일별 지하철 이용자 수에 대한 시계열 분석

2 조 김민규 노현경 유태윤

1. 개요

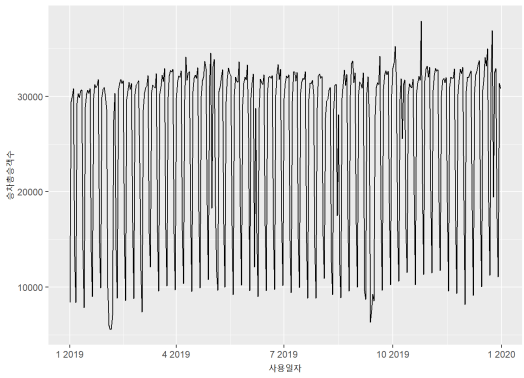
지하철은 서울에서 가장 많이 이용되는 대중교통 수단이다. 그 중에서도 2 호선은 서울의 중심부를 순환하면서 많은 사람들을 태운다. 특히 출퇴근 시간대에 2 호선을 이용하면 발 디딜 틈도 없이 사람들로 가득한 지하철을 경험할 수 있다. 이렇게 사람들 속에 끼어서 지하철을 타다 보면 사람들이 몰리는 시간대를 피해 지하철을 타고 싶다는 생각이 든다. 만약, 지하철의 승객 수 데이터를 분석하여 알맞은 시계열 model 을 fitting 한다면 사람들이 몰리는 시간대가 언제인지 파악할 수 있을 것이다. 이런 의미에서 지하철 승객 수 데이터는 굉장히 뚜렷한 패턴을 가지고 있다. 출퇴근 시간대에 사람들이 몰리는 것도 그렇고 평일에는 출퇴근하는 사람들이 많은 역에 사람들이 몰리고, 주말에는 유흥지 등 놀거리가 많은 역에 사람들이 몰리는 등 요일별로도 상당히 뚜렷한 시계열 패턴이 있다. 만약 알맞은 시계열 model 을 fitting 하여 지하철 승객 수를 예측할 수 있다면, 지하철 배차 간격 조정 및 내부 공사 시기 설정 등에 응용될 수 있을 것이다. 이에 프로젝트 주제를 지하철 승객 수에 대한 시계열 분석으로 설정하였고, '서울열린데이터광장'에 공개되어 있는 지하철 승객 수 데이터 셋을 이용하여 분석을 진행하였다.

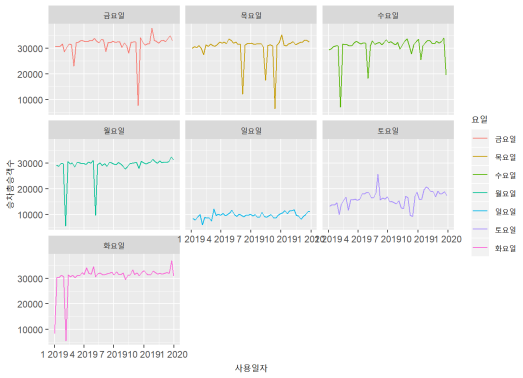
서울에 있는 역의 개수는 총 681 개로 모든 역을 대상을 분석하기에는 시간적 제약이 있었기 때문에 역들을 특징적인 몇 개의 group 으로 분류하고, 각 group 의 대표적인 역들을 하나씩 분석하는 것으로 프로젝트 목표를 설정하였다. Grouping 방법은 2 호선의 역들을 대상으로 사용일자 및 요일 별 승차총승객수 그래프를 그려 개형이 비슷한 역들끼리 묶었다. 이런 방법으로 grouping 하면 사용일자 별 승객 수에 대한 시계열 모델을 설정하였을 때 각 group 별로 유사한 시계열 model 을 얻을 수 있을 것이기 때문이다. 이에 본 프로젝트에서는 서울 지하철 2 호선에 해당하는 역들을 6 개의 group 으로 분류하였고, 각 group 에 대해 Seasonal ARIMA model 을 수립하였다. 또한, 주어진 data 를 training set 과 test set 으로 분류하여 training set 을 이용해 model 을 수립한 후 과적합진단, 잔차분석을 통해 model 을 검증하였다. 또한, 수립한 model 로 prediction 을 진행하여 이를 test set 과 비교, 분석하였다.

2. Grouping the subway stations

2 호선의 역들을 대상으로 사용일자 및 요일 별 승차총승객수 그래프를 그린 결과 비슷한 개형의 그래프를 가진 역들이 많았고, 이에 역들을 그래프의 개형에 따라 여섯 개의 group 으로 나누었다.

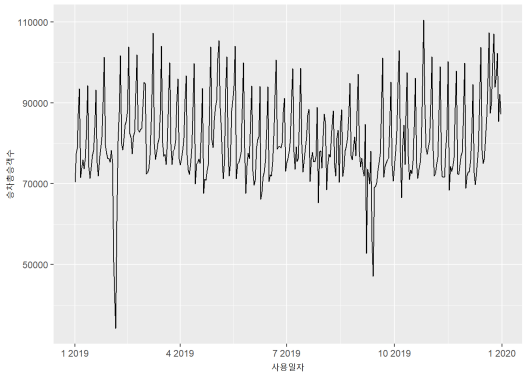
Group 1. 승차 인원이 평일에 많고 주말에 적은 group

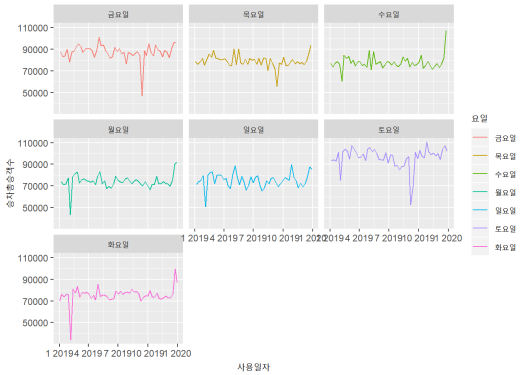




첫 번째 그룹은 주말에 비해 평일에 승차인원이 많은 역들이다. 대표적으로 시청, 영등포구청 등이 여기에 속하며, 출퇴근을 위해 이용하는 사람들이 많다는 특징이 있다.

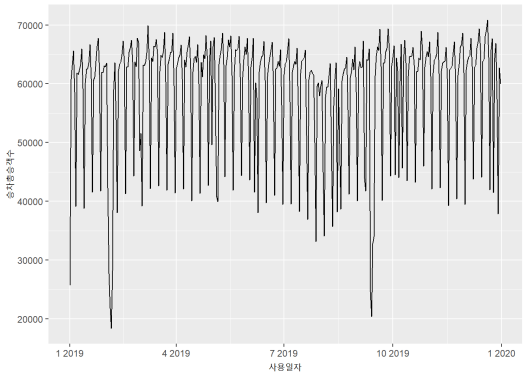
Group 2. 승차 인원이 평일보다 금, 토요일에 많은 group

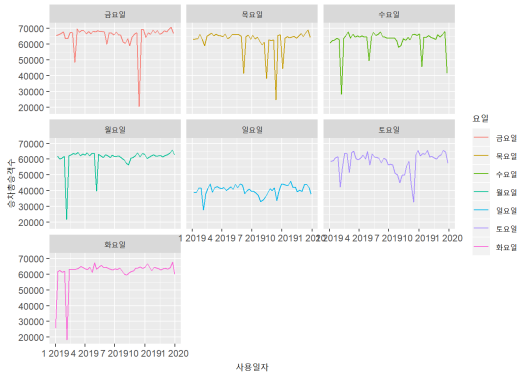




두 번째 그룹은 평일에 비해 금요일과 토요일에 승차인원이 많은 역들이다. 대표적으로 홍대입구, 동대문역사문화공원 등이 여기에 속하며, 서울의 대표적인 유흥지라는 특징이 있다.

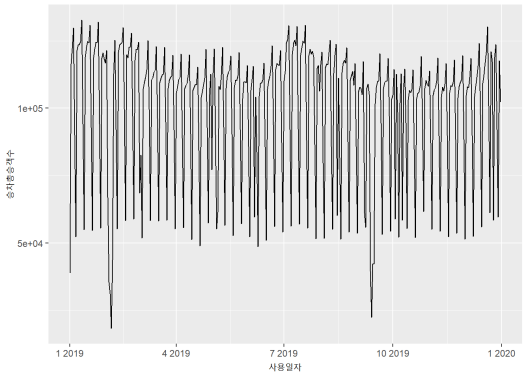
Group 3. 일요일만 승차인원이 적은 group

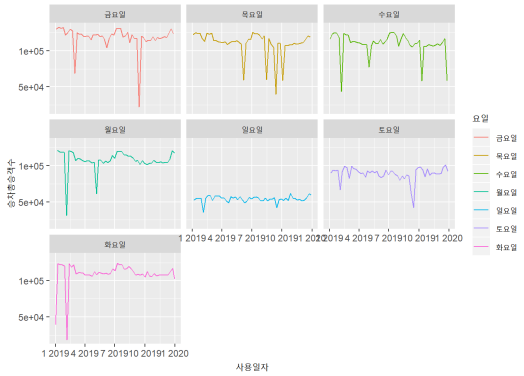




세 번째 그룹은 일요일에만 승차인원이 적은 역들이다. 대표적으로 신도림, 사당 등이 여기에 속하며, 승객수가 많은 환승역이라는 특징이 있다.

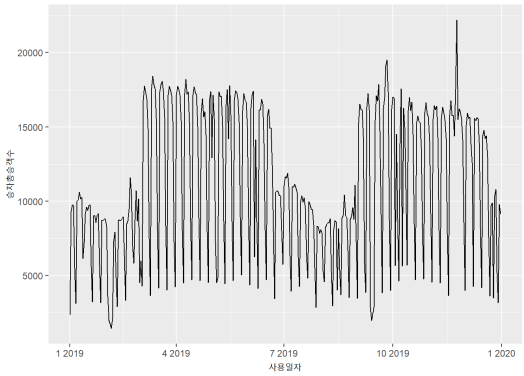
Group 4. 일요일의 승차 인원이 가장 적고 토요일이 중간 정도인 group

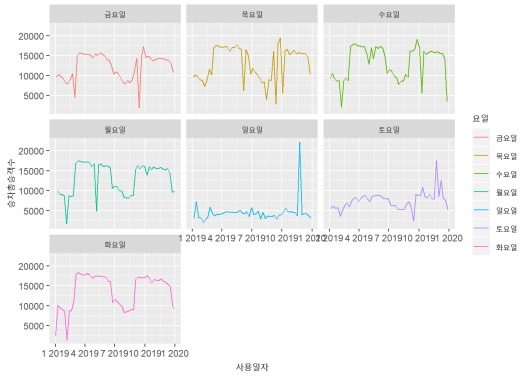




네 번째 그룹은 주말에 승차인원이 적지만, 일요일의 승차 인원이 토요일에 비해 특징적으로 더 적은 역들이다. 대표적으로 강남, 서울대입구 등이 여기에 속하며, Group 1 과 Group 2 의 특징을 모두 갖고 있다. 즉, 출퇴근을 위해 이용하는 사람들도 많으면서 유흥지로서의 성격도 있는 역들이 group 4 에 속한다.

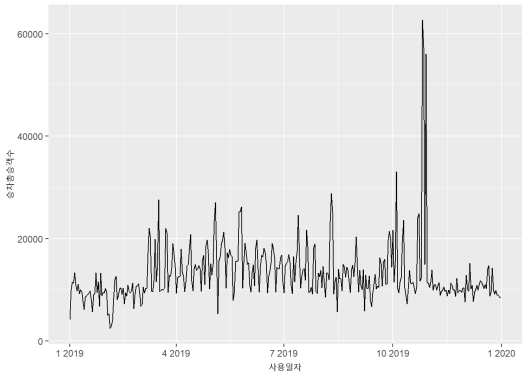
Group 5. 특정 기간에 승차인원이 많은 group

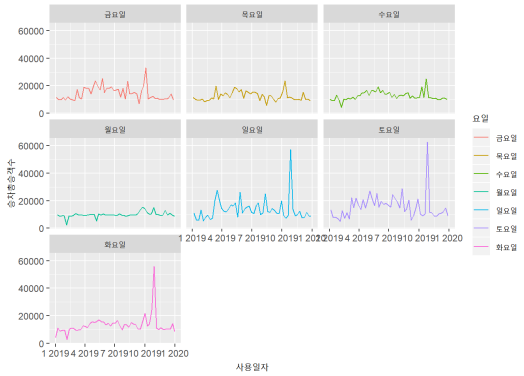




다섯 번째 그룹은 특정 기간에 승차인원이 더 많은 역들이다. 대표적으로 한양대 등이 여기에 속하며, 승차 인원의 대부분이 통학하는 학생들로 구성되었다는 특징이 있다.

Group 6. 특정 날에만 승차인원이 많은 group

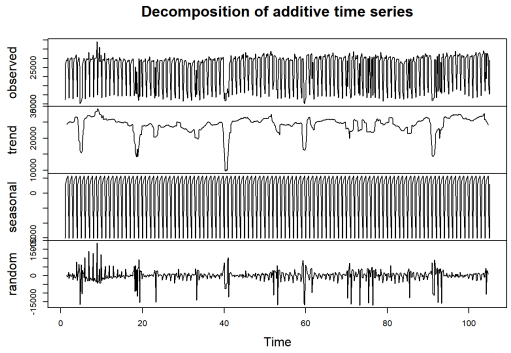


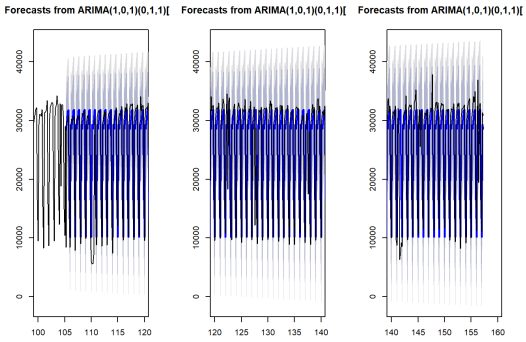


여섯 번째 그룹은 특정 날에 승차인원이 특징적으로 많은 역들이다. 대표적으로 종합운동장 등이 여기에 속하며, 콘서트 등 대형 행사의 개최지라는 특징이 있다.

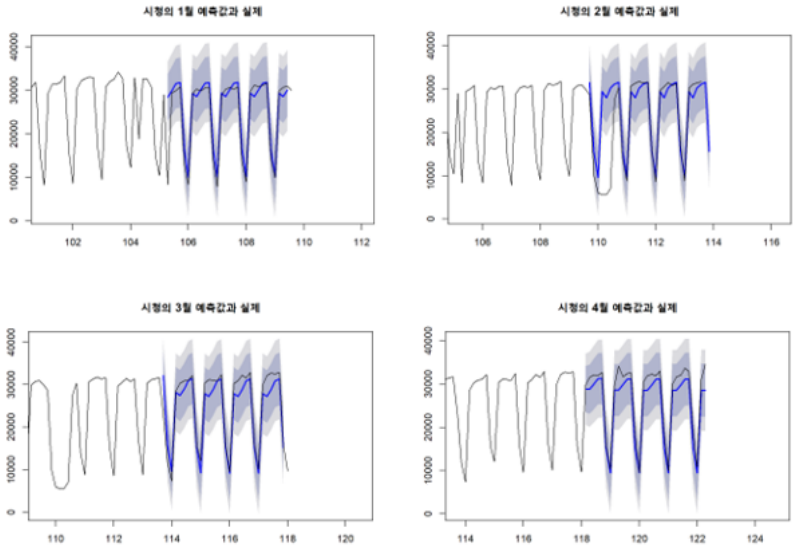
3. 모델 선택 및 예측

Group 1. 승차 인원이 평일에 많고 주말에 적은 group (시청)





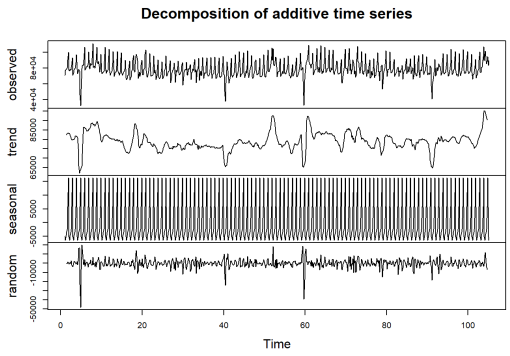
위 그래프를 확인하면, 시청의 경우 명확한 주기를 가지고 있는 것을 알 수 있다. 7일을 주기로 하였을 때, seasonal 성분의 분리가 잘 이루어졌고, ARIMA(1,0,1)(0,1,1)[7] 로 적합되었다. 17~18년 데이터를 가지고 19년 승차인원을 예측해보았다. 예측결과, 예측값의 분산은 많이 커지지 않았지만, 예측을 잘 하지는 않았다. 동일 모델로 예측 월 직전의 데이터로 그 다음월을 예측해보았을 때, 1년 예측보다 정확했다.

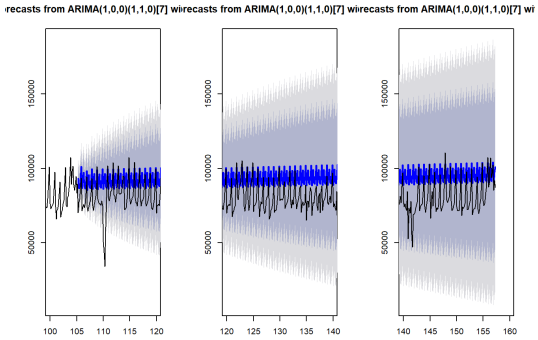


2월 예측을 보면 예측값보다 실제값이 확연하게 적은 날이 있었는데, 이는 설날 연휴로 인해 승차인원이 줄어든 것으로 보였다.

다음으로, 적합된 모델로 잔차의 평균과 acf 를 알아보고, Lyung-Box test, 과적합분석을 진행해보았다. 그 결과, 잔차의 평균은 35.22, acf는 0에 가까운 값이 나왔고, 잔차간의 자기상관이 없다는 결과가 나왔다. 또, AR(2), MA(2) 모델로 적합했을 때 보다 해당 모델의 AIC 가 작게 나와서 ARIMA(1,0,1)(0,1,1)[7] 모델은 본 데이터에 잘 적합된 모델이라고 판단하였다.

Group 2. 승차 인원이 평일보다 금,토에 많은 group (홍대입구)

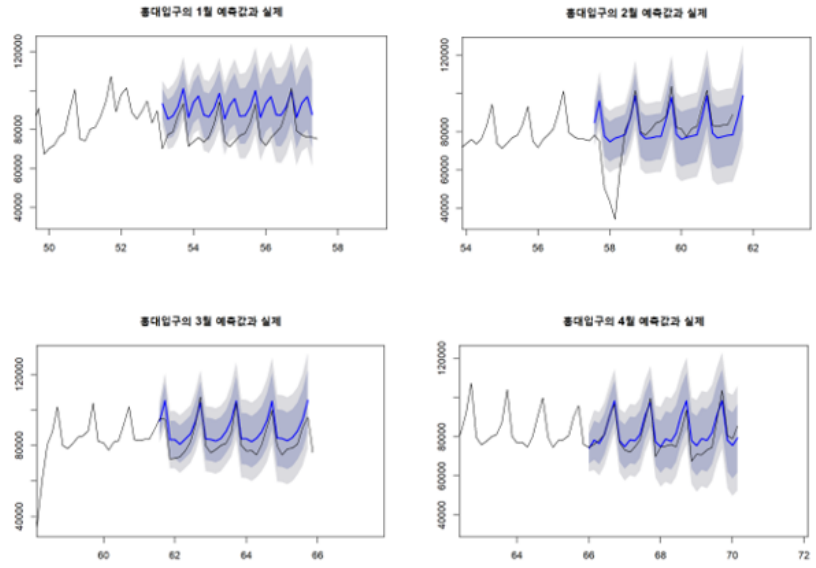




홍대입구의 경우도 명확한 주기를 가지고 있었다. 7일을 주기로 하였을 때, seasonal 성분의 분리가 잘 이루어졌고, ARIMA(1,0,1)(0,1,1)[7] 로 적합되었다.

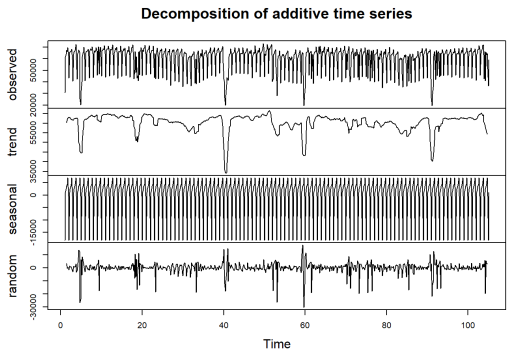
하지만, 1년치 예측을 하였을 때, 실제 데이터와 오차가 클 뿐더러 예측치의 분산이 점점 커졌다. 적합된 모델로 잔차분석을 한 결과, 잔차의 평균은 -4.89로 0에 가까웠지만, 잔차간의 자기상관이 있다는 결과가 나왔다.

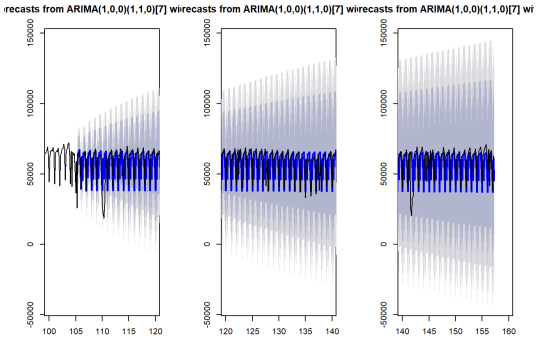
모델을 바꿔 ARIMA(2,0,2)(1,1,0)[7]으로 적합해보았다.



적합 결과, 그 전 모델보다 실제와 유사해진 것을 확인할 수 있다. 잔차의 acf 가 0에 가까운 값들이 나왔고, Lyung-Box test에서 잔차의 자기상관성이 없다고 나왔다.

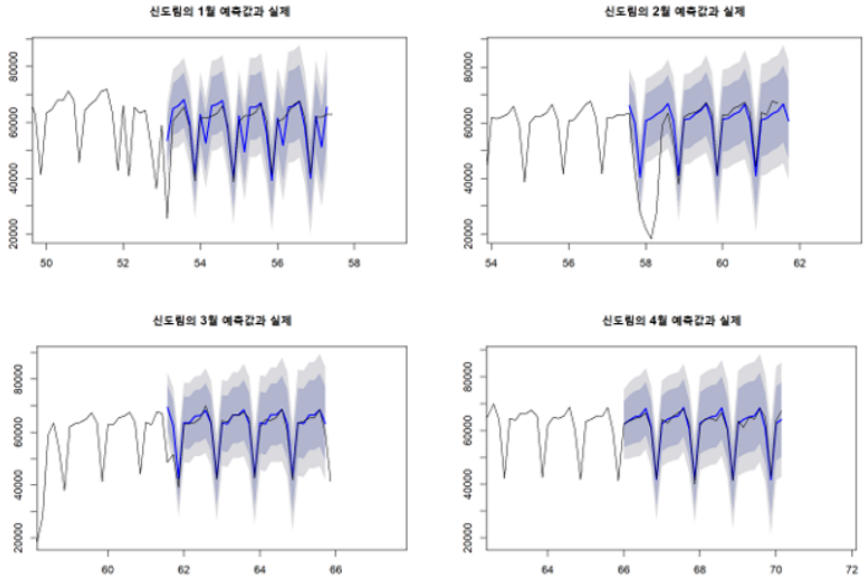
Group 3. 일요일만 승차인원이 적은 group (신도림)





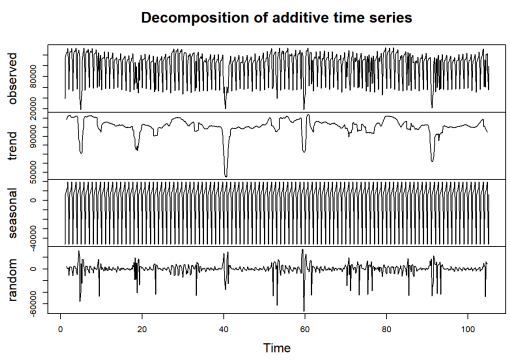
신도림의 경우도 1주일의 주기가 있었다. ARIMA(1,0,0)(1,1,0)[7] 로 적합하고 19년 승차인원을 예측해보았다. 그 결과, 예측도 잘 맞지 않았고 연말로 갈수록 분산이 커졌다.

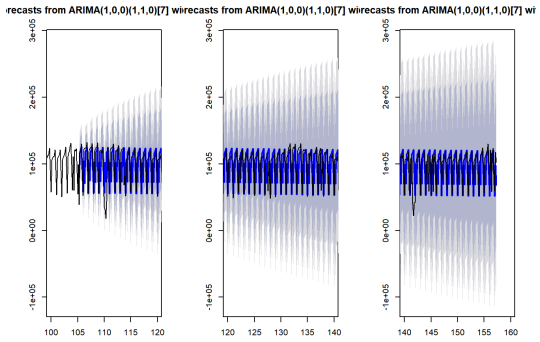
예측의 분산을 줄이기 위해, 월별로 예측해보았고, 모델도 ARIMA(2,0,0)(2,1,0)[7]로 바꾸어 적합했다.



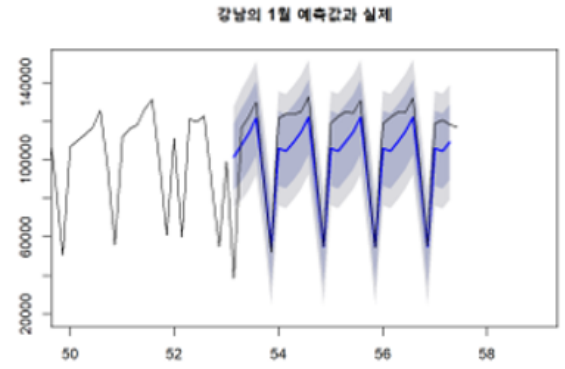
그 결과, 2월의 설날연휴를 제외하고 잘 예측했다. 잔차분석 결과, 잔차의 자기상관이 없다고 판단되었다.

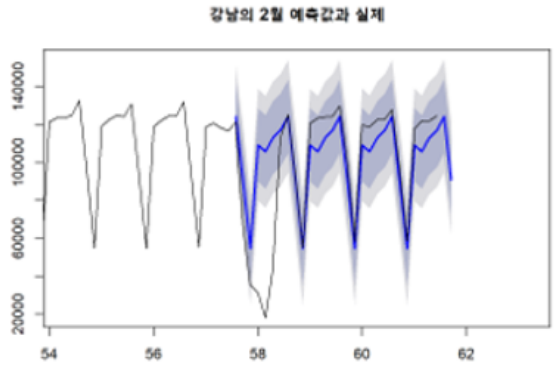
Group 4. 일요일의 승차인원이 가장 적고 토요일이 중간 정도인 group

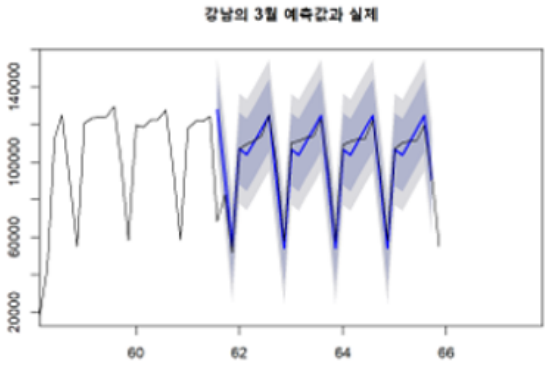


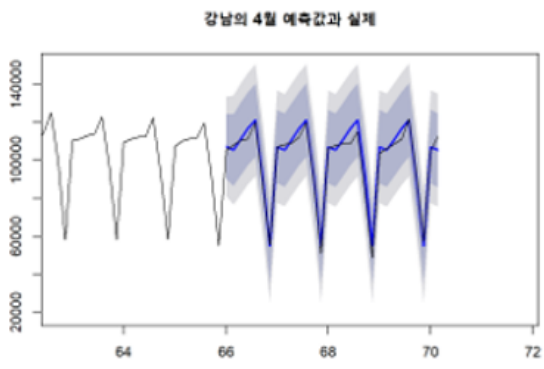


강남 역시 7일의 주기를 가졌다. ARIMA(1,0,1)(0,1,1)[7] 모델로 적합하였지만, 1년치의 예측은 좋지 못했다. 예측의 분산을 줄이기 위해 월별 예측을 진행하고, 모델을 바꾸어 ARIMA(2,0,1)(0,1,1)[7]로 적합하였다.



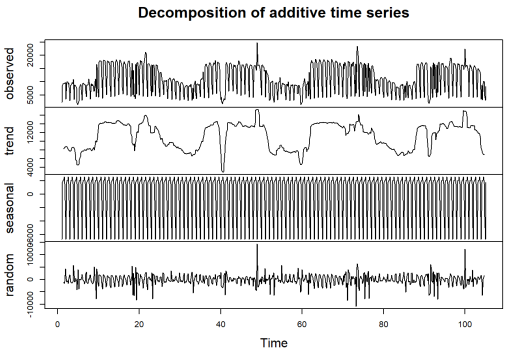


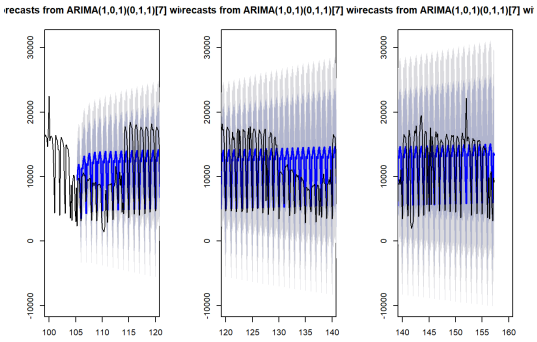




그 결과, 1,2월의 예측은 다소 부정확했지만, 3,4월의 예측은 정확했다. 잔차분석을 진행했을 때, 잔차의 자기상관이 없다고 판단되었다.

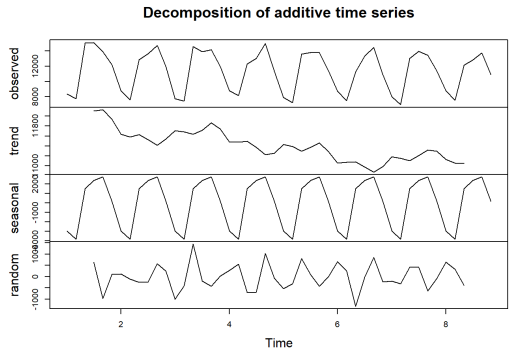
Group 5. 특정 기간에 승차인원이 많은 group (한양대입구)

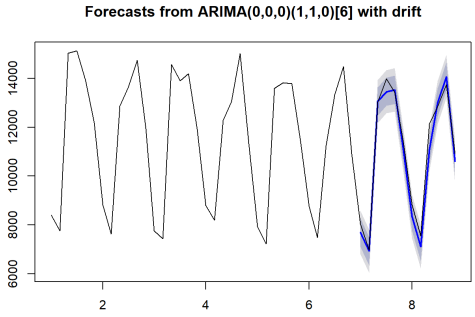




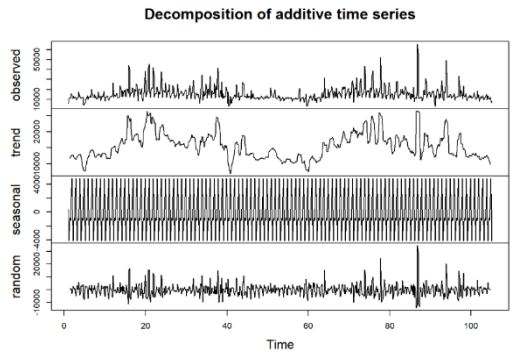
한양대입구의 경우, R의 findfrequency 함수를 사용했을 때 1주일이라는 주기를 찾아주기는 했지만, 그래프상 보이는 6개월의 주기 (한학기와 방학)은 찾지 못하였다. 6개월의 주기와 1주의 주기를 한꺼번에 적합하기에는 데이터가 부족하다고 판단하였다. ARIMA(1,0,1)(0,1,1)[7] 로 모델 적합을 했을 때, 예상대로 특정 기간동안 승차 인원이 많고 적음을 제대로 예측하지 못했다.

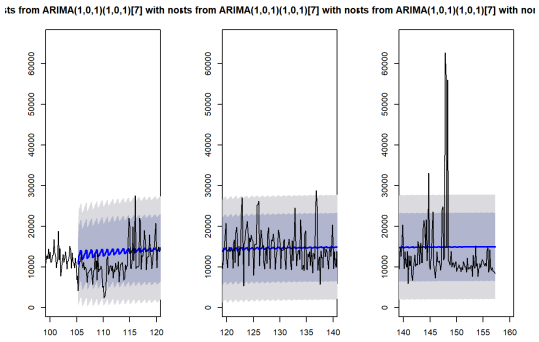
한양대입구의 승차인원 특성을 고려해 월별 평균 승차인원으로 데이터를 파악하고, 월 단위 예측을 시도해보았다. 그 결과, 6개월의 주기가 나왔고, ARIMA(0,0,0)(1,1,0)[6] 모델로 적합해보았다. 예측은 잘 이루어졌고, 잔차분석 결과 자기상관성이 없다고 나왔다.





Group 6. 특별한 날에 승차인원이 많은 group (종합운동장)





종합운동장의 경우 콘서트장의 영향으로 인해 특별한 경우에만 승차인원이 많았다. 따라서, 트렌드와 주기를 파악하기 어려웠다. ARIMA(1,0,1)(0,1,1)[7] 모델을 적합시켜 보았지만, 예측의 정확도가 크게 떨어졌다.

4. Conclusion

본 프로젝트에서는 2호선의 역들을 대상으로 사용일자 및 요일 별 승차총승객수 그래프를 그려 비슷한 개형의 역들을 6개의 group으로 분류하였다. 이후 각 group의 대표적인 역들에 대해 18년도 이전까지의 data를 training set으로 설정하여 Seasonal ARIMA model을 수립하였고, 과적합 진단과 잔차분석을 통해 fitting된 model을 검증하였다. 마지막으로, 수립한 model을 통해 prediction을 진행하여 19년도의 data와 비교, 분석하였다.

그 결과, 첫 번째 group 은 승차 인원이 평일에 많고 주말에 적은 group 으로 대표로 시청역을 선정하였다. ARIMA(1,0,1)(0,1,1)[7] model 을 fitting 하였고 prediction 결과 한 달 예측이 비교적 잘 맞는 경향이 있었다. 두 번째 group 은 승차 인원이 평일보다 금, 토에 많은 group 으로 대표로 홍대입구역을 선정하고 ARIMA(2,0,2)(1,1,0)[7] model 을 fitting 하였다. 세 번째 group 은 일요일만 승차인원이 적은 group 으로 대표로 신도림역을 선정하고 ARIMA(2,0,0)(2,1,0)[7] model 을 fitting 하였다. 네 번째 group 은 승차 인원이 평일에 많고 주말에는 일요일에 비해 토요일에 더 많은 group 으로 대표로 강남역을 선정하였다. ARIMA(2,0,1)(0,1,1)[7] 모델을 fitting 하였으며 prediction 결과 1, 2 월에 예측 결과에 비해 승객수가 더 많은 경향이 있었고, 이를 방학 때 성형, 라식 등 시술의 이유로 강남역을 찾는 사람이 많기 때문일 것이라 생각하였다. 다섯 번째 group 은 특정 기간에 승차인원이 많은 group 으로 대표로 한양대입구역을 선정하고 ARIMA(1,0,1)(0,1,1)[7] model 을 fitting 하였으나 주기가 7 인 seasonal model 로는 학기 중에만 승차 인원이 많아지는 특성을 분석할 수 없었다. 이에 승차인원의 월별평균을 계산하여 이를 통해 새로운 model 을 수립해 보았고, ARIMA(0,0,1)(1,1,0)[6] model 로 6 개월의 주기를 찾아낼 수 있었다. 마지막 group 은 특정 날에만 승차인원이 많은 group 으로 대표로 종합운동장역을 선정하였지만, 콘서트 날짜 등 변수가 너무 많아 modeling 을 할 수 없었다.

결론적으로, 본 프로젝트에서는 여러 group 에 속하는 역들에 대한 Seasonal ARIMA model 이 어느 정도 잘 맞는 것을 확인하였으며 이를 활용하여 사람들이 많이 다니는 날에는 배차를 늘리는 등 배차 조정에 이용하거나 지하철역 내부 또는 출구 공사를 진행할 때 사람들이 많이 다니는 시기를 피해 공사를 진행하는 등 여러 가지로 활용할 수 있을 것이다.

프로젝트 진행 후, 크게 두 가지 방면에서 프로젝트를 확장해보고 싶다는 생각이 들었다. 먼저 자동으로 지하철 역을 Grouping 하는 Algorithm 을 만들어보고 싶다. 본 프로젝트에는 grouping 을 할 때 각 역들에 대한 그래프를 일일이 그려가면서 group 에 분류하였는데, 만약 특정 역의 승객 수 데이터를 받은 후 각 요일별 승객 수 패턴을 분석하여 특징에 맞게 자동으로 classification 할 수 있다면 모든 역에 대한 classification 을 손쉽게 할 수 있을 것이다. 또한, 시간대별 승차인원을 분석해보고 싶다. 평일과 주말을 나누었을 때 시간대별 승차인원이 너무 다르기 때문에 이들을 나눠서 각각 다른 model 로 fitting 해보고, 이렇게 세운 model 이 사람들이 많이 몰리는 시간대를 잘 예측할 수 있는지 확인해보고 싶다.