ARIMA모형을 이용한

모기예보제 분석

응용통계학과

201932106 김수진

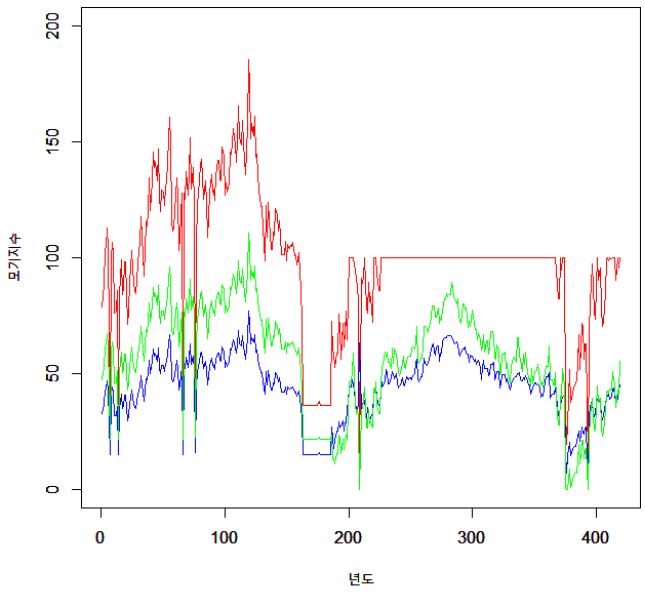
1. 서론

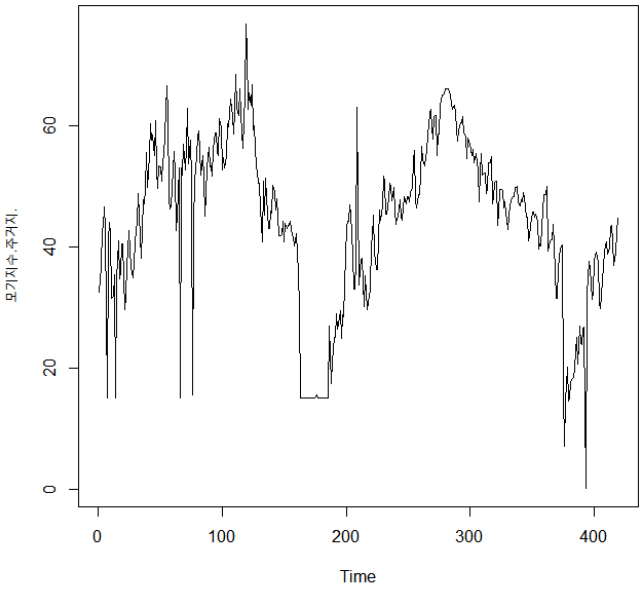
계절이 바뀔 때마다 우리는 계절이 주는 기쁨과 원하지 않는 것을 동시에 얻는다. 곧 다가오는 여름을 예로 들면 바다, 계곡 그리고 모기를 빼 놓을 수 없을 것이다. 우리나라는 말라리아 발생 국가이다. 말라리아란 '나쁜 공기'란 뜻의 이름을 가진 급성 열성 전염병으로, 말리라 원충에 감염된 모기에 물리는 것으로 2주부터 수개월 정도의 잠복기를 거쳐 증상이 발생하는 질병이다.(인천광역시교육청) 동남아시아와 같은 더운 나라에서만 발생한다고 흔히들 알고 있지만 25도 이상의 기온이 3개월 동안 지속되는 나라에서 발생할 수 있기 때문에 대한민국도 예외는 아니다. 말라리아에 걸리면 치아가 떨리는 심한 오한, 구토, 두통, 39도가 넘는 고열 등의 증상이 나타난다.

서울시에서는 모기예보 정보에 대한 일별 모기지수 발생일과 발생일의 모기지수 정보를 제공한다. 시민들은 모기예보제를 통해 모기지수를 확인함으로써 모기 발생 단계와 그에 따른 행동 규칙 및 방제법을 확인할 수 있다. 이를 활용하여 모기예방 활동을 할 수 있을 것으로 기대된다.

2. 본론

모기활동 지수는 모기가 활발히 활동하는 5월부터 10월까지 일별로 수집되었다. 기간에 해당하는 2020년 ~ 2022년 6월 1일까지 자료를 사용하자. 모기지수는 수변부, 공원, 주거지 3개 부분으로 나눠 측정되었다. 데이터는 서울 열린데이터 광장에서 얻었다. 2022년도부터 2022년도까지 3군데로 나눈 모기지수를 시각화해서 보자. 빨간색선은 수변부, 파란색은 주거지, 초록색은 공원이다.

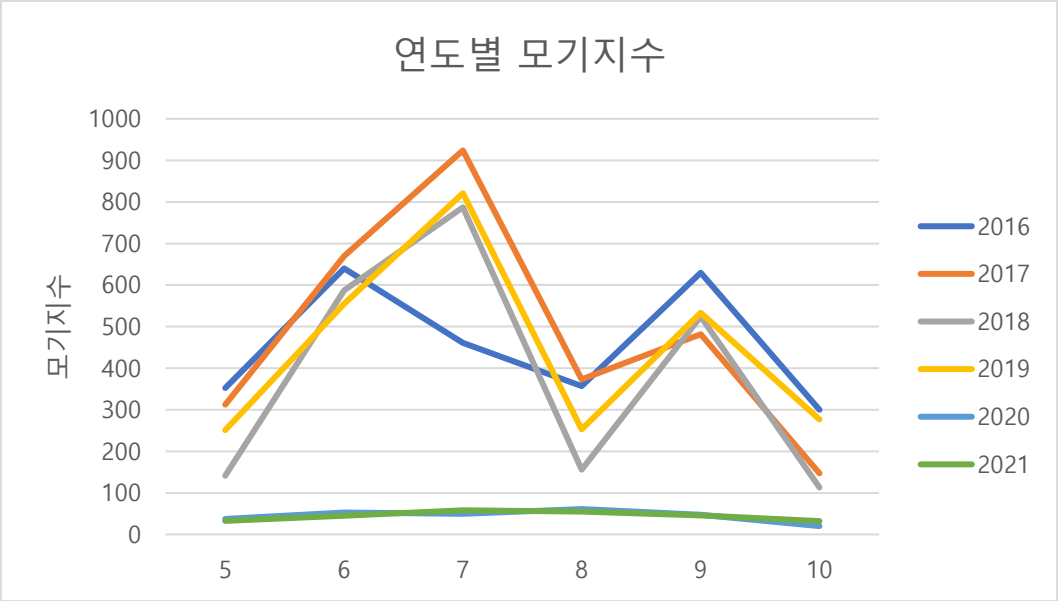




그래프를 보면 수변부 > 주거지 > 공원순으로 모기가 많이 나타나는 것으로 보인다. 이 중 직접적으로 물릴 수 있는 주거지를 선택하자. ts()함수를 이용하여 시계열 데이터로 변환한 후 100을 기준으로 시각해주면 오른쪽 그래프와 같다.

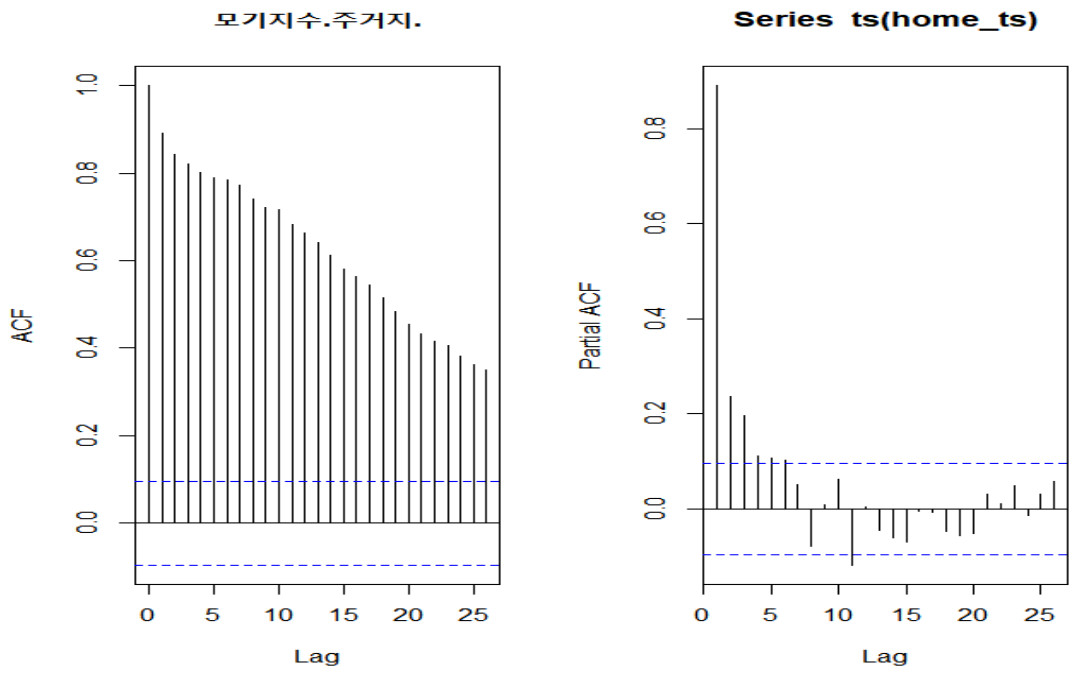
시계열 패턴을 설명하려면 추세, 계절성, 주기성을 봐야한다. 데이터가 장기적으로 증가하거나 감소할 때, 추세가 존재하며 이는 반드시 선형적일 필요는 없다. 계절성은 해마다 어떤 특정한 때나 일주일마다 특정 요일에 나타나는 것 같은 계절성 요인이 시계열에 영향을 줄 때 패턴이 나타난다. 계절성은 빈도의 형태로 나타나며 그 빈도는 일정해야 한다. 마지막으로 주기성은 고정된 빈도가 아닌 형태로 증가 혹은 감소할 때 나타난다. 이러한 지속기간은 최소 2 년 이상이여야 한다. 이 데이터는 모기가 활발한 늦은 봄, 여름, 초가을이 끼는 5 월 ~ 10 월을 포함하고 있어 계절성 요인을 띌 것이다. 기간에 해당하는 월들의 평균을 구해보면 다음 표와 같다.

	May	Jun	Jul	Aug	Sep	Nov
2016	352.60	640.71	461.44	357.40	629.57	299.32
2017	311.87	668.20	924.34	373.44	482.49	146.94
2018	141.54	587.31	787.63	156.01	524.76	113.89
2019	240.90	555.17	821.40	252.98	533.04	277.60
2020	36.75	52.60	50.45	61.07	47.55	20.60
2021	33.35	45.25	58.27	55.17	46.72	32.32



2020 년부터 모기의 수가 급감하기는 하지만 매년 5 월달부터 7 월까지 모기수가 늘어나는 것을 볼 수 있다. 그 수는 감소했다가 9 월 달에 다시 증가하고 하락하는 모양을 띈다. 즉, 계절성과 주기성을 갖고 있다고 판단할 수 있다. 2020 년부터 모기 수가 급감한 이유는 2020 년부터 모기 유충서식이 많거나 시민 생활과 밀접한 곳으로 구분하여 수변부, 공원, 주거지로 세분화하여 모기지수를 산출했고 폭염과 긴 장마로 인해 그 수가 감소하였다.

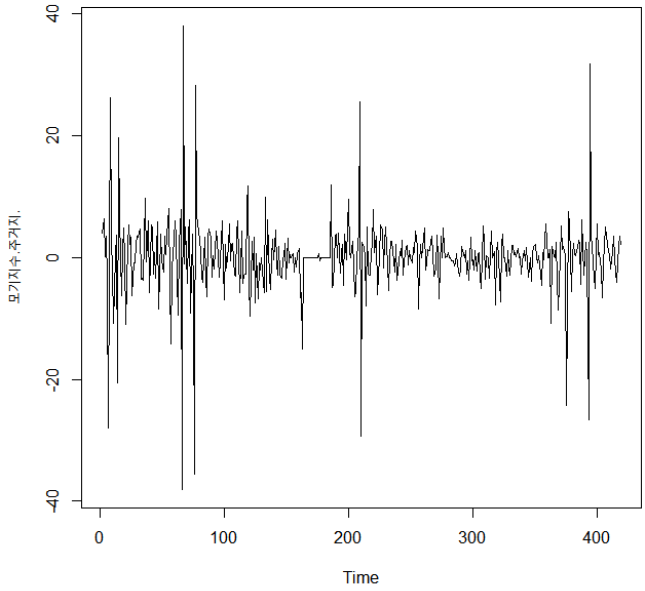
시계열의 정상성 여부를 판단하기 위해서 ACF 와 PACF 테스트를 사용해보자.

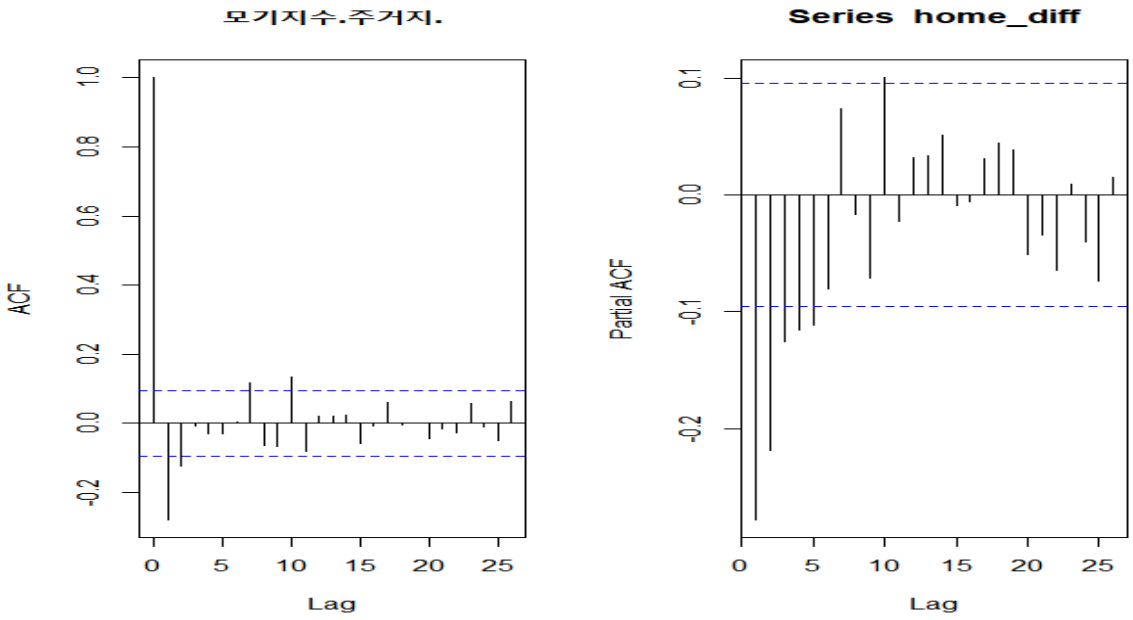


ACF의 값들이 파란선에서 크게 벗어난다. 따라서 비정상 시계열임을 유추할 수 있다. KPSS검정을 사용하는 ndiff()함수를 이용해 적당한 차분의 횟수를 알아보자.

```
> ndiffs(home_ts)
[1] 1
```

데이터가 정상성을 나타내려면 한 번의 차분이 필요하다. 처분 후 ACF, PACF 테스트를 다시 해주면 다음과 같다.





ACF 테스트가 많이 안정된 것이 느껴진다. ADF 테스트를 통해 정상성 여부도 해주자.

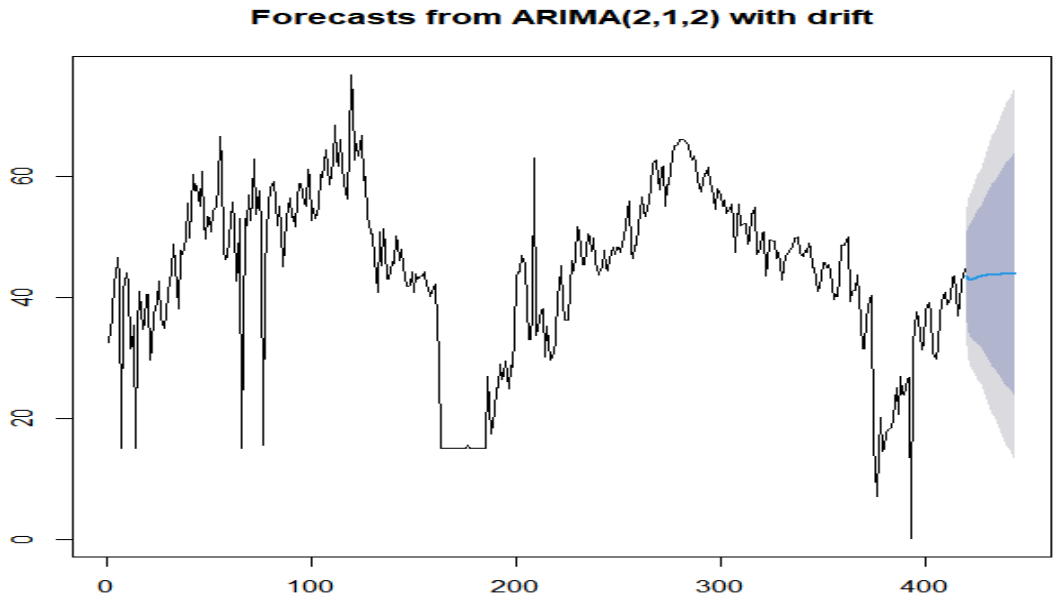
```
> adf.test(home_diff, alternative ="stationary", k=0)

        Augmented Dickey-Fuller Test

data:  home_diff
Dickey-Fuller = -27.077, Lag order = 0, p-value = 0.01
alternative hypothesis: stationary
```

p-value가 0.01<0.05으로 귀무가설을 기각한다. 즉, 모형이 정상성을 가지게 되었다. 시계열 분석을 위해 ARIMA 모형을 이용하여 예측해보자.

```
> summary(arima)
Series: home_ts
ARIMA(2,1,2) with drift

Coefficients:
         ar1      ar2      ma1     ma2   drift
      1.2588  -0.4068  -1.6586  0.7391  0.0132
s.e.  0.1170   0.0924   0.0988  0.0871  0.1547

sigma^2 = 34.29:  log likelihood = -1329.62
AIC=2671.23   AICc=2671.44   BIC=2695.45

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE     MASE          ACF1
Training set 0.02049155 5.813759 3.60358 -33.62126 42.06135 1.002327 -0.006800039
```

p : AR모형 차수, d : 차분 차수, q : MA모형 차수

ARIMA(2,1,2)모형이 최적임을 나타낸다. 마지막으로 예측을 위한 모형을 생성해보자. 식별된 모형과 파라미터를 이용하여 시계열 모형을 생성하는 것이다.



모형을 예측하면 위의 그래프와 같다. 크게 증가하거나 감소하는 모양을 띄고 있지 않다.

3. 결론

서울시 모기예보제 정보를 통해 연도별 모기의 증가와 감소, 어느 장소에 모기가 분포하여 사는지 알 수 있었다. 또 데이터가 계절성과 주기성을 갖는 것을 확인했다. 2019년도부터 급감한 모기수에 의해 말라리아병도 그 수가 급감할 것으로 예상된다. 올해에도 폭염이 이어진다면 그 수가 전과 같이 늘지는 않지만 가뭄으로 인해 전년보다는 늘어날 것이라고 생각된다.

참고자료

Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia.

코드

```
mosquito<-read.csv("C:/Users/zzxzz/Downloads/서울시 모기예보제 정보 (1).csv", header=TRUE)

mosquito<-na.omit(mosquito)

mosquito$모기지수.발생일<-as.Date(mosquito$모기지수.발생일)

library(dplyr)

library(forecast)

mos2020<-subset(mosquito, 모기지수.발생일>="2020-05-01" & 모기지수.발생일<"2020-10-31")

mos2021<-subset(mosquito, 모기지수.발생일>="2021-05-01" & 모기지수.발생일<"2021-10-31")

mos2022<-subset(mosquito, 모기지수.발생일>="2022-05-01" & 모기지수.발생일<"2022-06-01")

mos<-rbind(mos2020,mos2021,mos2022)

mos2<-rbind(mos2021,mos2022)

plot(mos$모기지수.수변부., pch=20, cex=0.5, type='l', xlab="년도", ylab="모기지수
",col='red',ylim=c(0,200))

par(new=TRUE)

plot(mos$모기지수.주거지., pch=20, cex=0.5, type='l', xlab="년도", ylab="모기지수
",col='blue',ylim=c(0,200))

par(new=TRUE)

plot(mos$모기지수.공원., pch=20, cex=0.5, type='l', xlab="년도", ylab="모기지수
",col='green',ylim=c(0,200))
```

```
home<-mos %>% select(모기지수.주거지.)

home_ts<-ts(home)

plot(ts(home_ts))

par(mfrow=c(1,2))

acf(ts(home_ts))

pacf(ts(home_ts))

ndiffs(home_ts)

home_diff<-diff(home_ts, differences = 1)

plot(home_diff)

par(mfrow=c(1,2))

acf(home_diff)

pacf(home_diff)

#정상성 검정

adf.test(home_diff, alternative ="stationary", k=0)

#모형 식별과 추정

arima<-auto.arima(home_ts, approximation = F)

summary(arima)

#예측

model<-arima(home_diff, order=c(2,1,2))

home_fcast<-forecast(arima,h=25)

plot(home_fcast)

#계절성 변동을 위한 데이터 조작

mos2016<-subset(mosquito, 모기지수.발생일>="2016-05-01" & 모기지수.발생일<"2016-10-31")%>%select(모기지수.발생일,모기지수.주거지.)

mos2017<-subset(mosquito, 모기지수.발생일>="2017-05-01" & 모기지수.발생일<"2017-10-31")%>%select(모기지수.발생일,모기지수.주거지.)

mos2018<-subset(mosquito, 모기지수.발생일>="2018-05-01" & 모기지수.발생일<"2018-10-31")%>%select(모기지수.발생일,모기지수.주거지.)

mos2019<-subset(mosquito, 모기지수.발생일>="2019-05-01" & 모기지수.발생일<"2019-10-31")%>%select(모기지수.발생일,모기지수.주거지.)

mos2020ave<-mos2020%>%select(모기지수.발생일,모기지수.주거지.)

y<-cbind(mos2020ave, month=substr(mos2020ave$모기지수.발생일,6,7))

a2020<-tapply(y$모기지수.주거지.,y$month,mean)

mos2021ave<-mos2021%>%select(모기지수.발생일,모기지수.주거지.)

y2<-cbind(mos2021ave, month=substr(mos2021ave$모기지수.발생일,6,7))

a2021<-tapply(y2$모기지수.주거지.,y2$month,mean)

mos2022ave<-mos2022%>%select(모기지수.발생일,모기지수.주거지.)

y3<-cbind(mos2022ave, month=substr(mos2022ave$모기지수.발생일,6,7))

a2022<-tapply(y3$모기지수.주거
```

```
지.,y3$month,mean)

y4<-cbind(mos2016,
month=substr(mos2016$모기지수.발생일,6,7))

a2016<-tapply(y4$모기지수.주거
지.,y4$month,mean)

y5<-cbind(mos2017,
month=substr(mos2017$모기지수.발생일,6,7))

a2017<-tapply(y5$모기지수.주거
지.,y5$month,mean)

y6<-cbind(mos2018,
month=substr(mos2018$모기지수.발생일,6,7))

a2018<-tapply(y6$모기지수.주거
지.,y6$month,mean)

y7<-cbind(mos2019,
month=substr(mos2019$모기지수.발생일,6,7))

a2019<-tapply(y7$모기지수.주거
지.,y7$month,mean)

season<-
rbind(a2016,a2017,a2018,a2019,a2020,a2021,a2
022)

seasonal<-round(season,2)

season<-as.data.frame(seasonal)
```