붙임 3 **분석 결과보고서 작성 양식**

접수번호	※작성하지 않음

「통계빅데이터 자료분석·활용대회」 분석 결과보고서

신 청 자 명	고명지	공동연구자	박정민, 유채연
분석주제명	서울특별시 버스 노선 혼잡도 예측을 통한 다람쥐버스 신규 노선 제안		

1. 주제 선정 및 자료분석 배경 (주제 선정의 독창성)

을지로, 여의도 같은 기업의 밀집도가 높은 지역들의 출퇴근 시간에는 대중교통, 자가용 할 것 없이 모든 이동수단의 포화상태를 볼 수 있다. 출퇴근 시간에서의 교통 혼잡도를 줄이기 위해서 대중교통의 이용을 적극 홍보하고 있지만, 사실상 버스와 지하철을 타지 못하고 몇 대씩 보내고 난 다음에야 탑승할 수 있을 정도로 이용객이 넘쳐난다. 이러한 문제를 해결하기 위해 서울시에서 다람쥐버스의 운행을 시작하였다. 다람쥐버스는 서울시에서 운행하는 출퇴근 맞춤버스로 출퇴근 시간대의 특정 구간에서 발생하는 차내 혼잡을 해소하기 위해 운행하는 순환형 셔틀버스이다. 다람쥐버스라는 이름은 다람쥐가 쳇바퀴 돌듯 짧은 구간을 반복 운행한다고 하여 붙여졌다.

2017년 6월에 4개의 노선으로 반년 간의 시범 운행을 한 결과 차내 혼잡을 39.5% 완화하는 효과와 동시에 해당 구간의 전체 버스 이용객이 16.5% 증가하는 효과를 볼 수 있었다. 이러한 결과를 반영하여 2018년에는 3개의 노선이 추가되어 현재는 총 7개의 노선이 운행 중이다.

다람쥐버스는 현재 이미 시행 중인 정책이지만 2018년에 국토교통부가 주관한 2018 지속가능 교통도시 평가에서 최우수정책으로 선정될 만큼 긍정적인 평가를 받고 있고 타 지역에서의 운행을 기대하는 시민들의 목소리도 높은 것으로 나타났다. 따라서 이번 과제에서 버스 정류소 및 노선별 이용객 수를 예측하고 버스 내 혼잡도를 계산함으로써 가장 큰 효과를 얻을 수 있는 신규 노선을 제안하고자 한다.

2. 분석 내용 (자료분석의 우수성, 데이터 활용성)

활용데이터

데이터명	Source	비고
서울시 버스노선별, 정류장별, 시간대별 승·하차 인원 정보	서울열린데이터광장	기본 데이터
stationList	공공데이터포털 API	정류장별 위·경도 자료 사용
busStationbyRouteList	공공데이터포털 API	노선 및 정류장 별 정보 이용
서울시 동별 사업체 및 종사자 밀도 통계	서울열린데이터광장	정류장 주변 인구 밀도 정보
구/동 별 주거 지역, 상업지역	서울열린데이터광장	정류장 주변 환경 정보
지하철 역별 시간대별 승하차 인원수	공공데이터포털(file)	가까운 지하철역 승하차 수
학교 정보(중·고·대학교·대학원)	K-ICT 빅데이터센터	학교 별 학생 및 교직원 수
기초정보_인구집중유발시설(2018.7)	통계빅데이터센터	대형유통점, 백화점, 숙박시설의 수

『**서울시 버스노선별, 정류장별, 시간대별 승·하차 인원 정보**』 데이터에서 일별, 시간대별 승·하차 인원을 구하기 위해 출근 시간대인 7시~9시의 승·하차 인원을 추출, 평균을 취하고 연도별 일수로 나누었다. 『**stationList**』, 『**busStationbyRouteList**』에서 얻은 버스정류장의 위·경도를 위에서 만든 버스 데이터에 합치고, (승차 인원 – 하차 인원)/(60 / 배차간격) 식을 이용하여 버스 한 대별 순승차인원을 구하고, (버스 별 누적 탑승인원 / 60) * 100 식으로 혼잡도지표를 생성하였다. 지역별 인구정보를 활용하기 위해 Python의 selenium을 이용하여 정류소별 행정동 주소를 네이버 지도에서 크롤링하고, 이 주소와 『**서울시 동별 사업체 및 종사자 밀도 통계**』 이용하여 행정동별 총 인구수, 사업체 수, 종사자 수 데이터를 합쳤다. 정류소번호의 위·경도 데이터와 『**구/동 별 주거 지역, 상업지역**』, 『**학교 정보**』 데이터를 이용하여 반경 500m 내 상업지구, 주거지구 수와 중·고·대학교·대학원 수, 그 학교에 다니는 학생(재적학생) 수, 교사(교수) 수를 합쳤다. 또한 『**지하철 역별 시간대별 승·하차 인원수**』 데이터를

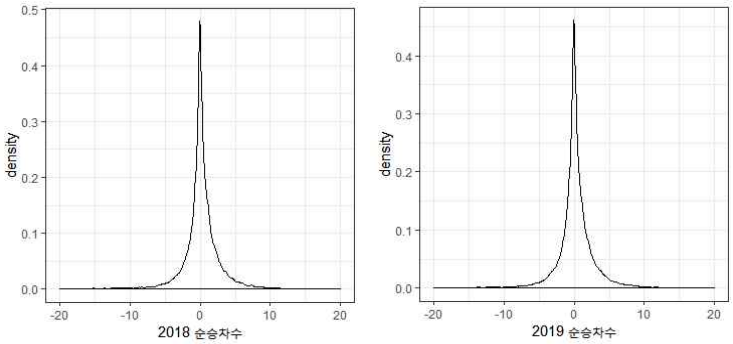
이용하여 가장 가까운 지하철역의 출근 시간대 평균 승, 하차 정보와 해당 지하철역을 다니는 호선 수를 합쳤다. 마지막으로 통계빅데이터센터『**기초정보_인구집중유발시설**』데이터에서 대형유통점, 백화점, 숙박시설의 자료만을 추출하여 버스정류장의 반경 500m 내에 있는 시설의 수를 추가하였다.

MODELING

모델링을 하기에 앞서 기준을 세우기 위해 다음과 같이 설명변수를 전혀 넣지 않은, 랜덤으로 순승차량을 예측했을 때의 RMSE값을 구했다. 또한 y변수인 순승차량의 분포를 그래프로 나타냈다.

기본	RMSE
random	3.761407

랜덤으로 순승차량을 예측했을 때의 RMSE값은 3.76 정도이다. 이를 기준으로 하여 아래 표에 나와있는 총 6 가지의 모델의 성능을 비교하고 가장 작은 test RMSE 값을 갖는 것을 우리의 최종 모형으로 선택하고자 한다.

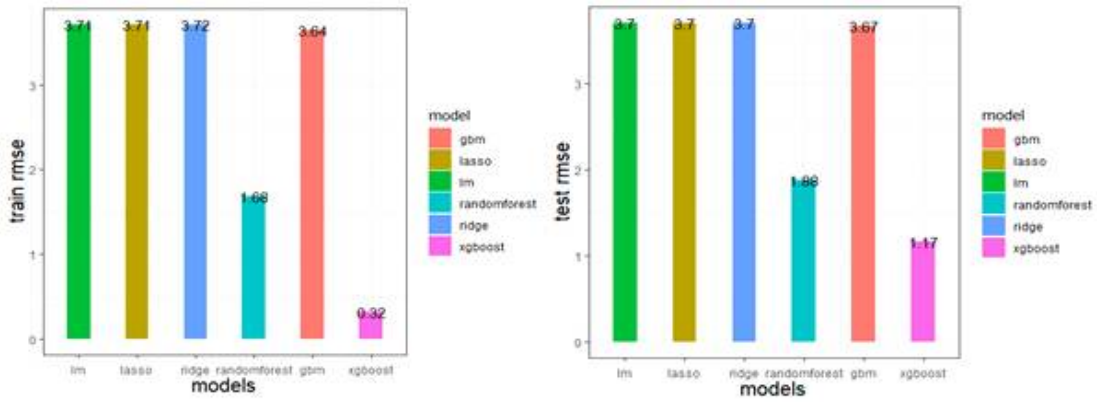


model	explanation
linear regression	종속변수(y)와 독립변수(x)들의 선형관계를 모델링하는 기법
Ridge & Lasso	회귀계수의 값을 0에 가깝게 만들어 분산을 줄인다. 이 중 Lasso 방법은 특정 변수의 회귀계수를 0으로 만들어 변수를 선택하기도 한다.
Random Forest	일부 변수만을 사용하여 여러 개의 tree (의사결정나무)를 생성한다. 모든 변수를 사용하지 않기 때문에 변수 사이의 상관관계를 줄일 수 있다.
GBM	gbm은 연속적인 tree에서 이전 tree의 예측오류를 gradient boosting algorithm으로 수정하며 예측 모델을 생성한다.
XGBoost	xgboost는 gbm과 마찬가지로 gradient boosting 알고리즘을 사용하지만 과적합을 방지하기 위해 변수의 정규화를 사용함으로써 gbm보다 정확도를 높인다.

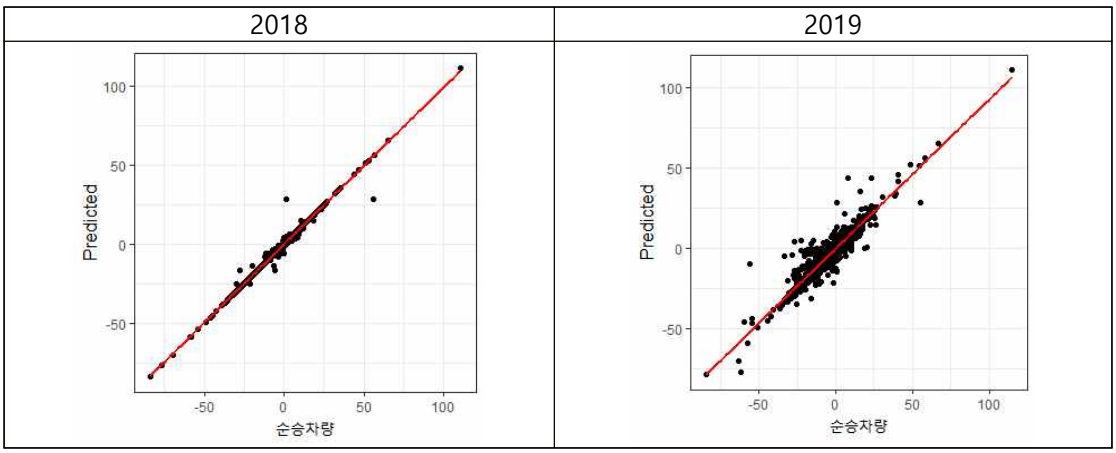
6 가지 모델 중에서 Random Forest, GBM, XGBoost 는 tuning parameter 를 바꿔가며 모델링을 실시했다. Random Forest 와 GBM 은 의사결정나무의 생성 수를 500, 1000, 1500 으로 바꿔가며 그 중 가장 최소의 test RMSE 를 갖는 경우를 아래 표에 기입하였다. XGBoost 에서는 tuning parameter 로 max.depth, eta 값을 조정해 주었다. Ridge 와 Lasso 회귀 모형에서는 tuning parameter 인 lambda 를 0.01 부터 10^{10} 까지 바꾸어가며 각 test RMSE 가 최소가 되는 값을 채택했다. max.depth 는 10, 30, 50 을 사용하였고, eta 값은 0 부터 1 까지 0.1 간격을 주어 사용하였다. 이렇게 총 30 개의 parameter 를 조정하며 test error 를 비교하여 가장 작은 parameter 를 선택하였다.

아래 표와 그래프는 6가지의 모델들의 train 및 test RMSE 값을 나타낸 것이다. 비교를 해보면 train과 test에서 모두 XGBoost의 성능이 가장 탁월한 것으로 나타났다. 따라서 순승차량을 예측하는 방법으로 XGBoost모델을 최종모형으로 채택하였고 이 때의 max.depth는 30, eta는 0.3이다.

	MODEL	TUNING	RMSE	RMSE.TEST
1	**Linear Regression**	-	3.71272	3.70241
2	**Lasso**	0.01	3.71313	3.70283
3	**Ridge**	0.87	3.71525	3.70500
4	**Random Forest**	n.trees = 500	1.68062	1.87812
5	**GBM**	n.trees = 500	3.64299	3.67059
6	**XGBoost**	Max.depth =30 eta = 0.3	0.31726	1.16769



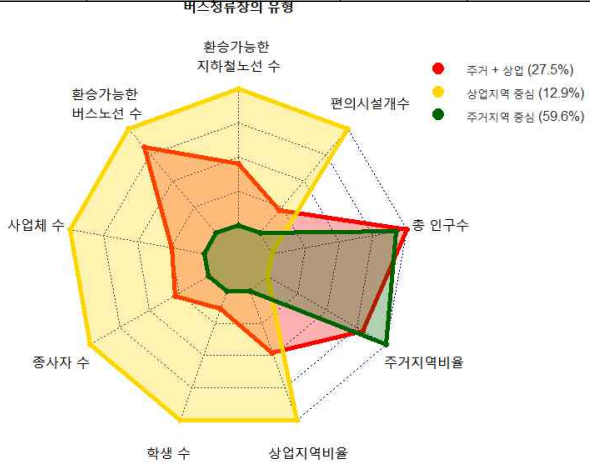
최종 모형을 이용하여 실제값(x 축) vs 예측값(y 축) 그래프를 그려보았다. Train data 인 2018 년 데이터에 대해서는 거의 일직선에 가까운 형태를 보인다. Test data 인 2019 년 데이터에 대해서도 어느 정도 오차를 보이지만 뛰어난 예측력을 가지고 있다고 해석할 수 있다.

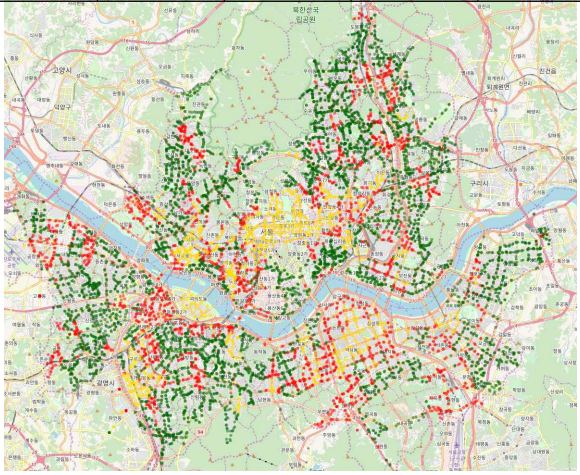


클러스터링

현재 운행 중인 다람쥐버스가 지나는 지역의 특성을 파악하기 위해 군집화 분석을 진행하였다. EM알고리즘을 기반으로 한 군집분석을 시행한 결과 3개의 그룹으로 나누는 것이 가장 적절하다고 판단하였다. 그 결과 각 그룹은 아래 표와 같은 특성을 보였고 이를 서울시 지도에 다음과 같이 나타낼 수 있었다.

class	학생수	사업체수	종사자수	총인구수	지하철	버스	상업지역	주거지역	편의시설
1(주거+상업)	3695	2400	18131	27446	1.37	3.99	1.33	13.7	2.91
2(상업)	12129	5224	45720	21989	1.81	4.27	4.27	13.5	8.15
3(주거)	2373	1505	7099	27035	1	2.67	2.67	13.8	1.48

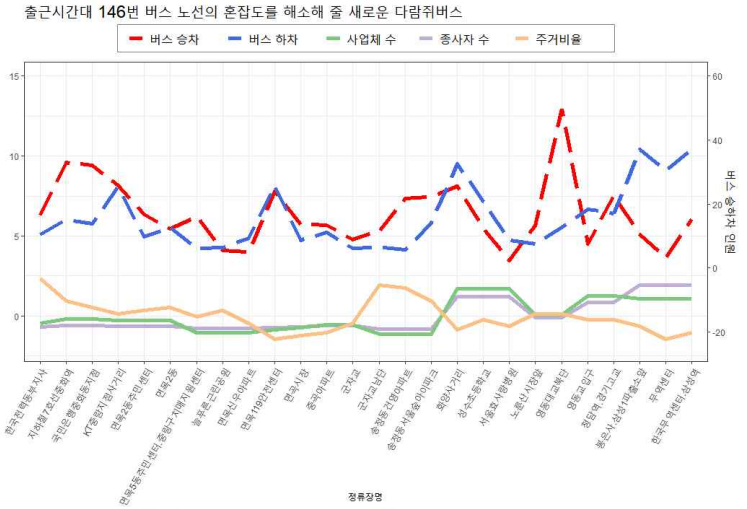


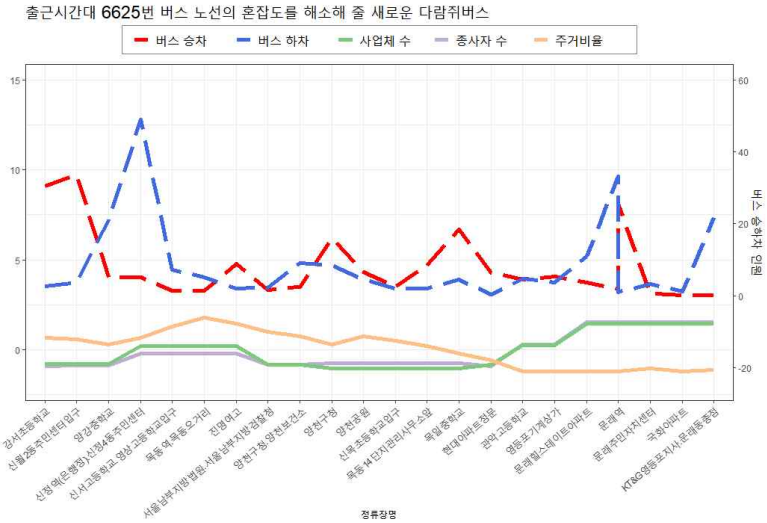


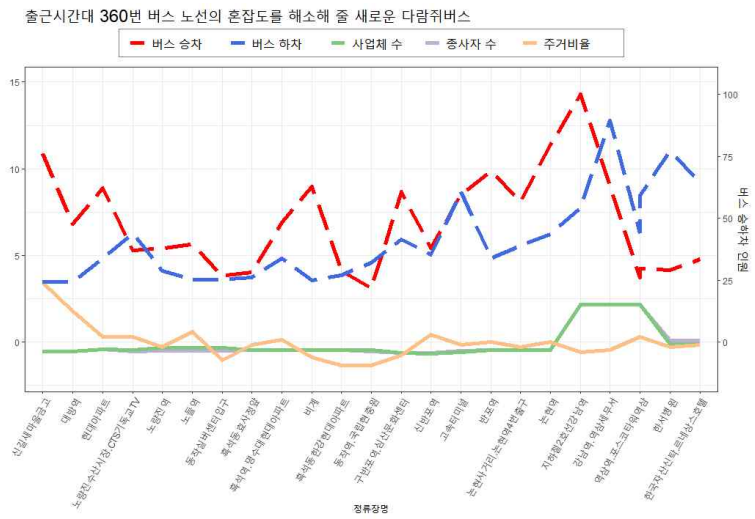
다람쥐 버스 추천

위에서 버스 정류장 주변에 있는 정보를 이용하여 해당 버스 정류장의 순 승차량을 예측하였다. 이제 예측한 순 승차량을 이용하여 새로운 다람쥐 버스를 추천하고자 한다. 혼잡도 지표 식을 이용하여 계산한 다음, 전체 혼잡도의 상위 5% 안에 있는 값들이 연속된 구간을 찾고 개수를 세었다.

노선버스	5616	**146**	**6625**	650	333	**360**
연속구간	89	69	56	44	42	41



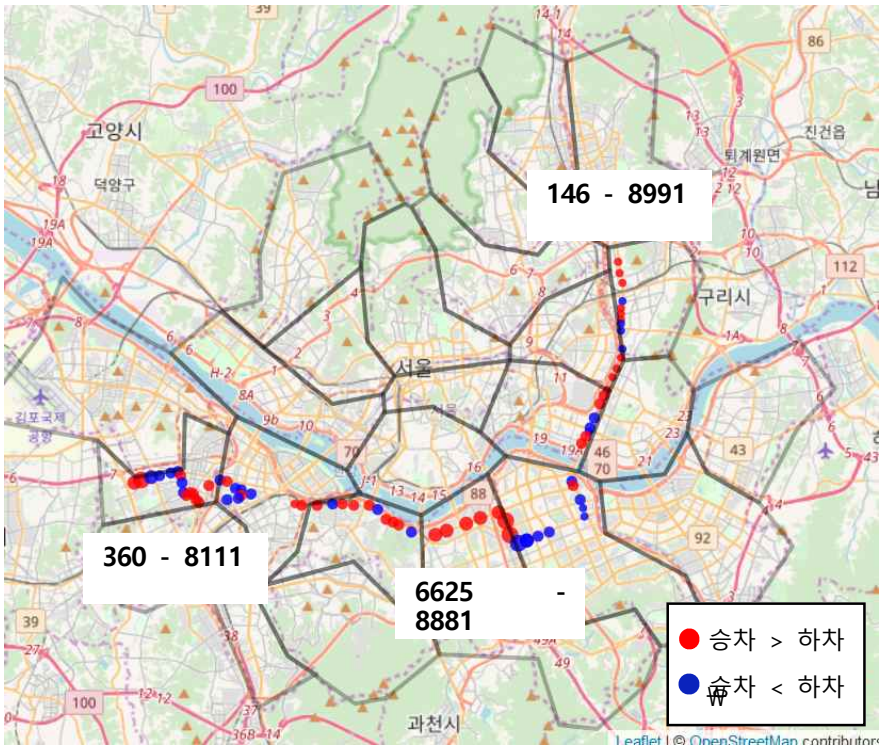




연속 구간이 가장 많은 버스는 5616인데, 혼잡 구간이 현재 다람쥐버스 8552가 다니고 있기 때문에 제외하였다. 650번 버스의 혼잡구간은 6625번 버스의 구간과 비슷했고, 333번 버스의 경우 혼잡구간이 상업지역(class 1,2)에서 주거지역(class 3)으로 운행하는 경향이 있어 추천 목록에서 제외하였다. 다람쥐 버스 노선 구간은 위의 노선버스에서 주거지역에서 상업지역으로 운행하는 구간을 차용하였다. 제안한 다람쥐 버스의 정류소별 지역정보를 시각화하여 나타내면 다음과 같다.

146번 버스를 차용한 다람쥐버스 추천 노선 구간을 살펴보면, 기점쪽에는 주거비율, 종점 쪽에는 사업체 및 종사자 수가 더 많음을 볼 수 있다. 6625번, 360번 버스도 마찬가지로 나타났다.

최종 다람쥐 버스의 노선 구간을 지도로 나타내면 다음과 같다.



최종 추천 다람쥐버스 노선 구간을 지도와 표로 나타내면 다음과 같다. 지역 정보에 대한 내용을 세 개의 클러스터로 나누어 추천하였기 때문에 위 구간을 반대 방향으로 이용하면 퇴근 시간에 운행하는 다람쥐 버스를 신규 노선으로 제안할 수 있을 것이다.

버스 노선	기·종점	대수	거리	배차간격	횟수	운행시간
8991	한국전력동부지사 ~ 한국무역센터.삼성역	5	12.3km	10~12 분	11	07:00~09:00
8881	강서초등학교 ~ KT&G영등포지사.문래동종점	4	9.18km	10~11 분	12	07:00~09:00
8111	신길새마을금고 ~ 한국자산식탁.르네상스호텔	5	13.7km	10~12 분	11	17:30~19:30

3. 분석 결과 활용

현재 출퇴근 시간에서의 교통 혼잡도는 매우 높으며 대중교통의 이용 또한 이미 포화상태이다. 전체 대중교통의 이용자 수는 늘리되, 차량 당 혼잡도를 줄이고자 일종의 셔틀 개념을 가진 다람쥐버스가 시행되었으며, 이는 충분히 서울 내, 외 다양한 지역에서 적용이 가능하다. 버스의 이용객 수 예측이 적절히 이루어진다면 최대 효과를 볼 수 있는 루트를 찾고 빠른 시행을 기대해볼 수 있다. 따라서 버스 정류소 및 노선별 이용객 수를 예측하고, 이를 토대로 혼잡도 계산을 하여 신규 루트를 제안해보았다.

다람쥐버스는 다시 말하자면 이미 시행되고 있는 정책이다. 학생들의 등교 및 직장인들의 출근시간에 맞춘 단거리 순환버스는 짧은 운행시간으로도 상당한 효과를 보여주고 있다. 이 프로젝트에서는 다양한 변수들을 이용하여 기존의 버스 노선들의 정류장별 순승차인원(승차 – 하차)을 예측하였다. 모델링을 통해 버스정류장 주변 지역 정보를 이용했기 때문에 객관적인 노선제안 방식이라고 말할 수 있다. 여기서 얻은 승차 인원 예측 정보를 통해 배차 간격 및 운행시간을 알맞게 조정하여 시행할 수 있을 것이다. 버스 정류장들을 군집화함으로써 신규 정류장이 건설되거나 기존의 것이 변경되었을 때 손쉽게 적절한 분류가 가능하다는 것이다. 따라서 다람쥐 버스의 신규 노선을 추천하는 작업에서 일반화된 클러스터를 이용할 수 있다는 장점을 가지고 있다.