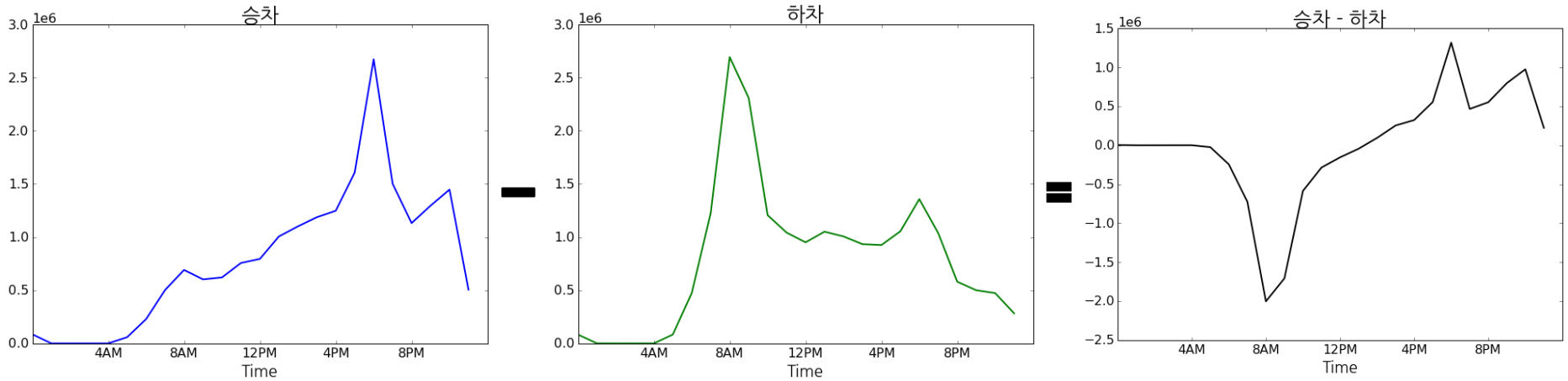






2. 연구방법 - 분석 데이터

- 승차, 하차 정보를 통합하여 이용패턴을 대표할 수 있는 벡터로 변환
  - 승차수와 하차수의 차이를 계산



  - 정규화
    - 지하철역별로 이용빈도가 상이
    - 패턴에 집중하기 위해서 정규화 진행

$$\bar{V}_{s,v} = \frac{V_{s,v}}{\max_t |V_{s,t}|}$$

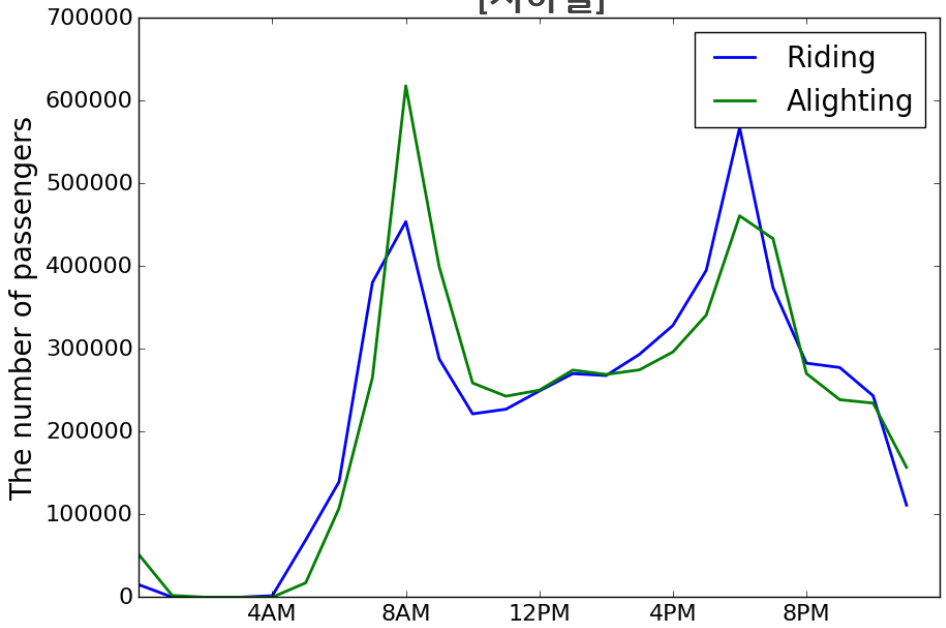
      - s: 지하철역, v: 시간

2. 연구방법 - 분석 데이터

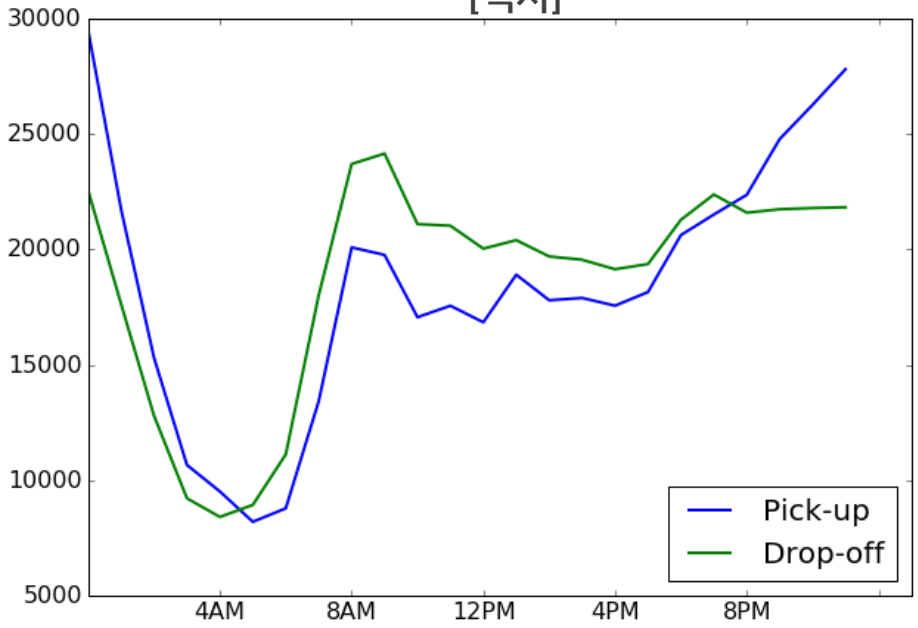
- 지하철역별로 정제 후 지하철, 택시 데이터의 기본 통계량($\times 10^5$)

통계량	지하철 (승차)	지하철 (하차)	택시 (승차)	택시 (하차)
평균값	72.20	72.28	12.53	13.15
표준편차	66.08	67.53	9.17	9.08
최소값	2.31	2.58	0.17	0.27
중앙값	50.50	48.54	9.95	10.89
최대값	370.52	375.93	52.14	53.80
합	19926	19949	3458	3630

[지하철]



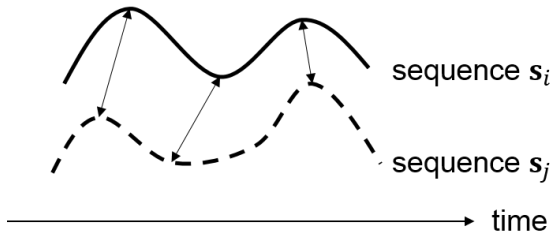
[택시]



2. 연구방법 - 분석 방법론

이용패턴 파악 → 군집 분석

- 비슷한 이용패턴을 보이는 지하철역으로 구성된 군집을 도출
- k-means 군집방법론을 지하철역의 차이벡터 데이터에 적용
  - 군집의 수를 2, 3, 4, 5로 변화시켜 가며 분석 수행
- 차이벡터간의 거리는 1시간의 편차를 허용하는 dynamic time warping (DTW) 거리척도 이용
- 지하철, 택시, 지하철+택시, 세 가지 경우에 대해서 분석 진행



이용패턴을 결정하는 요인 분석 → 분류, 회귀 분석

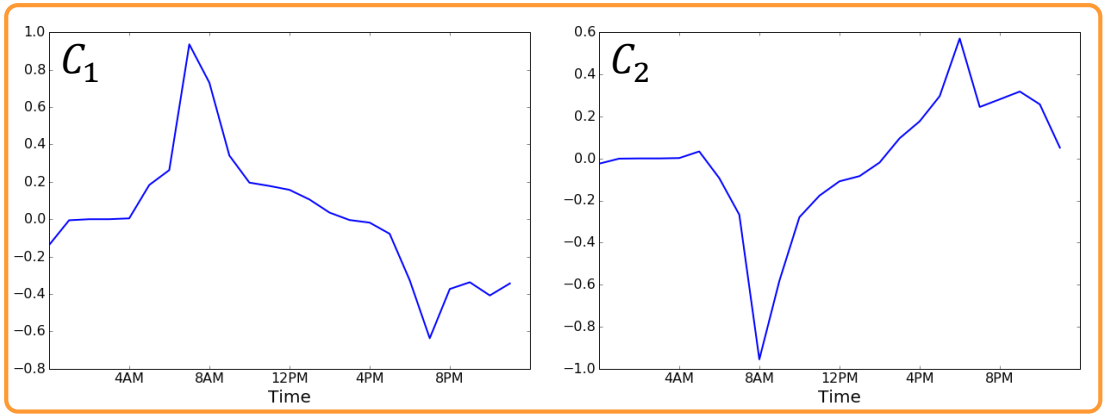
- 군집 분석의 결과를 활용하여 이용패턴을 예측하는 분류 모형을 여러 설명 변수를 이용하여 학습
  - '집계구별 인구·가구·주택·사업체 통계' 및 '추정매출액' 데이터로부터 설명 변수 도출
- 여러 설명 변수 중에서 이용패턴 예측에 유의미하게 활용되는 일부 변수만 선택하기 위해 로지스틱 회귀 모형에 L1 규제 적용한 Lasso 로지스틱 회귀 모형 이용

$$\log \mathcal{L} = \sum_{i=1}^{n} \log(\exp(-y_i(\mathbf{w}^T\mathbf{x}_i + c) + 1)) + \frac{\lambda}{2}\sum_{j=1}^{p}|w_j|$$

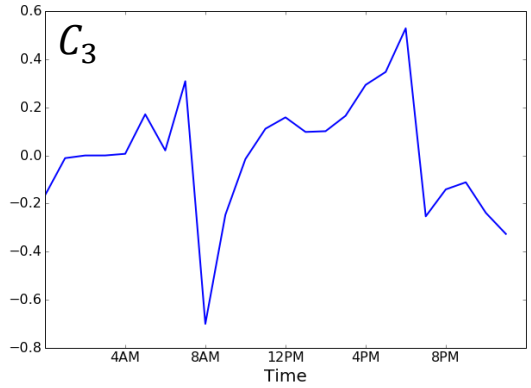
  - 최적의 λ는 5-fold cross-validation을 이용해서 결정 → λ 고정 후 전체 데이터로 재학습
- 이용패턴뿐 아니라 택시와 지하철의 이용을 유발하는 요인 도출을 위해 이용빈도에 대한 회귀 모형 학습
  - 분류 모형과 마찬가지로 선형 회귀 모형에 L1 규제를 추가한 Lasso 모형 이용

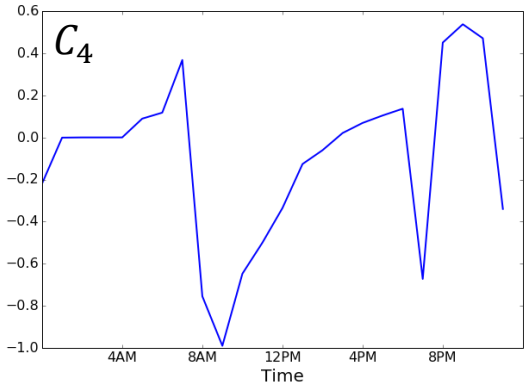
3.1 분석결과 - 택시·지하철 이용패턴 군집 분석 결과 (지하철)

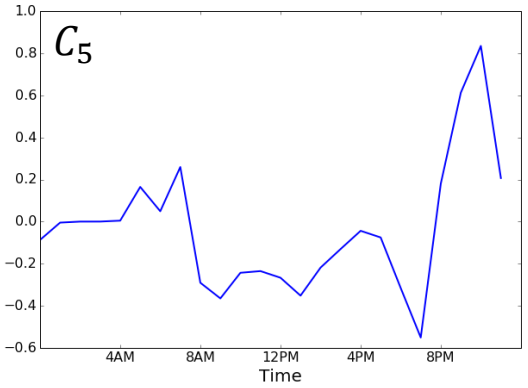
- 군집 수를 변화시켰을 때 형성되는 군집의 평균 이용패턴 (차이벡터)
  - C_1, C_2에서 나타나는 peak의 위치는 일반적인 통근 시간대
    - C_1은 주거지, C_2는 비즈니스 지구로 볼 수 있음
  - 군집 수가 5일 때 C_1, C_2의 비중이 75%



군집	평균 DTW 거리
C_1	0.74
C_2	0.84
C_3	1.49
C_4	1.44
C_5	1.49
전체	1.91

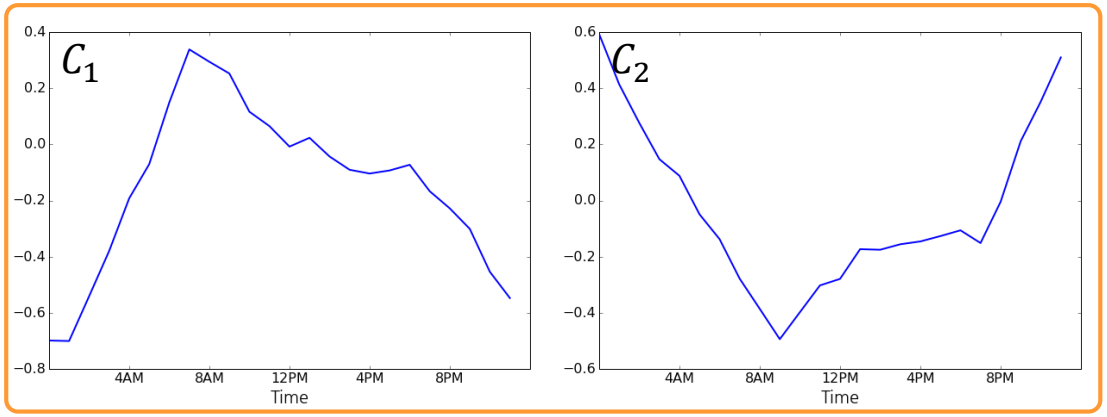




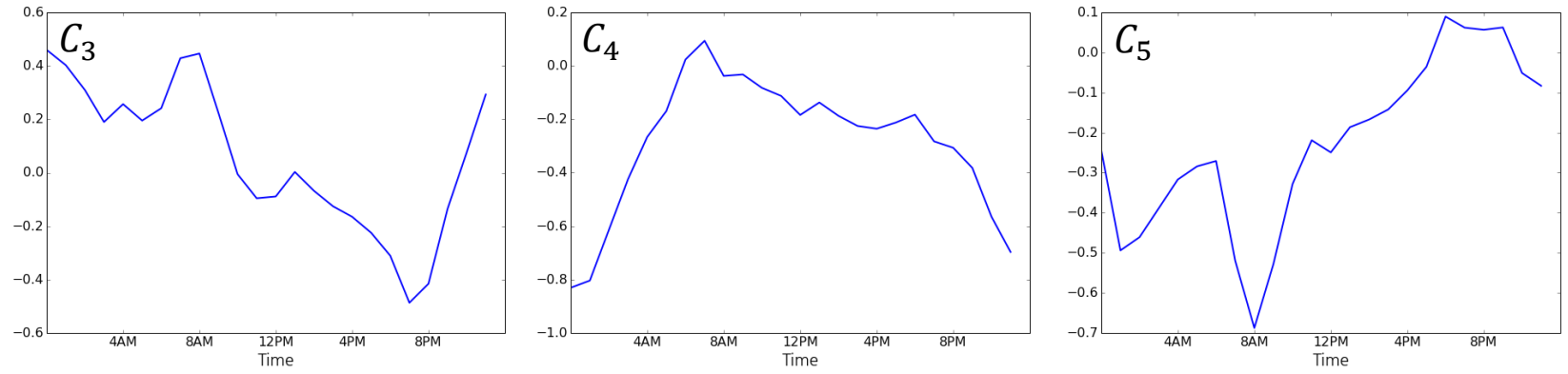


3.1 분석결과 - 택시·지하철 이용패턴 군집 분석 결과 (택시)

- 지하철 군집의 평균 이용패턴과는 다른 특징이 나타남
  - C_1, C_2에서 2개의 peak가 나타나지 않음
    - 출근시간대만
  - 군집 수가 5일 때 C_1, C_2의 비중이 75%



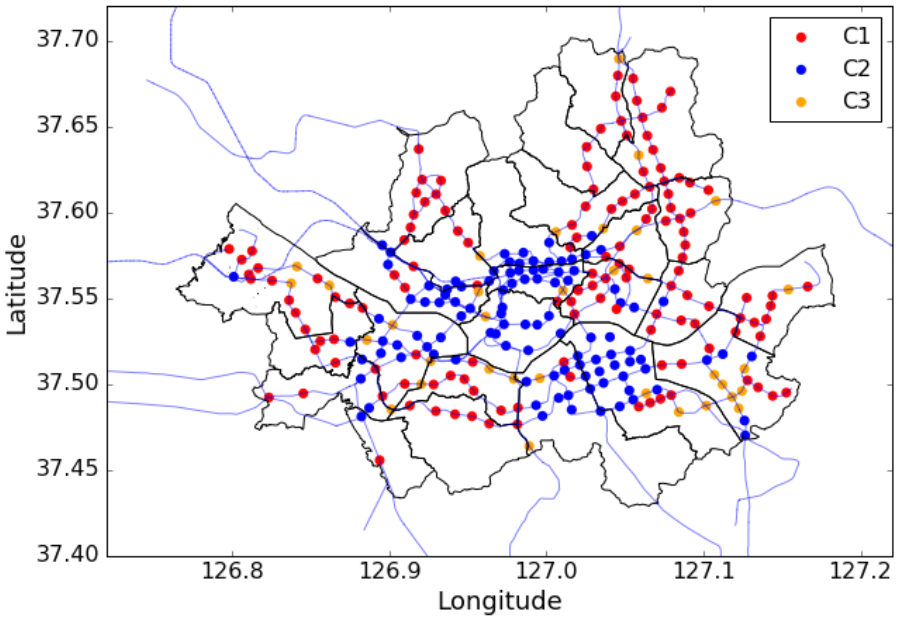
군집	평균 DTW 거리
C_1	1.21
C_2	1.59
C_3	1.93
C_4	1.68
C_5	1.65
전체	2.63



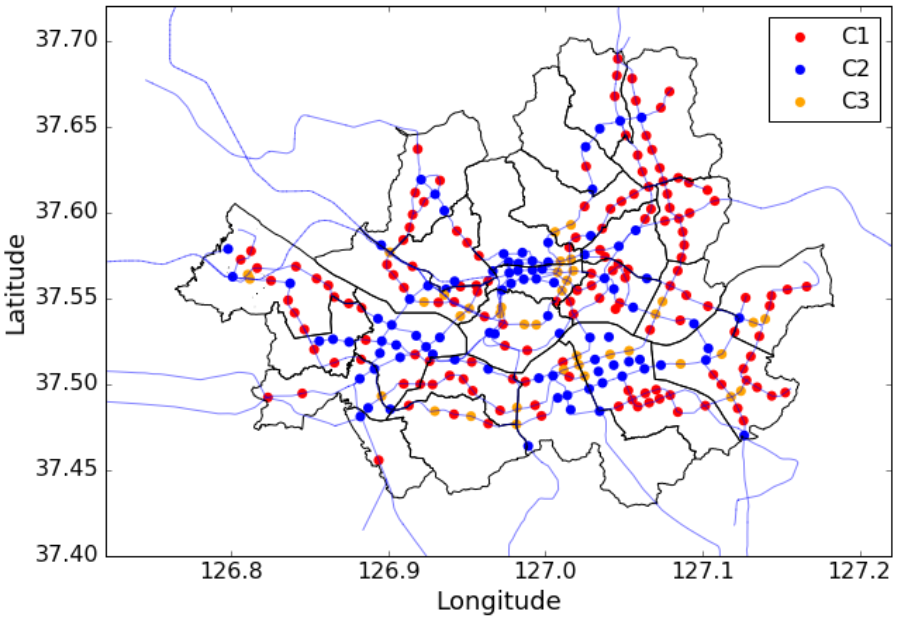
3.1 분석결과 - 군집의 분포 (지하철, 택시)

- 군집의 개수를 3개로 설정했을 때 형성된 군집의 지리적 분포

[지하철]



[택시]

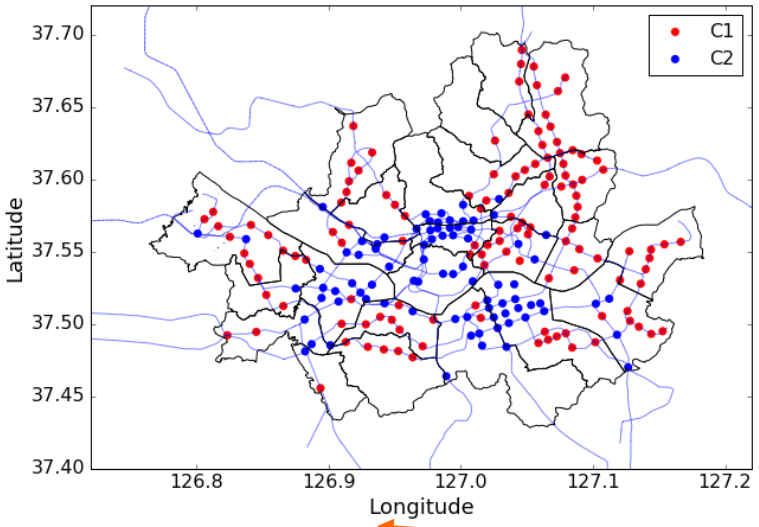


◦ 지하철과 택시 군집의 평균 DTW 거리 비교

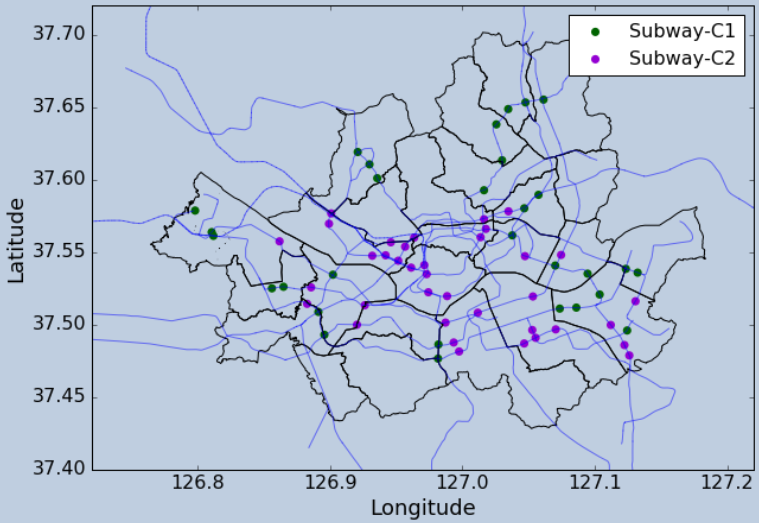
	지하철	택시
전체	1.91	2.63
인접	1.24 (35.1%↓)	2.05 (20.9%↓)

3.1 분석결과 - 군집의 분포 (지하철, 택시)

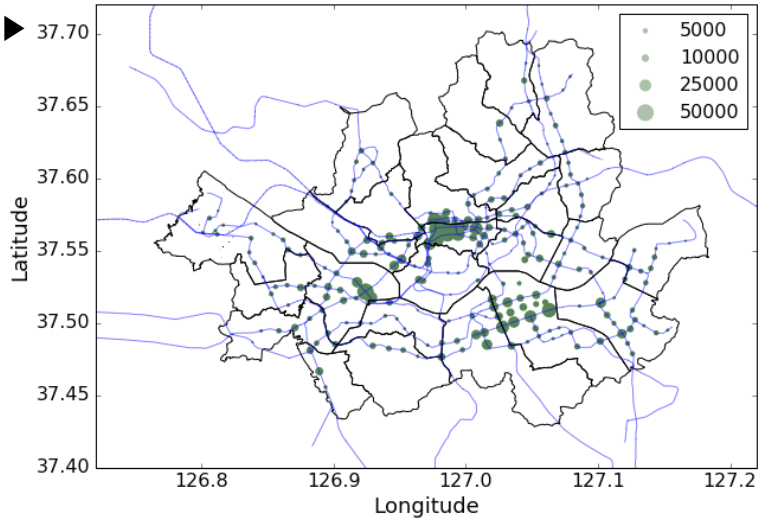
[동일]

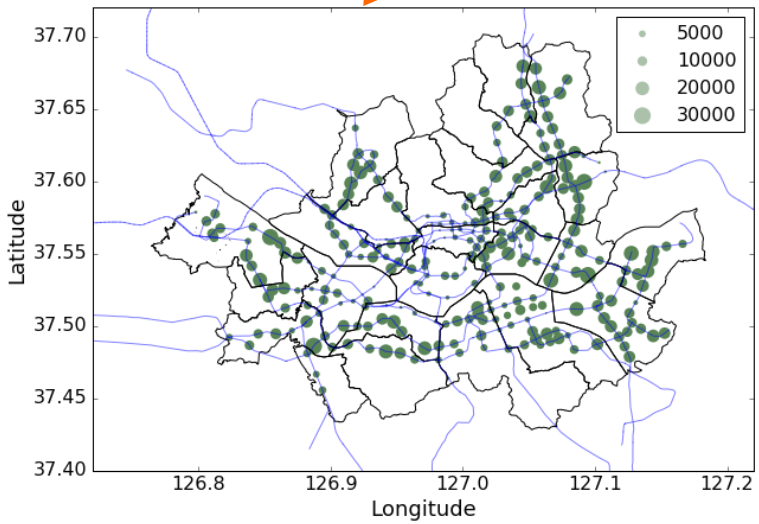


[반대]



직장인구▶





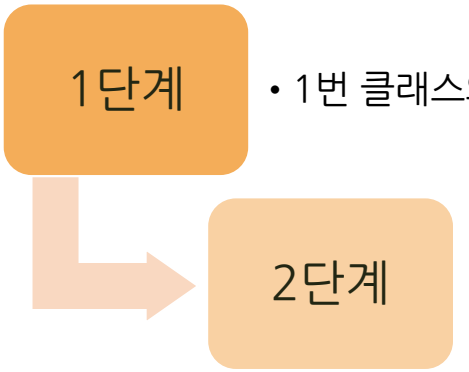
◀거주자수

3.2 분석결과 - 이용패턴 분류 모형 학습

- 군집 분석 결과를 활용하여 지하철 군집과 택시 군집의 조합으로 종속 변수 설정
  - 지하철은 군집 수가 2일 때, 택시는 군집 수가 3일 때를 기준
    - 택시의 이용패턴은 지하철에 비해 역별로 다양하게 나타나면 같은 군집 내에서 유사도가 적음
    - 지하철 이용패턴은 대부분은 통근 시간에 peak 나타나는 형태로 양분 가능

종속변수	1	2	3	4	5	6
지하철	C_1	C_1	C_1	C_2	C_2	C_2
택시	C_1	C_2	C_3	C_1	C_2	C_3
역 개수	121	28	20	69	15	23

  - 1번 클래스에 속하는 역이 전체의 약 44% 차지
    - 6개의 클래스를 하나의 모형으로 예측하려고 할 때 1번 클래스 이외에는 잘 예측하지 못할 가능성 존재



- 1번 클래스와 나머지 클래스를 분류하는 모형 학습
- 1번을 제외한 나머지 다섯 클래스를 분류하는 모형 학습

3.2 분석결과 - 이용패턴 분류 모형 학습

- 설명 변수는 '집계구별 인구·가구·주택·사업체 통계' 및 '추정매출액' 데이터로부터 도출
  - 집계구 데이터: 집계구 경계로부터 집계구의 중심이 지하철역 반경 670m 내에 있는 집계구를 추출하여 이용
    산업별 사업체수는 모든 역 주변의 사업체수가 200개 이하라면 제외
    일반항목의 값을 기준별로 비율로 변환한 변수를 추가
  - 추정매출액 데이터: 매출 금액, 매출 건수, 매출 비율 정보 중 매출 건수 이용
    생활밀착형 43개 업종을 외식업, 서비스업, 도소매업으로 기준으로 합산

데이터명	하위분류	변수
집계구 통계	인구	총인구, 평균나이, 인구밀도, 노령화지수, 총부양비, 19세이하 인구, 20대 인구, 30대 인구, 40대 인구, 50대 인구, 60대 이상 인구, 초중고 인구, 대학생(4년제), 대학생(4년제 미만), 대학원생, 미혼 , 기혼자, 사별 및 이혼
	가구	총가구수, 평균가구원수, 자가 가구, 전세 가구, 월세, 무상, 1인가구, 1세대 가구, 2세대 가구, 3세대 이상 가구, 비혈연 가구
	주택	총주택수, 40㎡이하 주택, 40㎡초과~85㎡이하 주택, 85㎡초과 주택, 아파트, 다세대·연립주택·상가주택
	사업체	총사업체수, 총종사자수, 산업별 사업체수
블록 매출 데이터	매출 건수	당월, 주중, 주말, 0~6시, 6~11시, 11~14시, 14~17시, 17~21시, 21~24시, 남성, 여성, 10대, 20대, 30대, 40대, 50대, 60대 이상
	매출 건수 비율	당월, 주중, 주말, 0~6시, 6~11시, 11~14시, 14~17시, 17~21시, 21~24시, 남성, 여성, 10대, 20대, 30대, 40대, 50대, 60대 이상

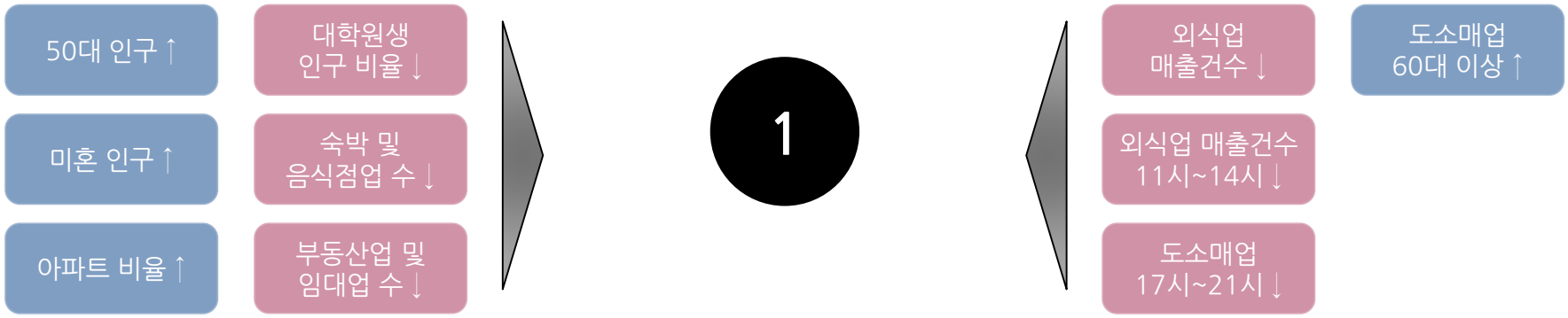
3.2 분석결과 - 이용패턴 분류 모형 학습 결과 (1단계)

- 1번 클래스를 다른 클래스로부터 분류하는 모형

집계구 변수	집계구만 사용	모두 사용	매출 변수	매출만 사용	모두 사용
인구밀도	-0.98	-1.25	외식업 전체	-10.10	-6.57
평균가구원수	0.84		외식업 10대	-1.81	
50대 인구	3.77	5.89	외식업 30대	-40.30	
초중고 인구	1.66	2.13	외식업 40대	1.79	
대학생(4년제) 인구	-2.06	-0.35	외식업 50대	10.10	
미혼 인구	12.3	10.4	외식업 0~6시	-18.10	
자가		-1.68	외식업 11~14시	-26.10	-30.90
무상 가구	-2.21	-2.30	서비스업 전체	-16.70	-2.65
3세대 이상 가구	1.02	0.45	서비스업 20대	-5.39	
다세대·연립주택·상가주택	0.17	1.15	서비스업 30대	3.90	
운수업		-2.59	서비스업 40대	5.89	0.17
숙박 및 음식점업	-25.20	-20.30	서비스업 0~6시	0.58	0.51
부동산업 및 임대업	-7.25	-8.15	서비스업 6~11시	-0.73	
예술, 스포츠 및 여가관련 서비스업	-1.84	-2.98	서비스업 14~17시	-1.55	
협회 및 단체, 수리 및 기타 개인 서비스업	0.41	2.47	도소매업 전체	-5.12	-1.40
대학원생 인구 비율	-12.7	-11.10	도소매업 10대	4.98	1.46
월세 비율	-0.32	-0.16	도소매업 30대	0.75	
40㎡초과~85㎡이하 주택 비율	0.87	0.28	도소매업 50대	6.51	
아파트 비율	6.03	6.16	도소매업 60대 이상	5.33	3.66
			도소매업 0~6시	0.72	1.02
			도소매업 6~11시	0.91	
			도소매업 11~14시	-2.86	
			도소매업 17~21시	-7.69	-3.17
정확도(%)	84.06	84.78		74.98	84.78

3.2 분석결과 - 이용패턴 분류 모형 학습 결과 (1단계)

- 1번 클래스 확률에 영향력이 큰 변수



- 음식점업과 관련된 변수가 1번 클래스(전형적인 주거지)를 분류하는 모형에 큰 영향력을 끼침
  - 숙박 및 음식점업 수, 외식업 매출 건수, 11시~14시 사이의 외식업 매출건수가 낮을 수록 1번 클래스로 분류
- 인구, 가구, 주택 중에서는 인구 변수가 모형에 큰 영향력을 끼침
  - 50대 인구나 미혼 인구가 많을 수록, 대학원생 인구 비율이 낮을수록 1번 클래스로 분류
  - 미혼 인구수는 연령에서는 20대 인구수, 세대구성별 가구에서는 2세대 가구수와 0.91, 0.90의 상관계수를 보여 가족단위의 인구가 높은 곳이 1번 클래스일 확률이 높음
- 주택 구성에서는 아파트 비율이 높을 수록 1번 클래스

3.2 분석결과 - 이용패턴 분류 모형 학습 결과 (2단계)

- 1번 클래스 이외의 클래스를 분류하는 모형
  - 집계구 변수

변수	2	3	4	5	6
평균 나이					-7.36
인구밀도	-3.64		2.64	-1.50	
30대 인구			-2.94		
40대 인구	-2.87		9.31		
초중고 인구					-9.43
대학생(4년제) 인구	-3.00			-1.02	
대학원생 인구		2.60			
미혼 인구			0.15		
자가 가구	7.53				
전세 가구			4.24	-14.30	
임대 가구			0.59		
무상 가구	-9.13		9.03	-0.72	-0.17
3세대 이상 가구	-1.12		-1.83		
비혈연 가구		1.10			3.17
40㎡이하 주택	-2.70		1.19	-1.30	
40㎡초과~85㎡이하 주택			-5.55		
아파트			-4.92		
건설업	-3.66		3.97	-0.42	
운수업		3.21		-9.93	
도매 및 소매업					12.80
숙박 및 음식점업		3.60		1.02	1.39
금융 및 보험업	-3.59	-8.10	-5.65	4.25	-12.80
부동산업 및 임대업		-3.25			3.97
예술, 스포츠 및 여가관련 서비스업	0.09		-13.10		-0.87
협회 및 단체, 수리 및 기타 개인 서비스업		1.24			-4.35

3.2 분석결과 - 이용패턴 분류 모형 학습 결과 (2단계)

- 1번 클래스 이외의 클래스를 분류하는 모형
  - 집계구 비율 변수

변수	2	3	4	5	6
미혼 인구 비율				3.57	
기혼 인구 비율				-2.47	-1.65
자가 가구 비율	3.42				-1.65
전세 가구 비율			9.02		
임대 가구 비율	-7.03		-3.91	10.10	
40㎡이하 주택 비율		-1.25			
40㎡초과~85㎡이하 주택 비율			0.46		
85㎡초과 주택 비율			-3.89		
아파트 비율		-4.26			-0.16
다세대·연립주택·상가주택 비율	4.46			-2.28	1.58
숙박 및 음식점업 비율			-3.58	1.64	6.70

  - 매출건수 변수

외식업 전체	-3.19		-1.32		9.82
외식업 여성				-5.69	
외식업 10대		-0.83		4.40	-8.26
외식업 0~6시	4.31	0.66		-0.70	
외식업 21~24시		6.02	-7.96	-3.02	
서비스업 여성			1.92		
서비스업 남성			-2.40	2.06	
서비스업 20대				-6.49	3.57
서비스업 30대				0.21	-1.26
서비스업 6~11시			2.17		
서비스업 21~24시	2.20		-0.65		
도소매업 전체 매출		-1.75	-0.19	-1.99	3.14
도소매업 10대	9.93	-7.17			
도소매업 30대	-4.74				
도소매업 60대 이상	-7.91		4.51	-2.22	-2.78
도소매업 0~6시	-0.92	-2.49			3.02
도소매업 14~17시				1.46	-0.34
도소매업 21~24시		1.57	-1.91		

3.2 분석결과 - 이용패턴 분류 모형 학습 결과 (2단계)

2 지하철: C_1 택시: C_2	• 자가 가구 ↑ • 다세대·연립주택·상가 주택 ↑ • 임대 가구 비율 ↓
3 지하철: C_1 택시: C_3	• 외식업 21시~24시 매출건수 ↑ • 금융 및 보험업 ↓ • 도소매업 10대 매출건수 ↓
4 지하철: C_2 택시: C_1	• 40대 인구 ↑ • 전세 가구 및 전세 가구 비율 ↑ • 외식업 21시~24시 매출건수 ↓
5 지하철: C_2 택시: C_2	• 임대 가구 비율 ↑ & 전세 가구 ↓ • 금융 및 보험업 수 ↑ • 외식업 여성 매출건수 ↓
6 지하철: C_2 택시: C_3	• 숙박 및 음식점업 수 & 비율↑ • 외식업 매출 건수 ↑ • 초중고 인구 ↓

3.2 분석결과 - 이용패턴 분류 모형 학습 결과

- 종속변수 클래스별 지하철역 인근 사업체 중 도소매업과 숙박 및 음식점업의 비중(%)*

업종	1	2	3	4	5	6
도소매업	33.83	36.95	30.55	34.95	32.93	37.47
숙박 및 음식점업	22.09	24.15	28.21	18.14	19.44	21.81

- 클래스별 목적별 택시 이용 비율(%)**

이동 목적	1	2	3	4	5	6
통근	31.59	32.55	33.90	32.17	31.30	33.24
통학	4.71	3.50	5.07	4.15	3.69	3.94
가정기반	42.77	45.24	40.91	39.50	39.42	37.97
비가정기반	20.91	18.91	20.11	24.17	25.58	24.85
종사자/거주인구	0.26	0.27	0.47	1.24	4.33	2.31

- 인구 대비 주간인구와 목적별 이용 비율 사이의 상관계수**

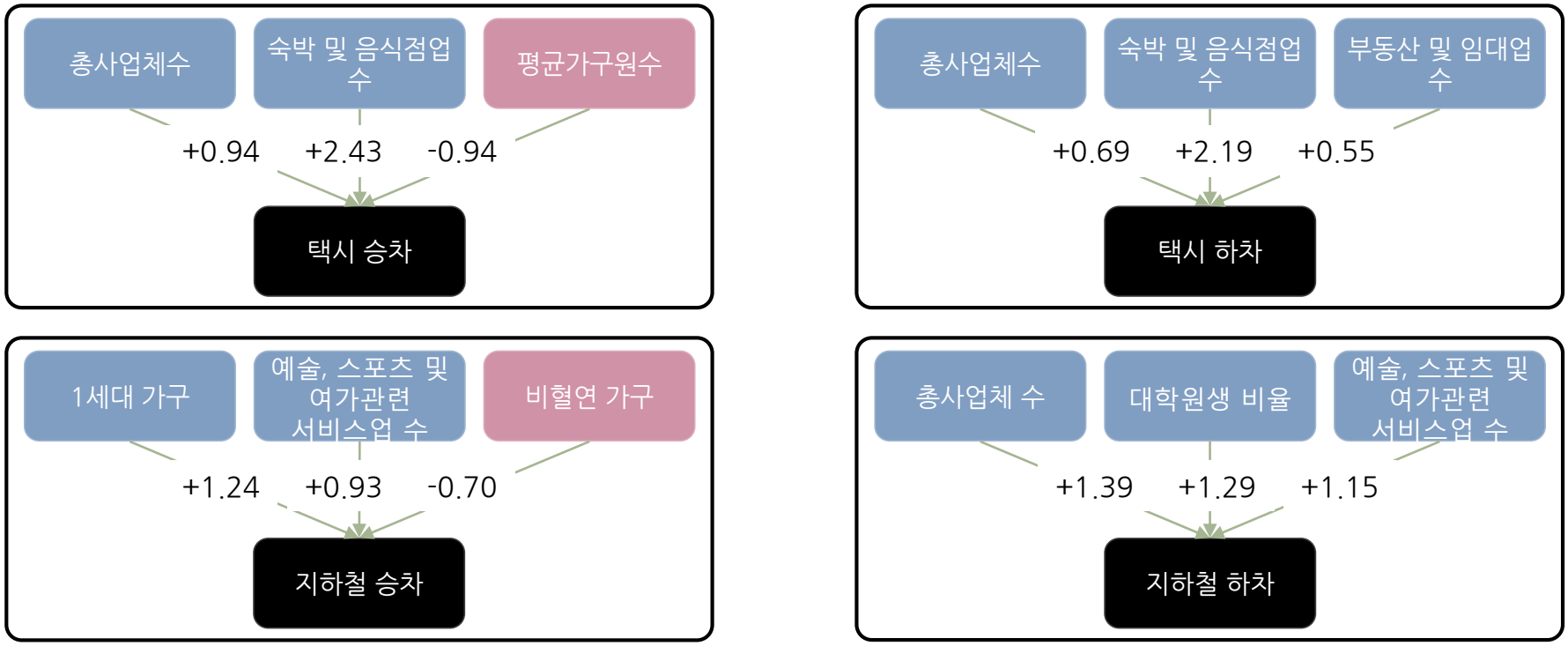
수단	이동 종류	통근	통학	가정기반	비가정기반
택시	전체	0.04	-0.29	-0.53	0.70
	구 내 이동	0.05	-0.37	-0.66	0.78
	다른 구 이동	-0.13	0.24	-0.20	0.37
지하철	전체	-0.26	-0.33	-0.26	0.86

* 2015년 사업체 통계 이용
** 2010년 가구통행실태조사 이용

3.2 분석결과 - 이용빈도 예측 모형 학습 결과

- 종속변수 설정 - 이용빈도는 정규분포를 따르지 않아 log를 취함
  - 지하철: 오전 6시에서 10시까지의 탑승, 하차
    - 지하철의 이용패턴은 크게 두 가지로 나눌 수 있으며 특히 출근 시간대에 정형화된 패턴을 보임
  - 택시: 전체 시간대에서의 탑승, 하차

[모형별 계수의 절대값 상위 3개 변수]



4. 결론 및 제언

- 지하철역을 중심으로 택시와 지하철의 이용패턴을 서로 비교
  - 군집 분석을 활용하여 승·하차 차이 벡터를 표현되는 이용패턴 중에서 대표적인 이용패턴을 추출
    - 지하철은 이용패턴이 다양하지 않고 크게 양분 가능: 오전 출근시간대에 탑승이 많은 지역, 하차가 많은 지역
    - 택시는 지하철에 비해 더 다양한 패턴을 보이고 오전 출근시간대에 강한 피크를 나타내지 않으며 인접 지하철역간의 이용패턴의 유사도도 더 낮음

- 이용패턴에 영향을 주는 요인을 파악하기 위해 분류 분석을 수행
  - 지하철과 택시의 군집 분석 결과를 이용해서 분석 모형의 종속 변수를 설정
    - 지하철 군집 C_1, C_2와 택시 군집 C_1, C_2, C_3의 조합

✓ 지하철 이용패턴이 주거밀집 지역을 더 잘 대변
→ 통근, 통학과 같이 매일 반복되는 이동에 지하철을 이용하는 빈도가 높기 때문

✓ 지하철 이용패턴은 주거지의 특징을 보이나 택시는 반대인 경우 실제로 그 지역은 주거지에 속하는 경우가 많으며 다른 주거지역보다 도소매업의 비중이 높음
→ 단거리 택시 이동에서 이동 목적은 주거지역에서는 가정기반목적인 경우가 많아 이런 지역은 주거지역 내에서 소비, 유흥의 중심지로 역할함을 유추 가능

✓ 지하철 이용패턴은 비즈니스 지구의 특징을 보이나 택시는 반대인 경우 실제로 그 지역은 비즈니스 지구에 속하는 경우가 많음
→ 그런 지역에서 출근 시간대의 탑승이 많은 것은 다른 대중교통과 연계해 통근에 택시를 이용하는 가능성을 시사

✓ 지하철과 택시 양쪽 오전에 탑승이 많은 지역과 아닌 지역을 판별하는 데 중요한 요인은 숙박 및 음식점업체 수, 외식업 매출 건수 등 음식점업과 관련된 요인
→ 점심 시간에 외식 건수와 음식점 수가 적으면 오전에 탑승이 많은 지역 (주거지의 특징)

4. 결론 및 제언

- 출근 시간대에 택시만 단독으로 이용하는지 여부를 확인하여 출근의 편의성 증대를 위해 출근 시간대에 대중교통과의 연계 방안을 검토
  - 출근 시간에 지하철 탑승이 많은데 택시 하차가 많은 지역이나 지하철 하차가 많은데 택시 하차가 많은 지역을 심층적으로 분석
- 지하철 이용내역만으로 알기 어려운 지역의 특성을 택시 이용패턴을 통해 세분화해 분류 가능
  - 택시는 통근, 통학 이외의 목적으로도 많이 활용하기 때문
- 택시 이용패턴의 특징을 설명하는데 요식업 관련 변수가 크게 작용하므로 택시 이용패턴 변화를 통해 요식업이 밀집되어 있는 상권의 이동을 간접적으로 파악
  - 노선이 고정되어 있는 대중교통과 달리 택시는 유동적

참고문헌

- 김관호·오규협·이영규·정재윤. 2013. “스마트카드 빅데이터를 이용한 서울시 지하철 이동패턴 분석.”「한국전자거래학회지」18(3): 211-22.
- 김기혁·이승철. 2011. “교통카드자료를 이용한 환승정류장의 유형별 입지특성에 관한 연구- 대구시를 중심으로.”「대한토목학회논문집」D31(4D): 519-26.
- 성현곤·노정현·김태현·박지형. 2006. “고밀도시에서의 토지이용이 통행패턴에 미치는 영향: 서울시 역세권을 중심으로.”「국토계획」41(4): 59-75.
- 신예철·김태호·장명준. 2013. “도시 및 교통계획의 합리적 연계를 위한 서울시 역세권의 개발특성과 대중교통서비스여건의 영향관계분석.”「한국도시설계학회지 도시설계」, 14(2), 99-111.
- 이연수·손동욱. 2012. “역세권의 적정 공간범위 설정 방법론을 통한 지하철 이용수요와 역세권의 도시공간구조간의 연관성 분석.”「한국도시설계학회지 도시설계」, 13(4), 23-32.
- 이연수·진창종·추상호. 2012. “공간계량분석을 이용한 대중교통 이용에 영향을 미치는 공간적 특성요인 분석에 관한 연구- 서울시 행정동을 중심으로.”「서울도시연구」13(4): 97-111.
- 임용택·박철·김대성·엄진기·이준. 2012. “교통카드 데이터를 이용한 통행분포 모형 추정.”「교통연구」19(2): 1-11.
- Cortes, Corinna, and Vladimir Vapnik. 1995. “Support-Vector Networks.” Machine Learning 20(3): 273-297.
- Cox, David R. 1958. “The Regression Analysis of Binary Sequences (with Discussion).” Journal of the Royal Statistical Society: Series B 20: 215-242.
- Kamga, Camille, M. Anil Yazici, and Abhishek Singhal. 2015. “Analysis of Taxi Demand and Supply in New York City: Implications of Recent Taxi Regulations.” Transportation Planning and Technology 38(6): 601-625..
- Lee, SI, H Lee, P Abbeel, and AY Ng. 2006. “Efficient L~ 1 Regularized Logistic Regression.” In Proceedings of the National Conference on Artificial Intelligence, 401-408.
- Liu, Yu, Fahui Wang, Yu Xiao, and Song Gao. 2012. “Urban Land Uses and Traffic ‘source-Sink Areas’: Evidence from GPS-Enabled Taxi Data in Shanghai.” Landscape and Urban Planning 106(1): 73-87.
- MacQueen, J B. 1967. “Some Methods for Classification and Analysis of Multivariate Observations.” 5th Berkeley Symposium on Mathematical Statistics and Probability 1(233): 281-297.
- Martin, J Kent, and D S Hirschberg. 1996. “On the Complexity of Learning Decision Trees.” International Symposium on Artificial Intelligence and Mathematics: 112-15.
- Peng, Chengbin et al. 2012. “Collective Human Mobility Pattern from Taxi Trips in Urban Area.” PLoS ONE 7(4): e34487.
- Salvador, Stan, and Philip Chan. "Toward accurate dynamic time warping in linear time and space." Intelligent Data Analysis 11, no. 5 (2007): 561-580.
- Quinlan, J. R. 1986. “Induction of Decision Trees.” Machine Learning 1(1): 81-106.
- Rish, I. 2001. “An Empirical Study of the Naive Bayes Classifier.” IJCAI 2001 workshop on empirical methods in artificial. 3(2): 41-46.
- Veloso, Marco, and Santi Phithakkitnukoon. 2011. “Exploratory Study of Urban Flow Using Taxi Traces.” In First Workshop on Pervasive Urban Applications (PURBA) in Conjunction with Pervasive Computing, San Francisco, California, 1-8.
- Zheng, Zhong, Soora Rasouli, and Harry Timmermans. 2016. “Two-Regime Pattern in Human Mobility: Evidence from GPS Taxi Trajectory Data.” Geographical Analysis 48: 157-175.

사용 데이터

- 가구통행실태조사 자료, 수도권교통본부, 2010
- 대중교통 환승데이터(교통카드 거래내역), 서울시(스마트카드사), 2016.04.01.~2016.04.07
- 도로명주소기본도, 행정자치부, 2015
- 서울시 상주 (야간) ·주간인구 (12세 이상) 통계, 통계청, 2010
- 서울시 지하철 호선별 역별 시간대별 승하차 인원 정보, 서울시, 2015
- 주민등록인구(동별), 서울시 통계데이터담당관, 2010
- 집계구별 인구통계, 통계청, 2010
- 집계구별 가구통계, 통계청, 2010
- 집계구별 주택통계, 통계청, 2010
- 집계구별 사업체통계, 통계청, 2015
- 추정매출액(블록단위), 서울시, 2015.01~2015.05
- 택시운행 통행량(택시운행분석데이터), 서울시, 2015