The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.1.111
32(1), 111–128

A study on the number of passengers using the subway stations in Seoul

Soojin Cho[a] · Bogyeong Kim[a] · Nahyun Kim[a] · Jongwoo Song[a,1]

[a]Department of Statistics, Ewha Womans University

(Received September 6, 2018; Revised October 17, 2018; Accepted November 30, 2018)

Abstract

Subways are eco-friendly public transportation that can transport large numbers of passengers safely and quickly. It is necessary to predict the accurate number of passengers in order to increase public interest in subway. This study groups stations on Lines 1 to 9 of the Seoul Metropolitan Subway using clustering analysis. We propose one final prediction model for all stations and three optimal prediction models for each cluster. We found three groups of stations out of 294 total subway stations. The Group 1 area is industrial and commercial, the Group 2 ares is residential and commercial, and the Group 3 area is residential districts. Various data mining techniques were conducted for each group, as well as driving some influential factors on demand prediction. We use our model to predict the number of passengers for 8 new stations which are part of the 3rd extension plan of Seoul metro line 9 opened in October 2018. The estimated average number of passengers per hour is from 241 to 452 and the estimated maximum number of passengers per hour is from 969 to 1515. We believe our analysis can help improve the efficiency of public transportation policy.

Keywords: subway, demand prediction, GMM, extreme gradient boosting, random forest, linear model

1. 서론

개통 전부터 서울 중심부를 가로지르는 '황금라인'으로 주목받았던 서울매트로 9호선은 극심한 혼잡도로 인하여 '지옥철'이라 불린다. 이는 초기 수요 예측 실패로 인한 것으로 초기 예상 승객 수는 일평균 24만 여명이었으나 실제 승객 수는 50만 명 규모로 2배에 달했다. 9호선의 예에서 알 수 있듯이 잘못된 수요 예측은 교통 혼잡과 서비스 질의 하락, 에너지 낭비 등 사회적 비용을 증가시키기 때문에 수요의 정확한 예측은 교통정책 시행에 필수적이다. 특히 지하철은 버스, 택시와 같은 교통수단에 비해 많은 승객들을 원거리까지 안전하고 신속·정확하게 원하는 지점으로 대량 수송할 수 있어 미래 지향적인 교통수단이라 할 수 있다 (Kim, 2016). 따라서 지하철의 공익성을 증대시키기 위해서는 정확한 승객 수요 예측이 이루어져야 하며, 이에 따른 적절한 운송 계획이 입안되고 시행되어야 한다 (Song, 1991).

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the ministry of Education, Science and Technology (No. NRF-2017R1D1A1B03036078).
The draft of this paper was awarded the prize in the 'Transportation Big Data Application Papers and Idea Contest' by the Korea Transportation Research Institute in 2018.
[1]Corresponding author: Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail: josong@ewha.ac.kr

지하철 이용 수요와 관련한 선행연구는 이미 오래전부터 진행되어왔다. Kim (2013)은 대구광역시의 지하철 자료를 이용하여 오전 첨두시간대의 승차수와 하차수를 지하철역 주변 정보들을 설명변수로 하여 회귀분석을 실시하였다. Lee 등 (2015)는 서울 도시철도 각 지하철역의 첨두시간과 비첨두시간의 승·하차 이용자수를 중심으로 군집분석을 이용하여 역세권을 유형화하고, 다항로짓모형을 통해 6개 유형의 역세권과 토지이용의 관계를 규명하였다.

이에 본 논문은 선행연구의 아이디어를 종합하고 확장하여 먼저, 모형 기반인 가우시안 혼합 모형(Gaussian mixture model; GMM) 군집분석을 실행하여 지하철역들을 3개로 유형화하였다. 그 후 각 그룹 별로 다양한 데이터 마이닝 기법을 이용해 지하철 승차인원 예측 모형을 제시하고, 수요 예측에 중요한 영향을 미치는 요인들을 도출하였다. 구체적으로 분석에 사용되는 모형은 다양한 변수 선택법(Stepwise, Ridge, LASSO)을 이용한 선형 회귀 모형, 포아송 일반화 선형 모형, 랜덤포레스트, 그래디언트 부스팅, 익스트림 그래디언트 부스팅, 서포트 벡터 기계이며 예측력 평가 지표로는 평균 제곱근 오차(root mean squared error; RMSE)를 이용하였다. 이를 기준으로 지하철 승차 인원에 관한 최적 예측 모형을 제시한 후, 최종적으로 2018년 10월에 개통되는 9호선 3단계 연장역인 8개의 신설역의 3개월 수요를 예측하고자 한다.

본 논문은 다음과 같은 순서로 구성된다. 2장에서는 분석 자료 수집 방법과 주요 변수들에 대해 설명하고, 3장에서 상기된 기법들을 이용한 분석 결과와 신설역 수요를 예측하는 최종 모형을 제시한 후 4장 결론에서는 결과를 요약하고, 분석 의의를 도출한다. 모든 분석은 통계프로그램 R을 활용하였으며, 사용한 기법들에 대한 설명과 패키지 목록은 Appendix Table 1.1에 있다. Appendix는 http://home.ewha.ac.kr/∼josong/subway에서 볼 수 있다.

2. 분석자료 설명

2.1. 자료수집 과정

분석에 사용된 자료는 2013년 1월부터 2018년 3월까지 서울시 1–9호선 246개 지하철역의 일별 시간대별 승차인원이다. 단, 6호선 연신내역과 3호선 충무로역의 경우 약 50% 이상의 자료가 손실되어 분석에서 제외하였으며 1호선의 경우, 서울교통공사가 관리하는 서울역에서부터 청량리역까지 10개의 역만을 대상으로 하였다. 일별 시간대별 지하철 승차 인원 자료는 서울 열린 데이터 광장(http://data.seoul.go.kr)을 통해 서울특별시 도시교통본부 교통기획관 교통정책과에 직접 요청하였다.

지하철 수요에 영향을 미치는 여러 요인은 크게 역 관련 정보(지하철 운영일, 호선, 노선개수, 출구개수, 위도, 경도 등), 역 주변 정보(역 주변 버스정류장 수, 백화점 수, 영화관 수, 500m 내 용도지역 등), 역의 행정동 별 정보(인구 수, 종사자 수, 사업체 수 등), 날씨 정보로 구분하였다. 이에 대한 자세한 설명은 다음 절에서 할 것이다. 역 관련 정보와 행정동별 정보는 서울 열린데이터 광장(http://data.seoul.go.kr) 및 국가통계포털(http://kosis.kr)에서, 역 주변 정보는 포털사이트 다음(https://www.daum.net)과 네이버(https://www.naver.com), 각 백화점 및 영화관 홈페이지로부터 얻을 수 있었다. 마지막으로 날씨 정보는 기상청(https://www.kma.go.kr)에서 수집하였다.

2.2. 변수 설명

본 연구의 목적은 지하철 승차인원에 영향을 미칠 것이라 예상되는 변수들을 이용하여 월평균 시간당 평균 승차인원과 월평균 시간당 최대 승차인원을 예측하는 것이다. 본 연구진은 지하철의 효과적인 운영에 중요한 것은 시간당 평균 승차인원수 뿐 아니라 지하철이 가장 혼잡한 시간대의 인원인 시간당 최대

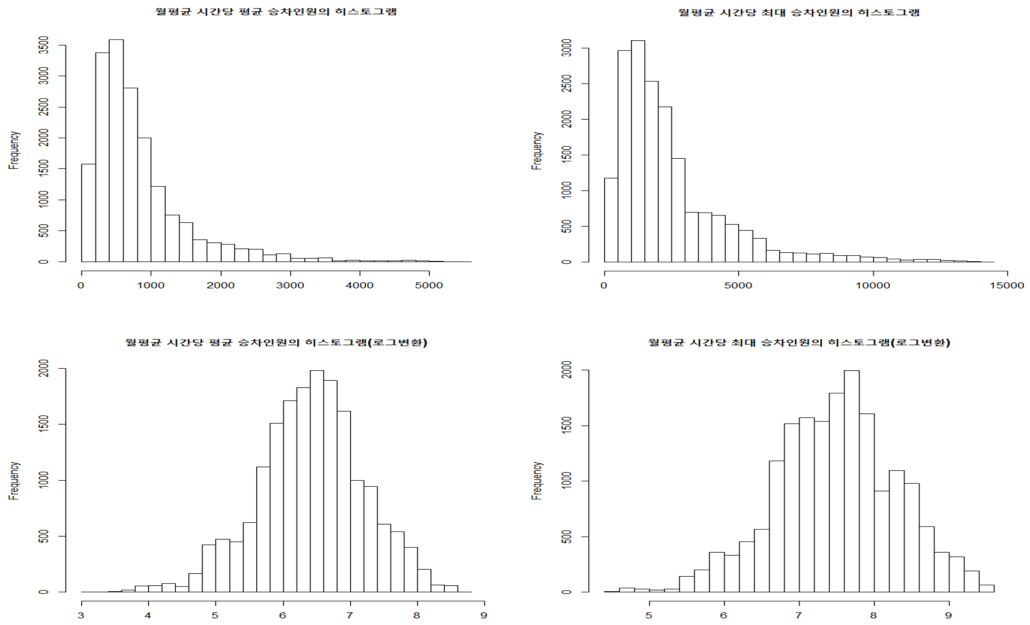


Figure 2.1. Histogram of each response variable.

승차인원수라고 판단하였다. 이에 고유 역들의 하루 중 시간당 평균 승차인원수와 시간당 최대 승차인원수를 계산하고, 이를 다시 월평균을 취하여 월평균 시간당 평균 승차인원수와 최대 승차인원수를 각각 반응변수로 설정하였다. 본 논문에서 시간당 평균 및 최대 승차인원 수를 월평균 취한 것은 Figure 2.3에서 볼 수 있듯이 월별로 시간대 평균 및 승차 인원수가 많이 다르기 때문이다. 두 반응변수의 히스토그램을 그려보면 오른쪽으로 꼬리가 긴 비대칭분포로 나타나므로 선형모형의 가정 및 모형의 예측력을 높이기 위해 반응변수들을 로그 변환하여 대칭분포로 만들었다. 변환 전후 반응변수들의 히스토그램은 Figure 2.1에서 볼 수 있다.

참고로 월평균 시간당 평균 승차인원과 최대 승차인원이 가장 많았던 역은 모두 2014년 7월의 2호선 강남역으로 시간당 평균 인원이 약 5,469명, 최대 인원이 약 14,207명이었다. 이를 예측하기 위하여 분석에 이용한 설명변수들은 다음과 같다.

- **날짜 관련 변수**: 반응변수를 월평균 시간당 평균 및 최대 승차인원수로 설정하였기 때문에 날짜 관련 변수는 연도와 월로 구성된다. 또한 주말 및 공휴일이 월평균 지하철 수요에 영향을 미칠 것으로 판단하여 해당 년도·월의 주말 및 공휴일 개수를 변수로 추가하였다.

  Figure 2.2와 같이 연도별로 월평균 시간당 평균 승차인원수와 최대 승차인원수를 살펴보면, 평균 승차인원의 경우, 2014년 840명으로 최대를 기록하고 감소하는 추세이나 최솟값과 최댓값의 차이가 50명 정도로 크지 않다. 이에 반해 최대 승차인원은 해가 갈수록 증가 추세이며 최솟값과 최댓값의 차이가 110명 정도로 평균 승차인원보다는 큰 편이다. 따라서 해가 지날수록 연도별 월평균 시간당 승차인원은 감소하나 특정 시간대에 집중적으로 승차인원이 집중되는 것을 알 수 있다.

  Figure 2.3을 통해 월별로 월평균 시간당 평균 승차인원수와 최대 승차인원수를 살펴보면, 평균 승차인원과 최대 승차인원 모두 월별 효과가 존재한다. 3월부터 5월, 9월부터 11월의 경우 학기 중으로

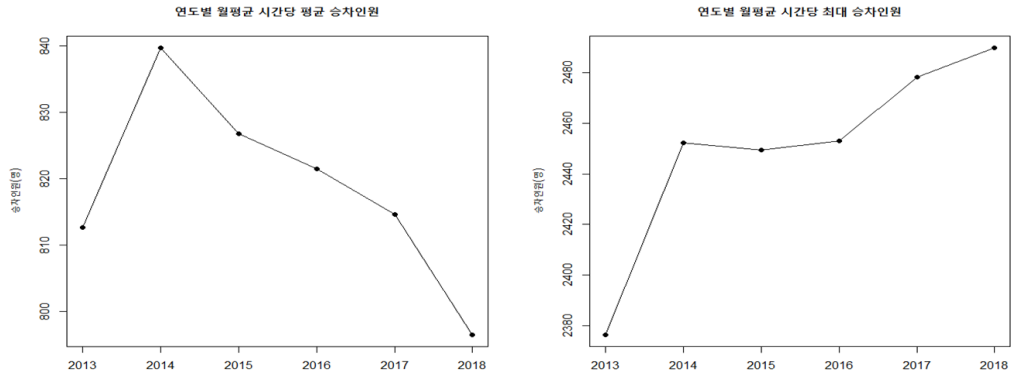


Figure 2.2. Yearly average of each response variable.

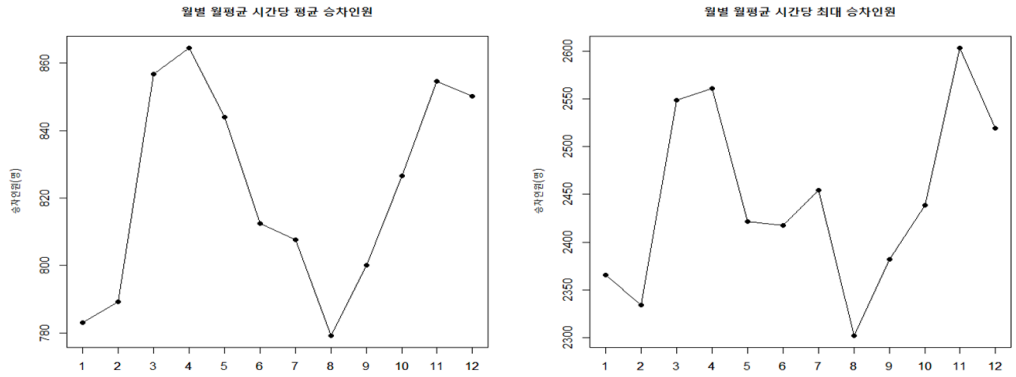


Figure 2.3. Monthly average of each response variable.

통학하는 학생 수가 증가하고 이에 반해 7–8월 1–2월의 경우 방학을 맞아 통학하는 학생 수가 감소하기 때문에 이러한 등락을 나타내는 것으로 보인다. 이에 따라 월의 경우 방학기간과 비방학기간으로 나누어 방학기간 1, 2, 7, 8월을 나타내는 더미변수를 생성하여 범주화하였다.

- **지하철 운영일**: 지하철 운영일이란 지하철 개통일로부터 현재까지의 운영일수를 말한다. 지하철 개통 초기에는 지하철 주변에 역세권이 형성되지 못하였기 때문에 수요가 낮으나 개통 4–5년 후에는 실제 수요가 계속적으로 증가하여 개통 후 약 10–13년 정도에 수요가 안정에 이르게 된다 (Shon 등, 2004). 따라서 지하철 운영일은 지하철 승차인원에 중요하게 영향을 끼치는 변수이므로 분석에 고려하였다. 지하철 운영일을 기준으로 4개의 그룹으로 나눈 다음, 그룹별 반응변수에 대한 상자그림을 그리면 Figure 2.4와 같으며, 운영일이 증가함에 따라 승차인원이 증가하는 것을 볼 수 있다.
- **호선**: 연도별 및 월별로 호선별 월평균 시간당 승차인원을 살펴보면 평균과 최대 승차인원 모두 호선에 상관없이 추세가 거의 비슷하나 호선별로 승차인원의 차이가 있었다. 2호선 - 1호선 - 4호선 - 3호선 - 7호선 - 5호선 - 8호선 - 9호선 - 6호선 순으로 승차인원이 높게 나타나므로 호선을 범주형 변수로 고려하였다.
- **노선 개수**: Figure 2.5를 통해 노선 수에 따른 승차인원을 살펴보면 노선 개수가 증가할수록 평균, 최대 승차인원이 모두 증가하는 추세이며 이는 역의 크기나 접근성과도 관련된 것이므로 예측의 변수

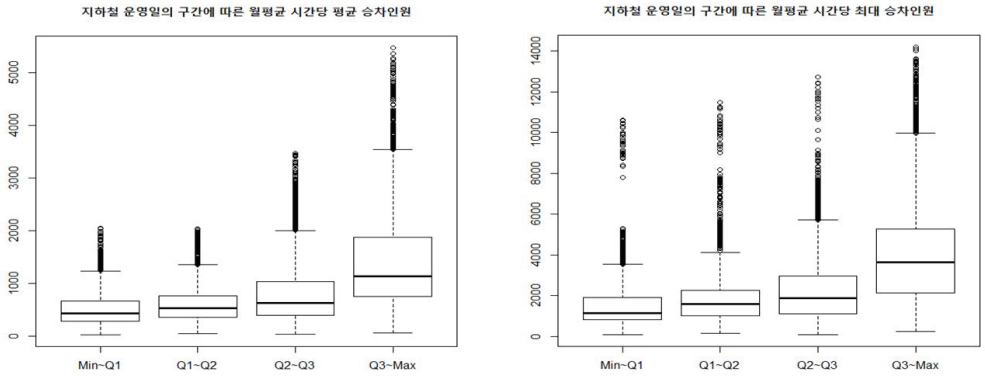


Figure 2.4. Response variable vs. Quantile of duration.

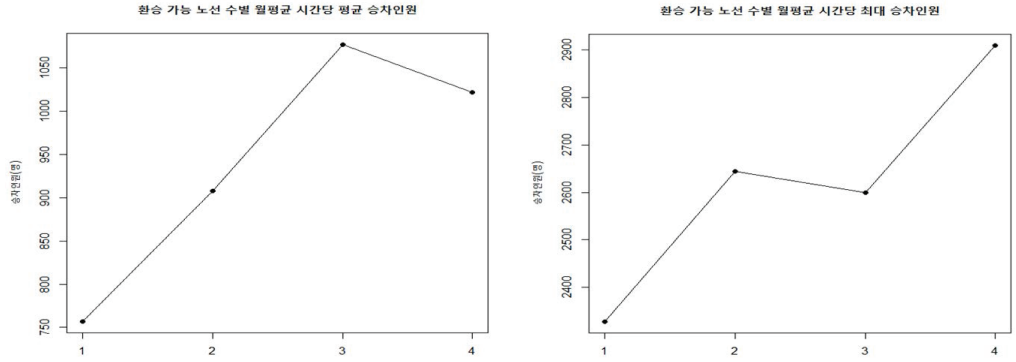


Figure 2.5. Response variable vs. Number of subway lines.

Table 2.1. Number of subway stations by the number of subway lines

노선 수	1	2	3	4
역 개수	183	53	7	3

로 고려하였다. Table 2.1은 환승 가능 노선 수별 역의 개수를 나타내며 참고로 환승 가능 노선 수가 4개인 역은 총 3개로 서울역, 왕십리역, 공덕역이었다.

- **출입구 개수**: 역의 출입구 수는 역사의 크기를 가늠할 수 있으며 해당 역의 접근성과 관련된 척도이므로 분석에 고려하였다. 그래프를 그려보았을 때에도 출입구 개수가 증가할수록 평균, 최대 승차인원 모두 증가추세이므로 지하철 수요에 영향을 끼치는 변수임을 알 수 있다 (Figure 2.6).
- **버스정류장 개수**: 역 주변 버스정류장 수는 지하철-버스 간 환승의 용이성을 나타내는 지표이다. 해당 역에 정차하는 버스 노선이 많다는 것은 역의 접근성이 좋다는 것이기 때문에 승차인원 예측의 변수로 고려하였다.
- **주변 백화점, 영화관, 종합병원 개수 및 기차역, 버스터미널 유무**: 역 주변 백화점, 영화관 개수, 종합병원 개수, 기차역 및 버스터미널 유무를 반영하여 쇼핑, 문화생활 등과 같은 해당역의 통행 목적 특성을 반영하였다.
- **토지용도**: 지하철역 주변의 토지이용은 지하철 수요를 결정짓는 중요한 요인 중 하나이다. 주거지역

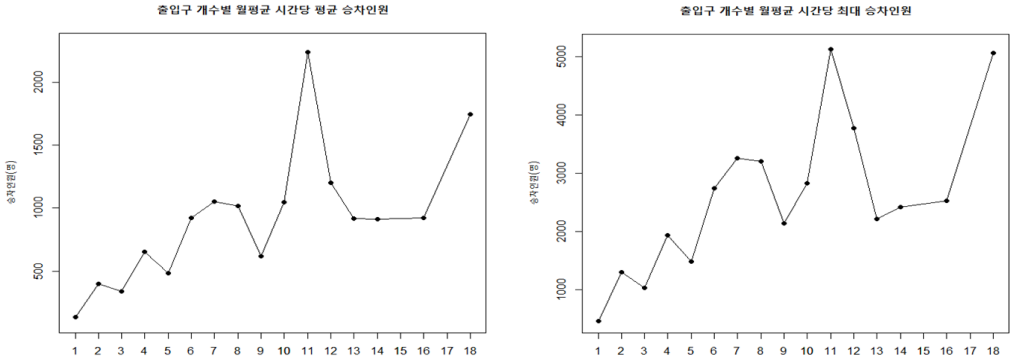


Figure 2.6. Response variable vs. Number of exits.

의 경우, 출근 및 등교 시간인 오전에 승차인원이 높을 것이고, 공업지역의 경우 퇴근 시간인 오후에 승차인원이 집중될 것이다. 본 분석은 역세권의 개념을 적용하여 지하철 역 주변 반경 500m 내의 토지이용을 주거지구, 상업지구, 공업지구, 녹지로 나누어 고려하였다. 토지용도는 지도에서 색으로 구분 되어 있는데 R에서 해당하는 색상 RGB코드를 추출하여 비율을 계산하였다.

- **학생 수, 행정동 별 인구수, 종사자수, 사업체수, 병원 수**: 최대 승차인원수를 기록하는 시간대의 통행목적은 주로 통근과 통학이기 때문에 해당 역의 인구수, 학생 수, 종사자 수, 사업체 수를 고려하였다. 학생 수의 경우, 역 주변 중학교, 고등학교, 대학교 재학생 수의 합을 구하였다. 그 외의 자료는 해당 지하철 역 주소의 행정동을 기준으로 삼았다.
- **날씨**: 날씨에 따라 사람들의 행동 패턴이 좌우된다. 맑고 화창한 날씨라면 외출을 많이 하는 경향이 있는 반면 흐리거나 비가 내리는 등 날씨가 좋지 않을 때는 외출을 삼가는 경우가 많으므로 지하철 수요가 감소할 수 있다. 한편 지하철은 지상교통수단에 비해 신속하며 정확하게 원하는 지점으로 이동할 수 있다. 흐린 날의 경우 지상교통수단 이용자들이 지하철을 타면서 지하철 수요가 늘어날 수도 있다. 두 가지 측면 모두 날씨가 지하철 수요에 영향을 미칠 것으로 판단되어 날씨정보를 설명변수로 삽입하였다. 분석에 사용된 변수를 정리하면 Table 2.2와 같다.

2.3. 결측치 처리 방법

분석에 필요한 결측값 처리는 크게 2가지로 나눌 수 있다. 첫째, 모형 적합 시 사용될 데이터들의 결측값 추정과 둘째, 본 분석의 최종 목표인 9호선 3단계 신설역 수요 예측에 필요한 설명변수들의 미래 값을 예측해야 한다.

먼저 분석 대상이 되는 2013년 1월부터 2018년 3월까지의 기한 중 결측값이 존재하는 설명변수는 월별 주민등록 인구 수, 연도별 행정동 기준 학생 수, 종사주 수, 사업체 수, 병원 수이다. 이러한 설명변수는 시간 t에 따른 추세가 존재하기 때문에 시간 t를 설명변수로 사용한 단순 회귀 분석을 통해 결측값을 대체하였다.

신설역 예측에 필요한 2018년 10월부터 12월까지 미래 설명변수 값은 월별 주민등록 인구 수, 연도별 행정동 기준 학생 수, 종사주 수, 사업체 수, 병원 수의 경우, 앞선 결측치 처리와 같은 방식으로 시간 t를 설명변수로 하여 예측하였고, 날씨 관련 변수의 경우, 분석 대상기간이었던 2013년 1월부터 2018년 3월까지 각 월의 평균으로 예측하였다.

Table 2.2. Description of variables

	Variable	Description	Type
Input variables	year	연도	Continuous
	holiday	월별 휴일(주말 및 공휴일) 개수	
	duration	지하철 운영일(일)	
	nline	환승 노선 개수	
	nexit	출입구 개수	
	bus	역 주변 버스 정류장 개수	
	ndpart	역 주변 백화점 및 복합쇼핑몰 개수 (롯데백화점, 현대백화점, 신세계백화점, NC백화점, IFC몰, 포도몰, 타임스퀘어, 코엑스, 엔터식스)	
	theat	역 주변 영화관 개수 (CGV, 메가박스, 롯데시네마)	
	nhos	역 주변 종합병원 개수	
	resi_area	역 반경 500m내 주거지구 비율	
	comm_area	역 반경 500m내 상업지구 비율	
	manu_area	역 반경 500m내 공업지구 비율	
	green_area	역 반경 500m내 녹지 비율	
	stu	역 주변 중, 고, 대학교 재학생 수 합	
	pop	해당 역이 속한 행정동 인구수(월별)	
	employ	해당 역이 속한 행정동 종사자수(연도별)	
	business	해당 역이 속한 행정동 사업체수(연도별)	
	sumhos	해당 역이 속한 행정동 병원 수(연도별)	
	area	해당 역이 속한 행정동 면적	
	avgtemp	평균 기온(℃) (1개월간 1일 8회 관측값 평균)	
	minhum	최소 상대 습도(%)	
	monsumrain	월합강수량(mm) (1개월간 지표면에 떨어진 강수량 합)	
	smalleva	소형총증발량 (1개월간 규격 용기에 채워진 물의 양이 변화한 정도의 총합)	
	maxsmalleva	소형일최대증발량 (1개월간 규격 용기에 채워진 물의 양이 변화한 정도의 최댓값)	
	avgwindsp	평균풍속(m/s) (1개월간 하루 중 임의 10분 평균풍속)	
	maxmonsp	최대순간풍속(m/s) (1개월간 하루 중 임의 10분 평균풍속의 최댓값)	
	shinesum	일조시간합 (1개월간 태양이 구름, 안개 등에 차단되지 않고 지표면을 비춘 시간의 합)	
	shineratio	일조율 (%) (1개월간 일조시간/가조시간 비율)	
	sunrad	전천일사합(MJ/m^2) (1개월간 태양복사량의 합)	
	monsumsnow	월적설량합 (1개월간 하루 동안 내린 눈의 총합)	
	m3	3.0m 평균지중온도(℃)	
	m5	5.0m 평균지중온도(℃)	
	vacation	방학기간 여부	Categorical
	line	호선 (1 9호선)	
	train	기차역 유무	
	bus_ter	버스터미널 유무	
Response variables	avg_daymean	월평균 시간당 평균 승차인원수	Continuous
	avg_daymax	월평균 시간당 최대 승차인원수	

3. 분석 결과

분석은 크게 3단계 과정으로 이루어진다. 첫 번째는 역별 군집화 과정이다. 역만의 고유한 특징을 나타

내는 설명변수들을 이용하여 모형 기반인 GMM 군집분석을 실행하여 294개의 지하철역들을 유형화하였다.

두 번째는 모형 선택 과정이다. 최종 모형의 성능을 평가하기 위해 전체 데이터 중 2013년 1월부터 2017년 3월까지 약 80%를 훈련데이터로 사용하고 나머지 2017년 4월부터 2018년 3월까지는 시험데이터로 사용하였다. 훈련데이터의 모든 역들을 포함 시켜 만든 전체 모형과 군집 분석을 통해 나누어진 각 그룹의 역들로만 학습시킨 그룹 별 모형을 비교하여 최적 모형을 찾기 위해 10겹 교차검정을 실시하였다. 이 때 모형 평가 기준은 평균 제곱근 오차로 하여 모형의 예측력을 살펴 최적 모형을 도출하였다.

마지막으로 최적 모형 평가 과정에서는 모든 훈련데이터로 최적 모형을 적합한 다음, 시험데이터로 예측오차를 계산하여 평가할 것이다. 그 후 모든 데이터를 사용해 최적 모형을 적합하여 수요 예측에 중요한 영향을 미치는 요인들을 도출할 것이다. 최종적으로 2018년 10월에 개통되는 9호선 3단계 연장역인 8개 신설역의 개통 후 3개월 수요를 예측할 것이다.

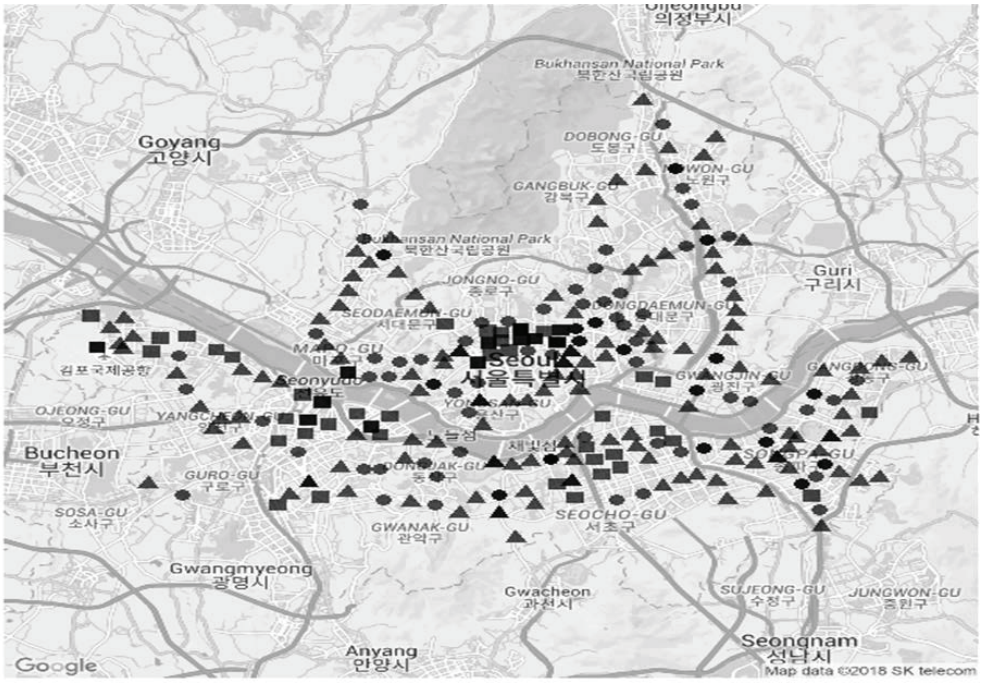
3.1. 가우시안 혼합 모형 군집화 분석

역관련, 역주변, 행정동 관련 정보들을 설명변수로 활용하여 GMM 군집분석을 실행하였다. 모형적합에 사용되는 설명변수들 중에서 시간과 관련한 year, vacation, holiday, duration 변수와, 날씨와 관련된 13개의 변수들은 군집화 시 제외되었다. 또한, 역 주변 정보 중 극히 일부 역만 1값을 갖는 지시변수 train, bus terminal과 주거지구, 상업지구 및 공업지구의 비율로 표현 가능한 green_area 변수는 사용하지 않았다. GMM 군집분석에 사용한 R 패키지는 mclust로 해당 패키지에서는 다양한 군집 개수 및 각 변수들의 분산 구조를 고려한 모형들의 Bayesian information criterion (BIC)를 계산하고 비교하여, 그에 맞는 최적 군집 개수를 제시한다. 그 결과 294개의 지하철역들이 3개의 그룹으로 유형화되었다. 단, 환승역의 경우 각 호선을 분리하여 독립적인 역으로 보았다.

3.1.1. 군집화 결과 GMM 군집화에 사용된 변수는 해당 지하철역의 호선, 환승 가능 노선 개수, 출입구 개수, 역 주변의 버스 정류장 개수, 영화관 개수, 종합병원 개수, 백화점 및 복합쇼핑몰 개수, 학교 학생 수와 해당 지하철역 반경 500m내 주거 지구, 상업 지구, 공업 지구의 비율, 역이 속한 행정동의 주민등록 인구 수, 종사자 수, 사업체 수, 병원 수이다. 이 15개의 변수로 분석한 결과 총 3개로 군집화 되었으며, 그룹 1은 1호선 서울역, 광화문, 을지로입구, 강남 등 63개의 역, 그룹 2는 왕십리, 고속터미널, 종합운동장 등 94개의 역, 그룹 3은 대림, 동작, 충정로 등 137개의 역이 포함되었다. 최종적으로 예측하고자 하는 9호선 3단계 신설역 8개는 그룹 1에 속한 보훈병원역을 제외하고 모두 그룹 3에 속해 있었다. 군집화된 역을 지도에 표시한 그림은 Figure 3.1과 같다. 각 그룹에 해당하는 지하철 역명은 Appendix Table 3.1과 같다.

3.1.2. 그룹 별 특징 Figure 3.2와 Figure 3.3을 통해 반응변수인 월평균 시간당 평균 승차인원 및 월평균 시간당 최대 승차인원을 그룹별로 살펴보면 그룹 1, 2, 3 모두 비슷한 추세를 보이고 있지만 그룹 1에 속한 역의 승차인원이 제일 많고, 그룹 3에 속한 역의 승차인원이 제일 적다. 또한 그룹 1, 2의 경우 연도별, 월별 월평균 시간당 평균 승차인원 값의 차이가 작지만, 최대 승차인원 값의 차이는 큰 것을 볼 수 있다. 월별 주기성은 그룹에 관계없이 비슷하며, 연도별 추세는 월평균 시간당 최대 승차인원의 그룹 1의 경우에만 증가하는 추세를 보이고 있다.

한편, 전체 지하철역 반경 500m의 용도지구별 평균 구성 비율을 살펴보면 주거지구가 70% 이상으로 대부분을 차지하며 그 다음은 상업지구와 녹지 순이고 공업지의 비율이 가장 낮다는 것을 알 수 있다.



그룹 1 (●) 그룹 2 (■) 그룹 3 (▲)

Figure 3.1. Location of subway stations by group.

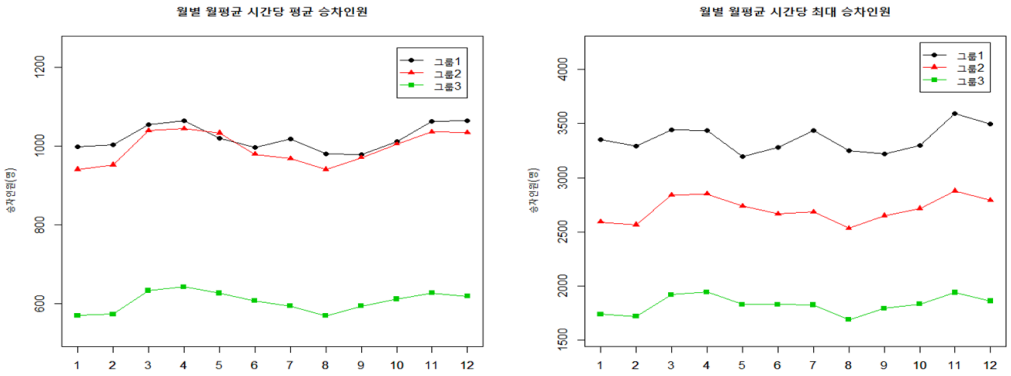


Figure 3.2. Monthly average of each response variable by group.

이를 그룹별로 비교해보면 다른 그룹은 공업지구가 1%도 차지하지 않음에도 불구하고 그룹 1의 경우 15% 이상을 차지하며, 상업지구 또한 전체나 다른 그룹에 비하여 월등히 높다. 뿐만 아니라 주거지구의 비율이 상대적으로 낮으므로 상업 및 공업의 중심인 지역이라고 생각할 수 있다. 그룹 2의 경우 전체보다 주거지구와 상업지구의 비율이 높아 주거와 함께 상업지구가 함께 발달한 지역으로 보인다. 반면 그룹 3은 주거지구가 다른 그룹에 비하여 높은 비율을 차지하지만 상업지구의 비율은 상대적으로 낮아 주거가 중심이 되는 곳으로 볼 수 있다 (Figure 3.4).

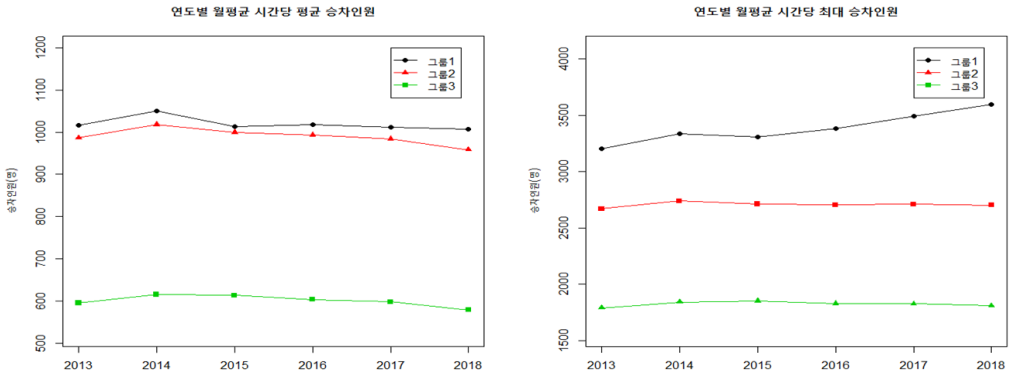


Figure 3.3. Yearly average of each response variable by group.

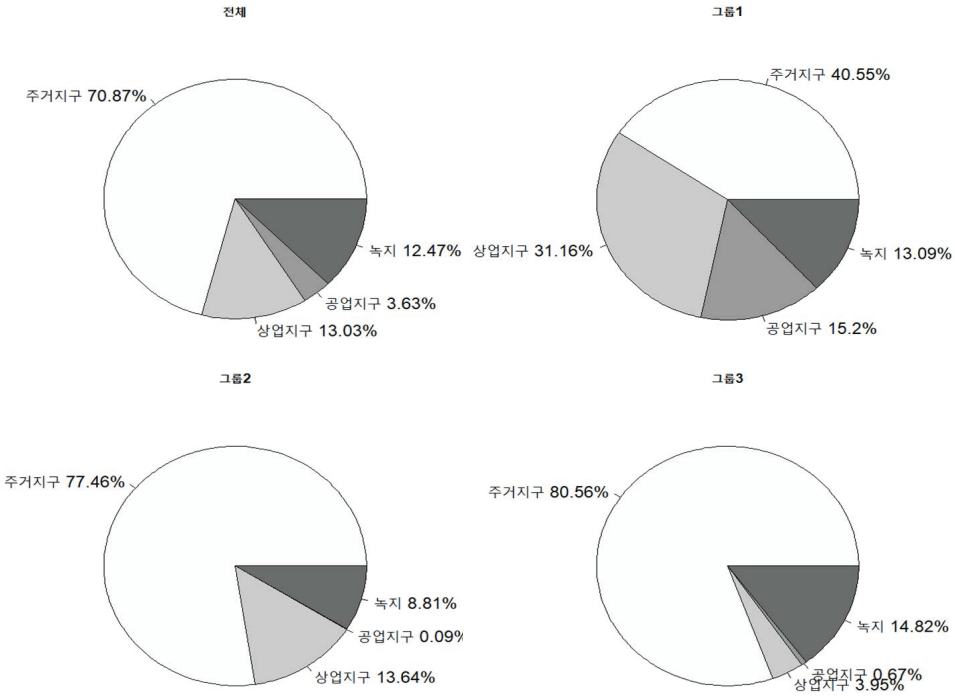


Figure 3.4. Usage area proportion within a radius of 500m around the station.

그룹별 역 주변 버스정류장수를 비교해보면 주거지구와 상업지구가 함께 발달한 그룹 2의 버스정류장수가 제일 많았고, 주거지구가 중심인 그룹 3의 버스정류장수가 가장 적게 나타났다. 그룹별 역 주변 학교의 학생 수를 살펴보면 주거지구와 상업지구가 함께 발달한 지역인 그룹 2에서의 학생수가 다른 그룹에 비해 압도적으로 높다. 이것은 대학교 학생수가 중, 고등학생수보다 월등히 많은데, 대학교가 있는 역들이 대부분 그룹 2에 분류되었기 때문이다. 그룹별 역 행정동의 인구수를 비교하면 주거지구 비율이 높은 그룹 2와 그룹 3의 인구수가 그룹 1의 인구수가 많다. 그룹 1의 경우 상공업지구이기 때문에 인구수

Table 3.1. Average of predictor by group

	그룹1	그룹2	그룹 3
버스정류장수 (개)	10.08	10.53	7.15
학교 학생수 (명)	1,050	7,629	2,063
행정동 인구수 (명)	18,846	27,000	26,290
행정동 종사자수 (명)	56,077	12,878	11,759

Table 3.2. Test root mean squared error of each model using 10-fold cross validation (Model 1)

		전체	그룹 1	그룹 2	그룹 3
Linear	Linear (stepwise)	0.5746	0.9920	0.7465	0.5990
	Ridge	**0.5688**	0.7440	0.6722	0.5864
	LASSO	0.5742	0.9698	0.7412	0.5945
	Poisson GLM (stepwise)	0.5888	0.8404	0.6846	0.6030
Non linear	Random forest	0.5977	**0.7434**	**0.5687**	**0.5598**
	Gredient boosting	0.6984	0.8127	0.6057	0.6756
	Support vector machine	0.6210	0.7757	0.6094	0.7406
	Extreme gradient boosting	0.6011	0.7628	0.5992	0.5654

가 그룹 2, 3보다 적게 관측이 된 것으로 보인다. 그룹별 역 행정동의 종사자수는 인구수와 반대의 결과를 보인다. 그룹 1의 경우 상업 및 공업지구의 비율이 높기 때문에 그룹 2, 3보다 종사자수가 많게 관측되었으며, 그룹 3의 경우 주거 중심의 지역이므로 종사자수가 가장 적게 관측되었다 (Table 3.1).

3.2. 훈련데이터 내 교차검정 비교

2013년 1월부터 2017년 3월까지의 데이터를 이용하여 선형모형(선형 회귀, Ridge 회귀, LASSO 회귀, 포아송 일반화 선형 회귀)과 비선형 모형(랜덤포레스트, 그래디언트 부스팅, 서포트 벡터 기계, 익스트림 그래디언트 부스팅)으로 10겹 교차검정을 100번 실시하였다.

3.2.1. 월평균 시간당 평균 승차인원의 예측 모형 (Model 1) 월평균 시간당 평균 승차인원의 예측 모형(Model 1)의 10겹 교차검정 결과는 Table 3.2와 같다. 시험 평균 제곱근 오차를 기준으로 최적 모형을 결정하면 전체 모형에서는 Ridge 회귀 모형이, 그룹 1, 2, 3에서는 랜덤포레스트가 평균 제곱근 오차 값이 최소가 되어 최적 모형이 된다. 전반적으로 그룹 1 모형들의 성능이 다른 모형들에 비해 떨어짐을 볼 수 있다. 이는 그룹 1에 속한 역의 수가 적어 정확한 예측에 필요한 정보량이 부족하고, 추정량의 분산 또한 크기 때문으로 생각된다. 전체 모형과 군집화 모형을 비교하면 전반적으로 전체 모형에서는 선형 모형이, 군집화 모형에서는 비선형 모형들의 성능이 더 좋다. 이때 전체 모형에서 선형 회귀 모형과 LASSO 모형, Ridge 회귀 모형의 시험 평균 제곱근 오차는 각각 큰 차이가 없는데, 해석의 용이성 및 선택된 변수의 개수를 고려하여 더 간단한 선형 회귀 모형을 선형 모형 중 최적 후보 모형으로 결정하였다. 비선형 모형 중에서는 랜덤포레스트와 익스트림 그래디언트 부스팅 모형을 최적 후보 모형으로 선택하였다.

3.2.2. 월평균 시간당 최대 승차인원 예측 모형 (Model 2) 월평균 시간당 최대 승차인원의 예측 모형(Model 2)의 10차 교차검정 결과는 Table 3.3과 같다.

전체 모형과 그룹 1에서는 Ridge 회귀 모형, 그룹 2와 그룹 3에서는 랜덤포레스트 모형이 시험 평균 제곱근 오차가 가장 작았다. 전체 모형과 군집화 모형을 비교하면 앞절과 같다. 따라서 동일한 이유로 최

Table 3.3. Test root mean squared error of each model using 10-fold cross validation (Model 2)

		전체	그룹 1	그룹 2	그룹 3
Linear	Linear (stepwise)	0.6185	1.0458	0.7919	0.6468
	Ridge	**0.6147**	**0.7612**	0.7219	0.6346
	LASSO	0.6181	1.0265	0.7857	0.6440
	Poisson GLM (stepwise)	0.6340	0.8690	0.7727	0.6605
Non linear	Random forest	0.6458	0.7842	**0.6027**	**0.6142**
	Gredient boosting	0.7328	0.8489	0.6181	0.7178
	Support vector machine	0.6911	0.8671	0.6999	0.7711
	Extreme gradient boosting	0.6454	0.8447	0.6511	0.6187

Table 3.4. Test root mean squared error of each group (Model 1)

	전체	그룹 1		그룹 2	그룹 3
		전체모형	군집화 모형		
Linear (stepwise)	0.5006	0.4916	0.4424	0.4066	0.4784
Random Forest	**0.0928**	**0.1269**	**0.1435**	**0.0726**	**0.0834**
Extreme Gradient Boosting	0.2342	0.2633	0.2111	0.1154	0.1843

적 후보 모형을 선형 회귀, 랜덤포레스트, 익스트림 그래디언트 부스팅 모형으로 선택하였다.

3.3. 시험데이터 최종 예측 결과

훈련데이터(2013년 1월–2017년 3월) 내 10겹 교차검증을 통해 결정된 각 반응 변수에 대한 최종 후보 모형은 변수 선택 방법론을 이용한 선형 회귀 모형, 랜덤포레스트, 익스트림 그래디언트 부스팅 모형이다. 이를 사용하여 시험데이터(2017년 4월–2018년 3월)의 월평균 시간당 평균 및 최대 승차인원을 예측하였다.

3.3.1. 월평균 시간당 평균 승차인원 예측 모형 (Model 1) 시험 평균 제곱근 오차를 기준으로 최종 후보 모형의 예측력을 평가한 결과는 Table 3.4와 같다. 월평균 시간당 평균 승차인원 예측 모형은 그룹 1, 2, 3 모두 랜덤포레스트 모형을 사용했을 때 시험 평균 제곱근 오차 값이 가장 작았다. 그룹 1은 전체 모형보다 군집화 모형에서 시험 평균 제곱근 오차가 크기 때문에 전체 모형으로 예측하는 것이 더 좋다. 그룹 2, 3은 군집화 모형에서의 시험 평균 제곱근 오차가 전체 모형보다 더 작기 때문에 군집화 모형의 예측력이 더 좋다. 따라서 그룹 1은 랜덤포레스트 전체 모형, 그룹 2, 3은 랜덤포레스트 군집화 모형을 최종모형으로 결정하였다.

최종 모형의 예측력 개선 정도를 살펴보기 위해 그룹별 최종 모형의 승차인원 오차를 그룹별 기본 평균 모형을 이용해 예측한 승차인원 오차와 비교하면 다음과 같다. 그룹 1의 시험 평균 제곱근 오차 값은 0.1269로 이를 지수 변환하여 월평균 시간당 평균 승차인원의 오차를 계산하면 약 66명이다. 만약 어떤 모형을 사용하지 않고, 그 그룹에 속한 역들의 평균 승차인원만을 이용한 평균 모형으로 예측했을 때 그룹 1의 월평균 시간당 평균 승차인원의 오차는 616명이다. 따라서 그룹 1은 랜덤포레스트 전체 모형으로 예측했을 때 평균 모형으로 예측했을 때보다 오차가 약 9.3배 줄었다. 그룹 2의 시험 평균 제곱근 오차는 0.0726으로 이를 지수 변환하여 오차를 계산하면 약 48명이다. 그룹 2 승차인원을 평균 모형으로 예측했을 때 오차는 521명이므로 그룹 2의 랜덤포레스트 군집화 모형은 평균 모형보다 오차가 약 10.9배 감소했다. 그룹 3의 시험 평균 제곱근 오차는 0.0834로 이를 지수 변환하여 오차를 계산하면 약 32명이다. 평균 모형에서 그룹 3 승차인원의 오차는 437명이므로 그룹 3의 랜덤포레스트 군집화 모형

Table 3.5. Test root mean squared error of each group (Model 2)

	전체	그룹 1		그룹 2	그룹 3
		전체모형	군집화 모형		
Linear (stepwise)	0.5427	0.5314	0.4635	0.4636	0.5107
Random forest	**0.1086**	**0.1252**	**0.1407**	**0.1039**	**0.1046**
Extreme gradient boosting	0.2640	0.2719	0.2079	0.1712	0.2037

은 평균 모형보다 9.5배 개선되었다.

3.3.2. 월평균 시간당 최대 승차인원 예측 모형 (Model 2) 최종 모형에 대한 평균 제곱근 오차 값은 Table 3.5와 같다. 월평균 시간당 최대 승차인원 예측 모형은 그룹 1, 2, 3 모두 랜덤포레스트 모형에서 평균 제곱 오차의 제곱근 값이 가장 작았다. 따라서 앞절에서와 같은 이유로 그룹 1은 랜덤포레스트 전체 모형, 그룹 2, 3은 랜덤포레스트 군집화 모형을 최종 모형으로 결정하였다.

그룹 1의 시험 평균 제곱근 오차 값은 0.1252로 이를 지수 변환하면 월평균 시간당 최대 승차인원의 오차는 약 234명이다. 만약 어떤 모형을 사용하지 않고, 그 그룹에 속한 역들의 평균 승차인원만을 이용한 평균 모형으로 예측했을 때 그룹 1의 월평균 시간당 평균 승차인원의 오차는 2,151명이다. 따라서 그룹 1은 랜덤포레스트 전체 모형으로 예측했을 때 평균 모형으로 예측했을 때보다 오차가 약 9.2배 줄었다. 그룹 2의 시험 평균 제곱근 오차는 0.1039이며 이를 지수 변환하여 오차를 계산하면 약 185명이다. 그룹 2 승차인원을 평균 모형으로 예측했을 때 오차는 1,439명이므로 그룹 2의 랜덤포레스트 군집화 모형은 평균 모형보다 오차가 약 7.8배 감소했다. 그룹 3에서의 시험 평균 제곱근 오차는 0.1046으로 이를 지수 변환하여 오차를 계산하면 약 115명이다. 평균 모형에서 그룹 3 승차인원의 오차는 969명이므로 그룹 3의 랜덤포레스트 군집화 모형은 평균 모형보다 8.4배 개선되었다.

3.4. 최종 모형

시험데이터에서 월평균 시간당 평균 및 최대 승차인원의 예측력을 바탕으로 그룹 1은 랜덤포레스트 전체 모형, 그룹 2, 3은 랜덤포레스트 군집화 모형을 최종 모형으로 선택하였다. 최종 모형을 통해 각 그룹별로 승차인원 예측에 중요했던 상위 6개의 변수를 살펴보겠다. 변수 중요도 그림은 Appendix Figures 3.1, 3.2에 있다.

3.4.1. 월평균 시간당 평균 승차인원 예측 모형 (Model 1) 그룹 1에서 중요한 변수는 지하철 운영일(duration)이다. 그 외에 호선(line), 출입구 개수(nexit), 역 주변 버스 정류장 개수(bus), 역 반경 500m내 녹지 비율(green_area), 상업지구 비율(comm_area) 순이었다. 상대적으로 중요한 변수였던 상위 4개의 수치형 변수에 대한 부분 의존도 그림은 Figure 3.5과 같다. 지면상 그 외의 부분 의존도 그림은 Appendix Figures 3.3–3.7에 포함시켰다. 그룹 1에 해당되는 역들은 지하철 운영일이 오래될수록, 역 주변 버스정류장 개수가 많을수록, 녹지지구 비율이 낮을수록 월평균 시간당 평균 승차인원이 많아진다. 출입구의 경우, 출입구가 많을수록 평균 승차인원이 많아지다가, 출입구가 5개 이상이면 평균 승차인원 예측에 영향을 주지 않는다.

그룹 2에서 예측에 중요한 주요 변수는 지하철 운영일(duration), 호선(line)이다. 그 외에 역 행정동 사업체 수(business), 역 주변 버스정류장 개수(bus), 역 행정동 면적(area), 역 반경 500m내 주거지구 비율(resi_area) 변수가 중요하였다. Appendix Figure 3.3을 살펴보면, 그룹 2에 해당되는 역들은 지하철 운영일이 오래될수록, 역 주변 버스정류장 개수가 많을수록, 행정동 면적이 작을수록 월평균 시간당 평

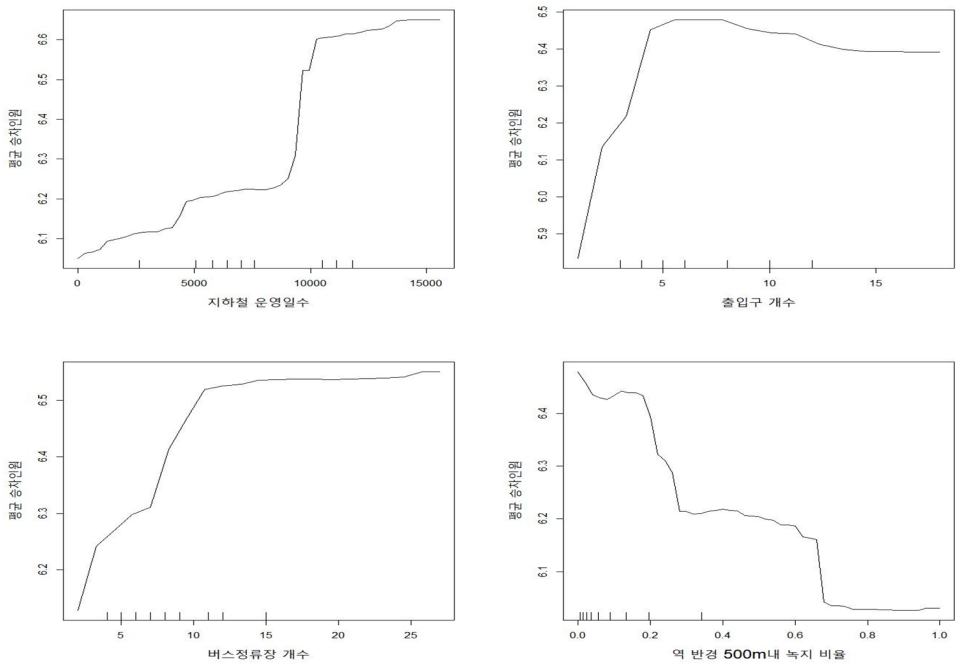


Figure 3.5. Partial dependence plot of all stations (Model 1).

균 승차인원이 많아진다. 역 행정동의 사업체수의 경우, 사업체수가 많을수록 평균 승차인원이 많아지다가, 사업체수가 4,000개 이상이면 평균 승차인원이 감소함을 알 수 있다.

그룹 3에서 중요한 변수는 역 주변 버스 정류장 개수(bus), 호선(line)이다. 그 외에 역 반경 500m내 상업지구 비율(comm_area), 주거지구 비율(resi_area), 출입구 개수(nexit), 지하철 운영일(duration) 순이었다. Appendix Figure 3.4를 보면, 그룹 3에 해당되는 역들은 버스정류장 개수와 출입구 개수, 역 반경 500m내 주거지구와 상업지구가 많을수록 월평균 시간당 평균 승차인원이 많아진다. 특히 주거 지구의 비율이 0.3–0.4사이에서 평균 승차인원의 증가량이 매우 크고, 출입구의 개수가 5개 이상인 경우에는 평균 승차인원에 대한 영향이 비슷하다.

3.4.2. 월평균 시간당 최대 승차인원 예측 모형 (Model 2) 그룹 1에서 중요한 변수는 지하철 운영일(duration), 호선(line), 출입구 개수(nexit), 역 반경 500m내 상업지구 비율(comm_area), 녹지 비율(green_area), 역 주변 버스 정류장 개수(bus) 순이었다. Appendix Figure 3.5의 부분 의존도 그림을 살펴보면, 그룹 1에 해당되는 역들은 지하철 운영일이 오래될수록, 상업지구 비율이 높을수록, 녹지지구 비율이 작을수록 월평균 시간당 최대 승차인원이 많아진다. 출입구의 경우, 개수가 많을수록 최대 승차인원이 많아지다가, 약 7개 이상이면 최대 승차인원이 약간 감소함을 알 수 있다.

그룹 2에서 중요한 변수는 지하철 운영일(duration), 호선(line)이 다른 변수들보다 예측에 더 중요함을 알 수 있다. 그 외에 역 행정동 사업체 수(business), 역 주변 버스정류장 개수(bus), 역 행정동 면적(area), 역 행정동 인구수(pop) 순으로 예측에 중요하다. 또한 그룹 2에서 상대적으로 중요한 상위 4개의 수치형 변수에 대한 부분 의존도 그림은 Appendix Figure 3.6과 같다. 그룹 2에 해당되는 역들은 지하철 운영일이 오래될수록, 역 주변 버스정류장 개수가 많을수록 월평균 시간당 최대 승차인원이

Table 3.6. Predicted number of passengers at 8 new stations (2018)

	월평균 시간당 평균 승차인원			월평균 시간당 최대 승차인원		
	10월	11월	12월	10월	11월	12월
9석촌	420	428	422	1233	1268	1232
9올림픽공원	321	319	315	1036	1028	1014
보훈병원	255	252	241	1065	1040	969
삼전	447	452	441	1483	1515	1448
석촌고분	364	363	354	1044	1051	999
송파나루	394	391	380	1094	1103	1047
오륜	324	323	318	1050	1048	1024
한성백제	430	427	415	1188	1189	1148

많아진다. 반면, 역 행정동의 사업체수는 최대 승차인원이 비슷하다가, 사업체수가 약 4,000개 이상이면 최대 승차인원이 감소하는 모습이 보인다. 행정동 면적의 경우 1km^2 이하에서는 해당 역이 속한 행정동 면적이 클수록 월평균 시간당 최대인원이 증가하며 1–4km^2 구간은 감소하다가 4–6km^2 구간은 다시 승차인원이 증가하고, 6km^2 이상인 경우에는 일정한 것을 볼 수 있다. 참고로 1–4km^2 구간에 속한 역이 152개로 가장 많다.

그룹 3에서 중요한 변수는 호선(line), 역 주변 버스 정류장 개수(bus), 역 반경 500m내 상업지구 비율(comm_area), 주거지구 비율(resi_area), 출입구 개수(nexit), 역 행정동 인구수(pop) 순이다. Appendix Figure 3.7을 보면, 그룹 3에 해당되는 역들은 버스정류장 개수와 출입구 개수, 역 반경 500m내 주거지구와 상업지구가 많을수록 월평균 시간당 최대 승차인원이 많아진다. 특히 주거 지구의 비율이 0.3–0.4사이에서 최대 승차인원의 증가량이 매우 크고, 출입구의 개수가 5개 이상인 경우에는 최대 승차인원에 대한 영향력이 비슷하다.

이제 전체적으로 중요한 변수를 살펴볼 것이다. 반응변수가 월평균 시간당 평균 승차인원인지 월평균 시간당 최대 승차인원인지에 관계없이, 해당 역이 어느 그룹에 속하는지에 관계없이 다른 변수들보다 더 중요한 변수는 호선 정보와 운영일수로 나타났다. 이는 지하철 역세권을 중심으로 상권이 시간에 걸쳐 발달하기 때문에 해당 지하철역의 개통 후 운영일수가 승차인원 예측에 중요한 것이다. 그룹별로 비교해보면 반응변수가 월평균 시간당 평균 승차인원인지 월평균 시간당 최대 승차인원인지에 관계없이 그룹 1은 출입구 개수가, 그룹 2는 사업체 수가, 그룹 3은 주거 지구의 비율이 중요한 변수임을 알 수 있었다. 이는 그룹 1의 경우 상공업 중심 지역이므로 역의 접근성과 관련된 출입구 개수가, 그룹 2는 주상복합지구이므로 상업지구의 발달 지표 중 하나인 사업체 수가 해당 그룹 내에서 승차인원의 많고 적음을 결정짓는 요소로 작용했다는 것을 짐작할 수 있다. 그룹 3은 주거 중심의 지역이기 때문에 주거 지구의 비율이 예측에 중요시된 것으로 보인다. 마지막으로 그룹과 상관없이 월평균 시간당 평균 승차인원은 역 주변 버스정류장 개수, 월평균 시간당 최대 승차인원은 인구수가 중요한 것으로 나타났다. 이는 지하철 이용객들이 지하철과 버스 간의 환승을 통한 대중교통 이용이 편리하기 때문에 버스 정류장 개수가 중요한 것으로 해석이 가능하다. 또한 인구수가 많을수록 통근 승객 등으로 인해 특정 시간대에 승차인원이 많아져 최대 승차인원 예측에 중요하게 작용한 것으로 보인다.

3.5. 신설역 예측

9호선 3단계 연장선은 2018년 10월에 개통 예정이다. 2013년 1월–2018년 3월의 서울시 지하철 승차인원 자료를 이용하여 신설역 8개역의 개통 후 3개월 동안의 월평균 시간당 평균 및 최대 승차인원을 예측하였다. 이때 신설역들의 지하철 운영일수에 대한 변수 값은 월의 중간인 15일을 기준으로 하여, 개통

첫째 달인 10월은 15일, 둘째 달인 11월은 45일, 셋째 달인 12월은 75일로 할당하였다. 보훈병원역은 랜덤포레스트 전체 모형으로, 그 외의 7개 역은 랜덤포레스트 군집화 모형을 사용하여 예측하였고 그에 대한 결과는 다음 Table 3.6과 같다.

4. 결론

지금까지 GMM 군집분석을 통해 지하철역들을 3개로 유형화한 다음, 다양한 데이터 마이닝 기법들 중 최적 수요 예측 모형을 제시하고, 수요 예측에 중요한 영향을 미치는 요인들을 도출하였다. 15개의 수치형 설명변수로 군집화한 결과, 294개의 역이 3개로 군집화 되었다. 그룹 1은 상공업지구, 그룹 2는 주상복합지구, 그룹 3은 주거지구 중심으로 발달한 곳이라 할 수 있다.

군집화 모형과 전체 모형에 다양한 데이터 마이닝 기법들을 적용하고 시험 평균 제곱근 오차를 기준으로 예측력을 비교한 결과, 각 그룹별 월평균 시간당 평균 및 최대 승차인원 예측의 최종 모형을 랜덤포레스트로 결정하였다.

최종 모형을 이용해 모형 적합에 사용되지 않았던 최근 1년(2017년 4월–2018년 3월)을 예측한 결과, 월평균 시간당 평균 승차인원의 경우 각 그룹들 중 가장 큰 오차가 66명(그룹 1), 월평균 시간당 최대 승차인원의 경우 234명(그룹 1)으로 기존 평균 모형으로 승차인원을 예측했을 때의 오차 616명(그룹 1), 2,151명(그룹 1)보다 개선되었다. 이러한 모형으로 모든 데이터(2013년 1월–2018년 3월)를 적합하여 신설역 개통 직후 3개월(2018년 10월–2018년 12월)동안의 승차인원을 예측하였다. 그 결과, 8개 역의 월평균 시간당 평균 승차인원은 241명에서 452명, 월평균 시간당 최대 승차인원은 969명에서 1,515명으로 추정되었다.

본 분석의 최종 모형을 활용한 신설역의 지하철 수요 예측은 대중교통 정책 결정을 위한 기초자료로 활용될 수 있을 것이다. 또한 이를 바탕으로 효율적인 지하철 운영 방안 수립이 가능하다. 예컨대 지하철을 운영하는 공사는 예상된 승차 수요에 맞춰 지하철 배차간격을 적절하게 조정하여 최소한의 비용으로 최대의 효율을 얻을 수 있을 것이다. 마지막으로 고객에게 편리한 서비스 제공으로 고객만족도를 극대화 할 수도 있을 것이다

References

Breiman, L., Friedman. J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*, Chapman and Hall, New York.

Breiman, L. (2001). Random forests, *Machine Learning*, **45**, 5–32.

Cortes, C. and Vapnik, V. (1995). Support-vector networks, *Machine Learning*, **20**, 273–297.

Horel, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55–67.

Douglas, R. (2015). Gaussian mixture models, Encyclopedia of biometrics.

Kim, J. I. (2013). The determinants of subway riderships at AM-peak in Daegu metropolitan city: focusing on the land use of station neighborhood areas, *Journal of Transport Research*, **20**, 15–25.

Kim, J. S. (2016). Subway congestion prediction and recommendation system using big data analysis, *Journal of Digital Convergence*, **14**, 289–295.

Lee, J., Go, J. Y., Jeon, S., and Jun, C. (2015). A study of land use characteristics by types of subway station areas in Seoul analyzing patterns of transit ridership, *The Korea Spatial Planning Review*, **84**, 35–53.

R Development Core Team (2010). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. http://www.R-project.org.

Ridgeway, G. (2007). Generalized Boosted Models: A guide to the gbm package, https://cran.r-project.org/

web/packages/gbm

Shon, E. Y., Kwon, B. W., and Lee, M. H. (2004). Modelling the subway demand estimation by station using the multiple regression analysis by category, *Journal of Korea Society of Transportation*, **22**, 33–42.

Song, J. (1991). A study on prediction of passenger demand in Seoul Subway, *Statistical Consulting*, **6**.

Tianqi, C. and Carlos, G. (2016). XGBoost: A Scalable Tree Boosting System, KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society B*, **58**, 267–288.

데이터마이닝 기법을 이용한 서울시 지하철역 승차인원 예측

조수진[a] · 김보경[a] · 김나현[a] · 송종우[a,1]

[a]이화여자대학교 통계학과

(2018년 9월 6일 접수, 2018년 10월 17일 수정, 2018년 11월 30일 채택)

요 약

지하철은 많은 승객들을 원거리까지 안전하고, 신속·정확하게 원하는 지점으로 대량 수송할 수 있는 친환경적인 교통수단이다. 지하철의 공익성을 증대시키기 위해서는 정확한 승객 수요 예측이 이루어져야 한다. 본 연구는 정확한 지하철 수요예측을 위하여, 군집분석을 통해 서울시 1–9호선 지하철역들을 군집화 하였다. 그 후, 전체 역과 각 군집 별 최종 예측 모형을 제시하였다. 군집화 결과, 294개의 역이 3개로 군집화 되었으며 그룹 1은 상공업지구, 그룹 2는 주상복합지구, 그룹 3은 주거지구가 중심이 되는 역들로 나타났다. 그 후 각 군집 별로 다양한 데이터 마이닝 기법을 이용해 지하철 승차인원 예측 모형을 제시하고, 수요 예측에 중요한 영향을 미치는 요인들을 도출하였다. 그리고 최종 모형을 바탕으로 2018년 10월에 개통될 서울시 9호선 3단계 연장역인 8개 신설역의 3개월 수요를 예측하였다. 8개 신설역의 월평균 시간당 평균 승차인원은 약 241에서 452명, 월평균 시간당 최대 승차인원은 약 969에서 1,515명으로 추정되었다. 본 분석의 최종 모형을 활용한 신설역의 지하철 수요 예측은 대중교통 정책 결정을 위한 기초자료로 활용되어 효율적인 지하철 운영 방안 수립에 기여할 수 있을 것이다.

주요용어: 지하철, 수요예측, GMM, 익스트림 그래디언트 부스팅, 랜덤 포레스트, 선형 모형

이 논문은 2018년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2017R1D1A1B03036078).

본 논문의 초본은 2018년 한국교통연구원 교통 빅데이터 활용 우수논문 및 아이디어 공모전에서 장려상을 받았음.

[1]교신저자: (03760) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과.

E-mail: josong@ewha.ac.kr