Journal of the Research Institute of Industrial Technology, Vol. 26, 2009

교통량 자료의 이상치 탐색에 관한 연구

임성한* · 허태영**
*한국건설기술연구원, **한국해양대학교 데이터정보학과

A Study on the Outlier Detection of Traffic Data

*Sung-Han Lim** · *Tae-Young Heo***
**Korea Institute of Construction Technology, Ilsan, Korea*
***Department of Data Information, National Korea Maritime University, Busan 606-791, Korea*

요 약 : 본 연구에서는 지능형교통체계(Intelligent Transportation System, ITS) 검지기로부터 수집되는 교통량 자료의 이상치를 판단하는 두 가지 통계적 방법을 제시하였다. 첫 번째 방법은 요일과 5분 단위를 설명변수로 하는 가변수회귀모형을 구축하여 이상치 여부를 95% 신뢰구간으로 판단하였다. 비슷하게 두 번째 방법은 교통자료의 특성인 순환성을 반영한 순환회귀모형을 구축한 후 역시 95% 신뢰구간을 통하여 이상치 여부를 판단하는 과정을 제시하였다. 실제 자료를 통한 분석 결과 두 방법 모두 교통량 자료에 대한 이상치 판단 여부에 활용될 수 있음을 보였다.

핵심용어 : 지능형교통체계, 교통량, 가변수, 순환성, 회귀모형

ABSTRACT : *In the paper, we suggested two statistical methods to select the outliers of traffic data collected from intelligent transportation system. First, we construct the dummy regression model and select the outliers using 95% confidence interval. Similarly, we construct the circular regression reflected by circularity of the characteristic of traffic counts and select the outliers using 95% confidence interval. In real application, we show that two methods can be applied to detect the outliers from traffic counts.*

KEY WORDS : intelligent transportation system, traffic counts, dummy variable, circularity, regression model

1. 서 론

지능형 교통체계(Intelligent Transportation System, ITS) 검지기에서 관측되는 교통 자료(교통량, 속도, 밀도 등)를 살펴보면 여러 가지 이유로 인해 전체 자료의 전반적인 추세에서 벗어난 관측값들이 자주 발견되며 이를 이상치(outlier)라 할 수 있다. 이러한 이상치들은 ITS 검지기가 관측 및 검지과정에서 큰 오차가 발생하였거나 자료를 기록하는 과정에서 실수가 있었을 가능성도 있는 등 전체적으로 품질 관리의 미비에 따른 문제로 인해 발생하고 있다. 보통의 경우 많은 연구자들은 이상치 자료들을 제외시키거나 가중치를 낮게 주어 처리하고 있다. 이와 같은 이상치는 전체 자료의 경향을 많이 벗어남으로써 전체 자료의 경향성을 왜곡시킬 수 있기 때문에 많은 연구자들은 적절한 이상치 판단을 위하여 다양한 방법들을 개발하여 왔다. 이상치 탐색에 대한 통계적 방법으로 탐색적 자료 분석을 이용하는 방법과 회귀모형, 시계열 모형과 같은 통계적 모형을 이용하는 방법이 연구되어 왔다 (강진기 외, 2002; Dion and Rakha, 2003; 장진환, 2004; 도명식 외, 2004; 허태영 외 2008). 특히, 최윤혁(2003)은 택시에 GPS 수신기를 설치하여 수집한 프로브 차량으로부터 수집되는 자료의 이상치 탐색 방법을 제시하였으며, 이지연 외 (2003)는 국도 3호선을 대상으로 검지기로부터 수집한 교통량 자료의 이상치 탐색 방법을 제시하였다. 그러나 기초통계량 방법은 시각적인 방법으로 이상치 탐색에 있어 정확도가 떨어지는 반면에 시계열 모형은 정확도는 높으나 모형이 복잡하면서 실제 문제에 적용이 어려운 단점이 있다. ITS 검지기로부터 실시간으로 자동으로 관측되는 교통자료는 단기간에 많은 자료를 생성하기 때문에 교통정보의 시간적 변화를 파악하는데 도움이 되며, 다량의 관측값

* atdaya@kict.re.kr
** heoty@hhu.ac.kr 051)410-4373

에 대하여 전산처리가 가능하기 때문에 교통자료 분석에 있어서 많은 도움을 준다. 그러나 ITS 검지기로부터 수집되는 교통자료는 관측값의 기계적 오류 또는 결함 등에 의한 변위는 큰 오차를 야기할 수 있다. 이러한 이유로 일련의 관측값과 동떨어진 값이 관측될 경우 이 값을 진정한 관측값으로 사용하여야 할지 이상치로 판단해야 할지 하는 문제가 대두되고 있다. 이에 본 연구에서는 ITS 검지기에서 관측되는 교통자료의 이상치를 판단하는 여러 가지 통계적 기법을 고찰하여 관측값의 선택에 도움을 주고자 한다. 따라서 본 연구에서는 실제 적용에 쉽고 교통자료의 특성인 순환성을 반영한 가변수 회귀모형(dummy regression model)과 순환회귀모형(circular regression model)을 기반으로 이상치를 탐색할 수 있는 방법론을 제시하였다. 본 논문의 구성은 다음과 같다. 2장에서는 교통량 자료에 대한 특성 및 기초통계량을 제시하고, 탐색적 자료 분석을 통하여 이상치 판단여부를 확인하였으며, 3장에서는 실제 자료를 통한 이상치 탐색을 위하여 가변수회귀모형을 구축하고 그에 따른 결과값을 제시하였다. 4장에서는 순환회귀모형을 구축하고 그에 따른 결과값을 제시하였으며, 마지막으로 5장에서 결론을 도출하였다.

2. 자료

2.1 자료

본 연구에서는 ITS 검지기로부터 수집되는 교통자료에 대한 특성을 파악하고 이상치 탐색에 대한 다양한 통계적 기법을 검증하기 위해서 검지기로부터 수집된 자료에 대한 기초자료 분석이 필요하다. 본 연구에서는 부산지역의 11개의 ITS 검지기로부터 수집된 자료를 이용하여 본 연구에 활용하였다. 자료 수집기간은 각 ITS 관측지점별 자료 품질의 일관성을 유지하기 위하여 2007년 1월 1일부터 2007년 12월 31일까지 1년간의 자료를 확보하였으나, 교통자료의 순환성(circularity) 특성으로 일주일 자료만을 가지고 분석을 하였다. 본 연구에서 사용된 ITS 자료의 집계시간 단위는 수집된 검지기의 원시 집계간격인 30초 대신 본 연구에서는 5분 자료를 이용하였다. 수집된 30초 검지기 자료의 차로별 교통량 자료를 5분 단위의 자료로 변환하였다. 결론적으로, 본 연구에서는 교통량자료에 대한 이상치 탐색을 위한 통계적 분석을 수행하기 위하여 모든 값이 관측된 교통 자료만을 이용하였으며, 교통자료의 가장 큰 특성인 순환성을 이유로 1년 자료 대신 3월 달의 한 주를 선택하여 월요일부터 일요일까지의 일주일 자료만을 가지고 분석을 시도하였다.

2.2 교통자료에 대한 탐색적 자료 분석

한 지점의 ITS 검지기에서 관측되는 교통자료는 단일 시계열자료로서 이상치를 발견하기 위하여 자료의 전체적인 패턴과 함께 기본적인 통계량인 평균값과 표준편차로부터 자료의 전체적인 분포도를 파악하기 위하여 기초통계량을 제시하였다.

Table 1과 Table 2는 일주일 간의 5분 자료에 대한 다음과 같은 기초통계량을 나타낸다. 자료의 수는 2016개이며, 각 지점별 5분 간격의 상행교통량에 대한 평균과 표준편차를 보여주고 있다. 지점 50802와 지점 51102의 경우 상대적으로 다른 지점보다 상행교통량이 작고, 지점 52802와 지점 53102의 경우 다른 지점보다 상행교통량이 많은 것으로 보아 검지기가 지리적으로 도심에 가까운 곳에 위치하고 있음을 알 수 있다.

Table 1. Basic statistics for upstream traffic counts of ITS monitoring stations

ITS 검지기 고유번호	자료수	평균	표준편차
50302	2016	10.424	7.971
50402	2016	10.993	8.508
50702	2016	12.756	9.791
50802	2016	9.823	7.995
51102	2016	8.514	6.806
51202	2016	10.458	8.277
51702	2016	12.590	10.291
52102	2016	15.911	12.217
52502	2016	18.916	14.705
52802	2016	23.705	17.627
53102	2016	23.667	16.147

Table 2는 하행교통량에 대한 기초통계량을 보여주고 있으며, 지점별 평균 교통량과 표준편차에 있어 상행교통량과 비슷한 양상을 보여주고 있다.

Table 2. Basic statistics for downstream traffic counts of ITS monitoring stations

ITS 검지기 고유번호	자료수	평균	표준편차
50302	2016	11.1756	7.2779
50402	2016	11.4543	7.4449
50702	2016	13.7152	9.5855
50802	2016	10.5575	8.0875
51102	2016	9.5679	7.4718
51202	2016	11.7167	9.0314
51702	2016	13.9350	11.2776
52102	2016	16.7073	12.8696
52502	2016	18.7929	14.7929
52802	2016	25.3065	19.4814
53102	2016	26.4747	19.7889

Fig. 1에서 보는 것과 같이 ITS 수집 자료의 경우 하루 24시간, 또는 5분 단위로는 288분을 주기로 교통량 자료가 주기성을 가지고 있음을 보여주고 있다. Fig. 1의 위는 상행을 나타내고 아래는 하행의 교통량을 나타낸다.

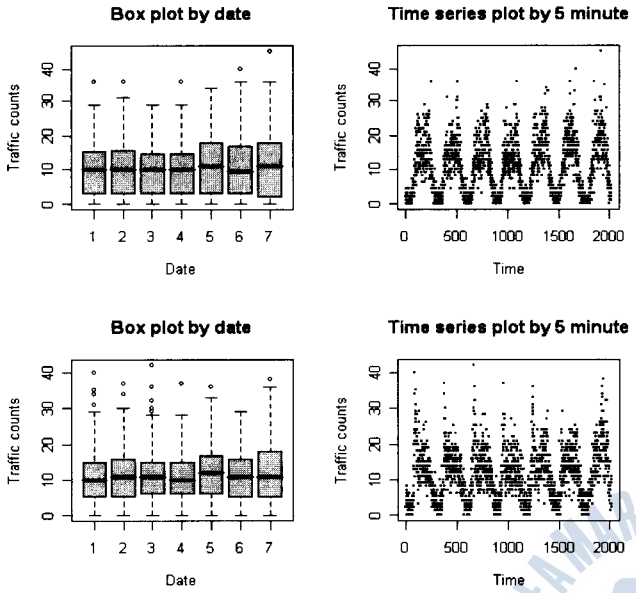

Fig. 1 Box plot and time series plot for upstream(above) and downstream traffic counts(below). Plot for 50302 monitoring station and then number represents as 1:Monday, 2:Tuesday, 3:Wednesday, 4:Thursday, 5:Friday, 6:Saturday, 7:Sunday)

상행과 하행선에 대한 교통량에 대하여 상자그림(box plot) Fig. 1을 통해 살펴보았다. 먼저, 요일별로 살펴보면 교통량은 금요일, 토요일, 일요일의 교통량이 다른 요일보다 약간 교통량이 많은 것으로 나타나며, 5분 단위의 교통량이 40을 넘는 이상치(outlier)가 많이 나타났다. 상행선과 하행선 모두 월요일, 수요일, 목요일은 교통량이 좀 낮게 나타났고 작게 퍼져 있는 반면, 일요일은 사분위수범위(IQR)가 넓게 퍼져 있다.

3. 이상치 탐색을 위한 통계적 방법

이상치 판단을 위한 통계적 모형을 구축하기 위하여 간단한 가변수 회귀모형과 주기성을 반영한 회귀모형을 통하여 이상치 판단을 위한 통계적 모형을 구축하기로 한다.

3.1 선형회귀모형

선형회귀분석이란 관찰된 연속형 변수들 사이의 관계에 있어서 한 변수를 원인으로 하고 다른 변수들을 결과로 하여 원인변수들과 결과변수 사이의 선형식을 구하고 그 식을 이용하여 원인변수들의 변수값들이 주어졌을 때, 결과변수의 변수값을 예측하는 통계적 분석방법이다. 선형회귀분석에서 원인의 역할을 하는 변수를 설명변수 또는 독립변수라고 하고, 결과를 관측하는 변수를 반응변수 또는 종속변수라고 한다. 하나의 종속변수와 하나의 독립변수 사이의 선형모형을 단순선형회귀모형이라고 하고, 하나의 종속변수와 둘 이상의 독립변수들 사이의 선형모형을 다중선형회귀모형이라 한다.

3.2 회귀모형의 추정법

종속변수 또는 반응변수 Y를 설명하는데 k개의 독립변수 또는 설명변수인 $X_1, X_2, \cdots, X_k$를 도입할 때 다중회귀모형은 다음과 같이 정의 된다.

$$Y_i = \beta_o + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i, \quad i = 1, 2, \cdots n \quad (1)$$

여기서, $\beta_0, \beta_1, \cdots, \beta_k$ 는 추정되어야 할 회귀계수들이며, ε_i는 서로 독립이고, 정규분포($Normal(0, \sigma^2)$)를 따르는 오차항을 의미한다. $\beta_j (j = 1, 2, \cdots k)$는 j번째 독립변수 X_j의 회귀계수(기울기)를 의미하며, X_j를 제외한 다른 모든 독립변수들의 값이 고정된 상태에서 X_j의 값이 한 단계 증가할 때 Y값의 평균변화량을 나타낸다. 그리고 X_{ji}는 독립변수 X_j의 i번째 관측치를 의미한다. 위의 회귀모형은 각 독립변수들에 대한 1차항 만을 포함하고 있는데 경우에 따라서는 2차항 이상의 고차항이나 교호작용(interaction)항이 모형에 포함될 수도 있다. 다중회귀모형은 행렬과 벡터를 통해 간단한 형태로 나타낼 수 있다.

3.3 모수추정방법

최소제곱법이란 관측값 Y_i와 추정값 $\hat{Y}_i$간의 편차인 잔차 $\varepsilon_i = Y_i - \hat{Y}_i$의 제곱합을 최소로 하는 표본회귀계수들을 구하는 방법이다. 이렇게 구해진 표본회귀계수는 모형의 가정을 충족한다면 모집단 회귀계수의 최량 선형불편추정량(Best Linear Unbiased Estimator: BLUE)이 된다. 이 때 최량선형불편 추정량이란 관측값 Y_i들의 선형결합으로 나타나는 (선형) 불편추정량들 중에서 최소 분산을 갖는 추정량을 의미한다. 잔차는 회귀모형이 올바로 설명되었다는 가정 하에 관찰되는 오류 또는 회귀식을 통해서 설명할 수 없는 편차라고 한다. 중회귀모형의 중요한 가정 중에 하나는 오차항의 정규성, 즉 $\varepsilon \sim Normal(0, \sigma^2)$이다. 잔차의 정규성을 검토함으로써 이를 생각해 볼 수 있는데, 만일 잔차들이 정규분포로부터 많이 벗어나 있다면 설정된 회귀모형은 본질적으로 타당하지 않다. 잔차들의 형태가 가정들에 위배되는지의 여부를 알아보는 가장 편하고 일반적인 방법은 그림으로 표시해 보는 것이다. 먼저 추정된 회귀선이 직선성인지와 동분산성인지를 알아보기 위해 다중회귀분석에서는 잔차도를 그려보는 것이 좋은 방법이다.

잔차에 대한 이분산성이 발생하면 가설검정이 의미가 없게 된다. 즉, 회귀분석이 의미가 없는 것이 된다. 통계학적으로 정의하면, 이분산성이 존재하면 최소제곱추정량은 여전히 불편추정량이되지만 유효추정량은 되지 않고, 분산의 추정량은 편의추정량이 되어 가설검정은 의미가 없다. 이러한 점 때문에 동분산성의 가정이 회귀분석에서 제일 중요하다.

3.4 가변수 회귀모형(Dummy Regression Model)

일반적으로 회귀분석이란 관찰된 연속형 변수들 사이의 관계에 있어, 한 변수를 원인으로 하고 다른 변수를 결과로 하여 설명변수와 종속변수 사이의 선형식을 구하고 그 식을 이용하여 원인(설명)변수의 변수값들이 주어졌을 때, 결과(종속)변수의 변수값을 예측하는 통계적 분석 방법을 의미한다.

본 연구에서는 우선, 분산분석 결과, 5분 교통량이 요일과 5분 자료에 의존한다는 사실을 확인하였으며, 이를 기반으로 5분 교통량 추정을 위한 회귀모형을 구축하였다. 이때, 각각의 요일과 5분 자료를 가변수(dummy variables)로 변환하여 이를 설명변수로 하고 5분 교통량을 종속변수로 설정하였다. 5분 교통량 추정을 위한 가변수 회귀모형의 기본식은 아래와 같다.

$$Y_{ij}=\mu+\text{요일}_i+5\text{분}_j+\varepsilon_{ij},\ \begin{cases} i=1,2,\cdots\ 6 \\ j=1,2,\cdots\ 287 \end{cases} \quad (2)$$

여기서, 요일$_i$과 5분$_j$ 변수는 가변수로 표현되었으며, 이들 가변수에 대하여 개략적으로 설명하면 다음과 같다. 요일$_i$ 변수는 월요일부터 금요일까지 7개가 있으나, 6개의 가변수로 구성(나머지 1개는 자동으로 결정되기 때문에 모형에 포함되지 않음)되며, 5분$_j$ 변수는 00:00~00:05 분부터 23:55~00:00 분까지 288개가 있으나, 요일 변수와 마찬가지로 287개의 가변수로 구성된다. 교호 효과(interaction effect)인 (요일×5분)$_{ij}$는 분산분석결과 통계적으로 유의하지 않은 것으로 나타나 이를 제거하였다. 따라서 최종적으로 구축된 일교통량 추정을 위한 가변수 회귀모형은 다음과 같다.

$$\begin{aligned} Y_t=\mu+\alpha_1\text{월요일}_t+\alpha_2\text{화요일}_t+\cdots+\alpha_5\text{금요일}_t \\ +\alpha_6\text{토요일}_t+\alpha_7 M_{1,t}+\cdots+\alpha_{287}M_{287,t}+\varepsilon_t, \\ t=1,2,\cdots,366 \end{aligned} \quad (3)$$

여기서, α_i는 요일과 5분 단위에 대한에 대한 계수를 나타내며, M_i는 5분 시간간격을 나타낸다. 여기서, 각 변수는 0 또는 1의 값을 가진다. 즉, 5분 교통량이 월요일에 관측되었다면 월요일$_t$는 1 그렇지 않으면 0을 가진다. 이렇게 구축된 17개의 가변수를 통해 최종 모형을 구축하였으며, 모형 구축과 변수 선택을 위해 상용 통계패키지인 SAS를 이용하였다.

3.5 가변수 회귀모형의 신뢰구간을 통한 이상치 탐색

일반적으로 많이 다루어지는 선형회귀모형은

$$y=X\beta+\varepsilon,\ \varepsilon\sim Normal(0, I\sigma^2) \quad (4)$$

과 같이 표현되고, $Var(\varepsilon)=I\sigma^2$으로 오차들은 각각 동일한 분산 $Var(\varepsilon_i)=\sigma^2$을 갖고 오차항들은 서로 독립이라고 가정한다. 그리고 위의 모형을 통한 어떤 주어진 독립변수들에 대한 종속변수 하나의 값인 y_s의 예측값을 $\hat{y}_s$라 하면, $\hat{y}_s$가 가지는 분산은

$$Var(\hat{y}_s)=\sigma^2\left[1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] \quad (5)$$

이며, 이의 추정값인 $\widehat{Var}(\hat{y}_s)$은 σ^2대신 MSE를 쓰면 된다. y_s에 대한 점추정 값은 추정된 회귀선 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$를 사용하기 때문에 $100(1-\alpha)\%$ 신뢰구간은

$$\hat{y}\pm z_{\alpha/2}\cdot A \quad (6)$$

이 된다. 여기서

$$A=\sqrt{MSE\left[1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]} \quad (7)$$

를 나타낸다.

따라서 주어진 설명변수의 값에서 하나의 예측값 y를 얻으려 할 때, 이 예측값의 95%또는 99% 신뢰구간을 통하여 자료가 신뢰구간 안에 포함되면 정상적인 자료로, 포함되지 않으면 이상치로 판별할 수 있다.

95% 신뢰구간은 다음과 같이 구할 수 있으며,

$$-1.96\cdot A<y<1.96\cdot A$$

99% 신뢰구간은 다음과 같이 구할 수 있다.

$$-2.58\cdot A<y<2.58\cdot A$$

본 연구에서는 일반적으로 많이 통용되는 95% 신뢰구간을 통하여 이상치 판단여부를 결정하였다.

3.6 결과

본 연구에서 사용된 11개의 ITS 검지기에서 수집된 5분 간격의 교통량, 속도, 점유율 간의 가변수 선형회귀모형을 구축한 후 선택되는 이상치를 판별하였다.

Table 3. The results from dummy regression model for upstream traffic counts

ITS 검지기 고유번호	이상치	정상치	전체 갯수	비율
50302	40	1976	2016	1.98%
50402	47	1969	2016	2.33%
50702	56	1960	2016	2.78%
50802	53	1963	2016	2.63%
51102	53	1963	2016	2.63%
51202	53	1963	2016	2.63%
51702	54	1962	2016	2.68%
52102	47	1969	2016	2.33%
52502	54	1962	2016	2.68%
52802	49	1967	2016	2.43%
53102	72	1944	2016	3.57%

Table 4. The results from dummy regression model for downstream traffic counts

ITS 검지기 고유번호	이상치	정상치	전체 갯수	비율
50302	67	1946	2016	3.32%
50402	67	1946	2016	3.32%
50702	71	1945	2016	3.52%
50802	70	1946	2016	3.47%
51102	70	1946	2016	3.47%
51202	69	1947	2016	3.42%
51702	64	1952	2016	3.17%
52102	81	1935	2016	4.02%
52502	80	1936	2016	3.97%
52802	82	1934	2016	4.07%
53102	87	1929	2016	4.32%

4. 주기성(periodic cycle)을 반영한 회귀모형을 통한 이상치 탐색

분석대상 시계열의 순환주기(cycle length or period)가 이미 잘 알려져 있는 경우에는 시계열의 순환성분(cycle component)을 모형화 하기 위해 조화분석(harmonic analysis)을 시행할 수 있으며 사인(sine)항과 코사인(cosine)항을 모두 포함한 식은 다음과 같이 표현될 수 있다.

$$Y_t = \mu + \alpha sin(wt) + \beta cos(wt) + \epsilon_t \quad (8)$$

여기서, Y_t는 시점 t에서의 변수 Y의 관찰값을 의미하며, μ는 시계열의 평균값, $w = 2\pi/\tau$는 각의 빈도수(angular frequency)이며, τ는 순환주기(period)를 나타내며, ϵ_t는 평균이 0이고 분산이 σ^2인 백색잡음(white noise)을 의미한다. 본 연구에서 t는 5분 단위를 나타내며, 순환주기(period)인 τ는 288을 사용하였다.

Table 5. The results from circular regression model for upstream traffic counts

ITS 검지기 고유번호	이상치	정상치	전체갯수	비율
50302	81	1935	2016	4.02%
50402	93	1923	2016	4.61%
50702	96	1920	2016	4.76%
50802	87	1929	2016	4.32%
51102	96	1920	2016	4.76%
51202	92	1924	2016	4.56%
51702	98	1918	2016	4.86%
52102	117	1899	2016	5.80%
52502	121	1895	2016	6.00%
52802	113	1903	2016	5.61%
53102	113	1903	2016	5.61%

Table 6. The results from circular regression model for downstream traffic counts

ITS 검지기 고유번호	이상치	정상치	전체갯수	비율
50302	154	1862	2016	7.64%
50402	161	1855	2016	7.99%
50702	152	1864	2016	7.54%
50802	99	1917	2016	4.91%
51102	100	1916	2016	4.96%
51202	97	1919	2016	4.81%
51702	92	1924	2016	4.56%
52102	110	1906	2016	5.46%
52502	99	1917	2016	4.91%
52802	112	1904	2016	5.56%
53102	124	1892	2016	6.15%

Table 3과 Table 4는 주기성을 반영한 회귀모형을 통한 이상치 탐색 결과이다. 주기성을 반영한 회귀모형의 결과값이 가변수 선형회귀모형보다 이상치를 더 많이 제공해 주고 있다. 그 이유는 본 연구에서 사용한 주기는 288로서 봉우리가 하나인 단봉형 주기를 사용한 반면에 실제 자료는 출근시간대와 퇴근시간대에 교통량이 많아지는 쌍봉형이기 때문이다. 그러나 실제로 주기분석에 쌍봉형 주기를 통한 주기모형을 구축하기 어렵기 때문에 가변수선형회귀모형과 주기성을 반영한 회귀모형에서 동시에 이상치로 판별되는 관측값을 이상치로 구분하면 된다. 결과적으로 본 연구에서는 이상치의 처리 방법에 대한 두 가지 통계적인 방법론을 제시하였으며, 현실적으로 적용 가능한 이상치 처리방법론을 체계화시켜 제시하고자 하였다.

이를 위해 본 연구에서는 부산광역시 ITS 사업에서 활용되고 있는 ITS 검지기로부터 얻어진 교통자료를 활용하여 이상치를 분류하였다.

5. 결 론

ITS 구축에 따라 시스템 상호간의 연계가 확대되면서 획득되어진 교통자료의 품질에 관한 관심이 고조되고 있는 실정이다. 교통자료에 대한 품질관리는 ITS 검지기에서 얻어지는 원시자료의 불량 여부 및 결측치에 대한 검사 방법을 기초로 한다. 이상치의 확인을 위해서는 우선적으로 자료를 도표화하여 잠재적인 극단값이나 이상치가 존재하는지를 우선적으로 판단하고, 다음으로 확인된 자료가 통계적인 방법에서도 이상치로 분류되는지를 확인한 뒤 적절한 처리방법이 적용되어져야 한다.

본 연구에서는 교통자료의 품질관리를 위하여 ITS 검지기에서 관측되는 교통자료의 이상치를 판단하는 여러 가지 통계적 기법 중 간단하고 실시간 교통량에 적합한 가변수회귀모형과 순환회귀모형을 구축하여 이상치 탐색에 적용할 수 있도록 하였다. 이러한 이상치 탐색을 위한 통계적 기법을 통하여 현재 ITS 검지기로부터 관측되는 다양한 교통자료에 대하여 실시간으로 이상 상황 및 패턴분석을 실시하여 ITS 관리자가 탐색된 이상치에 대하여 즉각적인 조치를 할 수 있는 이상패턴 인식 및 경보 기능을 구축할 수 있다는 장점을 가지고 있다.

참 고 문 헌

[1] 강진기, 손영태, 윤여환, 변상철, 2002: 비매설식 자동차량인식장치를 이용한 구간교통정보 산출방법 연구, 한국ITS학회 논문집, pp. 22-32.

[2] 도명식, 김성현, 배현숙, 김종식, 2004: 국도의 동질구간 선정과 이상치 제거 방법에 관한 연구, 대한교통학회지, 제22권, 제7호, pp. 7-16.

[3] 이지연, 도명식, 김성현, 류승기, 2003: 교통량 데이터의 실시간 보정로직 - 국도 3호선을 중심으로, 응용통계연구, 제16권, 2호, pp. 203-215.

[4] 장진환, 2004: 자동차량인식장치 자료 필터링 알고리즘 개발, 서울시립대학 석사학위논문.

[5] 최윤혁, 2003: 택시 GPS Probe 자료의 실시간 이상치 제거 알고리즘 개발, 아주대학교 석사학위 논문.

[6] 허태영, 엄진기, 박희문, 박찬근, 2008: 공간회귀모형을 이용한 수시교통량자료의 통계적 검정, Journal of the Korean Data Analysis Society, 제10권, 4호, pp. 1837-1848.

[7] Dion, F. and Rakha, H., 2003: Estimation spacial travel time using automatic vehicle identification data, TRB Annual Meeting.

원고접수일 : 2008년 12월 22일
원고채택일 : 2009년 01월 05일