시계열 모형을 이용한 공공자전거 이용 패턴 분석:

서울시 '따릉이' 이용자 수 데이터를 중심으로

서울대학교 2024-1 시계열분석 및 실습 (001)
팀1
2017-10483 이정민
2017-12906 홍준성
2020-11842 강지수

1 서론

1.1 연구 목적

공공자전거는 효율적이고 친환경적인 교통수단 중 하나로서, 2010년대 중반부터 여러 지자체에서 시민의 호응 속에 운영되기 시작했다. 특히 서울시에서 운영하는 공공자전거 '따릉이'는 2015년 도입 이후 수요가 꾸준히 상승하였다 (남궁옥 외, 2021). 실제로 2022년에는 회원 수 41만여 명, 이용자는 4,000만여명에 육박할 정도로 서울시의 친환경 교통수단으로 잘 자리 잡았다 (김두일, 2023). 그러나 낮은 이용요금 탓에 따릉이의 운영적자도 함께 증가하고 있다 (김지현, 2024). 공공자전거 서비스의 지속성을 위해서는 이용 패턴에 관한 분석이 선행되어야 한다. 그러나 지금까지의 여러 선행 연구는 대체로 특정 짧은 기간 동안의 이용 패턴만을 분석하거나 공공자전거의 수요를 예측함에 통행 목적을 고려하지 않는 등의 한계를 보여왔다. 본 프로젝트에서는 2022년~2023년 2년간의 서울시 공공자전거 대여 이력 데이터를 정기권과 당일권으로 나누어 시계열 분석하는 방법론을 적용하여 공공자전거 이용 각 유형의 추세를 비교하였다.

1.2 선행 연구 검토

공유 자전거는 최초 출발지(최종 목적지)인 집이나 회사에서 대중교통 시점(종점)인 역, 정류장까지의 이동을 매개하는 first/last mile 교통수단으로 주로 분석됐다 (서울특별시, 2021). 그러나 따릉이 대여 데이터를 분석한 여러 선행 연구에서는 다른 목적으로 이용하는 패턴도 발견되었다.

곽민정 외(2022)는 공공자전거의 통행 유형을 세 가지로 구분하여 각 유형 간 상호관계를 분석하였다. 구체적으로 지하철 역 간 통행(Type 1), 대여 또는 반납지점 중 한 곳만 지하철역 근처인 지하철 연계통행(Type 2), 그 외 기타 지역 간 통행(Type 3)으로 구분하였다. 분석 결과 유형별로 평균 이용시간 및 이용 거리에 차이가 있으며, 평일-주말 여부에 따라 통행 유형의 비중이 달라지는 현상도 확인하였다. 그러나 2020년 10월 한 달에 한정한 데이터 분석 때문에 통행 유형의 변화 추세에 관해서는 확인하지 못했다는 한계를 가진다.

이선재(2024)는 도로 수준에서 따릉이가 어떻게 이용되고 있는지를 진단하고자 2021년 10월 한

달간의 따릉이 이동 궤적 데이터를 분석하였다. 분석 결과, 따릉이가 first/last mile 교통수단 외에도 자가용을 운전하지 않는 청년 세대를 중심으로 주요 교통수단으로 활용되는 이용 패턴을 확인하였다. 이는 개인 소유 자전거가 전제로 하는 집/직장에서의 왕복 통행 패턴이 공유 자전거 이용에서도 전체 대여 중 약 9%에 해당하는 비율로 발생했다는 사실을 근거로 한다. 일방향 통행은 출퇴근 시간대(7~10시, 17~20시)에 집중되는 반면 왕복 통행은 점심 시간대(11~12시)에 두드러지는 등 시간대 분포 분석을 통해서도 고유한 특성을 발견할 수 있었다. 또한, 통행거리와 목적지 점 방문 등의 지표를 이용해서 레저 목적 이용인지 일상생활 목적 이용인지를 구분하였다.

김민혁(2018)은 2016년 9월부터 2017년 9월까지의 따릉이 대여 이력 자료에 시계열 군집 분석과 역 간 벡터 회귀모형을 이용하여 수요를 예측하였다. 이를 위해 수요 패턴이 같은 대여소들을 L2-norm 거리함수를 기준으로 하여 총 18개의 군집으로 묶었으며 그 결과 군집 분석을 하지 않은 모형에 비해 평균 제곱근 오차(RMSE)가 12.8%, 평균 절대오차(MAE)가 12% 감소한 높은 예측력의 모형을 얻을 수 있었다. 그러나 기상, 주변 시설 등의 외부 요인을 배제한 채 시계열 분석을 진행했다는 점과, 거리 함수와 군집 수를 결정한 근거가 뚜렷하지 않다는 점에서 한계를 가진다.

1.3 연구 가설

본 프로젝트는 "따릉이의 이용 패턴에 따라 수요가 서로 다른 추세를 보일 것이다"는 가설을 검증하고자 한다. 이를 위해 이용 패턴을 구분할 수 있는 변수로 '정기권 여부'를 설정하였다. '정기권 여부'는 따릉이 대여에 사용한 이용권 유형을 뜻한다. 회원 전용 정기권과 일일권 두 가지로 구성되어 있으며, 이러한 특성 때문에 워싱턴 DC의 자전거 대여 서비스를 분석한 Younes, H.(2020)의 연구에서도 회원과 비회원 사이의 이용 특성 차이를 포착할 수 있었다. 이와 같은 연구 사례를 통해 정기권 이용자는 공공 자전거를 주로 출퇴근 등 일상생활 목적으로 이용하고, 일일권 이용자는 레저 목적으로 이용할 것이라는 가설을 세워볼 수 있다.

그러나 정기권 여부 외에 다른 요인으로도 이처럼 공공 자전거 대여 목적을 구분할 수 있다는 점을 고려해야 한다. 예를 들어 앞서 언급한 이선재(2024)의 연구에서는 '왕복 여부', 통행 거리, 경로 상 정차 구간의 유무 등을 따져 이에 따른 이용 패턴을 분석한 바 있다.

본 연구에서는 모형을 단순화하기 위해 독립 변인을 '정기권 여부'로 한정하였다. 선행연구에서 지적한 모든 변인을 한 번에 다루는 자료가 없었다. 여러 변인 중 정기권 여부는 이산형 변수여서 통행 거리와 같은 연속형 변수와는 달리 군집화를 위한 임의의 기준을 세워야 할 부담이 없고, 정기권 여부는 다른 요인에 비해 운영 주체인 서울시가 요금제를 변경하는 방식을 통해 직접 관여하기 상당히 유리하다는 장점이 있다. 실제로 서울시는 2024년 '공공자전거 요금 현실화 방안 학술용역'을 공고하고 요금 분석에 나서고 있는 만큼 (김지현, 2024), 정기권 여부에 근거한 시계열 분석 결과는 그 실용적 가치도 높을 것으로 기대된다.

2 분석 방법

2.1 역할 분담

본 프로젝트의 역할 분담은 <표 1>과 같다. 주기적인 회의를 통해 각자 주로 담당하는 부분의 진행 상황을 공유하고, 다른 팀원에게 피드백을 받는 과정을 성실히 수행하였다. 이를테면 '시계열 모형 적합' 과정이 홍준성 학생에게 배정되었지만, 다른 두 팀원이 R 코드 작성에 전혀 관여하지 않았다는 뜻은 아니다.

<표 1> 프로젝트 역할 분배

역할	주요 담당자
주제 선정 및 선행연구 조사	강지수, 이정민, 홍준성
데이터 수집 및 전처리	이정민
탐색적 데이터 분석	강지수
시계열 모형 적합	홍준성
발표자료 제작	강지수
보고서 작성	강지수, 이정민, 홍준성
발표	이정민

2.2 분석 자료

본 프로젝트에서 분석 기간은 2022년 1월 1일~2023년 12월 31일(총 730일)로 한정하였다. 해당 기간은 2024년 현시점에서 서울 열린데이터광장에서 제공하는 자료 중 가장 최신의 자료를 마지막으로 하여 시계열 분석에 충분한 수준의 데이터인 2주기(2년) 이상을 확보하고자 하는 취지에서 결정했다.

서울 열린데이터광장(https://data.seoul.go.kr/)에서 제공하는 CSV(Comma-Separated Values) 파일 데이터를 내려받아 서울시 공공자전거의 이용정보를 분석하였다. 구체적으로 대여소, 정기권 유무, 성별, 연령대별로 총계 처리(aggregation)한 일별 이용정보인 '서울시 공공자전거 이용정보(일별)'을 사용하여 일별 공공자전거 대여량을 분석하였다. 대여 유형은 단체권 이용 정보를 배제하였고, 일일권 사용은 회원 사용자와 비회원 사용자를 구분하지 않았다.

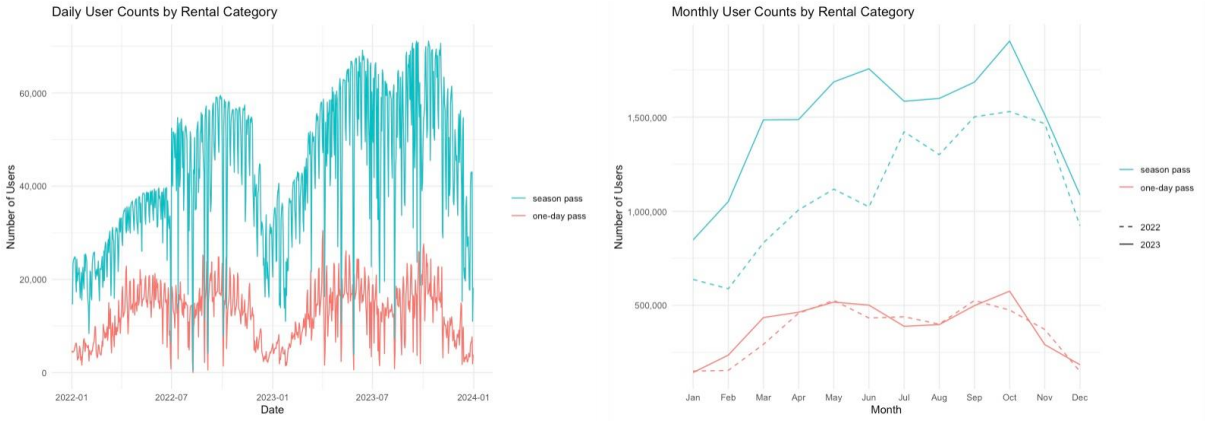
추가로 자전거 이용량에 상당한 영향을 미치는 것으로 널리 알려진 기상인자(기온, 강수량)를 수집하기 위해서는 기상청 기상자료개방포털(https://data.kma.go.kr/)의 기후통계분석 서비스를 이용하였다. 서울특별시에 존재하는 108개 지점에서 관측한 최저기온, 평균기온, 최고기온, 강수량 정보를 수집하였다. 본 분석에 사용하지는 않았지만, 한국천문연구원에서 제공하는 공휴일 정보와 요일 정보를 취합해 7일 주기에서 더 나아가 휴일 여부가 자전거 이용에 미치는 영향을 확인하였다.

2.3 탐색적 자료 분석

자료 분석 및 모형 적합 과정 전반에는 통계 분석용으로 널리 쓰이는 프로그래밍 언어인 R을 사용하였다 (R Core Team, 2024).

이용자 수 자료에 이상치가 존재하는지 확인하기 위하여 1.5 IQR 방식을 사용하여 탐색하였다. 그 결과 이상치가 나타나지 않았으므로 추가적인 처리는 진행하지 않았다. 이는 분석 자료가 이미 총계 처리된 자료이기에 처리 과정에서 이상치가 나타날 가능성이 떨어지기 때문으로 추정할 수 있다.

<그림 1> 서울시 공공자전거 '따릉이' 이용자 수 일별(좌) 및 월별(우) 추이

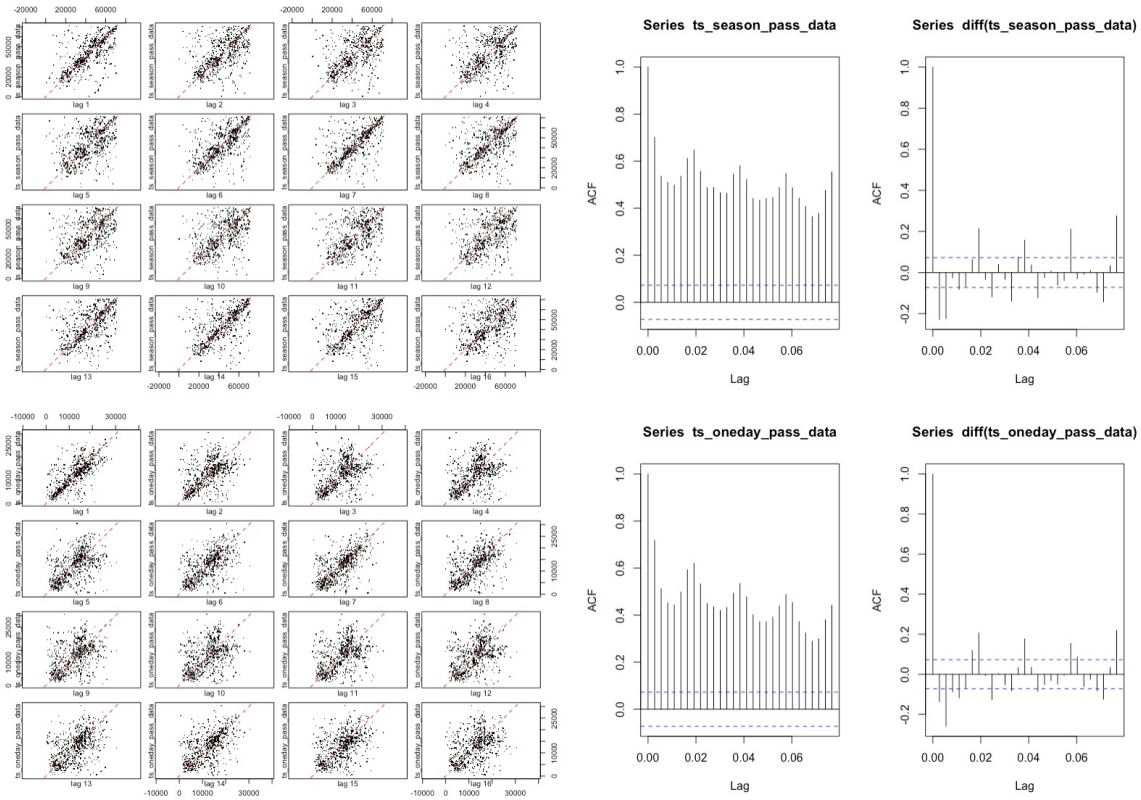


'서울시 공공자전거 이용정보' 데이터를 토대로 시계열도를 작도한 결과, 계절성이 뚜렷하고 분산이 매우 큰 시계열 자료를 확인할 수 있었다(그림 1). 계절과 시기에 따라 큰 변동폭을 보이지만, 분석 기간 동안의 일일 평균 따릉이 이용자 수는 약 5.5만 명이며 월평균 이용자 수는 약 166만 명으로 집계되었다. 따릉이 이용자 중 정기권(Season Pass) 이용자의 비중은 평균적으로 약 75%로 나타났는데, 2023년 정기권 이용자가 일일권(One-Day Pass) 이용자와 비교하면 두드러지게 증가한 점을 고려한다면 대여권 종류에 따라 서로 다른 시계열 모형을 적합하는 것이 타당하다.

스펙트럼 분석을 통해 도출한 두 시계열 자료(정기권, 일일권)의 주기는 7일이었으며, <그림 2>의 Lag plot 및 ACF plot의 결과 또한 이를 뒷받침했다. Lag plot은 lag 7일 때 산점도가 가장 선형적인 형태를 보이고, 차분된 시계열의 ACF plot에서는 0.02의 배수 lag에서 spike가 관찰된다. $(365 \times 0.02 \simeq 7)$

일별 자료의 특성상 365.25의 주기로 시계열을 바라보는 것이 일반적이지만, 본 프로젝트에서는 2년 치의 데이터만 사용하기에 과적합이 우려되고 자료의 추세를 뚜렷하게 확인할 수 없을 것으로 예상하였다. 실제로 적합 시 과적합되었음을 확인하였다. 따라서 모형 복잡도 및 노이즈가 많은 자료의 특성을 고려하여 시계열의 주기를 앞서 도출한 7일로 고정하였다

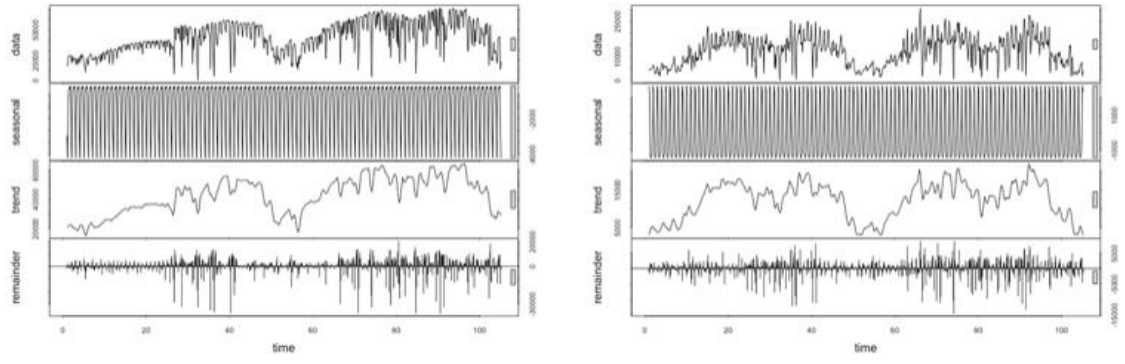
<그림 2> 이용자 수 시계열 자료 계절성 주기



3 분석 결과 및 비교

3.1 정상성 확인

<그림 3> 시계열 STL 분해 (좌: 정기권, 우: 일일권)



LOESS(Locally Estimated Scatterplot Smoothing)의 방법을 사용하여 각 시계열을 분해한 결과 정기권 이용자와 비교하면 일일권 이용자의 수의 변동이 불규칙적이지만, 전체적인 따릉이 이용자 수의 추세는 비슷한 것으로 드러났다 (그림 3). 또한, 추세와 계절성을 제외한 나머지 시계열 성분이 정상성을 만족하는지 확인하기 위하여 정상성 검정(Augmented Dickey-Fuller test)을 시행하

였다. 정기권은 검정 통계량과 p-value는 각각 -17.149와 0.01로 나타났고, 일일권은 검정 통계량과 p-value는 각각 -17.444와 0.01로 나타났다. 따라서 두 시계열 모두 유의수준 0.05 하에서 정상 시계열임을 확인했다.

3.2 모형 적합 및 진단

모형 적합에 앞서 분산이 매우 크고 불안정했기 때문에 분산 안정화를 위해 Box-Cox 변환을 진행했다. 이후 추세와 계절성을 고려하여 적합 시에는 승법 계절(multiplicative seasonal) ARIMA 모형을 우선으로 고려하였다 (이상열, 2023). 모형 선택 과정은 `forecast` 패키지의 `auto.arima()` 함수로 자동화하였다 (Hyndman, Rob, et al., 2024). 이때, `stepwise` 파라미터를 `FALSE`로 설정하여 order 5 이내의 모든 모형 중 AIC가 가장 작은 모형을 선택하였다.

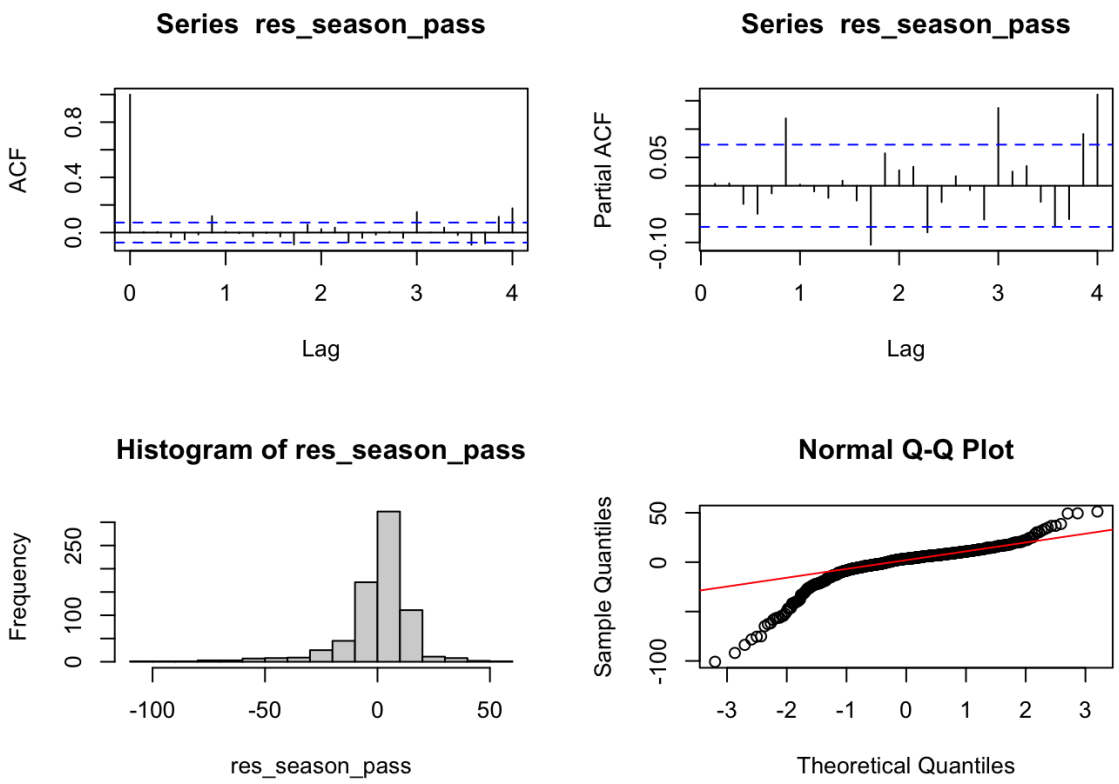
<표 2> 정기권, 일일권 ARIMA 모형 추정 결과

	변수	계수	표준 오차	
정기권	ma1	-0.5834	0.0352	sigma^2 = 256.6
	ma2	-0.3092	0.0343	log likelihood = -3055.17
	sma1	0.1446	0.0386	AIC=6120.33
	sma2	0.0324	0.0324	AICc=6120.42
				BIC=6143.29
일일권	ar1	0.3133	0.0549	sigma^2 = 8.284
	ar2	-0.1759	0.0429	log likelihood = -1803.15
	ar3	-0.0325	0.0428	AIC=3618.29
	ar4	-0.1244	0.0442	AICc=3618.41
	ma1	-0.8154	0.0441	BIC=3645.84

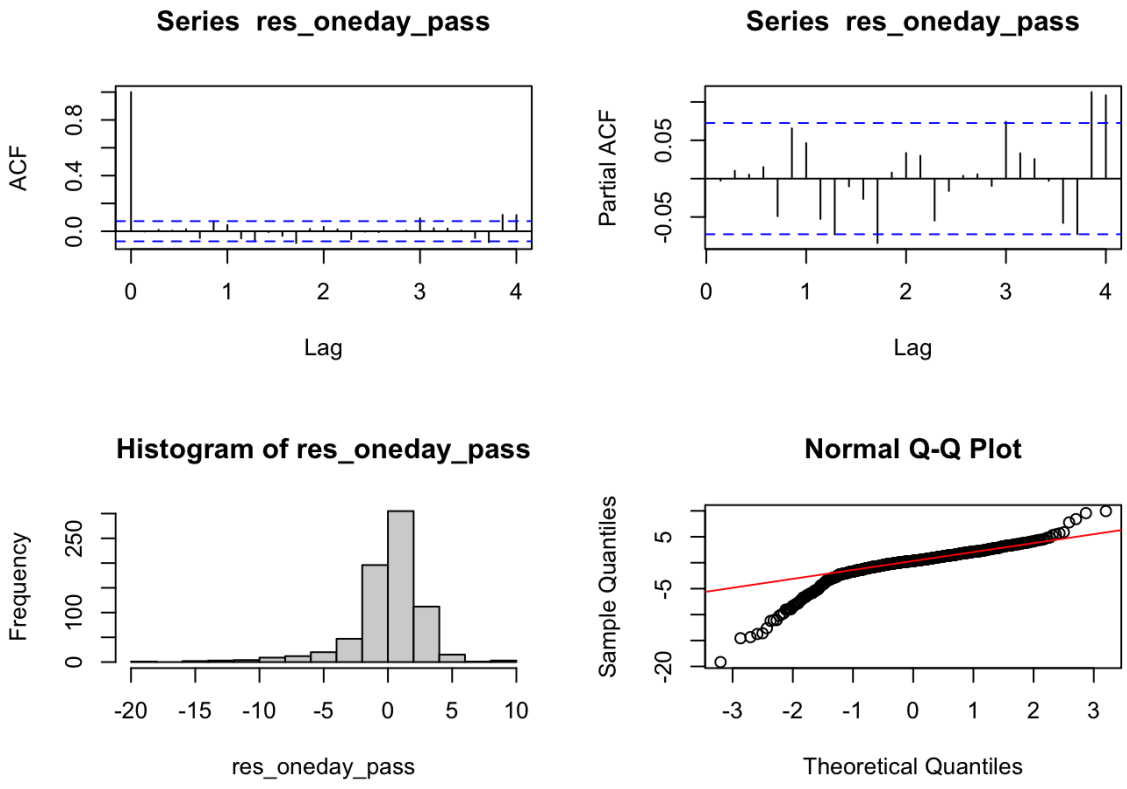
정기권 시계열에서는 SARIMA(0,1,2)(0, 0, 2)$_7$ 모형이, 일일권 시계열에서는 ARIMA(4,1,1) 모형이 가장 낮은 AIC를 보였으므로 해당 모형을 선택하였다(표 2). 모든 모수가 유의수준 0.05 하에서 유의하게 나왔으며 오차항의 분산은 각각 256.6, 8.3으로 추정할 수 있었다. 이후 두 모형을 각 시계열의 잠정적 모형으로 설정하고 잔차 분석 및 과적합 분석을 진행하였다.

잔차 분석 결과 두 모형의 ACF 플롯에서 시차 1, 3, 4지점을 제외하고 기준선 아래로 절단된 형태가 나타났다. PACF 플롯의 경우에도 시차 1, 3, 4지점 근처에서 기준선을 넘는 spike가 발견되었다. 잔차의 자기상관성을 확인하는 Box-Ljung 검정 결과 두 시계열 모두 잔차가 독립적이라는 귀무가설을 기각하지 못하였다. 반면 정규성에 대한 Jarque Bera 검정을 실행한 결과 두 시계열 모두 유의확률이 작아 잔차들의 분포가 정규성을 가진다고 할 수 없었고, 잔차도의 첨도(kurtosis)가 높게 나타났기에 과적합을 의심해볼 여지가 있다. 따라서 p, q의 차수를 1씩 높인 모형으로 다시 적합하고, 모수의 변화 양상과 추가된 모수의 추정량을 확인해보았다. 그 결과, 정기권과 일일권 시계열의 기존 모수에서는 0.02 이내에서 변화했으며, 추가된 모수는 유의수준 0.05 하에서 유의하다고 나타나지 않았다. 이를 근거로 두 모형은 과적합되지 않았다고 결론지었다.

<그림 4> 정기권 시계열 잔차 분석 결과



<그림 5> 일일권 시계열 잔차 분석 결과



3.3 외생변수 도입

모형의 설명력을 높이고 예측 오차를 줄이기 위해 자전거 이용자 수와 직관적으로 가장 밀접한 관련이 있다고 판단한 기상인자(일 평균 기온, 강수량)를 외생변수로 도입하였다. ARMA 모형에 외생변수를 공변량(covariates)으로 반영하는 방법에는 크게 ARMAX 모형과 Regression with ARMA errors의 두 가지 방법이 있다 (Hyndman, R., 2010).

ARMAX 모형은 단순히 ARMA 모형의 우변에 회귀항을 추가한 것이다. ARMAX 모형에 차분 계수와 계절 성분을 추가하면 아래 식과 같이 표현되는 SARIMAX 모형을 만들 수 있다. 그러나 이 모형에서의 회귀 계수 β는 autoregressive term $\phi(B)\Phi(B^s)\nabla^d\nabla_s{}^D$의 영향이 섞여 나타나기 때문에 종속변수와 외생변수 사이의 관계를 제대로 표현하지 못 한다는 한계가 있다.

$$\phi(B)\Phi(B^s)\nabla^d\nabla_s{}^D Y_t = \beta X_t + \theta(B)\Theta(B^s)\epsilon_t$$

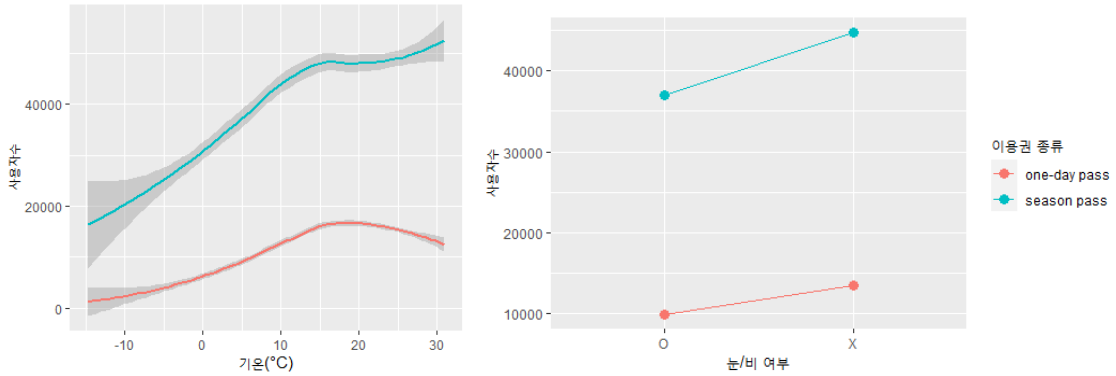
Regression with SARIMA errors 모형은 종속변수(따릉이 이용자 수)를 외생변수(일평균 기온, 강수량)로 선형회귀를 진행하고 오차항에 SARIMA를 가정하는 모형으로써 본 시계열 자료에서 불규칙적으로 이용자 수가 변화하는 부분을 효과적으로 설명할 수 있다. Regression with SARIMA errors 모형 식은 아래와 같다.

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \epsilon_t \; where \; \epsilon_t \sim SARIMA(p,d,q)(P,D,Q)_s$$

3.4 외생변수 도입 모형 적합 및 진단

기상 지표를 외생변수로 도입하기 전에 이용자 수와의 관계를 <그림 6>에서 확인하였다. 이용권 종류와 관계없이 대체로 기온이 증가하면 이용자 수도 함께 증가했다. 그러나 20도 이상의 고온에서 일일권 이용자 수는 감소하는 경향을 보였다. 정기권 이용자 수는 고온에 큰 영향을 받지 않았는데, 워싱턴 DC에서의 자전거 대여 서비스를 연구한 Younes, H. (2020)도 비회원이 회원보다 기상 요인에 더 예민하다는 같은 경향을 발견하였다.

<그림 6> 이용자 수와 일평균 기온, 이용자 수와 눈/비 여부의 관계



기온과 더불어 강수량이 이용자 수에 미치는 영향을 확인하려 했지만, 강수량이 0에서 조금만 증가해도 이용자 수가 감소하며 분산이 크게 나타났다. 이 때문에 제대로 된 비교가 어려웠으므로 강수량 자체가 아닌 강수량이 0인 날과 0이 아닌 날로 나누어 그 관계를 확인하였다. <그림 6>을

보면 일일 강수량이 이용자 수에 영향을 주는 것을 확인할 수 있다. 외생변수 도입 이전 모형과 동일하게 `auto.arima()` 함수를 이용하여 정기권, 일일권 이용자 수에 대한 잠정적인 모형을 도출하였고, 이후 잔차 분석과 더불어 과적합 여부를 판단하였다.

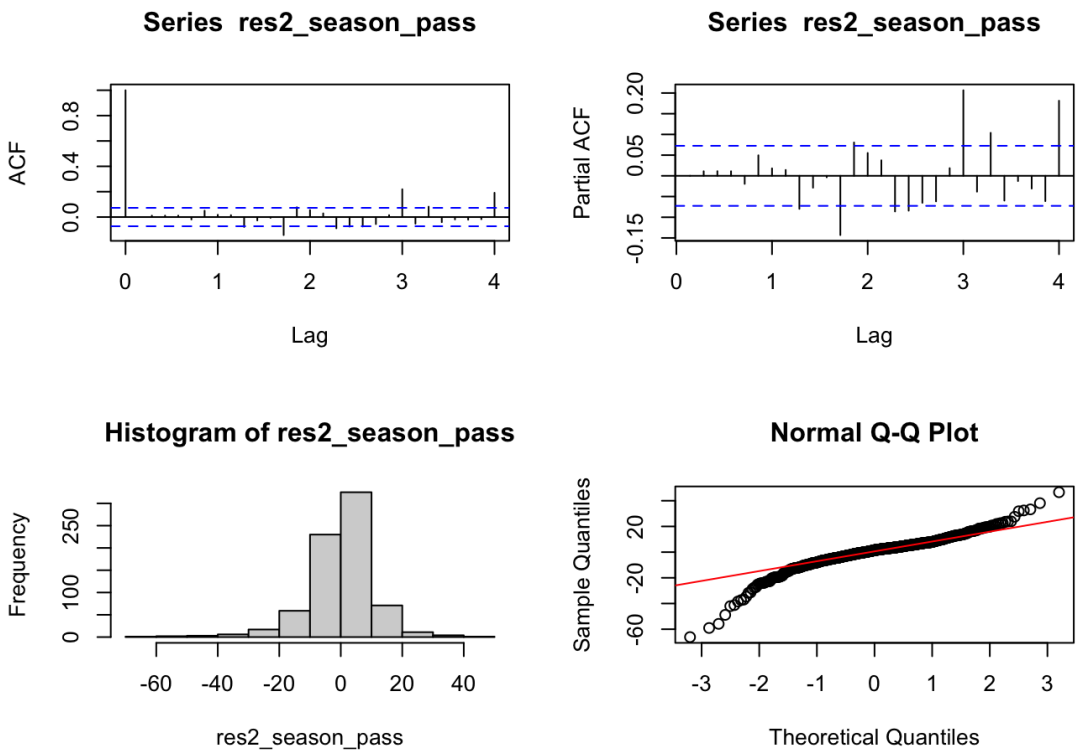
적합 결과 앞서 예상한 바와 같이 회귀 계수는 두 개의 모형 모두 유의하게 나왔지만, 상대적으로 정기권 이용자가 기상인자의 영향을 더 크게 받는다는 것을 회귀계수로부터 알 수 있다 (표 3). 다만, 정기권 및 일일권의 일일 평균 이용자 수와 σ^2의 추정량을 고려해보았을 때 유의미한 차이라고 볼 수는 없다. AIC와 BIC를 토대로 두 모형을 잠정적 모형으로 설정하고 이후 모형 진단을 잔차 분석을 통해 진행하였다.

<표 3> 정기권, 일일권 regression with ARIMA 모형 추정 결과

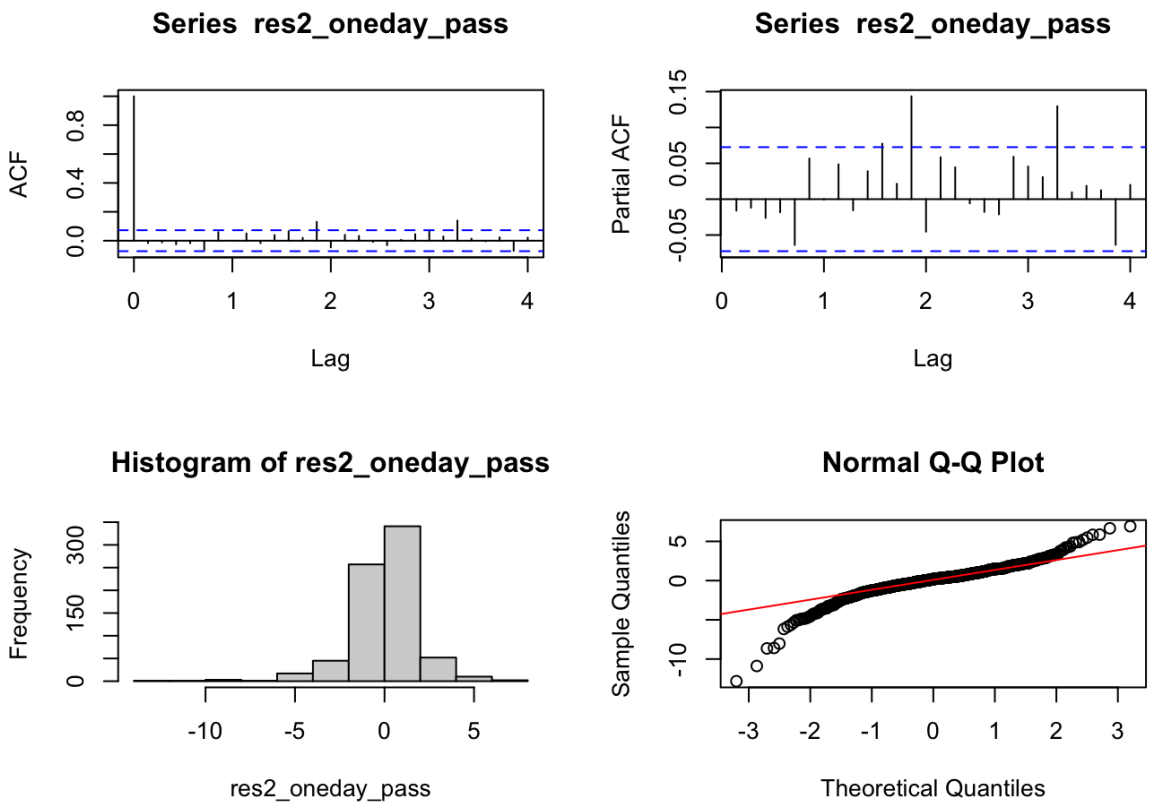
	변수	계수	표준 오차	
정기권	ar1	0.2320	0.0454	sigma^2 = 120.3
	ar2	-0.1021	0.0437	log likelihood = -2776.8
	ar3	-0.0582	0.0417	AIC=5573.6
	ar4	-0.1375	0.0434	AICc=5573.91
	ma1	-0.8826	0.0307	BIC=5619.52
	sma1	0.2201	0.0432	
	sma2	0.1024	0.0338	
	평균기온(℃)	1.0683	0.1236	
	강수량(mm)	-0.7024	0.0262	
일일권	ar1	0.3306	0.0393	sigma^2 = 3.479
	ar2	0.0434	0.0394	log likelihood = -1487.11
	ar3	0.0838	0.0391	AIC=2996.22
	ar4	0.0280	0.0394	AICc=2996.59
	ar5	0.1543	0.0375	BIC=3046.75
	sar1	0.8882	0.0391	
	sma1	-0.7060	0.0659	
	intercept	24.4998	0.5349	
	평균기온(℃)	0.285	0.020	
	강수량(mm)	-0.1238	0.0044	

자기상관성에 대한 Box-Ljung 검정을 실행한 결과 모든 시계열이 귀무가설을 기각하지 못해 잔차가 독립적인 것으로 나타났고, 정규성에 대한 Jarque Bera 검정을 실행한 결과 두 시계열 모두 유의확률이 낮아 잔차들의 분포가 정규성을 가진다고 할 수 없었다. 그렇기에 다시 한번 과적합 여부를 판단하게 되었다. 잠정 모형에서 AR 부분과 MA 부분의 모수를 하나씩 늘렸을 때 두 모형 모두 기존 모수 추정량은 큰 변화가 없었으며 추가된 모수 또한 유의하다고 볼 수 없었다.

<그림 7> 정기권 시계열 잔차 분석 결과 (외생변수 도입 모형)



<그림 8> 일일권 시계열 잔차 분석 결과 (외생변수 도입 모형)



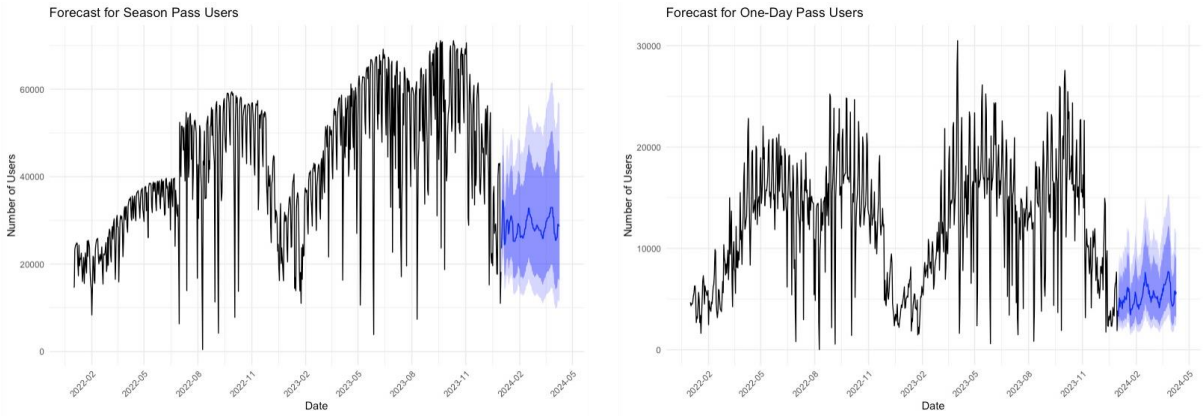
3.5 최종 모형 선택 및 예측

분석 결과 강수량과 일평균 기온이라는 외생변수를 고려하지 않은 모형과 외생변수를 고려한 모형 중 공공자전거 이용자 수를 예측하는 데 더 뛰어난 모형을 최종적으로 결정하였다. 결정 기준으로는 AIC, AICc, BIC 모두를 고려하였다. 해당 모형들의 지표는 <표1>에서 확인할 수 있다. 그 결과 정기권과 일일권 자료 모두 외생변수를 고려한 regression with ARIMA errors 모형이 더 좋은 값을 가졌으므로 해당 모형을 최종 모형로 선택하였다.

<표 4> 정기권과 일일권 자료 시계열 모형의 AIC, AICc, BIC

	정기권		일일권	
	외생변수 X	외생변수 O	외생변수 X	외생변수 O
모형	ARIMA(0,1,2)(0,0,2)$_7$	ARIMA(4,1,1)(0,0,2)$_7$	ARIMA(4,1,1)	ARIMA(5,0,0)(1,0,1)$_7$
AIC	6120.33	5573.6	3618.29	2996.22
AICc	6120.42	5573.91	3618.41	2996.59
BIC	6143.29	5619.52	3645.84	3046.75

<그림 9> 정기권과 일일권 자료 시계열 모형의 예측 시계열도



4 결론 및 향후 연구 과제

4.1 결론

본 프로젝트는 최근 2년간의 서울시 공공자전거 '따릉이'의 수요에 대한 시계열 자료를 분석하여 다양한 특성을 확인하였다. 구체적으로 스펙트럴 분석과 lag plot 분석을 이용하여 따릉이 자료는 7일의 주기성을 가짐을 검증하였고, 데이터의 비정상성과 분산의 불안정함에 대응하기 위해 Box-Cox 변환과 차분 변환을 도입하였다. 또한, 서울시 평균 기온과 강수량을 외생 변수로 도입한 모형으로 자료를 적합해 았고, 그 결과 두 요인 모두의 유의성을 확인할 수 있었다.

AIC와 잔차도를 기준으로 정기권과 일일권 자료 모두 regression with ARIMA errors 모형을 최종

모형으로 선택하였다. 분석 모형과 모수에 차이가 있는 것으로부터 정기권 여부에 따라 대여량의 변화 추세에 차이가 있음을 검증할 수 있었다. 따라서 서울시 '따릉이'의 지속 가능성을 위해서는 더 많은 서울 시민들이 반복적으로 따릉이를 이용하는 정기권 사용자가 될 수 있도록 전환을 장려하는 정책 등의 도입을 검토해볼 필요가 있다.

4.2 한계점과 발전 방향

외생변수를 도입하기 위해 기상 요인은 대여량과 선형 관계라고 가정했는데, 실제 자료에서는 그렇지 않아 오차가 발생할 수 있다. 평균 기온이 섭씨 20도를 넘어가는 범위에서부터는 오히려 음의 상관관계를 보인다는 비선형적 증거를 3.5절에서 확인하였다. 비선형성을 다루기 위한 추가적인 방법으로는 신경망 기반 모델 등 비선형 관계를 더 잘 포착할 수 있는 회귀 모형을 도입하는 것이 제안될 수 있다.

적합된 모형의 잔차를 분석했을 때 Ljung-Box 검정에서는 잔차의 자기상관성을 확인할 수 없었으나 Jarque-Bera 검정의 결과 잔차가 정규분포를 따르지는 않음을 확인하였다. 이는 예측 신뢰구간의 추정치의 신뢰성에 영향을 줄 수 있으므로 수요 예측에서 주의해야 한다. 이와 같은 잔차의 비정규성에는 원본 자료의 불안정한 분산이 영향을 주었을 것으로 의심된다.

본 연구 자료의 이분산성을 다루는 한 가지 해결책으로는 SARIMAX 또는 regression with ARIMA errors 모형을 사용하되 오차항의 분산을 상수로 가정하지 않고 GARCH로 적합하는 SARIMAX-GARCH 모형이 있다. 이처럼 ARIMA와 GARCH 모형을 결합하여 조건부 평균과 이분산성을 한 번에 다루려는 시도는 특히 주가를 예측하는 경제학 연구를 중심으로 활발히 이루어져 왔으나, 교통량 예측 문제에 적용하였을 때 일반적인 ARIMA 모형을 사용했을 때에 비해 예측 성능이 크게 차이가 나지 않는다는 연구 사례가 있다 (Chen, 2011). 따라서 만약 SARIMAX-GARCH 모형의 적용을 추후 고려하게 될 경우 모형의 복잡성이 증가한 것에 비하여 예측 성능의 향상이 두드러지는지 등을 엄밀히 검증할 필요가 있다.

본 연구에서는 일일권 사용자와 정기권 사용자의 자료를 완전히 분리하여 각각 적합하고, 최종 적합된 모형의 형태와 공유하는 계수의 값이 다르다는 근거로 두 시계열의 변화 추세에는 차이가 있다고 결론을 내렸다. 그러나 이렇게 서로 다른 그룹의 시계열 자료(Time-Series -Cross-Section Data)를 그룹 간의 계수 차이를 허용하되 하나의 모형에서 적합하는 random coefficient model로 해석하는 방법론이 제안되어 있다 (Beck & Katz, 2006). 향후 연구에서는 random coefficient model에 대한 이론적 배경을 연구하여 더 명확하게 그룹 간 패턴 차이를 보일 수 있을 것으로 기대한다.

시계열분석 및 실습수업 시간에 배운 내용을 현실의 데이터에 적용하는 팀 프로젝트를 진행하면서, 현실 데이터 분석의 어려움을 체감하였다. 원본 데이터를 정제하는 과정에서 주관성을 배제하고 이상치의 영향을 줄이기 위해 많은 요인을 분석에서 배제하여야 했던 점이 아쉬웠다. 그럼에도 당초 세웠던 가설을 시계열 분석 방법론을 통해 검증하고 팀원끼리 적절히 역할을 분배하여 목적을 달성해냈다는 점에서 이번 프로젝트는 큰 의의를 가진다.

4.3 피드백 반영 사항

(Q1) 모델 적합 과정에서 학습 자료에 과적합되는 것을 방지하기 위해 train-test split 등의 방법을 사용하였는지 묻는 질문이 있었다. 본 연구에서는 잠정적 모형 선택 이후 진단 부분에서 모형의 차수를 증가시키고 모수 변화 양상을 관측하는 것으로써 과적합 여부를 판단했다. 또한 모형을 선택할 때 AIC 기준을 사용하여 모수가 많은 모형을 선택하는 경향을 억제하였다. 보다 구체적인 설명은 3.2절 모형 적합 및 진단, 3.4절 외생변수 도입 모형 적합 및 진단에 보충하였다.

(Q2) 사계절의 변화에 따른 주기성이 시계열도에 드러나는 것 같음에도 적합 과정에서 주기를 7일로 선택한 이유에 대한 질문이 있었다. 이는 시계열 주기를 1년(365.25일)으로 가정한다면 전체 자료가 2년 치로 주어진 만큼 과적한 문제가 발생할 것을 우려하였다. 보다 구체적인 근거는 2.3절 탐색적 자료 분석 단계에서 논하였다. 사계절 요인은 기온과 강수량 등의 외생변수를 이용하여 간접적으로 반영하는 것으로 문제를 해결하였다.

(Q3) 정기권과 일일권 사용자의 사용 목적을 출퇴근용과 레저용으로 나누어 가정한 부분에 대한 구체적 근거를 묻는 질문이 있었다. 공공자전거 이용 목적과 관계를 갖는 변수는 정기권 여부 이외에도 여러 가지가 있음을 선행 연구를 통해 검토했지만, 본 연구의 취지와 자료 처리의 용이성을 고려하였을 때 정기권 여부에 따른 이용 패턴의 차이를 시계열 분석하는 것이 적합하다고 보았다. 구체적인 내용은 1.3절 연구 가설에서 보충하였다.

(Q4) 모형 적합 과정에서 mean을 고려하였는지, `auto.arima()`를 사용한 경우 과적합 진단을 진행하였는지 확인하는 질문이 있었다. `auto.arima()`를 이용하여 모형을 적합할 때 `allowdrift`, `allowmean` 매개변수는 기본값인 `TRUE`로 두었으므로 평균이 0이 아닌 경우를 고려하였다고 볼 수 있다. 외생변수를 고려하여 최종 선택한 모형을 기준으로 정기권 모형(d = 1)은 drift term을 고려하지 않았고, 일일권 모형(d = 0)은 intercept term을 고려하였다. 또한 각 모형의 과적합 진단 결과를 3.2절 모형 적합 및 진단(외생변수 도입 전)과 3.4절 외생변수 도입 모형 적합 및 진단에 추가하였다.

(Q5) SARIMAX 모형 적합에 사용한 패키지 함수의 구체적인 구현을 검토해보라는 피드백이 있었다. SARIMAX 모형은 기존의 SARIMA 모형에서 회귀 변수를 추가한 것이지만, 본 연구에서 사용한 모형은 Regression with SARIMA errors임을 확인하였다. 이 모형은 외생 변수를 독립변수로 하여 선형 회귀를 추가하고 회귀 오차항에 대하여 SARIMA를 적합하는 것으로 구현되며 구체적인 내용은 3.3절 외생변수 도입에 추가하였다.

5 참고 문헌

1. Beck, Nathaniel, and Jonathan N. Katz. "Random Coefficient Models for Time-Series—Cross-Section Data: Monte Carlo Experiments." Political analysis 15.2 (2007): 182–195. Web.
2. Chen, Chenyi, et al. "Short-time traffic flow prediction with ARIMA-GARCH model." 2011 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2011.
3. Hyndman, Rob, et al. "Forecast: Forecasting Functions for Time Series and Linear Models." 2024, https://pkg.robjhyndman.com/forecast/.

4. Hyndman, Rob. “The ARIMAX model muddle.” 2010년 10월 4일, https://robjhyndman.com/hyndsight/arimax/.
5. R Core Team. “R: A Language and Environment for Statistical Computing.” R Foundation for Statistical Computing, 2024, https://www.R-project.org/.
6. Younes, Hannah, et al. "Comparing the temporal determinants of dockless scooter-share and station-based bike-share in Washington, DC." Transportation Research Part A: Policy and Practice 134 (2020): 308-320.
7. 곽민정, 추상호, and 김상훈 “서울시 공공자전거와 공유 전동킥보드의 통행유형별 상호관계 분석: 서초 · 강남 · 동작구를 중심으로.” 대한교통학회지 40.6 (2022): 832–846.
8. 김두일. "'서울시 따릉이' 이용자만 연간 4000만명…안전교육 강화". 아주경제, 2023년 5월 2일, https://www.ajunews.com/view/20230502094535097.
9. 김민혁. 시계열 군집분석 기반 서울시 공공자전거 수요예측. Diss. 한양대학교, 2018.
10. 김지현, “[단독]15년째 1000원..서울 공공자전거 따릉이 적정요금 따져본다”. 머니투데이, 2024년 1월 30일, https://news.mt.co.kr/mtview.php?no=2024013009201542610.
11. 남궁옥, 박종한, and 고준호 “서울시 공공자전거 따릉이의 정성적 이용 행태 분석.” 교통기술과정책 18.6 (2021): 52–59.
12. 서울자전거 따릉이, “이용권 사용안내”. 2024년 5월 19일 수집. https://www.bikeseoul.com/info/infoCoupon.do.
13. 서울특별시, “서울시, 작년 ‘따릉이’ 이용 2,300만건 돌파…코로나시대 교통수단 각광”. 2021년 1월 21일, 서울특별시, https://news.seoul.go.kr/traffic/archives/504919.
14. 이상열. "시계열 분석 이론 및 SAS 실습." 파주: 자유아카데미 (2023).
15. 이선재 “서울시 공공자전거 이용 행태 기반의 도로구간 특성 해석 : 따릉이 이동 궤적 데이터를 중심으로 / 이선재 [electronic resource].” 서울대학교 대학원, 2024. Print.