군집분석을 통한 공공자전거 수요예측

군집분석을 통한 공공자전거 수요예측

이창환, *김경옥

서울과학기술대학교 데이터사이언스학과, 서울과학기술대학교 글로벌융합산업공학과

lch1181@ds.seoultech.ac.kr, kyoungok.kim@seoultech.ac.kr

목차

1. 서론

□ 연구배경

- 서울시 공공자전거 시스템은 2015년 10월 성수, 상암 등 5곳에서 2000여 대로 시작하여 3년 차인 2018년 11월 기준, 서울 전국에 1697개의 자전거 대여소와 2만여 대의 자전거 대수로 성장.
- 이용건수는 2015년(9~12월) 11만 건에서 2017년 503만 건으로 급증.

  → 유럽, 중국 등 세계적으로 수요가 늘고 있는 추세. [4] Shaheen 외, 2011, [2] Midgley 외, 2009
- 공공자전거 시스템에 대한 기존 연구.
  - 공공자전거 시스템의 수요 요인 및 이용특성 분석 [5] García-Palomares외, 2012, [8] 노윤승 외, 2014 [12] 장재민 외, 2016
  - 공공자전거 시스템의 정류소의 타입에 대한 연구. [9] Junming Liu 외, 2015
  - 공공자전거 시스템의 수요 예측 [5] Faghih-Imani 외, 2014 [13] 민지원 외, 2017
- 수요 예측 연구는 자전거 정류소의 대여 개수를 예측하여, 자전거 정류소들에 최적의 자전거 개수를 배치할 수 있도록 만들어 공공자전거 시스템의 효율성을 증진.

1. 서론

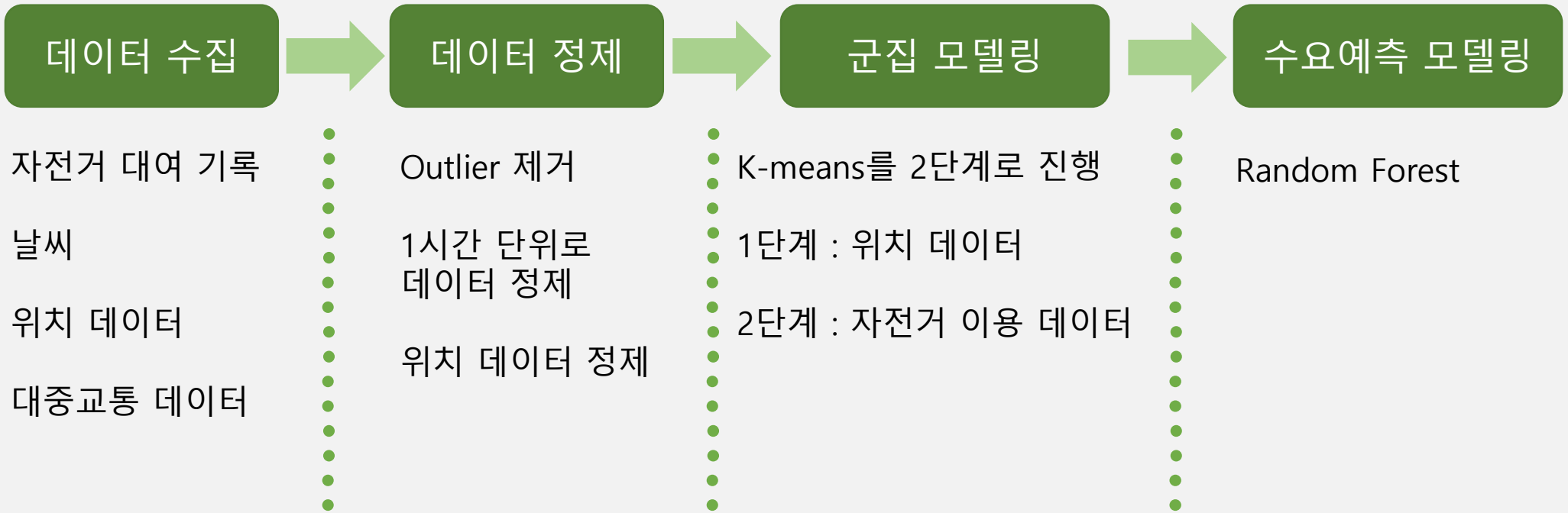
□ 연구배경

- 수요예측은 군집분석을 통해 정류소들의 최적의 군집을 찾아 군집화 한 후 군집 단위로 할 때 정류소별 예측에 비해 수요예측의 변동성을 낮춤으로써 수요예측의 정확도를 높임.
- 군집분석은 군집을 형성하는 기법과 이용하는 데이터에 따라 차이 발생.
  - 자전거 이용 데이터를 이용한 시계열 군집 분석 [6] Etienne 외, 2014 [15] 김민혁 2018
  - 자전거 이용 데이터, 위치 데이터를 이용하며, 반복적인 군집분석을 통해 최적 군집을 찾는 방법. [10] *Li, Yexin 외, 2015, [14] Jia 외, 2018, [16] Jia 외, 2019*
  - 자전거 이용 데이터, 환경적 특성 등을 이용한 군집 분석. *[11] Chen 외, 2016 , [17] 이다영 2019*
- 위 방법들은 아래와 같은 한계점 존재.
  - 시계열 군집 분석이나 환경적 특성을 이용하여 군집 분석하는 기존의 논문은 일별 시간이 아닌 일별 또는 월별 시간 당 수요량을 예측.
  - 반복적인 군집분석의 경우 계속 반복해도 최적의 군집을 못 찾거나 군집분석 시 최초의 값에 따라 결과가 달라짐.
  - 정류소 간 거리가 달라서 군집 시 이를 고려해야 하나 공통된 거리 이상 떨어진 정류소 간에는 군집이 거의 불가능 함.
- 이번 연구를 통해 위의 한계점을 극복하여 서울시 공공자전거 시스템의 효율성을 증진시킴.

2. 본론

□ 연구방법

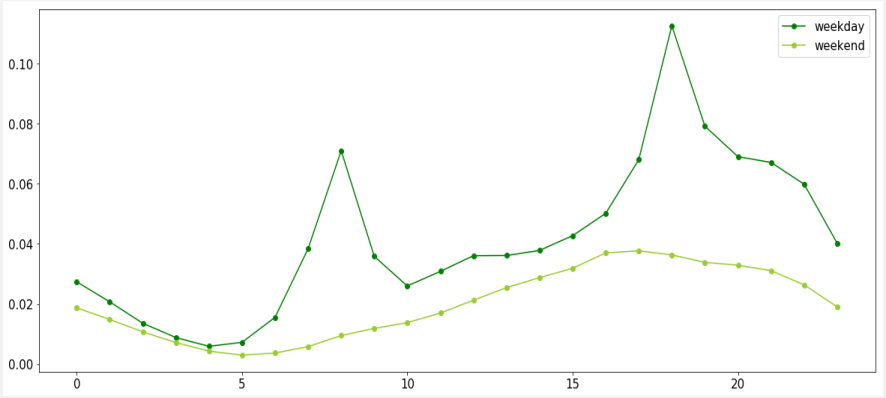
- FrameWork



2. 본론

□ 연구방법

- 군집 모델링
  - 특징
    - K-means 모델 사용.
      - Initial Centroid 값을 데이터 특성에 맞게 설정함으로써 예측력이 높아지는 군집을 형성.
    - Input data : 위도, 경도 데이터와 자전거 이용 데이터.
      - 위도, 경도 데이터와 자전거 이용 데이터를 한번에 군집하면 이용량은 전혀 다르나 위도, 경도가 비슷한 곳끼리 묶이는 현상 가능성 존재. → 2단계 군집화 방법을 적용.
    - 평일과 주말에 자전거 이용패턴이 다르기 때문에 평일, 주말을 기준으로 2개 모델로 나누며, 평일을 기준으로 모델링 함.

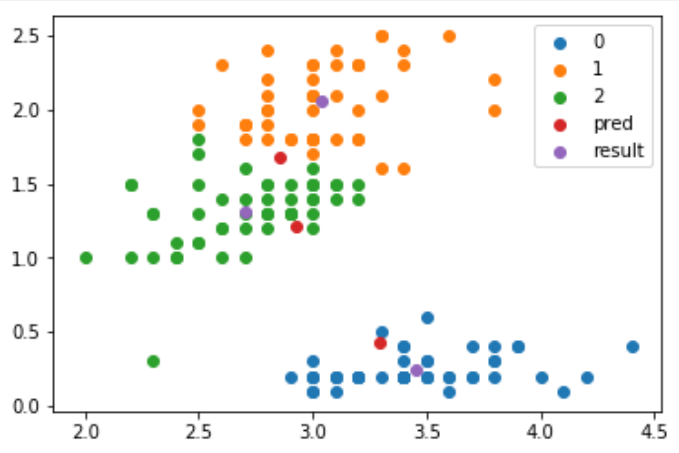


2. 본론

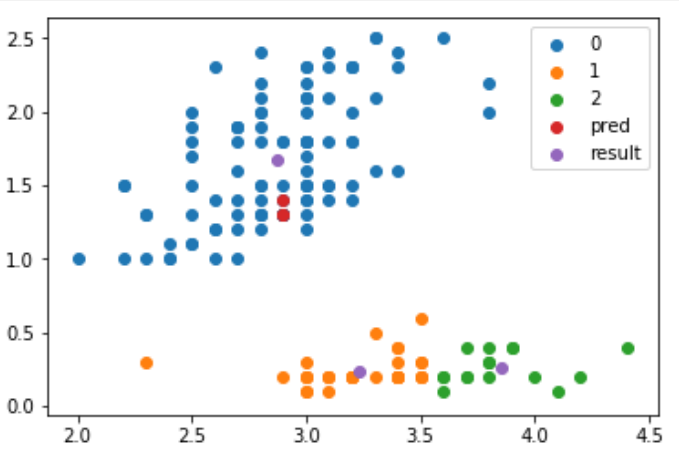
□ 연구방법

- 군집 모델링
  - Initial Centroid 정의
    - K-means의 Initial Centroid는 K-means에서 Input data로 사용하는 값들의 밀도 [1] Nazeer 외, 2009, 중간값 [3] Hae-Sang Park 외, 2009 등으로 구함.
    - 밀도 기준으로 하면 데이터가 서로 많이 떨어진 경우, 중앙값은 데이터가 서로 가까운 경우 군집을 잘할 것으로 추정.

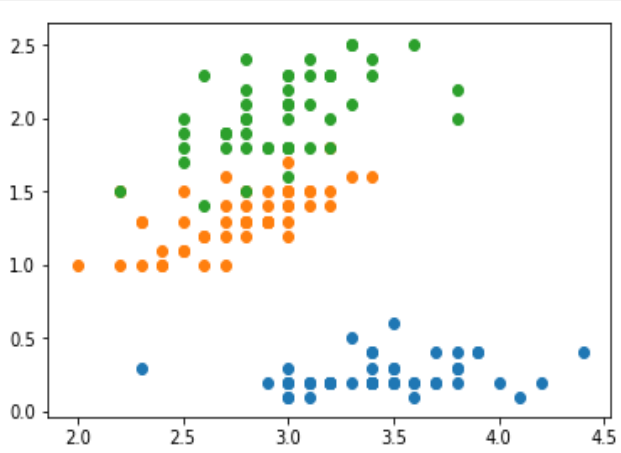
밀도 기준



중앙값 기준



정답

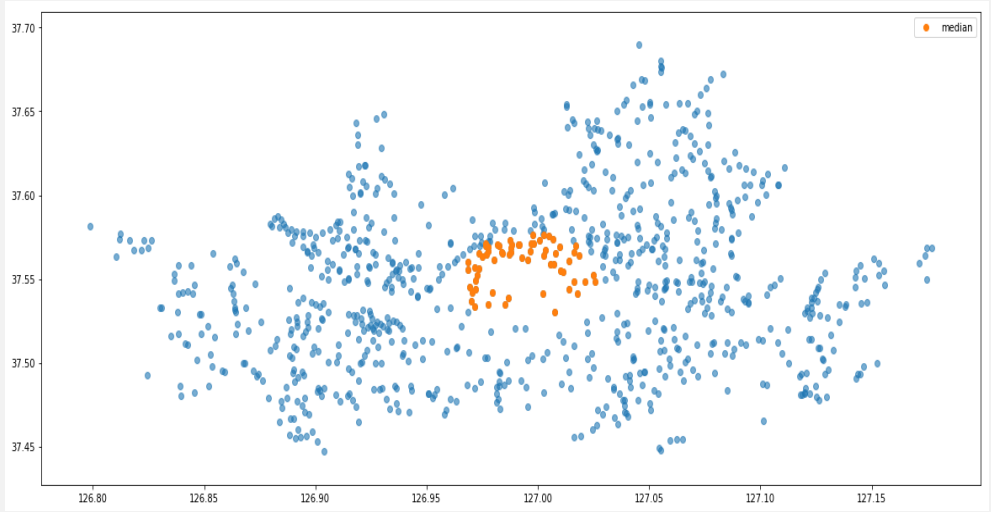


2. 본론

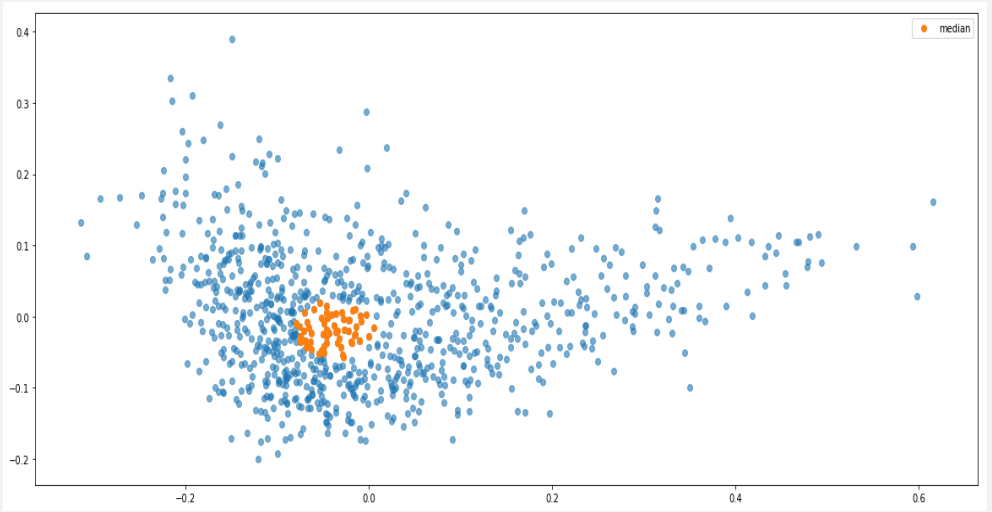
□ 연구방법

- 군집 모델링
  - Initial Centroid 정의
    - 위도,경도 데이터와 PCA를 적용한 대여, 반납 데이터를 보면 데이터 분포가 동떨어지지 않고 가운데 부근에서 퍼져 나가는 것을 볼 수 있기 때문에 중앙값을 기준으로 Initial Centroid 값을 설정하는 방법 이용.

위,경도



대여,반납 데이터



2. 본론

□ 연구방법

- 군집 모델링
  - Initial Centroid 정의
    - 알고리즘
      - 각 정류소의 위도, 경도 데이터를 기준으로 정류소 간 Euclidean 거리 비유사도 행렬(=D)을 구함.
      - $V_j = \sum_{i=1}^{n} \frac{d_{ij}}{\sum_{l=1}^{n} d_{il}}$, j = 1, …, n 계산. 이는 모든 속성을 기준으로 Point(정류소) 간 거리를 구하는 것을 의미.
      - V_j 값이 제일 낮은 값으로 정렬하고 1번째부터 군집화 개수인 K1번째까지 V_j에 해당하는 각각의 Point들을 추출.
  - Step 1 : 위도, 경도 데이터를 이용한 K-means 군집화
    - Input data : 자전거 정류소의 위도, 경도 데이터
    - Initial Centroid 정의 : Input data를 기준으로 앞에서 정의한 방식으로 구함.
    - 위 단계에서 구한 Initial Centroid를 사용하여 Input data 기준, K1개의 K-means로 군집화.

2. 본론

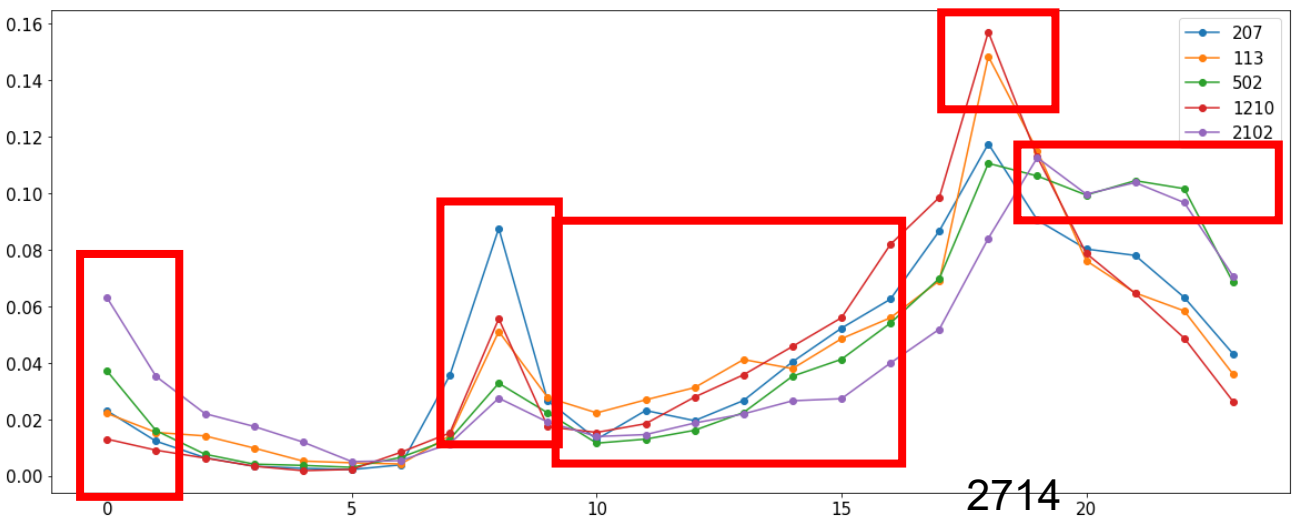
□ 연구방법

- 군집 모델링

  - Step 2 : Step 1의 각 군집 내 정류소의 자전거 대여, 반납 데이터를 24시간 기준, 시간 별로 나눈 뒤 여러 시간을 1개 그룹으로 묶은 'time slot'으로 만들어 Step 1의 각각의 군집에 대해 K-means로 군집화

    - 대여, 반납 시간 : 1년치 데이터를 기준으로 평일, 주말을 0~23시까지 각 시간대의 비율 값을 time slot으로 구성.
      → 출퇴근 시간대 등 특정 시간 그룹에 정류소간 대여 차가 크므로 차이가 큰 부분을 추출하여 군집 변수로 함.

    - 그래프를 보면(평일 기준, 대여 상위 5개 정류소) 빨간 박스 친 부분이 정류소마다 차이가 발생함. 이를 기준으로 오른쪽의 표로 time slot을 만듦.



slot	시간
1	7~9
2	10~16
3	17~19
4	20~23
5	0~6

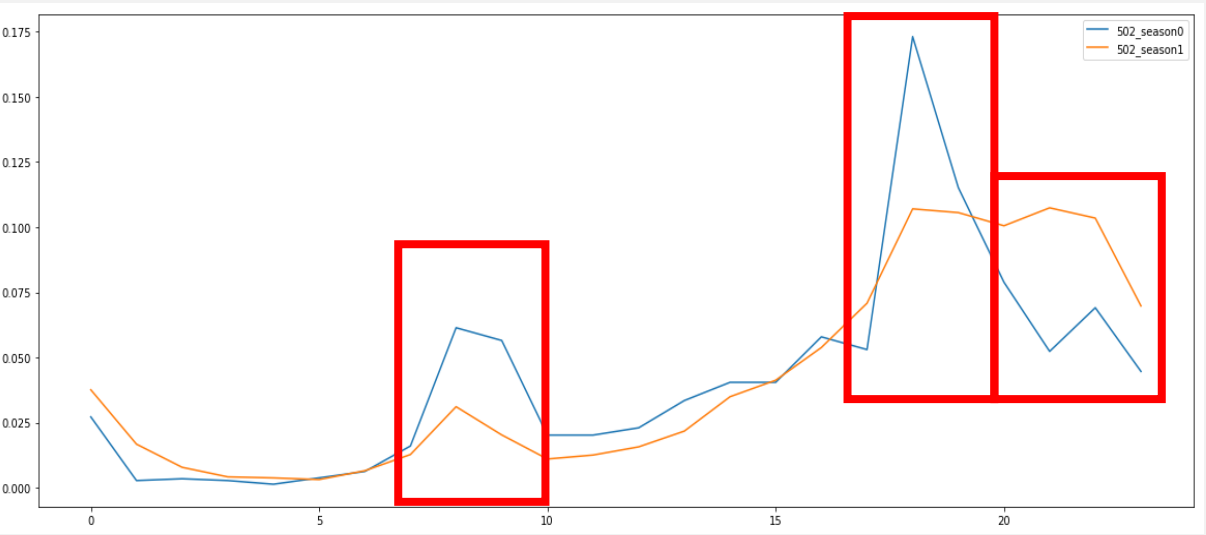
2. 본론

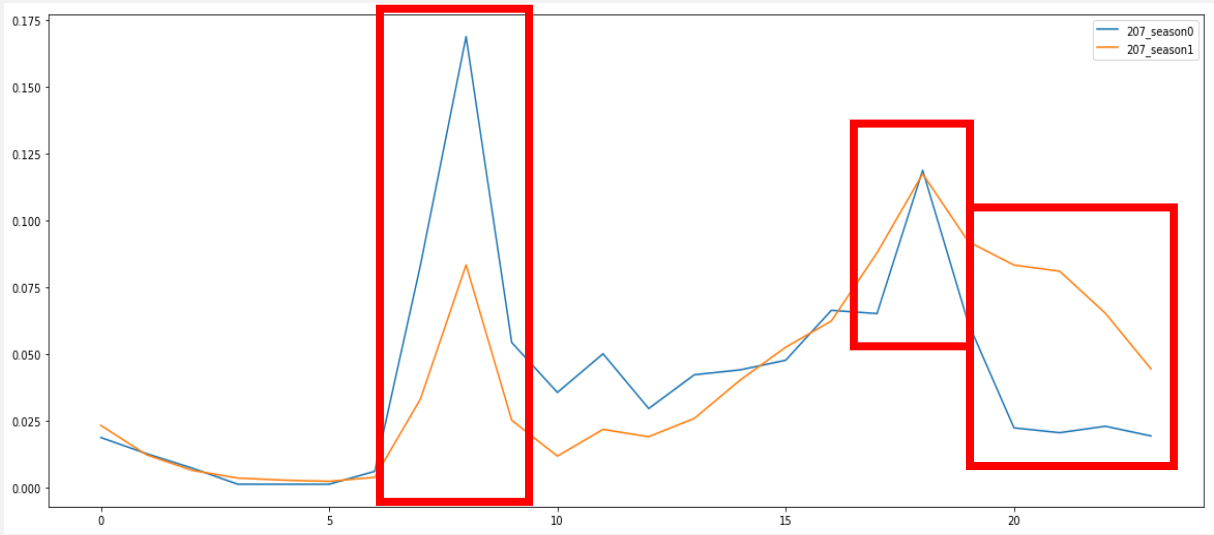
□ 연구방법

- 군집 모델링
  - Step 2
    - 겨울과 겨울이 아닌 계절에 시간대 별로 자전거 타는 비율이 다르기 때문에 12, 1, 2월을 Season0, 3~11월까지 Season1로 time slot을 한 번 더 나눔.

      -> 반납은 똑같은 slot으로 하여, 정류소 당 군집화 할 차원은 대여 10, 반납 10으로 총 20개.

slot	시간
1	(S0) 7~9
2	(S0) 10~16
3	(S0) 17~19
4	(S0) 20~23
5	(S0) 0~6
6	(S1) 7~9
7	(S1) 10~16
8	(S1) 17~19
9	(S1) 20~23
10	(S1) 0~6





2. 본론

□ 연구방법

- 군집 모델링
  - Step 2
    - Step 1에서 나눠진 군집끼리 각각 군집화를 진행하여 총 K2개의 군집 형성.
      - 여기서도 Step 1과 같은 방법으로 Initial Centroid를 정의함.
    - 개수 정하는 방법 : Step 1의 각 클러스터를 아래의 공식 개수로 나눔. 예를 들어 첫 번째 클러스터인 C_1은 N개로 나눈 뒤 K2개를 곱한 개수로 K-means 군집화 실행. (N : 총 정류소 개수, CN_1 : C_1 군집에 속한 정류소 수)

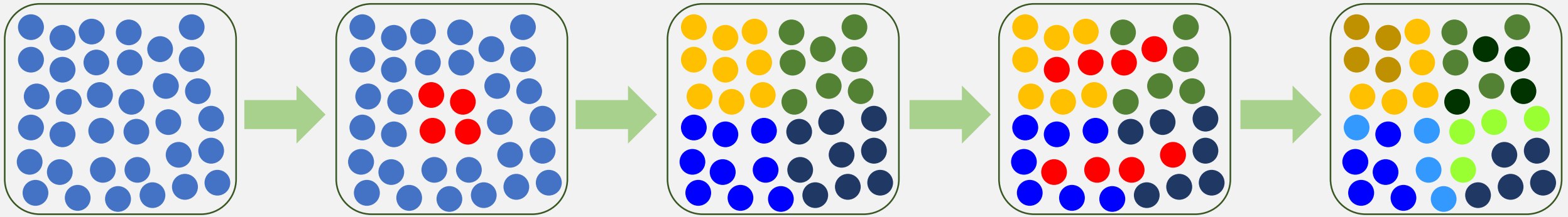
      -> $[\frac{CN_1}{N}]$ x K2, $[\frac{CN_2}{N}]$ x K2, $[\frac{CN_3}{N}]$ x K2, ..., $[\frac{CN_{k1}}{N}]$ x K2

      - [] : 반올림 기호, 만약 C_i에서 K-means 군집화 할 k 값이 0으로 나오면 k 값이 제일 큰 곳에서 1개를 빼서 해당 군집에 할당하여 군집화 실행.

2. 본론

□ 연구방법

- 군집 모델링
  - FrameWork



군집화 전의 정류소들

Step 1 이전에 초기 Centroid를 찾음

초기 Centroid를 기준으로 Step 1을 실행하여 군집화

Step 2 이전에 초기 Centroid 찾음

초기 Centroid를 기준으로 Step 2를 실행하여 최종 군집을 만듦

2. 본론

□ 실험구성

- 데이터 수집
  - 기간 : 2018년 1월 1일 ~ 2019년 3월 31일
    - Train : 2018년, Test : 2019년
  - 서울시 공공데이터
    - 서울시 공공자전거 이용내역 및 공공자전거 정류소 정보 데이터
    - 전통시장, 대형마트, 공공체육시설 등의 시설 데이터
    - 공원, 강 등 환경 데이터
  - 기상청 : 미세먼지, 강수량, 적설량, 기온 등의 기상 데이터
  - SGIS(통계 지리정보 서비스) : 주거 인구, 종사자 수 등 인구 통계 데이터
  - 서울시(스마트카드사) : 대중교통 환승 데이터
  - 다음 카카오 API : 공공자전거 정류소 주변 카페, 학교 등의 데이터

2. 본론

□ 실험구성

- 데이터 정제
  - 서울시 공공자전거 이용내역 및 정류소 정보 데이터
    - 자전거 이용시간이 4시간 초과시, 분실로 간주됨 -> 4시간 이전 데이터만 이용.
    - 최소 이용시간이 3분 이하 거나, 이동거리가 0m인 경우 제외.
    - 정류소 : 2018년 1월 2일부터 대여 기록이 있는 정류소를 기준으로 함 -> 893개
    - 이용내역 데이터 : 정류소 별로 대여, 반납을 1시간 기준으로 만듦
      ex) 정류소 101, 1월 2일 13시 대여량(13~14시) : 2대
  - 기상청 데이터
    - 온도, 풍속, 습도 : null값을 과거 3시간, 미래 3시간 값을 기준으로 평균값으로 넣음. 일부가 없는 경우 존재하는 값만 기준으로 처리.
    - 강수 : 비가 온 경우 1, 안 온 경우 0으로 처리.
    - 나머지 데이터 : 없는 경우 0으로 처리.

2. 본론

□ 실험구성

- 수요예측 모델링 : Random Forest
  - 설명변수 : 38개
    - 기상 관련 변수 : 먼지, 강수량, 적설량, 일사량, 온도, 풍속, 습도, 운량
    - 환경 변수
      - 자전거 정류소와 가장 가까운 특정 시설과의 거리 : 공원, 하천, 자전거 도로, 도서관, 대학교, 공공 운동시설
      - 자전거 정류소에서 반경 250m 내 특정 시설의 개수 : 카페, 학원, 음식점, 은행, 초중고등학교, 문화시설, 관광명소, 전통시장, 자전거 대여소
    - 인구통계 변수 : 250m 근방 값 기준. 250m 근방의 인구 중 20~40대 비율, 사업체 종사자 수, 사업체 수, 전체 거주자 수, 250m 근방의 전체 가구 중 1인 가구 비율
    - 대중교통 변수 : 일별 특정 시간대의 버스, 지하철 승차 및 하차 객 수
    - 기타 : 자전거 거치대 개수, 계절 변수(봄, 여름, 가을), 시간 변수(새벽, 출근, 낮 시간, 퇴근 시간)
    - 종속변수 : 일별 시간 당 군집에서 대여한 자전거 대수

2. 본론

□ 실험구성

- 수요예측 모델링
  - 비교 모델 : 자전거 정류소의 위경도와 대여 및 반납 데이터 사용
    - K-means을 이용한 군집 (이하 **Bipart-Clus**라 함)
      - 위경도 데이터로 최초 K-means 군집화 이후, 특정 정류소에서 어떤 time slot에 해당 군집으로 반납하는 비율 데이터와 위도, 경도 데이터를 번갈아가며 2단계로 나눠서 수행하는 모델. [7] Li, Yexin 외, 2015
      - K-means의 초기값을 특정 값으로 정하지 않음.
    - Gaussian Mixture Model을 이용한 군집 : 위경도 데이터로 최초 군집화 이후 위경도 데이터와, 특정 정류소에서 어떤 time slot에 해당 군집으로 반납하는 비율 값의 Frobenius norm을 이용하여 Gaussian Mixture Model로 군집화. [] Wenzhen 외, 2019
    - 비교 모델들의 경우 군집을 반복하여 특정 지점에 수렴하거나 L번 실행하는 모델.
      → 계속 값이 변화하면서 수렴하지 않아 100번을 클러스터링 하여 평균값을 기준으로 Metrics 값을 구함.

2. 본론

□ 실험구성

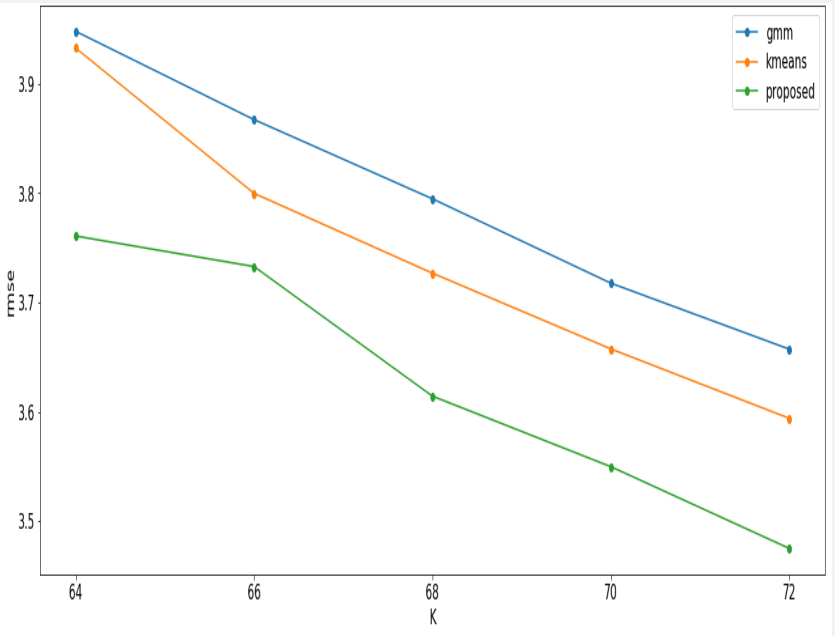
- 평가 기준
  - RMSE : 예측 값과 실제 값의 차이가 큰 쪽에 패널티를 많이 주는 지표 $= \sqrt{\frac{\sum(y-\hat{y})^2}{N}}$
  - RMSLE : 군집별로 오차를 측정하는 지표로 예측 값과 실제 값의 차이가 큰 쪽에 패널티를 작게 줌.

    $= \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\log(\hat{X}_{C_i,t} + 1) - \log(X_{C_i,t} - 1))^2}$
  - ER : 군집별로 오차를 측정하는 지표로 예측 값과 실제 값의 차이에 절대값을 씌움 $= \frac{\sum_{i=1}^{m}\left|\hat{X}_{C_i,t} - X_{C_i,t}\right|}{\sum_{i=1}^{m} X_{C_i,t}}$
- 군집 개수 : 기존 논문에서 제시한 군집 개수와 정류소 개수에 맞춰 64 ~ 72까지 군집 개수를 비교함. (64, 66, …, 72)

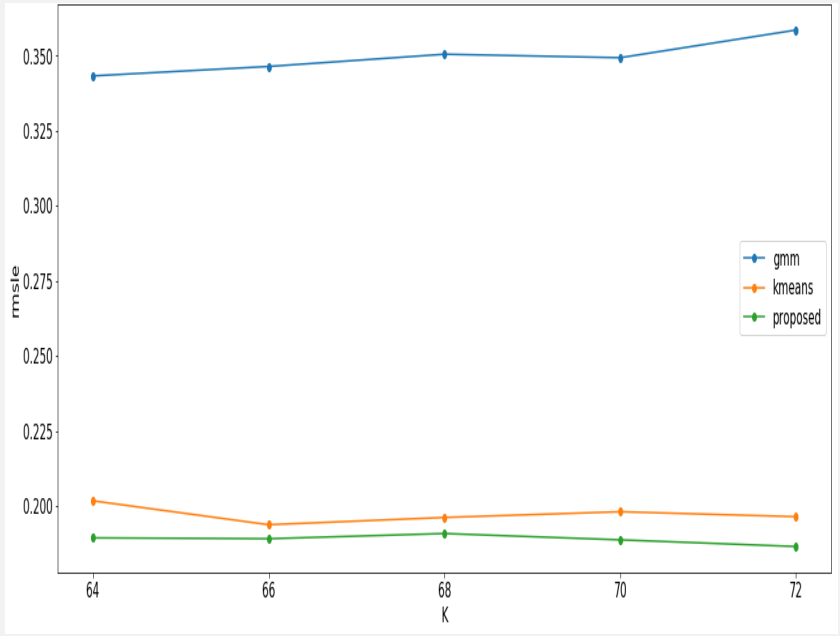
2. 본론

□ 결과

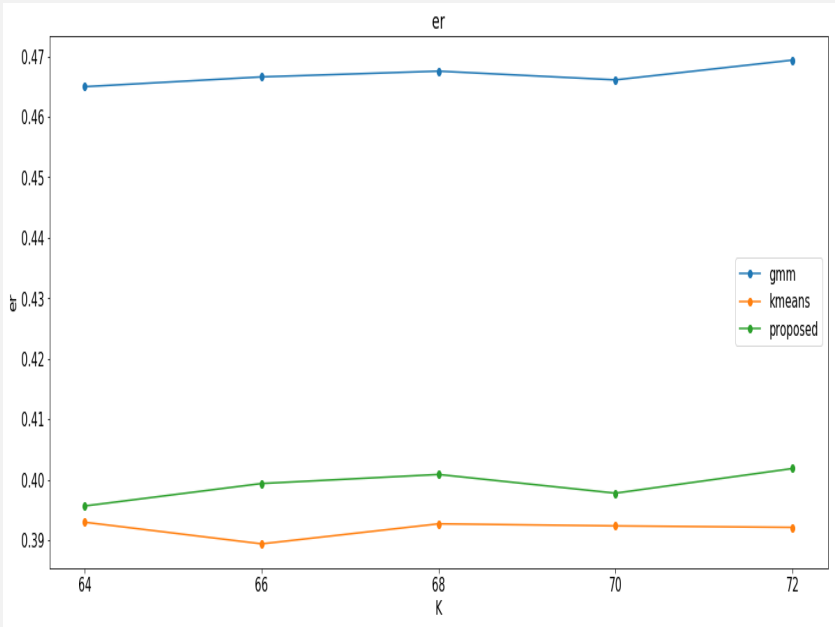
RMSE



RMSLE



ER



2. 본론

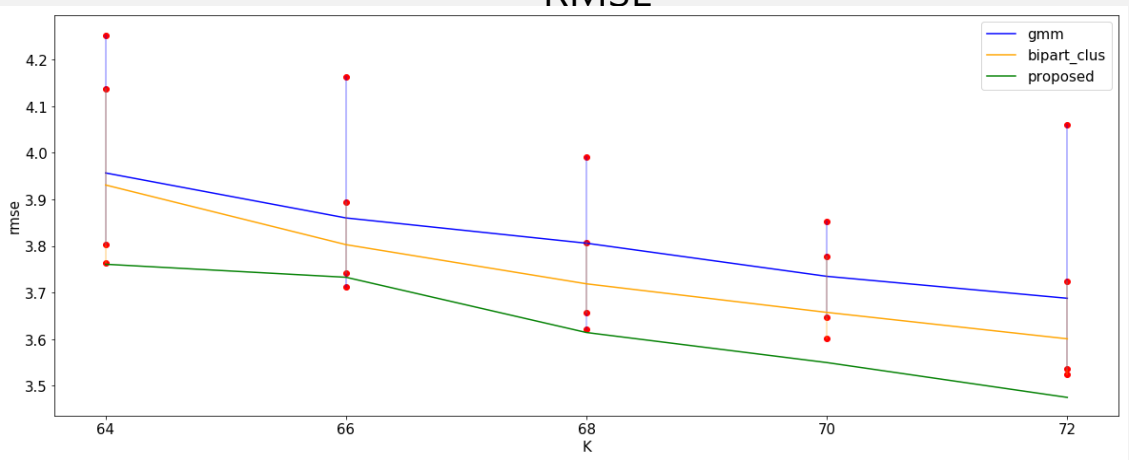
□ 결과

K	RMSE			RMSLE			ER		
	제안한 모델	Bipart-Clus	GMM	제안한 모델	Bipart-Clus	GMM	제안한 모델	Bipart-Clus	GMM
64	**3.761**	3.933(0.072)	3.948(0.083)	**0.189**	0.202(0.007)	0.343(0.019)	0.396	**0.393(0.005)**	0.465(0.006)
66	**3.733**	3.800(0.037)	3.867(0.078)	**0.189**	0.194(0.010)	0.346(0.020)	0.399	**0.389(0.005)**	0.467(0.007)
68	**3.614**	3.727(0.037)	3.795(0.070)	**0.191**	0.196(0.008)	0.350(0.016)	0.401	**0.393(0.005)**	0.468(0.005)
70	**3.550**	3.658(0.032)	3.718(0.071)	**0.189**	0.198(0.010)	0.349(0.012)	0.400	**0.392(0.006)**	0.466(0.005)
72	**3.475**	3.594(0.042)	3.657(0.097)	**0.186**	0.196(0.008)	0.359(0.017)	0.402	**0.392(0.006)**	0.469(0.006)

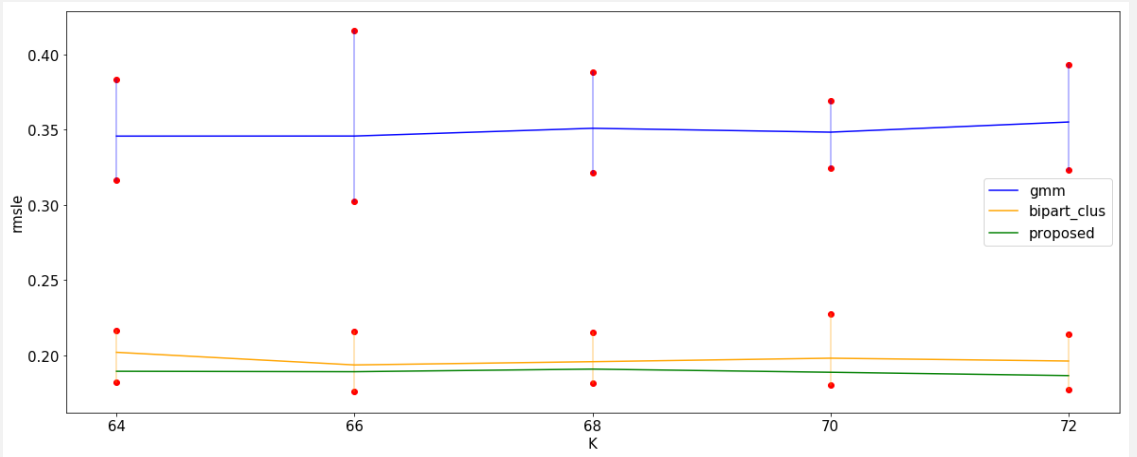
2. 본론

□ 결과

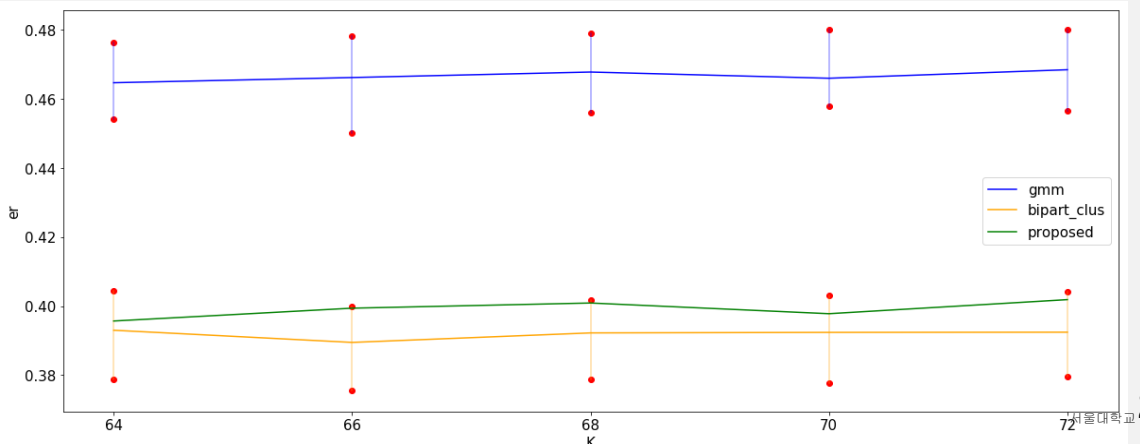
RMSE



RMSLE



ER



3. 결론

□ 한계점

- 날씨 예보와 대중교통 이용객 추정을 활용하여 검증하지 못함.
- 평일 데이터만 분석.

□ 기여점

- 자전거 정류소를 군집할 때 기존 논문에 비해 군집 되는 속도가 훨씬 빠름.
  - 비교 방식 : Bipart-Clus의 경우 100번 기준, 약 27분(iteration 1회 기준), GMM의 경우 100번 기준, 약 80분(iteration 1회 기준)
  - 제안된 방식 : **4분**
- 비교 방식에 비해 RMSE, RMSLE가 개선됨.
- 일별, 시간별 예측을 함으로써 기존의 월별, 시간별로 예측한 논문에 비해 결과값을 더 많은 곳에 활용 가능.

4. 참고문헌

[1] Nazeer, KA Abdul, and M. P. Sebastian. "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm." Proceedings of the world congress on engineering. Vol. 1. London: Association of Engineers, 2009.

[2] Midgley, Peter. "The role of smart bike-sharing systems in urban mobility." *Journeys* 2.1 (2009): 23-31.

[3] Park, Hae-Sang, and Chi-Hyuck Jun. "A simple and fast algorithm for K-medoids clustering." Expert systems with applications 36.2 (2009): 3336-3341.

[4] Shaheen, Susan A., et al. "China's Hangzhou public bicycle: understanding early adoption and behavioral response to bikesharing." *Transportation Research Record* 2247.1 (2011): 33-41.

[5] García-Palomares, Juan Carlos, Javier Gutiérrez, and Marta Latorre. "Optimizing the location of stations in bike-sharing programs: A GIS approach." Applied Geography 35.1-2 (2012): 235-246.

[6] Etienne, Côme, and Oukhellou Latifa. "Model-based count series clustering for bike sharing system usage mining: a case study with the Vélib'system of Paris." ACM Transactions on Intelligent Systems and Technology (TIST) 5.3 (2014): 39.

[7] Faghih-Imani, Ahmadreza, et al. "How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (BIXI) in Montreal." Journal of Transport Geography 41 (2014): 306-314.

[8] 노윤승, and 도명식. "대전시 공공자전거 이용패턴 분석 및 이용수요예측." *2014 년도 봄학술대회 논문집*. 2014.

[9] Liu, Junming, et al. "Station site optimization in bike sharing systems." 2015 IEEE International Conference on Data Mining. IEEE, 2015.

[10] Li, Yexin, et al. "Traffic prediction in a bike-sharing system." Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2015.

4. 참고문헌

[11] Chen, Longbiao, et al. "Dynamic cluster-based over-demand prediction in bike sharing systems." Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, 2016.

[12] 장재민, 김태형, and 이무영. "서울시 공공자전거 이용특성에 관한 연구: 여의도 및 상암 지구를 사례로." *서울도시연구* 17.4 (2016): 77-91.

[13] 민지원, 문현수, and 이영석. "랜덤 포레스트를 이용한 대전시 공공 자전거 ('타슈') 수요 예측." *한국정보과학회 학술발표논문집* (2017): 969-971.

[14] Jia, Wenzhen, Yanyan Tan, and Jing Li. "Hierarchical Prediction Based on Two-Level Affinity Propagation Clustering for Bike-Sharing System." IEEE Access 6 (2018): 45875-45885.

[15] 김민혁. 시계열 군집분석 기반 서울시 공공자전거 수요예측. Diss. 한양대학교, 2018.

[16] Jia, Wenzhen, et al. "Hierarchical prediction based on two-level Gaussian mixture model clustering for bike-sharing system." Knowledge-Based Systems 178 (2019): 84-97.

[17] 이다영. "시공간적 패턴에 기반한 서울시 공공자전거 수요예측." (2019)