한국ITS학회논문지 J. Korea Inst. Intell. Transp. Syst.

SARIMA 알고리즘을 이용한 교통량 보정 및 예측

A Study on the Traffic Volume Correction and Prediction Using SARIMA Algorithm

한 대 철* · 이 동 우** · 정 도 영***

* 주저자 : 한국건설기술연구원 도로교통연구본부 수석연구원
** 공저자 : 인천대학교 도시과학대학 도시행정학과 조교수
*** 교신저자 : 한국건설기술연구원 도로교통연구본부 전임연구원

Dae-cheol Han* · Lee, Dong Woo** · Do-young Jung***

* Dept. of Highway & Transportation Research, Korea Institute of Civil Eng, and Building Technology
** Dept. of Urban Policy and Administration, Incheon National University
*** Dept. of Highway & Transportation Research, Korea Institute of Civil Eng, and Building Technology

† Corresponding author : Do-young Jung, jdy@kict.re.kr

Vol.20 No.6(2021)
December, 2021
pp.1~13

pISSN 1738-0774
eISSN 2384-1729
https://doi.org/10.12815/kits.2021.20.6.1

Received 17 August 2021
Revised 2 September 2021
Accepted 2 November 2021

요 약

본 연구에서는 도로교통분야의 계획, 설계, 유지관리, 연구 등 다양한 목적으로 활용되고 있는 교통량 데이터의 정확도 확보를 위해 시계열 분석 기법을 적용하여 교통량 데이터의 보정 및 예측을 수행하였다. 기존 알고리즘의 경우 주기성 및 계절성이 강하거나 불규칙한 데이터에 한계를 보이고 있어 교통량 데이터와 같은 자료에 적용하기에는 한계가 있다. 이러한 한계점을 극복하고 보완하기 위해 ARIMA 모형에 자기상관 모형인 SAR(Seasonal Auto Regressive)과 계절 이동평균 모형인 SMA(Seasonal Moving Average)가 결합된 분석 기법인 SARIMA 모형을 적용하였다. 분석결과 최적 파라미터 조합인 SARIMA(4,1,3)(4,0,3) 12 모형을 활용한 교통량 예측 결과 평균 85% 정도의 우수한 성능을 보였다. 본 연구를 통해서 교통량 데이터의 결측 발생 시 교통량 보정 및 예측의 정확도를 높일 수 있으며, 교통량 데이터 외에도 계절성에 영향을 받는 시계열 데이터에 적용이 가능하다.

핵심어 : 교통량, ARIMA, SARIMA, 교통량 보정, 시계열

ABSTRACT

In this study, a time series analysis technique was applied to calibrate and predict traffic data for various purposes, such as planning, design, maintenance, and research. Existing algorithms have limitations in application to data such as traffic data because they show strong periodicity and seasonality or irregular data. To overcome and supplement these limitations, we applied the SARIMA model, an analytical technique that combines the autocorrelation model, the Seasonal Auto Regressive(SAR), and the seasonal Moving Average(SMA). According to the analysis, traffic volume prediction using the SARIMA(4,1,3)(4,0,3) 12 model, which is the optimal parameter combination, showed excellent performance of 85% on average. In addition to traffic data, this study is considered to be of great value in that it can contribute significantly to traffic correction and forecast improvement in the event of missing traffic data, and is also applicable to a variety of time series data recently collected.

Key words : Traffic volume, ARIMA, SARIMA, Traffic calibration, Time series

Ⅰ. 서 론

1. 연구의 배경 및 목적

우리나라는 1994년부터 교통체증을 완화하기 위한 목적으로 고속국도 등 주요 간선도로를 중심으로 지능형교통시스템(Intelligent Transport Systems, 이하 ITS)을 도입하여 지속적으로 확충 및 운영하고 있다. 교통량은 「도로법」 제102조(도로에 관한 조사)에 의거하여 도로관리청은 도로와 관련된 계획의 효율적인 수립과 도로의 보수, 도로의 유지관리 등을 위하여 필요하면 구간별 교통량, 도로의 구조, 그 밖에 도로에 관한 사항을 조사할 수 있다고 정의되어 있다. 또한, 교통량 조사는 「도로법」 제23조(도로관리청)에 의거하여 고속국도와 일반국도는 국토교통부, 국가지원지방도와 지방도는 도지사 및 특별자치도지사가 교통량을 조사하는 주체로 되어있다. 이러한 ITS에서 수집된 자료 중 교통, 환경, 도시, 사회 등 여러 분야에서 가장 활용가치가 높은 자료는 교통량이라고 할 수 있다. 「도로법」 제102조에 의거하여 매년 교통량을 조사하여 국가승인 통계인 도로교통량 통계연보를 발간하고 있으며, 「국가교통체계효율화법」 제88조에 따라 교통량과 관련 교통정보들을 수집 및 가공하여 제공하고 있다. 특히, 교통량은 도시 및 교통계획, 도로 설계, 도로 유지관리 및 운영 등 교통관련 분석의 기초 자료로 폭넓게 활용되고 있다.

교통량은 도로의 특정 지점 또한 구간을 단위 시간 동안 통과하는 차량 수의 총합을 의미하며, 도로를 이용하는 차량들의 단위 시간 당 통과 대수를 차종 및 방향별로 구분하여 수집하는 것으로 교통량 조사 혹은 교통량 데이터 수집이라고 한다. 현재 도로교통량 통계연보의 경우 교통량 조사는 수시조사와 상시조사로 구분하여 수행되고 있다. 수시조사는 기본 교통량 자료가 필요하다고 판단되는 도로 구간 전체에 대하여 광범위하게 실시하는 조사로 전체적인 도로 이용 상황을 파악하기 위해 수행하고 있으며, 조사 주체와 방법에 따라 일반국도를 대상으로 하는 수시조사와 고속국도와 지방도를 대상으로 하는 10월 수시조사로 분류하여 년 1회 교통량을 조사하고 있다. 상시조사는 특정 지점에 자동차종분류 조사장비(Automatic Vehicle Classification, 이하 AVC)를 설치하여 1년 365일 24시간 연속으로 통과 차량의 차종별, 방향별, 시간대별로 측정하고 기록하고 있다. AVC의 경우 매설식 장비의 특성상 중차량에 의한 도로노면 파손 등의 원인으로 고장이 잦고 노후화가 빨리 진행되는 경우가 많다. 이러한 경우에 관측된 교통량이 결측치 및 이상치를 가지게 되며, 통신이상, 장비의 제어기 문제 등 기계적인 결함의 정도에 따라 유지보수에 소요되는 시간이 상이하여 결측 및 이상 기간이 길어질 수 있다. 따라서 상시조사를 통해 수집된 교통량은 기술적인 한계, 기계적인 결함, 검지 오류, 외부 환경 요인, 기타 오류 등으로 인한 불확실성(uncertainty)을 피할 수 없다.

이렇게 불확실성이 내포된 교통량 자료가 적절한 가공 없이 다양한 분석의 기초자료로 활용된다면 신뢰성 있는 결과 도출이 불가능하다. 따라서 도로교통량을 활용하기 전에 원시자료의 정확성을 측정하고 적절한 방법론을 활용한 보정이 매우 중요하다. 실제 교통량 조사를 통한 방법도 가능하지만 많은 시간과 비용이 소요되기 때문에 현실적으로 불가능한 경우가 많다. 현재 상시조사 교통량의 보정은 크게 단기 결측과 장기 결측 보정으로 구분할 수 있다. 단기 결측 보정 시에는 차종별, 방향별 과거 자료를 바탕으로 단순한 통계기법(예 : 선형 보간)을 활용하여 보정 및 배분하는 방식으로 교통량에 대한 보정 작업을 수행하고 있다. 장기 결측에 대해서는 Factor법이 적용되고 있다고 제시되어 있지만 단기 결측 보정 방식과 크게 다르지 않는 것으로 확인되고 있다.

이러한 교통 부문 자료의 불확실성 문제는 차세대 교통시스템에서도 피할 수 없는 이슈가 될 것으로 보인다. 특히, 센서 기반으로 수집되는 시계열 형식의 교통량 자료는 실시간으로 방대한 규모의 자료를 축적하지만, 이러한 첨단 장비를 통해 수집된 자료 내에도 이상 검지 및 결측값이 상시 존재하기 때문에 수집된 자료

의 불확실성을 판단하고 적절한 보정 작업을 수행하는 것은 매우 중요하다. 방대한 교통량 자료, 이른바 교통 빅데이터가 곧 유용한 정보를 보장하는 것은 아니기 때문에 시계열 자료가 가진 불확실성을 진단하고 정확한 예측 및 보정을 제공하는 것은 스마트 시스템을 구축하는 기본 요소라 볼 수 있다.

본 연구의 목적은 AVC 장비에서 1시간 단위로 기록되는 측정 교통량을 보정하기 위한 교통량 보정 및 알고리즘을 개발하는데 있다. 전통적인 통계 분석 기법들이 가지고 있는 한계점을 파악하고, 결측이 다수 존재하고 주기가 짧으며 계절성이 복잡한 교통량 자료에 활용 가능한 알고리즘을 적용해보고자 한다. 교통량 자료는 관/산/학/연의 다양한 수요처에서 도로교통분야의 계획, 설계, 유지관리, 연구 등을 비롯하여 사회 각 분야에서 다양한 목적으로 활용되고 있다. 따라서 분석가들이나 도시 정책 분석에서의 활용도를 높이기 위하여 본 연구를 통해 개발 및 적용된 방법론을 실제 보정교통량과 비교 및 검증하여 모형 구축에서 흔히 발생할 수 있는 out-of-sample 이슈를 최소화하고자 하였다.

Ⅱ. 선행 연구고찰

IT 기술의 발달로 첨단의 교통량 조사 장비 도입을 통해 기존의 교통량조사 장비가 가지고 있는 단점을 보완하고 결측 및 이상 검지 비율을 줄일 수는 있겠지만, 앞서 언급한 자료의 불확실성을 해결하는 문제는 여전히 난제로 남아 있다. 이를 해결하고자 국내외에서 다양한 연구가 수행되어져 왔으며, 다음과 같이 크게 두 가지 형태인 (1) 과거 자료의 일관성에 기반한 교통량 보정 (2) 통계학습 알고리즘에 기반한 교통량 보정 연구들로 구분할 수 있다.

1. 과거 자료의 일관성에 기반한 교통량 보정 연구

2020년도 기준 일반국도 상시 교통량 조사장비(AVC)는 총 549대로 확인되고 있다. 예를 들어 AVC 장비의 경우 일정 빈도(frequency)에 따라 시계열 형식의 자료로 1일 24시간 연속 통과 차량 데이터를 생성하여 센터 서버로 전송하고 있다. 센터시스템에서는 현장수집 장비로부터 전송된 교통량 자료를 가공 및 보정 과정을 거쳐 최종적으로 교통량 정보를 구축하고 있다. 이렇게 수집/가공/보정된 교통량 자료는 다양한 교통분야의 기초 분석자료로 활용될 수 있기 때문에 교통량 자료의 품질을 확보하고 향상시키는 것이 매우 중요하다. 하지만 이렇게 수집된 방대한 시계열 자료는 불확실성(예 : 장기 및 단기 결측, 이상 검지 등)으로 인해 자료의 신뢰성에 대한 의문이 지속적으로 제기되고 있는 실정이다.

국내 사례를 살펴보면 서울시의 경우 2004년 이후 단순히 인력에 의한 조사 방식과 더불어 다양한 첨단장비와 조사 방식을 개선하여 교통정보 수집체계를 크게 개선하였다. 하지만 교통량 결측률은 여전히 30%에 달하는 것으로 확인되고 있다. 고속국도의 경우 2018년 기준 교통량이 수집되지 않는 지점이 약 15%를 초과하는 것으로 조사된 바 있다.(Lee, 2019) 국내 사례 외에도 국외에서도 이러한 불확실성 문제가 확인되고 있다. 미국의 텍사스주와 조지아주의 교통량 수집장비의 경우 결측률이 각각 약 15~90%, 5~14%로 조사되었으며, 캐나다 알버타주의 교통량 수집장비는 과거 7년간의 교통량의 신뢰성을 분석한 결과 평균 약 10 ~40% 정도의 결측률을 보이는 것으로 나타났다.(Turner et al., 2001; Satish et al., 2003) 이를 해결하고자 다양한 연구들이 수행되어왔지만, 실제 교통량 측정 시스템에 적용 중인 보정 방법은 과거 자료와의 일관성(consistency)에 의존한 방법을 고수하고 있는 경우가 대부분이다<Table 1 참고>. 어떠한 통계적 기법이나 모형도 불확실성의 문제에서 자유로울 순 없으나, 과거 자료에 기반한 방법론의 경우 보정 시 많은 시간이 소

요되며 교통량 패턴의 변화가 있는 경우 한계가 존재한다. 또한, 과거의 보정 자료의 신뢰성이 확인되지 않은 상황에서 분석가의 주관과 경험에 따라 과거에 기반한 임의 보정을 할 경우 전체 시계열 자료의 특성이 왜곡되거나 변질될 가능성이 매우 크다. 따라서 최근에는 통계학습 기반 알고리즘을 활용한 예측 기법이 주로 활용되고 있다. 관련 연구들은 다음과 같다.

<Table 1> Correction method based on consistency with domestic and foreign historical data

Item		Methodology
Lee and Shin(2013)		Utilize average traffic at the same time before a week
Kim et al.(2010)		Calculate the weight based on historical correction data and real-time data and apply it to correct the traffic volume
Zhong and Sharma(2009)		Utilize traffic at the same time before a week
Overseas Research Institutions	Alabama, US, Indiana, US South Dakota, US	Utilize traffic volume at the same time in the previous year and month
	Vermont, US, France, EU, Alberta	Presented by replacing the average monthly traffic volume data
	London, UK, Oklahoma, US	Utilize traffic at the same time before a week
	Delaware, US	The average value of the traffic volume at the same time in the month before and after the missing traffic volume is used
	Alberta, US	Utilize monthly average traffic data
	Montanan, US	Utilize of similar area traffic volume based on spatial consistency
	Netherlands, EU	Utilize nearest point data on the same axis

2. 통계학습법에 기반한 교통량 예측 연구

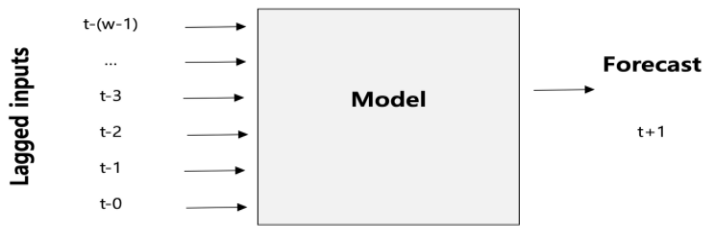
일반적으로 교통량을 예측하기 위해 통계기법에 기반한 데이터 중심의 접근법(data driven approach)이 적용되고 있으며, 모수(parametric) 추정 기법과 비모수(nonparametric) 추정 기법으로 나눌 수 있다. 먼저, 모수 추정 기법에는 전통적인 선형 및 비선형 회귀분석(linear and nonlinear regression)(Steffen and Bartz-Beielstein, 2017; Korea Institute for Health and Social Affairs, 2018; Korea Institute of Civil Engineering and Building Technology, 2008; Zhou and Meng, 2019), 과거 평균기법(historical average algorithms) 혹은 이동평균기법(moving average), 스무딩 기법(smoothing techniques)(Li et al., 2015), 자기 회귀 선형 프로세스(autoregressive linear processes)(Steffen and Bartz-Beielstein, 2017; Zhou Meng, 2019; Hamed et al., 1995; Lee and Fambro, 1999) 등이 존재한다. 데이터의 분포 형태나 특성(예 : 주기, 변동성)에 따라 모형 알고리즘 선택이 달라지지만, 모수 추정법 기반 선형 모형 중 대표적으로 많이 사용되는 모델은 자기회귀 이동평균(autoregressive moving average, 이하 ARMA)과 자기회귀 통합 이동평균(autoregressive integrated moving average, 이하 ARIMA)이다.

ARIMA 모형은 차분 시계열에 ARMA 모형을 적용한 모형이다. ARIMA 모형은 시계열 자체로는 정상성(stationary)이나 평균 회귀 특성이 없어도 이를 차분한(indifference) 시계열은 평균회귀 특성이 있다. 즉, ARIMA의 경우 ARMA 모형과 달리 비정상성(nonstationary)을 가진 데이터 형태에도 적용이 가능하며, 이를 위해 관측점들의 차분(difference)을 사용한다는 점이 특징이다. 즉, 시계열 데이터의 경향 및 계절성을 제거하여 ARMA 모델을 적용 시킨 후 이를 예측 시 재반영하는 것이 ARIMA의 추정 과정이다. ARIMA의 경우 많은 연구들을 통해 모형의 성능이 검증되었다. 교통량 시계열의 경우는 일반 시계열 데이터보다 계절성(seasonality 혹은 seasonality trend)이 강하게 나타나므로, 최근에는 이러한 주기 패턴을 추가적으로 반영한 Seasonal

ARIMA(SARIMA) 모형을 활용하는 것이 효과적이라는 연구가 발표된 바 있다.(Castro-Neto et al., 2009) 교통량 결측에 대한 추정을 위해서 1주일 교통량으로 모형에 대한 검증을 실시한 결과 단수회귀모형의 경우 AADT 추정오차는 4.85%, 단순 신경망 모델의 AADT 추정오차는 2.85%, 개량된 회귀모형의 평균 AADT 추정오차는 0.02%, 개량된 신경망 모델의 경우 평균 AADT 추정 오차는 약 1% 제시하고 있다.(Satish et al., 2003) 교통 시계열 분석의 네 가지 주요 문제인 비정상 데이터 탐지, 데이터 압출, 누락 데이터 보정, 교통 예측에서 추세 모델링이 수행하는 역할을 순차적으로 검토하였고, 비정상 데이터 탐지 및 추세 모델링, 교통 데이터 압축, 교통 데이터 누락 및 추세 모델링을 통해서 시간적 패턴을 구체화하는 기법일 뿐만 아니라 교통 시계열의 공간적 관계와도 관련이 있음을 제시하고 있다.(Li et al., 2015) 이에 본 연구에서는 교통량 시계열의 특성상 계절성을 반영한 SARIMA 모형을 적용하여 교통량에 대한 보정과 예측을 수행하고자 하였다.

Ⅲ. 연구 방법론

교통량 자료는 시간적 순서를 지닌 관측 자료로 횡단면 자료와는 다른 특징을 지니고 있다. 흔히 횡단면 자료(cross-sectional data)의 경우 시간의 특성을 고려할 수 없다. 반면 시계열 자료(time series)의 경우 특정 대상의 시간에 따라 변하는 관측자료를 수집한 데이터로 이를 활용하여 미래의 추이나 값을 예측하는데 주로 활용되고 있다. 시계열 데이터는 데이터의 생성 특성에 따라 연속적으로 생성되는 연속시계열(continuos time series)과 이산적으로 생성되는 이산시계열(discrete time series)로 구분된다. 교통량 자료의 경우 일정한 관측 시차를 나타내어 관측되므로 이산시계결의 형태를 지니게 되며, 관측 시점들 간의 시차(time lag)가 분석 시 독립 변수와 같은 역할을 한다. 본 절에서는 이산형 시계열인 교통량 데이터가 가지는 특징을 바탕으로 본 연구에서 활용 가능한 방법론인 SARIMA 알고리즘에 대해 살펴보고자 한다.



<Fig. 1> Time series analysis method(Model building and prediction)

1. 시계열 분석의 개요

시계열 자료(time series data)는 특정 변수 혹은 변수들을 시간에 따라 관측하여 얻는 자료이다. 예를 들어 주가, 통화량, 연간 지표의 변화율, 판매량 등 사회과학 분석에 사용되는 다양한 변수들이 시계열 자료의 형태로 수집 및 구축될 수 있다. 이러한 데이터를 활용한 시계열 분석이란 관측된 변수들이 시간에 따라 변하는 추이를 분석하고 이를 기반으로 장래를 예측하는 것이다. 즉, 추세를 파악하거나 향후 전망 등을 예측하기 위한 용도로 시계열 분석이 주로 활용되고 있다. 시계열 형태(the components of time series)는 데이터 변동 유형에 따라 <Table 2>와 같이 불규칙 변동, 추세 변동, 순환 변동, 계절 변동으로 구분할 수 있다.

<Table 2> The component of time series

Item	Contents
Irregular variation	Unlike regular movements over time in time-series data, it is an unpredictable and accidental variation due to the absence of any regularity.
Trend variation	It refers to the long-term changing trend of time series data. ※ At this time, trends mean a tendency to continue to increase or decrease over a long period of time, or to maintain a constant state, so there are disadvantages that it is difficult to find trends in a short period of time.
Cyclical variation	Usually, it refers to fluctuations that appear cyclically over a fixed period of time.
Seasonal variation	In time series data, usually according to seasonal influences and social customs. It means that it occurs in a certain cycle. It usually cycles and fluctuates with the seasons.

시계열 형태를 기준으로 시계열 데이터는 규칙적 시계열과 불규칙적 시계열로 나눌 수 있다. 규칙적 시계열은 트렌드와 분산이 불변하는 데이터이며, 불규칙 시계열은 트렌드 혹은 분산이 변화하는 시계열 데이터를 말한다.

2. 시계열 분석 기법

시계열 데이터의 효율적인 분석은 불규칙성을 갖는 시계열 데이터에 특정한 기법이나 모델을 적용하여 규칙적 패턴을 찾거나 예측하는 것을 말한다. 불규칙적 시계열 데이터에 규칙성을 부여하는 방법으로는 AR, MA, ARMA, ARIMA 모델을 적용하는 것이 가장 널리 알려져 있으나, 최근에는 머신기반 학습법인 가법적 회귀(additive regression, 딥러닝 등을 이용하여 시계열 데이터의 연속성을 기계 스스로 찾아내도록 하는 방법이 상대적으로 매우 우수한 예측 성능을 보여주고 있다.

1) ARIMA

ARIMA는 전통적인 시계열 분석 기법으로 추세성과 계절성이 동시에 나타나는 시계열 데이터를 (식1)과 같이 계절 차분, 일차 차분하여 정상 시계열로 변환시킨다. 일반적으로 식(1) 중 T는 추세성, S는 계절성, E는 임의성을 나타낸다.

$$\gamma_t = T_t + S_t + E_t \quad \cdots\cdots (1)$$

과거 관측 데이터 (t-1)로 인한 영향을 반영한 자기상관모형(Auto Regression, AR)과 오차항(E)으로 인한 영향을 반영한 이동평균모형인 MA(Moving Average, MA)이 결합된 모형이다. 적절한 AR과 MA를 구하기 위해 자기상관함수(Autocorrelation Function, ACF)와 편자기상관함수(Partial Autocorrelation Function, PACF)를 이용하여 차분하는 것이 주된 분석 목표이다.

$$ARIMA(p, d, q) \quad \cdots\cdots (2)$$

- p is the order of nonseasonal autoregression
- d is the degree of differencing
- q is the order of nonseasonal moving average of the error

ARIMA는 Box-Jenkins 모델(ARMA)의 일반적인 형태를 나타내는 모형으로 만약 일차 차분이 0이 된다면 정상 시계열이 보장된다는 의미이며, 이는 ARIMA와 ARMA는 같은 모형임을 나타낸다. 일차 차분과 ARMA 모형을 결합하여 ARIMA 모형의 추정식을 표현하면 다음 식(3)과 같다.

$$\gamma'_t = I + \alpha_1 y'_{t-1} + \alpha_2 y'_{t-2} + \cdots + \alpha_p y'_{t-p} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} \quad \cdots\cdots (3)$$

여기서 e는 에러를 의미한다.

그러나 ARIMA 모델은 지수평활 모델과 마찬가지로 특정 주기 단위(예 : 일, 월, 분기)처럼 계절성 주기(seasonal period, 주기성 관찰에 필요한 단일 주기)가 짧은 경우에만 적용이 용이하다는 단점이 있다. 1950년 이후 전통적으로 활용되어 온 모수 기반 선형 모형은 대체로 단기적인 패턴 분석에 장점을 가지며, 그 구조적 한계로 인해 최근에 길어진 계절성 주기와 다중 계절성 모델 반영에 한계가 있다. 따라서 짧은 시간 단위로 저장되는 교통량 관측 데이터의 경우 기존 ARIMA 모형과 지수 평활화법 모형 등의 전통적 시계열 기법을 통한 분석은 추정 과정이 매우 번거로우며 높은 예측 성능을 보장하기 어렵다.

2) SARIMA

ARIMA 모형은 주기성 및 계절성이 강하거나 불규칙한 데이터에 한계를 보이고 있어 교통량 데이터와 같은 자료에 적용하기에는 한계가 있다. 이러한 한계점을 극복하고 보완하기 위해 ARIMA 모형에 자기상관 모형인 SAR(Seasonal Auto Regressive)과 계절 이동평균 모형인 SMA(Seasonal Moving Average)가 결합된 분석 기법이 SARIMA 모형이다. 식(2)에 계절성(m)과 같은 확률적 특성이 반영된 모형으로 볼 수 있다.

$$ARIMA(p,d,q)(P,D,Q)_m \quad \cdots\cdots (4)$$

- p is the order of nonseasonal autoregression
- d is the degree of differencing
- q is the order of nonseasonal moving average of the error
- P is the order of the seasonal autoregression
- D is the degree of seasonal differencing
- Q is the order of seasonal moving average of the error
- m is the number of observations in the year(for yearly seasonality)

ARIMA 모형과 가장 큰 차이를 보이는 부분은 계절성 주기(m)가 반영되어 있다는 점이다. 추정식은 다음 식(5)와 같이 표현할 수 있다.

$$\Phi_p(B^s)\,\Phi_p(B)\,(1-B)^d\,(1-B^s)^D\,Y_t = (1+\theta_Q(B^s)\theta_q(B)\varepsilon_t \quad \cdots\cdots (5)$$

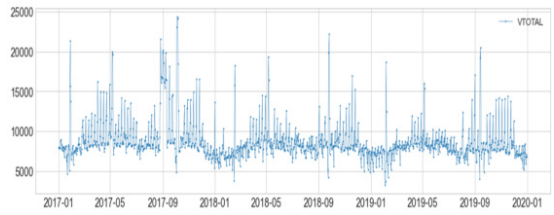
식(5)에서 ε_t는 오차항(a.k.a., white noise), B 후항 연산자이며, 만약 교통량 데이터 주기가 월별이라면 식은 다음 식(6)과 같이 표현할 수 있다.

$$(1-\Phi_p)(1-\Phi_1 B^{12})\,(1-B)(1-B^{12})\,y_t = (1+\theta_1 B)\,(1+\theta_1 B^{12})\varepsilon_t \quad \cdots\cdots (6)$$

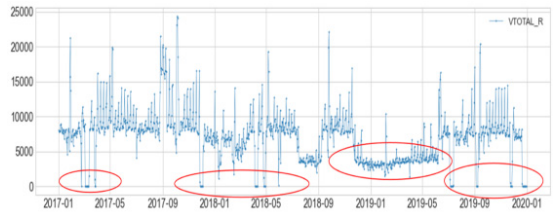
Ⅳ. 연구 데이터

1. 일별 데이터

본 연구에서는 연속된 3개 지점에 대하여 2017년, 2018년, 2019년 시간대별 원시 및 보정 교통량 데이터를 활용하여 일변량 시계열 예측 및 보정 모형을 구축하였다. 모형 구축 시 기본 가정은 보정 후의 데이터를 기준값(ground truth)으로 설정하고, 이를 바탕으로 원시 자료(raw data)의 이상 감지, 결측 보정, 장래 교통량 예측 과정을 수행하였다. 이상치는 원시 데이터에서 교통량이 0으로 표현된 것을 정의하였으며, 교통량의 패턴을 분석하여 이상치를 제거하여 교통량 데이터를 보정하였다.

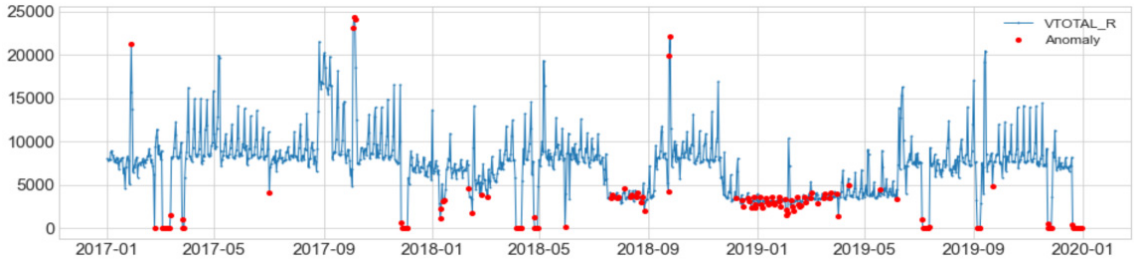


<Fig. 2> Results of the 2017~2019 traffic volume survey(After calibration data)



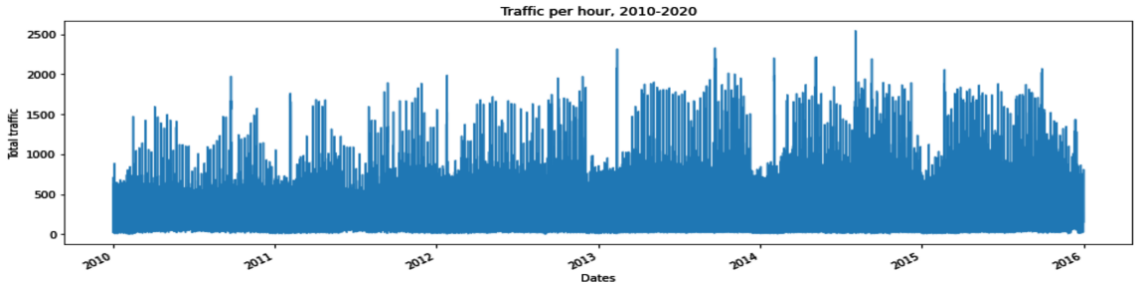
<Fig. 3> Results of the 2017~2019 traffic volume survey(Before calibration data)

<Fig. 2>의 경우 'VTOTAL'이 보정 후의 일별 시계열을 나타내며, <Fig. 3>은 'VTOTAL_R'은 보정 전의 데이터를 의미한다. 보정 전의 데이터의 단기 결측 구간을 살펴보면 2017년과 2018년도에는 결측점이 다수 존재하고 2019년은 이상치가 많은 것으로 나타났다. 단기 및 장기 결측이 불규칙하게 존재하는 것으로 나타났으며(<Fig. 3>의 적색 타원 참고), 전통적인 통계 분석법을 적용하기 위하여 결측을 모두 결측(교통량=0)으로 처리하여 분석을 수행하였다.



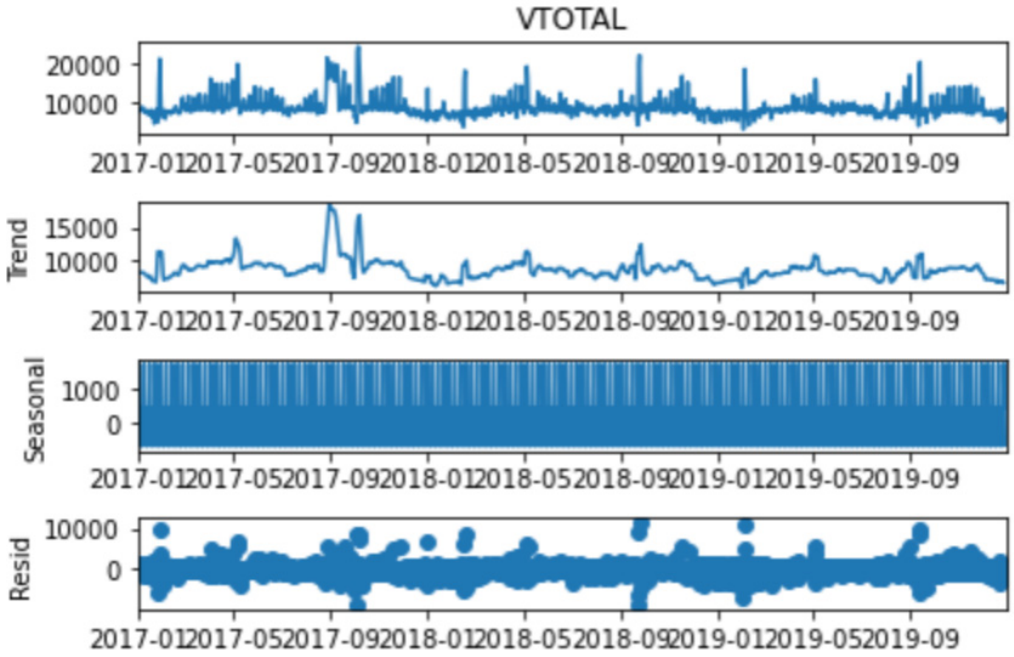
<Fig. 4> Detection of anomaly data

이상치의 경우 기준값이 있는 경우에는 판단이 가능하지만 현장 관측 시에 즉각적인 판단이 어려운 경우가 많다. 따라서 과거의 실제값을 바탕으로 패턴을 학습하여 이상치를 검지하는 알고리즘을 적용하여 다음과 같이 이상치들을 선별하는 알고리즘을 개발하고 적용하였다. 적용된 알고리즘은 seasonal decomposition, Quantile AD, aggregation 기법을 결합하여 활용하였으며, 실제값과 비교를 통해 91% 이상의 검지 성능을 나타내는 것을 확인하였다. 이를 기반으로 선별된 이상치 또한 결측치와 마찬가지로 결측(교통량=0)으로 처리하여 분석을 수행하였다.



<Fig. 5> Traffic volume survey results by time slot from 2010 to 2019 (data after calibration)

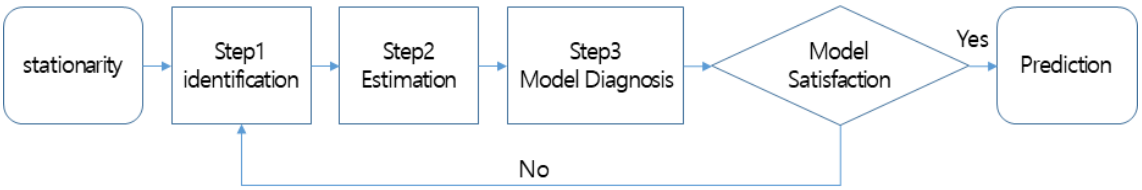
본 연구에서는 모형의 활용성과 적용 가능성을 확장 검증하기 위하여 2010년부터 2019년 시간대별 교통량 데이터도 활용하였으며, 일별 데이터와 마찬가지로 모형 구축 시 기본 가정은 보정 후의 데이터를 기준값(ground truth)으로 설정하고, 이를 바탕으로 원시 자료(raw data)의 이상 감지, 결측 보정, 장래 교통량 예측 과정을 수행하였다.



<Fig. 6> Identify trends, seasons, and residuals in traffic data

Ⅴ. SARIMA 분석 결과

Box-Jenkins의 3단계 방법론을 통해 주어진 데이터에 적합한 시계열 모형을 채택하였다. 먼저, 교통량 데이터의 주기성, 계절성, 추세성, 잔차 특성을 <Fig. 6>과 같이 확인해 보았다.



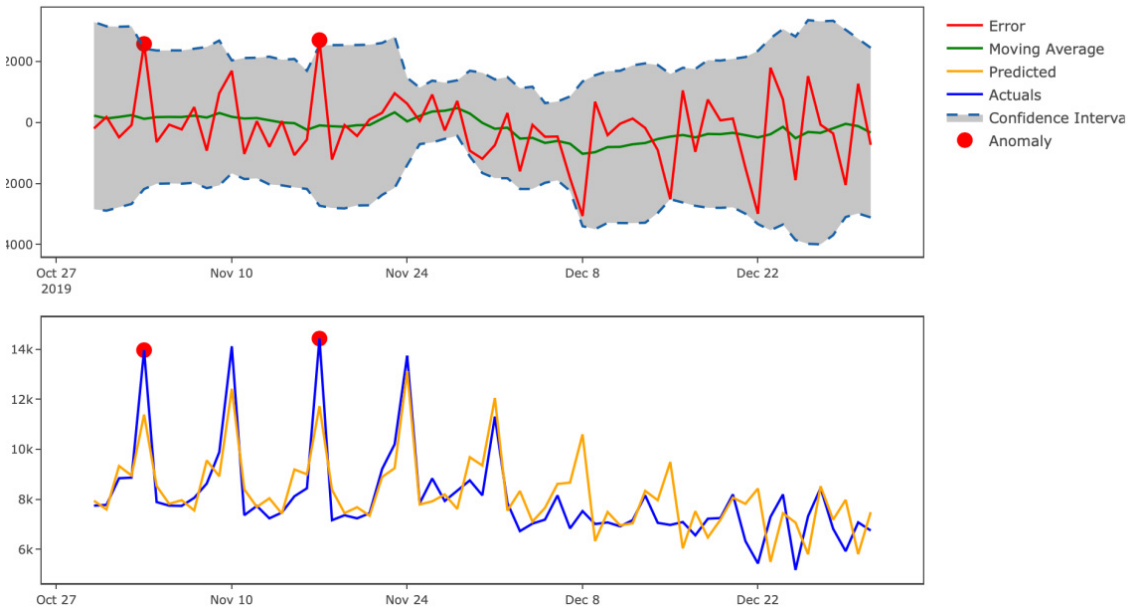
<Fig. 7> SARIMA Time Series Model Estimation Process

빈도수에 대한 원시 시계열 그림(Fig. 6의 2번째)을 보면 2017년 여름을 제외하면 시차에 따라 교통량이 비교적 안정적으로 나타나고 있는 것을 볼 수 있다. 또한, 시계열 데이터의 잔차 그림(Fig. 6의 4번째)에서 확인할 수 있듯이 특정한 패턴이 나타나지 않는 것을 확인되어 Heteroskedastic(이분산성) 문제도 없는 것으로 판단되었다, 1차분 수행 후에도 큰 차이를 보이지 않아 정상 시계열임을 확인할 수 있었다. 하지만 그림으로 파악되지 못하는 자기상관문제(정상성 여부)를 파악하기 위하여 자기상관함수(Autocorrelation funtion, ACF), 편자기상관함수(Partial Autocorrelation Function, PACF)를 관찰하여 신뢰수준 안에 값들이 위치하는지를 확인하였다.

정확한 판단을 위해 Ljung-Box(LB) 테스트도 수행하였으며 유의 확률이 0.1 이하인 0.04로 나타나 자기상관은 크게 없는 것으로 판단하였다. 따라서 추가적인 로그 변환, 차분 과정을 수행하지 않고 SARIMA의 최적 파라미터를 최우추정법을 활용한 Auto grid searching을 통해 도출하였다. 모형의 설명력을 판단하기 위하여 AIC(Akaike's Information Criterion) 통계량, BIC(Baysian Information Criterion) 통계량, 유의성 검정 등을 고려하여 모형의 최적 파라미터를 결정하였고 결과는 <Table 3>과 같다.

<Table 3> The Optimal Parameters Estimation Results for the SARIMA Model

Model: SARIMAX (4,1,3)(4,0,3)12			
Type	Factor	SE Factor	P>|z|
AR (Time lag1)	-1.1503	0.041	0.000
AR (Time lag2)	-0.1593	0.051	0.002
AR (Time lag3)	0.0759	0.061	0.210
AR (Time lag4)	-0.2939	0.040	0.000
MA (Time lag1)	0.7886	0.035	0.000
MA (Time lag1)	-0.6714	0.034	0.000
MA (Time lag1)	-0.7927	0.034	0.000
AR Season (Time lag1)	0.2791	0.309	0.367
AR Season (Time lag2)	0.0555	0.353	0.875
AR Season (Time lag3)	-0.7049	0.296	0.017
AR Season (Time lag4)	-0.1349	0.049	0.006
MA Season (Time lag1)	-0.3960	0.307	0.197
MA Season (Time lag1)	-0.0459	0.372	0.902
MA Season (Time lag1)	-0.7397	0.316	0.019
AIC	6983	Ljung-Box(Q)	0.04
BIC	7068	Log likelihood	-3201



<Fig. 8> Predictive Results for Model SARIMA (4,1,3) (4,0,3)12

최적 파라미터 조합인 SARIMA(4,1,3)(4,0,3)12 모형을 활용한 교통량 예측 결과는 <Fig. 8>와 같다. 모형의 평가지표는 Mean Absolute Percentage Error(MAPE)를 활용하여 측정하였으며 이는 실제값과 예측값을 비교하는 지표라고 볼 수 있다.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n} abs\left(\frac{y_i - \hat{f}(x_i)}{y_i}\right) \quad (7)$$

- $\hat{f}(x_i)$ is predicted value
- y_i is actual value
- x_i is independent variable (time)
- n is number of observation

전체적인 예측 결과는 평균 15% 정도의 MAPE를 보이며 우수한 성능을 보였다. 평균적인 예측 성능이 저하된 근본적인 이유는 장기 결측이 발생된 2019년이며, 이는 장기 결측으로 인해 주기성과 추세성이 제대로 반영되지 못했기 때문이다.

Ⅵ. 결 론

데이터 수집 및 저장 성능이 증가되면서 다양한 분야에서 데이터 기반 분석이 활발히 이루어지고 있다. 특히 도시 시스템이 실시간으로 수집되는 시계열 데이터의 빈도가 잦아지고 종류도 매우 다양해지고 있다. 이에 따라 시계열 데이터의 계절성이나 주기성이 길어지거나 복잡해지는 다중 계절성을 나타내기도 한다.

교통량 관측 자료의 경우 위와 같은 특성을 매우 잘 반영하고 있는 데이터라고 볼 수 있다. 하지만 교통량 데이터의 경우 결측 및 이상치가 다수 포함되어 있는 문제가 존재한다. 이렇게 불확실성이 내포된 교통량 정보가 적절한 가공 없이 다양한 분석의 기초 자료로 활용된다면 신뢰성 있는 결과 도출이 불가능하다.

이러한 교통 부문 자료의 불확실성 문제는 차세대 교통시스템에서도 피할 수 없는 이슈가 될 것으로 보인다. 특히, 센서 기반으로 수집되는 시계열 형식의 교통량 자료는 실시간으로 방대한 규모의 자료를 축적하지만, 이러한 첨단장비를 통해 수집된 자료 내에서도 이상 검지 및 결측값이 상시 존재하기 때문에 수집된 자료의 불확실성을 판단하고 적절한 보정 작업을 수행하는 것은 매우 중요하다. 따라서 본 연구에서는 이러한 교통량 데이터의 새로운 특성을 반영하고 예측 정확도를 높일 수 있는 시계열 추정 기법 중 기존 모델이 가진 한계점을 극복한 SARIMA 모형을 제시하였다.

분석결과 분석 주기가 비교적 긴 경우(1일 단위)에서 SARIMA가 약 85% 이상의 우수한 예측 성능을 보이는 것으로 나타났다. 이러한 교통량 보정 및 예측 향상은 실제 교통량 조사에서 운영 효율성 향상에 크게 기여할 수 있을 것으로 기대된다. 또한, 교통량 데이터 외에도 최근 수집되는 다양한 시계열 데이터에도 적용 가능하다는 점에서 이 연구는 큰 가치가 있다고 생각된다.

본 연구에서는 SARIMA 모형을 활용하여 교통량 데이터의 보정과 예측을 수행하였으나, 향후 Facebook의 Prophet, RNN, LSTM 등 다양한 모형을 대상으로 교통량 보정 및 예측에 대한 알고리즘을 추가로 검증하여 정확도 측면에서 SARIMA 모형과 비교해볼 필요가 있다고 판단된다. 또한, 교통량 예측을 위해 SARIMA 모형에서 반영되지 못하는 변수들의 영향을 추가하는 통계기법도 함께 고려하여 교통량의 예측 정확도를 높이는 연구가 추가로 필요할 것으로 판단된다.

REFERENCES

Castro-Neto M., Jeong Y. S., Jeong M. K. and Han L. D.(2009), “SVR for short-term traffic flow prediction under typical and atypical traffic conditions,” *Expert Syst. Appl.*, vol. 36, no. 3, pp.6164-6173.

Hamed M. M., Al-Masaeid H. R. and Bani Said Z. M.(1995), “Short-term prediction of traffic volume in urban arterial,” *J. Transp. Eng.*, vol. 121, no. 3, pp.269-1254.

Kim B. S., Kim S. W., Fang Y. and Wong T. F.(2010), “Feedback-assisted MAC protocol for real time traffic in high rate wireless personal area networks,” *Wireless Networks*, vol. 16, no. 4, pp.1109-1121.

Korea Institute for Health and Social Affairs(2018), *A Study on anomaly detection based on Machine Learning*, pp.81-84.

Korea Institute of Civil Engineering and Building Technology(2008), *A Study on the Improvement and Evaluation of Information Reliability*, pp.45-55.

Lee S. and Fambro D. B.(1999), “Application of subset autogressive integrated moving average model for short-term freeway traffic volume forecasting,” *Transport. Res. Part C: Emerg. Technol.*, vol. 1678, no. 1, pp.179-188.

Lee S. B.(2019), *Development of Dynamic Traffic Volume Estimation Model Using DSRC Prove Data*, Doctoral Dissertation, Seoul National University, p.12.

Lee S. H. and Shin J. M.(2013), “A Study on Imputing the Missing Values of Continuous Traffic

Counts," *Journal of the Korean Society of Civil Engineers*, vol. 33, no. 5, pp.2009-2019.

Li L., Su X. N., Zhang Y., Lin Y. and Li Z. H.(2015), "Trend modeling for traffic time series analysis:An integrated study," *IEEE Trans. Intell. Transport. Syst.*, vol. 16, no. 6, pp.3430-3439.

Li L., Xiaonan S., Yi Z., Yuetong L. and Zhiheng L.(2015), "Trend Modeling for Traffic Time Series Analysis: An Integrated Study," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, p.3430.

Satish S., Pawan L. and Ming Z.(2003), "Effect of Missing Value Imputations on Traffic Parameters Estimations from Permanent Traffic Counts," *Transportation Research Board 80th Annual Meeting, Washington, D.C.*, pp.1-33.

Satish S., Pawan L. and Zhong M.(2003), "Effect of Missing Value IMputations on Traffic Parameters Estimations from Permanet Traffic Counts," *TRB 2003 Annual Meeting CD-ROM*, p.1.

Steffen M. and Bartz-Beielstein T.(2017), "Time Series Missing Value Imputation in R," *Contributed Research Article, The R Journal*, vol. 9, no. 1, pp.207-218.

Turner S. M., Lomax T. and Margiotta R.(2001), *Monitoring Urban Roadway in 2000:Using Archived Operations Data for Reliability and Mobility Measurement*, Texas Transportation Institute, College, TX.

Zhong M. and Sharma S.(2009), "Development of improved models for imputing missing traffic counts," *The Open Transportation Journal*, vol. 3, no. 1, pp.35-48.

Zhou Y. C. and Meng P. C.(2019), "Diagnosis of causes for high railway traffic based on Bayesian network," *Math. Model. Eng. Probl.*, vol. 6, no. 1, pp.136-140.