실험실 프로젝트 보고서

# ARIMA분석을 통한 서울로7017 방문자 수 예측

Estimation of Seoullo7017 visitors through ARIMA analysis

지도 교수 나스디리노프 아지즈

충북대학교 소프트웨어학과

구 경 민

정 예 원

김 희 주

**(1) 서론** (연구의 내용이 무엇이고, 왜 이 연구가 필요한가?)

최근 기술의 발전으로 다양한 종류의 데이터를 만들어 내고, 수집한 데이터를 분석하며, 다양한 시각으로 정확하게 미래의 데이터를 예측하여 효율적으로 서비스를 제공할 수 있는 것이 일반화 되고 있다. 이것을 데이터 마이닝(Data Mining)기법이라고 하는데 이는 대규모 데이터를 체계적으로 통계적 규칙이나 패턴을 찾아내는 것을 의미한다. 예컨대, 여러 데이터간의 유용한 상관관계를 찾고 필요한 정보를 추출하거나 분석을 하는 과정에서 기대이상의 정보를 얻을 수 있는 장점을 가지고 있다. 때문에 많은 정보들을 통해 보다 쉽게 의사결정을 할 수 있어 하려는 것의 이익을 극대화 시킬 수 있다.

데이터 마이닝(Data Mining)의 기법에는 다양한 종류가 있다. 그러한 기법 중 본 연구에서는 '연속성(Sequencing)'과 예측(Forecasting)'을 통해 데이터의 미래 행동을 분석해 보고자 한다. 본 연구에 사용되는 데이터의 경우 특정 구역을 입장하거나 퇴장한 시간의 데이터로 이루어져 있다. 이 데이터를 ARIMA(Auto-regressive Integrated Moving Average)모형을 통해 분석한다. 분석을 진행하게 되면 특정 구역에 사람들이 자주 출입하는 시간을 알 수 있게 되며, 그러한 시간에 구역을 통과하는 사람을 위해 다양한 콘텐츠를 더 준비할 수 있다. 또는 특정 계절에 그 구역을 통과하는 사람이 많다면 계절과 관련한 서비스를 제공하여 사람들이 더 편하게 이용할 수 있어 기업에게 이익과 직결될 수 있다.

**(2) 관련 연구 분석** (연구 주제와 관련된 기존 기술 동향은 무엇인가?)

다양한 데이터 마이닝(Data Mining) 기법들이 존재하고 사용되기 때문에 다양한 분야에서 각각의 방식으로 데이터에 대한 분석들이 이루어지고 있고 지속적으로 발전하고 있다. 특히나 주가 전망 예측이나 수요 예측 등 다양한 경제, 산업 분야에서 그 분석들이 이루어지고 있다.

ㅇ국제 유가 예측 : 에너지를 수입하는 국가인 우리나라에서는 국제 유가의 변동에 따라 국가 경제에 미치는 영향이 크다. 때문에 국제 유가를 예측하여 국가는 시장 경제를 안정시키기 위해 보다 먼저 대책을 세우고 안정화시킬 수 있으며, 기업들도 국제 유가가 기업의 이익에 손해가 되지 않도록 보다 빠르게 대처할 수 있다.

ㅇ저가항공 수요 예측 : 최근 저가항공을 이용하여 해외여행을 자주 다니는 추세이다. 때문에 항공사가 계절에 맞게 고객들이 자주 수요 하는 상품을 예측하고 급변하는 환경에 대처하기 위해 단기적인 수요예측을 한다. 이는 증가하는 수요에 발맞춰 저가항공을 운영하기 힘들다고 판단된다. ARIMA를 통해 분석을 진행하여 항공노선, 직원 수를 조절할 수 있으며 이는 기업의 이익에 직결하여 효율적인 운항을 할 수 있다.

ㅇ국내 지역별 전력사용량 예측 : 전력의 경우 발전과 동시에 소비가 이루어지는 자원으로 안정적인 전력공급뿐만 아니라 낮은 원가에 전력을 공급하기 위해 정확한 예측을 필요로 한다. 특히, 계절에 영향을 받으므로 ARIMA모형이 적합하다. 예측을 통해 중장기적 수급정책 및 수급안정화 방안에 도움이 되며 지역별 산업구조 변화, 인구변화와 같은 지역경제에 대한 예측 지표로서 사용이 가능하다.

**(3) 세부 연구 내용**

● 연구 주제 범위

본 연구에서 사용한 데이터는 구역별 사람의 통과 시간으로 싱가폴 based-DFRC에서 제공하는 서울로 방문 데이터를 사용하였다. 본 연구는 6월달의 출입현황 데이터로 한정하고 구역은 제 0구역과 제 1구역으로 한정하였다. 본 데이터를 통해 사람이 자주 출입하는 시간을 예측하고 이를 바탕으로 효율적인 운영방식을 생각해 볼 수 있다.

● 세부 연구 내용별 기술 설명

1) 자기회귀(Auto Regressive)모형

어떠한 랜덤 값에 대해서 이전의 값이 이후의 값에 영향을 미치고 있는 상황으로 시점 t에서 얻게 될 $Z_t$의 평균값은 t-1에서 얻었던 $Z_{t-1}$에 $\theta_0$(자기상관계수)를 더하고 $a_t$(백색잡음)을 더한 값으로 $\phi$(자기회귀계수)이다. 자기회귀계수는 ACF를 통해서 구할 수 있다. $AR(p): Z_t = \theta_0 + \phi_1 Z_{t-1} + \ldots + \phi_p Z_{t-p} + a_t$

2) 이동평균(Moving Average)모형

시간이 지날수록 어떠한 랜덤 값의 평균값이 지속적으로 증가하거나 감소하는 상황으로 AR과 다른 점은 이전에 발생한 에러가 중요하지 $Z_{t-1}$가 무엇인지는 중요하지 않다. 여기서 $\theta$는 이동평균계수이고 이는 PACF를 통해서 구할 수 있다.$MA(q): Z_t = \mu + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q}$

3) ARIMA(Auto-regressive Integrated Moving Average)모형

시계열 데이터 기반 분석기법으로 과거 지식이나 경험을 바탕으로 한 행동에 따라 경제가 움직이고 있음을 기초로 한다. ARIMA 모형은 과거의 관측값과 오차를 사용해서 현재의 시계열 값을 설명하는 ARMA 모델을 일반화 한 것으로, 분기/반기/연간 단위로 지표를 예측할 수 있다. 안정적 시계열의 경우 시간의 추이와 관계없이 평균 및 분산이 불변하거나 시점 간의 공분산이 기준 시점과 무관한 형태의 시계열이다. ARMA에 자기 자신의 추세를 고려하여 나타낸 것으로 밑의 식으로 나타낸다. $W_t + \phi_1 W_{t-1} + \ldots + \phi_p W_{t-p} = a_t + \theta_1 a_{t-1} + .. + \theta_q a_{t-q}$ ARIMA(q, d, q)로 분석을 진행 할 수 있다. p의 경우 자기회귀를 한 횟수이고, d의 경우 비 계절적 차이로 차분을 한 횟수이고, q의 경우 이동평균을 한 횟수이다.

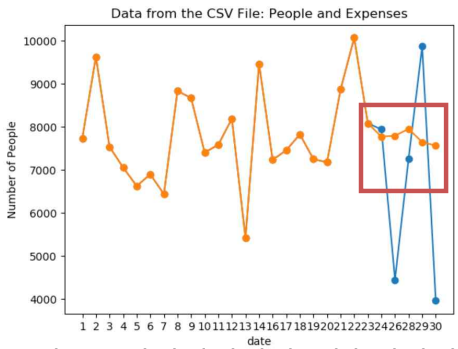
(4) ACF(Autocorrelation function)

정상 확률 과정을 표현하는 대표적인 특성으로 자기상관계수 함수라고 하

며 $p_k$로 표현한다. $p_k = \frac{\gamma_k}{\gamma_0} = \frac{\text{시차}k\text{의 공분산}}{\text{분산}}$

● 문제 개선 전략 및 개선 방법

본 데이터를 한달치로 30개의 구간을 나누고 x축으로 나타내고, 날짜별 방문한 인구수를 y축으로 나타내어 분석을 진행하였다. 그 중 80%는 실제 데이터의 값을 사용하고 20%는 테스트 데이터로 오차율에 대한 검증을 진행한 결과 평균 오차율이 40.12106%가 나왔다.




[그래프1 실제데이터와 예측데이터]

| 날짜 | 예측값 | 실제값 | 오차율 |
|---|---|---|---|
| 24 | 7776 | 7943 | 2.1024% |
| 25(26) | 7795 | 4447 | 75.2867% |
| 27(28) | 7957 | 7266 | 9.51% |
| 29 | 7646 | 9875 | 22.5721% |
| 30 | 7567 | 3959 | 91.1341% |

전체 오차율

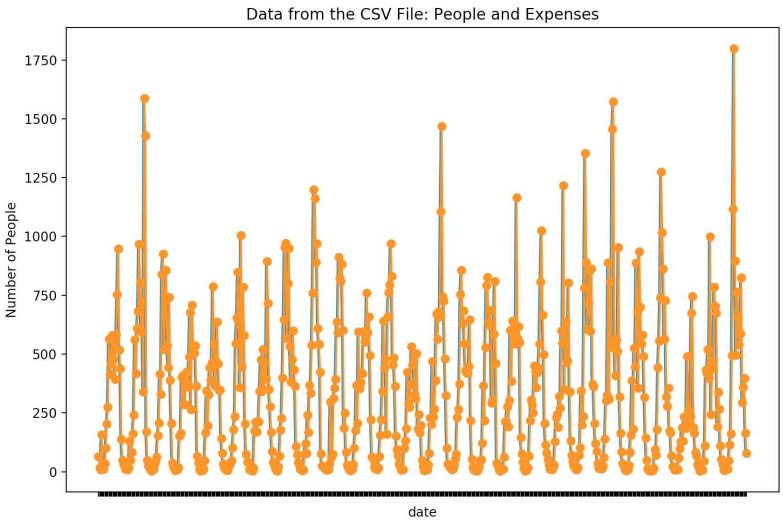
40.12106%

[표1 오차율 계산]

데이터의 수가 패턴을 분석하기에 적었고 특정한 구간에서 방문자 수가 대폭으로 감소하거나 증가하여 평균 오차율이 높아졌다. 이런 오차율을 개선하기 위해 데이터를 일별 시간으로 720구간으로 나누어 패턴이 나올 수 있는 충분한 양의 데이터를 만들어 진행하였다.

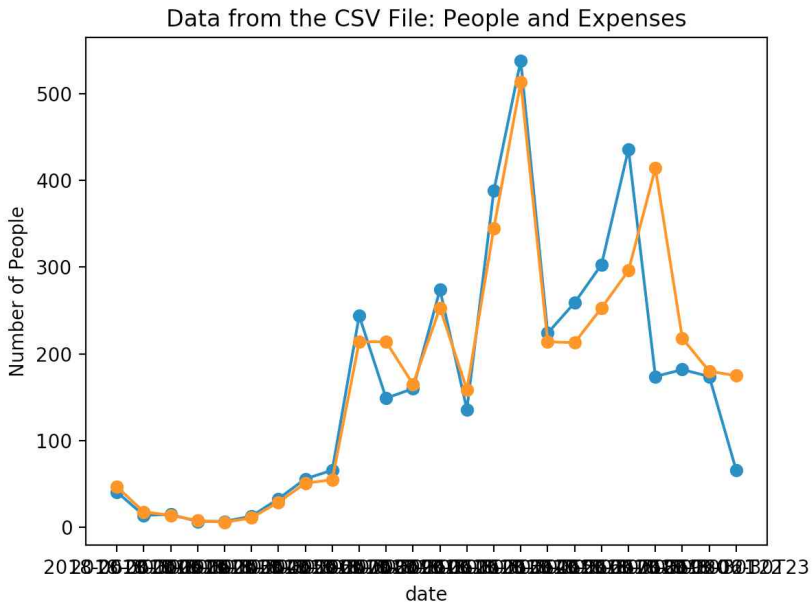
**(4) 실험 및 적용** (연구 주제 관련 실험 내용 또는 사례 적용 내용)

본 데이터를 일별시간인 24시간의 데이터로 잘게 나눈 것을 바탕으로 총 720개의 데이터를 만들어 보다 정확한 ARIMA 분석을 진행 하였다. 본 데이터의 80%인 647개의 데이터를 그래프로 표현하여 안정적인 시계열인지를 확인하였다. 그래프가 안정적인 형태를 띄고 있으므로 차분을 하지 않고 ARIMA 분석을 하였다.




[그래프2 실제 데이터]

ACF가 완만하게 감소하는 형태를 하고 PACF가 급격하게 감소하는 형태를 하고 있다. 이 경우 AR형태에서 차분을 진행하여 ARIMA 분석을 한다. 때문에 p=0, q=1이 되고 차분 d의 경우 본 데이터에서는 하지 않고 비슷한 결과 값을 예측할 수 있어 ARIMA(0, 0, 1)로 분석을 했다.




[그래프3 예측 결과]

| 날짜T시간대 | 실제 수 | 예측한 수 |
|---|---|---|
| 2018-06-30T00 | 41 | 47 |
| 2018-06-30T01 | 14 | 18 |
| 2018-06-30T02 | 15 | 14 |
| 2018-06-30T03 | 7 | 8 |
| 2018-06-30T04 | 7 | 6 |
| 2018-06-30T05 | 13 | 11 |
| 2018-06-30T06 | 33 | 29 |
| 2018-06-30T07 | 56 | 51 |
| 2018-06-30T08 | 66 | 55 |
| 2018-06-30T09 | 244 | 214 |
| 2018-06-30T10 | 149 | 214 |
| 2018-06-30T11 | 160 | 165 |
| 2018-06-30T12 | 274 | 253 |
| 2018-06-30T13 | 136 | 159 |
| 2018-06-30T14 | 388 | 345 |
| 2018-06-30T15 | 538 | 513 |
| 2018-06-30T16 | 224 | 214 |
| 2018-06-30T17 | 259 | 213 |
| 2018-06-30T18 | 303 | 253 |
| 2018-06-30T19 | 436 | 296 |
| 2018-06-30T20 | 174 | 414 |
| 2018-06-30T21 | 182 | 218 |
| 2018-06-30T22 | 174 | 180 |
| 2018-06-30T23 | 66 | 175 |

[표2 실제 데이터와 예측데이터 값]

ARIMA를 통해 위와 같은 예측 결과를 도출했다. 총 24개의 데이터 구간으로 다음날의 방문할 사람의 수를 예측하였고, 그 결과 조금의 오차를 보이지만 대부분의 구간에서 일치하는 것을 보였다. 이 결과를 통해 실제 데이터와의 오차율을 계산하였다. 아래의 표를 보면 오차율이 26.17%인 것을 확인 할 수 있다. 처음에 했던 실험과는 다르게 오차율이 확실하게 줄었고 보다 정확한 데이터를 얻은 것을 확인할 수 있었다.

| 날짜T시간대 | 오차율 | 평균 오차율 |
|---|---|---|
| 2018-06-30T00 | 14.63 | 26.17% |
| 2018-06-30T01 | 28.57 | |
| 2018-06-30T02 | 6.67 | |
| 2018-06-30T03 | 14.29 | |
| 2018-06-30T04 | 14.29 | |
| 2018-06-30T05 | 15.38 | |
| 2018-06-30T06 | 12.12 | |
| 2018-06-30T07 | 8.93 | |
| 2018-06-30T08 | 16.67 | |
| 2018-06-30T09 | 12.30 | |
| 2018-06-30T10 | 43.62 | |
| 2018-06-30T11 | 3.13 | |
| 2018-06-30T12 | 7.66 | |
| 2018-06-30T13 | 16.91 | |
| 2018-06-30T14 | 11.08 | |
| 2018-06-30T15 | 4.65 | |
| 2018-06-30T16 | 4.46 | |
| 2018-06-30T17 | 17.76 | |
| 2018-06-30T18 | 16.50 | |
| 2018-06-30T19 | 32.11 | |
| 2018-06-30T20 | 137.93 | |
| 2018-06-30T21 | 19.78 | |
| 2018-06-30T22 | 3.45 | |
| 2018-06-30T23 | 165.15 | |

[표3 데이터별 오차율과 평균 오차율]

### (5) 결론

ARIMA의 경우 다양한 분야에서 활발하게 사용되고 있고, 데이터의 예측에 있어 매우 비슷한 방향으로 결과 값을 나타내어 이와 같은 방문객 예측 실험에 적합하다는 것을 보았다. 결과적으로 위의 ARIMA 분석을 통해 나온 예측 값이 작은 오차율을 보이고 매우 비슷한 그래프를 나타내는 것을 도출 했다. 본 연구를 통해 새벽시간에는 방문객이 적고 오후시간대에 평균적으로 방문객이 많았으며 오후3시쯤이 최고치를 찍은 것을 알 수 있다. 또한 오전9시, 오후7시와 같은 아침 출근시간 또는 퇴근시간에 급격히 증가하는 것을 확인 할 수 있었고, 이것은 예측에서도 비슷한 추세를 보였다. 본 연구에서는 많은 양의 데이터가 없어 비교적 오차율이 큰 것처럼 보이지만 보다 많은 양의 데이터를 수집하여 월별, 년별, 요일별, 계절별 등 다양한 데이터 예측을 도출함으로써, 향후 증가하거나 감소하는 방문객 수에 대한 지침이 될 수 있을 것으로 기대된다. 이것을 통해 사전에 방문객 수를 알고 이것에 대비하여 본 시설을 관리하는 사람이나 주변 상권과 같은 곳에서 손실이 적게 발생할 수 있다.

**(6) 참고 문헌**

[1] 송경재, 양희민, "시계열 분석에 의한 국제 유가 예측", 통계청 [통계연구] 제10권 제 1호,2005
[2] 위키피디아 "데이터 마이닝"
[3] 김영주, "계절 ARIMA 모형을 활용한 저가항공 수요 예측", 관광연구논총 제 26권 제1호,2014
[4] 안병훈, "계절 ARIMA 모형을 이용한 국내 지역별 전력사용량 중장기수요예측",고려대학교, 2015 http://jkais99.org/journal/Vol16No12/p54/9h/9h.pdf
[5] "ARIMA 모형을 통한 미래 추세 예측",
http://www.dodomira.com/2016/04/21/arima_in_r/
[6] 정재원, "ARIMA 모형과 에너지 예측", 산업수학혁신센터,2017
[7] 김성민, "계절형 ARIMA 모형을 이용한 인천국제공항 LCC 이용여객 수요예측 연구", 인하대학교,2016