▣ 論 文 ▣

TCS데이터를 이용한 이상치제거 및 결측보정 알고리즘 개발

Outlier Filtering and Missing Data Imputation Algorithm using TCS Data

도 명 식 (한밭대학교 도시공학과 교수)
이 향 미 (한밭대학교 도시공학과 석사과정)
남 궁 성 (한국도로공사 도로교통기술원 교통연구팀 수석연구원)

목 차

Key Words : TCS데이터, 이상치제거, 결측보정, 시간처짐 현상, 알고리즘 , 통행시간
Toll Collection System, outlier filtering, missing data, imputation, travel time

요 약

지능형 교통체계구축과 교통 혼잡이 증가하면서 이용자는 과거보다 양질의 통행시간정보를 요구하고 있다. 기존 연구에서는 단속류, 연속류 모두 AVI검지기 자료를 이용한 이상치제거 및 통행시간 산출에 대한 연구가 많이 이루어져왔다. 현재 한국도로공사에서는 TCS(Toll Collection System)를 기반으로 정보제공을 준비 중에 있으며, TCS 데이터는 운전자가 실제교통상황을 경험한 동적특성을 가진 통행시간이 수집된 자료로 통행시간 추정자료로 잠재력이 크다. 그러나 '시간처짐현상'이 발생하고 속도위반, 휴게소, 고장 등으로 인해 평균통행시간보다 작거나 큰 이상치와 결측데이터가 존재하여 기존 방법을 적용하는데 효과적이지 못한 것으로 나타났다. 따라서 본 연구에서는 TCS 데이터에 맞는 이상치제거 및 결측보정 알고리즘을 개발하였다. 기존알고리즘과 비교한 결과 개발 알고리즘이 더 효과적인 것으로 나타났다.

With the ever-growing amount of traffic, there is an increasing need for good quality travel time information. Various existing outlier filtering and missing data imputation algorithms using AVI data for interrupted and uninterrupted traffic flow have been proposed. This paper is devoted to development of an outlier filtering and missing data imputation algorithm by using Toll Collection System (TCS) data. TCS travel time data collected from August to September 2007 were employed. Travel time data from TCS are made out of records of every passing vehicle; these data have potential for providing real-time travel time information. However, the authors found that as the distance between entry tollgates and exit tollgates increases, the variance of travel time also increases. Also, time gaps appeared in the case of long distances between tollgates.

Finally, the authors propose a new method for making representative values after removal of abnormal and "noise" data and after analyzing existing methods. The proposed algorithm is effective.

Ⅰ. 서론

최근 우리나라는 반나절 생활권에 들어오면서 급증하는 교통수요가 도로시설용량에 달하는 수준에 이르렀다. 이로 인해 교통 혼잡이 발생하면서 통행시간증가와 함께 시간적/물질적 손실이 크게 발생하고 있는 실정으로, 그 어느 때 보다도 고속도로 교통정보의 중요성이 높아지고 있다. 또한 2개 이상의 경로가 존재하는 고속도로 다중경로환경에서 "통행시간"정보에 대한 요구는 매우 높아지고 있다.

일반적으로 통행시간정보는 지점정보를 구간정보로 변환하여 이용하는 방법과 해당구간을 주행한 차량이 경험한 통행시간자료를 이용하는 것으로 구분되는데, 해당구간을 주행한 차량이 경험한 통행시간자료는 출발지점에서 도착지점까지 "소요된" 시간으로써 출발시각으로부터 도착시각까지의 기간 동안 시공간상의 정체상태 변화를 반영하는 것으로 실제 참값이며, 지점정보를 구간정보로 변환하여 얻는 통행시간과는 차이가 있다.

현재 고속도로상에 설치하여 운영중인 루프검지기, 영상검지기와 같은 차량검지장치를 통해 수집되는 자료는 '지점정보'로서 시간단면마다 고속도로의 정체상황 변화를 즉시적으로 파악할 수 있는 장점이 있으나, 고속도로 이용차량이 실제 경험 하게 될 통행시간 추정에는 다소 복잡한 과정이 요구된다. 또한 시시각각으로 통행상황이 변화하는 경우, 그 추정이 매우 어려운 단점을 가지고 있다.

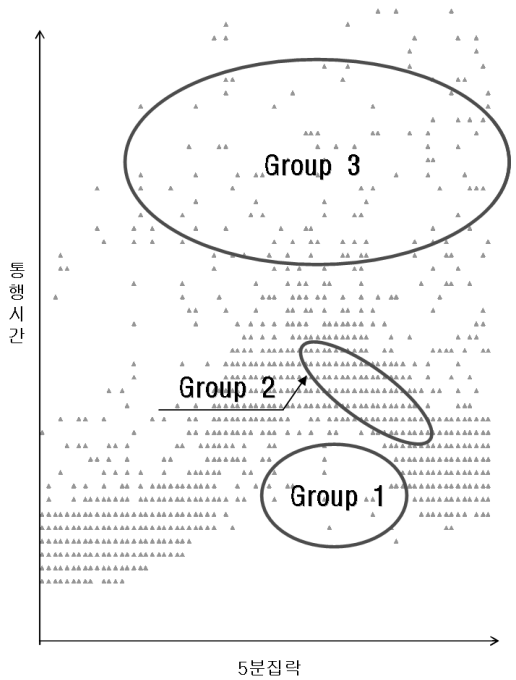
반면 고속도로 통행료수납시스템(TCS: Toll Collection System)에서 수집되는 자료(이하 TCS자료)는 앞에서 언급한 주행차량이 경험한 통행시간을 포함하고 있어 통행시간 정보에 아주 유용하다. 즉 TCS자료는 고속도로 이용차량이 경험한 실제통행시간이며 '구간정보'이다. 일반적으로 차량이 경험한 통행시간은 해당구간을 통과하기까지의 정체상황의 변화가 반영된 결과로서 최적경로 선택을 위한 가장 직접적인 정보가 되며 이용자 입장에서 유용한 교통정보 산출에 적합하다는 특징을 가지고 있다.

그러나 수집된 TCS자료를 실시간 교통정보로 이용하기 위해서는 '시간 처짐(time lag) 현상'을 해결할 수 있는 추정 및 예측 알고리즘이 필요하다. TCS자료는 '시간처짐 현상' 외에도 고속도로 통행시간에 포함될 수 있는 휴게소 체류시간 등의 노상 외 통행시간(고속도로 본선 외의 통행시간), 시스템 오류나 실제 통행이 없는 결측자료가 포함될 수 있다. 따라서 상기와 같은 비정상적으로 통행한 차량의 통행시간(이하 이상치)과 결측자료까지 포함하여 통행시간 정보를 산출할 경우 부정확한 교통정보제공의 우려가 있다.

이상치는 고장 및 사고로 인한 갓길 정차, 과도한 휴게소 방문 및 체류, 과속차량, 갓길 주행 등으로 인하여 다른 차량들에 비해 특히 크거나 작은 통행시간을 보이는 데이터를 의미한다. 일반적으로 이상치는 〈그림 1〉과 같이 3가지 그룹으로 구분할 수 있다.

Group1의 경우는 과속 등으로 인하여 타 차량에 비해 통행시간이 상당히 적은 차량 또는 버스전용차로를 이용한 차량의 통행시간을 나타내며, Group2의 경우는 개별 운전자 통행형태를 포함한 보편적인 차량들의 통행시간(휴게소 체류시간 포함) 특성을 나타낸다.

Group3은 휴게소 등에서 장기 휴식을 취하거나 고장차량 등 노상외 시간이 포함된 통행시간을 경험한 그룹이다.



〈그림 1〉 TCS데이터와 이상치

그리고 TCS 데이터를 통해 획득하는 통행시간 정보는 해당 구간의 실제 통행시간이며, off-line에서 DB로 활용하면 결측치 보정에 활용하거나, 실제로 제공된 통행시간 예측정보의 검정을 위한 기준값으로 활용할 수

있을 뿐만 아니라 요금정산을 위해서도 매우 유용한 정보로 활용할 수 있을 것이다. 그러나 고속도로를 이용하는 이용자에게 제공되어야 할 교통정보의 기준값으로 사용해야 할 경우 고장차량이나 휴게소에서의 장기간 휴식시간이 포함된 TCS 데이터를 통행시간 추정이나 예측을 위한 정보제공의 기초자료로 활용하기에는 무리가 있다고 판단된다.

따라서, 본 연구에서는 TCS자료를 이용하여 보다 정확하고 신뢰성있는 교통정보 제공을 위한 대푯값 자료 확보를 위한 이상치제거와 결측보정 알고리즘의 개발방안을 제시하는 것을 목적으로 한다.

Ⅱ. 기존문헌고찰

먼저 국내·외 기존의 이상치 제거기법 및 필터링 알고리즘을 고찰하고자 한다. 통행시간 정보를 제공하기 위해서는 관측되는 자료의 정도(精度)가 매우 중요하다. 이는 모든 교통정보가 관측 데이터로부터 가공되거나 알고리즘을 이용하여 처리 된 후 제공되기 때문이다.

이상치제거 연구는 단속류인 일반국도의 경우 최윤혁(2003)은 통계적 방법은 문제가 있다고 판단하고 택시에 GPS수신기 장착을 통하여 교통정보를 생성할 경우 주행과 관계없는 정보를 실시간으로 검지하여 제거하는 휴리스틱한 이상치 제거 알고리즘을 개발하여 가능성을 제시하였고, 이지연 외(2003)는 국도 3호선을 대상으로 검지기로부터 수집한 교통량 자료의 보정방안을 제시하였다. 또한, 장진환 외(2005)는 AVI자료를 이용하여 국내·외 기존의 이상치 제거기법 및 필터링 알고리즘을 고찰하고 그 중 단속류 일반국도 패턴에 적합한 부분만을 추출하여 새로운 이상치제거 알고리즘을 개발 제시하였다.

국도 3호선을 대상으로 한 연구에서 도명식 외(2004)는 대부분의 연구에서 기존의 이상치 제거방법으로 통계적인 방법을 이용하고 있으며, Box-Plot법, Shapiro-Wilk 검정법 등 통계적인 방법은 편향되거나 중심에서 많이 벗어난 자료를 이상치로 취급하는 과정에서 분포의 정규성을 임의로 가정하지만, 구간소요시간 등의 교통 데이터는 정규분포 대신 우 편향된 분포를 가지는 등, 기존의 통계적인 방법을 그대로 적용하기에는 많은 문제가 있다고 지적하였다.

연속류인 고속도로를 대상으로 한 연구로는 남궁성외(2000)은 자료의 분포를 파악하기 위하여 표준편차 대신에 자주 고려되는 통계량인 중위절대편차를 이용하여 이상치를 제거한 후 통행시간을 산출하였고, 오세창 외(2003)은 차량검지기 자료를 이용하여 고속도로 구간의 차량통과속도가 70km/h이상일 때는 기존 차량검지기 속도데이터를 이용한 통행시간 산출방식을 적용하고 혼잡시에는 교통량을 이용한 추정모형에 의한 통행시간 산출방식을 병용하여 적용하는 통행시간 산출 알고리즘을 적용하였다.

한편, Tanaka 외(1992)은 AVI 데이터를 지수평활방법으로 이상치제거를 하였고, Guo 외(2007)는 야간에는 교통량이 주간보다 적고 대부분의 차량이 낮 시간대 보다 빨리 달리지 않아, 통행시간은 최소통행시간보다 길다는 것을 확인하고 야간에는 야간 특성에 맞는 이상치 제거방법을 제안하였다. 또한 Southwest 연구소(이하, SWRI)에서는 이동평균알고리즘을 기반으로 하는 TransGuide알고리즘을 개발하였고, 뉴욕과 뉴저지에서는 평활화 기법을 이용하는 Transmit 알고리즘을 개발하여 운영 중이다.

결측데이터의 처리에 대한 연구 분야는 통계적으로 결측을 해결하는 방법에 대한 연구가 대부분이며(Little and Rubin, 2002; Barnett and Lewis, 1994), 교통분야에서의 결측에 대한 연구를 살펴보면 장진환 외(2004)는 누락되는 교통량 자료에 대해 전·후기간 평균, 회귀모형, EM, 시계열 모형들을 활용한 대체기법들을 적용, 평가하였고, Chen 외(2001)은 신경망과 ARIMA 모델을 이용하여 결측데이터의 퍼센트를 비교한 결과 신경망이 결측데이터에 대해 덜 민감하다는 결론을 얻었고, Whitlock 외(2000)는 복잡한 분기점에서 교통흐름을 예측하는데 결측데이터가 존재하는 것을 확인하고 Markov chain Monte Carlo방법으로 결측부분을 예측하였다.

그러나, 연속류를 대상으로 한 대부분의 연구에서 AVI검지기에서 획득한 교통류 데이터의 이상치와 결측에 대한 보정에 대한 연구가 이루어졌으나, 실제 통행시간의 결과인 TCS 자료를 통행시간의 분포나 정보의 제공을 위한 기초자료로 사용하기 위한 이상치나 결측치 보정에 대한 연구는 상대적으로 없는 실정이다.

Ⅲ. 기존 이상치제거 알고리즘 고찰

본 절에서는 연속류를 대상으로 한 이상치 제거 알고

리즘의 기존 연구를 고찰해 보고, 본 연구에서 대상으로 하는 구간의 평일, 휴일 및 특송기간 데이터를 이용하여 문제점을 분석해 보기로 한다.

먼저 시간적, 공간적 범위는 거리별 분산을 알아보기 위한 범위와 개발알고리즘 적용 범위로 나누워 경부고속도로를 중심으로 데이터를 수집, 분석하였다(〈표 1〉, 〈표 2〉 참조).

〈표 1〉 자료 분석 및 검증을 위한 대상기간 및 대상구간

시간적	공간적
8/14, 9/4~9/10, 9/25	서울~수원, 서울~천안, 서울~대전, 서울~대구, 서울~부산

〈표 2〉 개발알고리즘 적용 대상기간 및 대상구간

시간적	공간적
9/25(특송)	서울~대전(상행)

한편, 기존 알고리즘과의 비교 고찰을 위해 5분 단위로 집계한 데이터를 이용하였으며, 이상치 제거 및 결측보정에 대한 프로세서 가운데 문제점 위주로 언급하고자 한다.

1) 중위절대편차를 이용한 이상치제거 알고리즘

MAD는 자료의 분포를 파악하기 위하여 표준편차 대신에 Robust 추정에서 종종 고려되는 방법으로 중위절대편차(Median Absolute Deviation)를 이용하는 방법은 자료의 분포를 가정하지 않아도 되는 이상치제거 방법으로 식(1)과 같다.

$$MAD = 1.4826 \times median|x_i - x_{med}| \quad (1)$$

$$Z_i^{MAD} = \frac{|x_i - x_{med}|}{MAD}$$

여기서, MAD : 중위절대편차

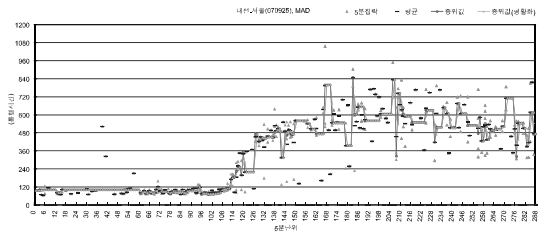
Z_i^{MAD} : MAD에 의한 표준점수(Z score)

x_i : i번째 데이터

x_{med} : x데이터 계열의 중위값

1.4826 : MAD를 정규분포의 표준편차와 동일하게 만드는 조정계수

중위절대편차를 이용한 이상치제거 알고리즘의 경우, 특송기간인 9월 25일 대전-서울 5분 단위 데이터를 대상으로 이상치를 제거한 결과 〈그림 2〉와 같이 몇몇 통행시간이 짧은 데이터로 인해 대표값이 과소추정이 되는 결과를 보였으며, 구간 거리가 길어져 분산이 큰 경우 더욱 왜곡된 대표값을 도출 하는 문제가 발생하였다.



〈그림 2〉 특송기간(대전-서울, 9월 25일)

2) 신뢰구간 추출법

이 방법은 강진기 외(2002)가 제시한 방법으로 상한값과 하한값을 초과한 값을 제거한 후 신뢰도 95%(신뢰구간 68%)의 범위를 초과하는 값을 제거하는 방법으로 상한값은 「구간의 설계속도의 2배를 초과하는 구간통행시간, 하한값은 「해당구간을 10km/h 통행할 때 통행시간」으로 정하였다. 단, 이러한 하한값을 보이는 구간 통행시간 값들의 갯수가 전체 구간 통행시간 값들의 50% 이상을 초과할 경우에는 포함시키는 반면, 신뢰구간 68%를 초과한 값은 이상치로 간주하여 제거하는 방법이다.

신뢰구간 68% 범위의 자료를 추출하는 기본 식은 (2),(3) 및 (4)와 같다.

$$\overline{T} = \frac{\sum T_i}{n} \quad (2)$$

$$\sigma = \sqrt{\frac{\sum T_i^2 - n\overline{T^2}}{n-1}} \quad (3)$$

$$T_{is} = |T_i - \overline{T}| \le \sigma \quad (4)$$

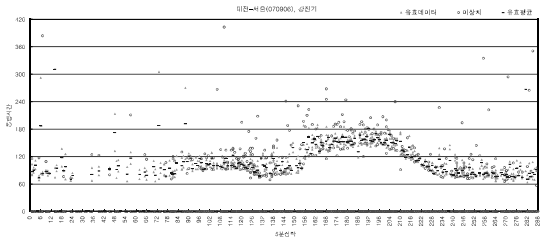
여기서, $\overline{T}$: 구간교통정보수집장치로부터 수집된 개별차량들의 구간통행시간 산술평균

T_i : 구간교통정보수집장치로부터 수집된 개별차량들의 구간통행시간

σ : 구간교통정보수집장치로부터 수집된 개별차량들의 구간통행시간 표준편차

n : 구간교통정보수집장치로부터 수집된 개별차량 중 상한값/하한값 제외한 차량수

T_{is} : 구간교통정보수집장치로부터 수집된 개별차량 중 유효한 차량



〈그림 3〉 평일(대전-서울, 9월 6일)

이 알고리즘을 대전-서울 구간을 대상으로 평일인 9월 6일 TCS 데이터를 대상으로 분석한 결과, 샘플의 정규성 가정의 문제점과 상대적으로 큰 폭의 범위를 유효데이터로 처리하는 문제점이 있었으며, 평균값이 과대산정되는 단점이 있었다(〈그림 3〉 참조).

3) TransGuide 알고리즘

TransGuide는 미국의 샌안토니오 고속도로교통관리시스템에서 운영 중인 알고리즘(Dion and Rakha, 2003)으로 AVI 자료를 이용한 통행시간 추정은 해당 수집주기 내에서 사용자가 정한 범위를 초과하는 통행시간 값들을 자동적으로 제거하는 이동평균 알고리즘을 도입하고 있다. Southwest 연구소(SwRI)에서 개발한 알고리즘으로, (식 6)은 (식 5)에 의해 생성된 유효 통행시간의 평균을 구하는 방법이다.

$$Stt_{ABi} = \begin{Bmatrix} t_{Bi} - t_{Ai} | t - t_w < t_{Bi} < t \text{ and } tt'_{ABi}(1-l_{th}) \\ \le t_{Bi} - t_{Ai} \le tt'_{ABi}(1+l_{th}) \end{Bmatrix} \quad (5)$$

$$tt_{ABi} = \frac{\sum_{i=1}^{|Stt_{AB}|}(t_{Bi} - t_{Ai})}{|Stt_{ABi}|} \quad (6)$$

여기서, Stt_{ABi} : 시간 t에서 AB구간에서 관측된 유효차량 수

t_{Ai} : A 지점에서 관측된 차량 관측시각

t_{Bi} : B 지점에서 관측된 차량 관측시각

t : 수집주기

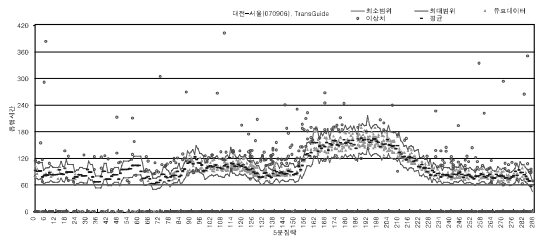
t_w : 이동평균 창(rolling average window)

l_{th} : 링크 초기 통행시간 파라미터(0.2)

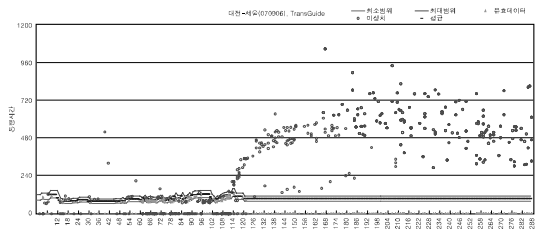
tt_{ABi} : 시간 t에서 AB구간에서 관측된 유효차량의 평균통행시간

tt'_{ABi} : $t-1$시간에서 AB구간에서 관측된 유효차량의 평균통행시간

TransGuide 알고리즘에서 주요한 파라미터는 t_w와 l_{th}이며, t_w는 현재 평균통행시간을 추정할 때 고려되어야 할 수집주기를 의미하고, l_{th}는 이전 수집주기와 현재 수집주기간의 유효통행시간 차이를 의미하는데, TransGuide 알고리즘에서는 t_w를 2분, l_{th}는 0.2로 설정하고 있다.



〈그림 4〉 평일(대전-서울, 9월 6일)



〈그림 5〉 특송기간(대전-서울, 9월 25일)

TransGuide 알고리즘은 각 현재 수집주기의 정상치를 판정하기 위해 바로 이전 수집주기의 평균통행시간만을 이용해 최대·소 값으로 유효 범위를 정한 후, 유효범위에 속하는 유효 데이터의 평균을 구하는 방법으로 평일인 9월 6일 대전-서울 구간과 특송기간인 9월 25일 대전-서울 구간을 대상으로 분석한 결과 이전 주기에서 수집된 유효 통행시간의 데이터 수와 값이 다음 주기의 데이터에 영향을 미치게 되는 특성을 가져 데이터가 유

효범위 안에서 증감할 경우 이상치제거가 가능하지만 〈그림 5〉의 경우처럼 통행시간이 갑자기 유효범위를 넘어 증가하게 되는 경우 정체로 인해 통행시간이 증가한 데이터를 이상치로 모두 제거해 버리는 문제가 발생하는 것을 알 수 있다. 또한 5분 집락을 한 데이터의 수가 적은 경우 이상치(outlier)의 영향을 크게 받아 대푯값이 왜곡될 가능성이 큰 것이 단점으로 나타났다.

4) Transmit 알고리즘

Transmit 알고리즘(Mouskos 외, 1998)은 뉴욕과 뉴저지에서 운영 중인 방법으로 15분 수집주기 동안의 자료의 평활화 기법을 이용한다.

이 방법은 해당 수집주기의 평균통행시간을 식(7)과 같이 평균한 후, 갱신된 평균통행시간을 구하기 위해 과거의 동(同)요일·동(同)시간대 자료를 이용해 평활화하게 되며, 이렇게 구해진 갱신된 평균통행시간과 이전 수집주기 동안 평활화 된 통행시간 값을 이용해 현재 수집주기의 평활화 된 평균통행시간을 식(8)과 같이 구하는 방법이다.

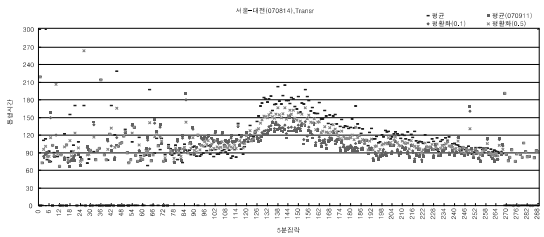
$$tt_{ABk} = \frac{\sum_{i=1}^{n_k}(t_{Bi} - t_{Ai})}{n_k} \quad (7)$$

여기서, tt_{ABk} : k시간대의 AB구간에서 관측된 차량의 평균통행시간
t_{Ai} : A지점에서 검지시각
t_{Bi} : B 지점에서 검지시각
n_k : k시간대에서 관측된 차량의 평균통행시간

$$tth''_{ABk} = (\alpha) \times tth_{ABk} + (1-\alpha) tth''_{ABk-1} \quad (8)$$

여기서, tth_{ABk} : 평활화된 시간대의 평균 통행시간
tth''_{ABk} : 갱신된 시간대의 평활화된 평균통행시간
α : 평활화 계수(0.1)

그러나 이 방법은 평활화 계수의 산정근거가 매우 모호하고 교통류 특성에 따라 상이한 값을 부여하는 별도의 과정을 거쳐야 하는 등 비정상적인 연속류의 특성을 가진 구간에 적용하기에는 어려움이 많은 것으로 판단된다.



〈그림 6〉 8월14일(서울-천안) 구간

평활화 알고리즘인 Transmit 알고리즘을 8월 14일 서울-천안 구간의 데이터를 대상으로 분석한 결과 〈그림 6〉과 같이 정상치와 차이가 큰 이상치가 많이 포함된 분석구간에서는 교통특성을 반영하는 대푯값의 산출에 한계가 있음을 확인하였다.

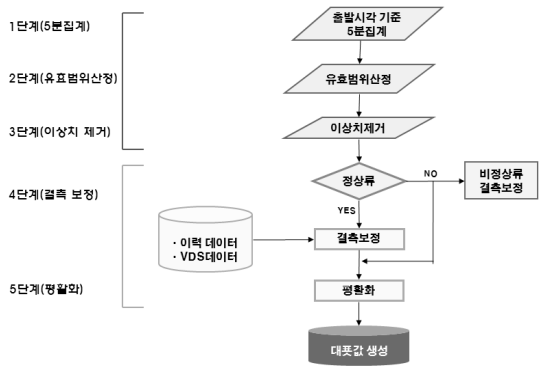
Ⅳ. 알고리즘 개발

앞 절에서 살펴본 바와 같이 기존 이상치제거 기법은 각 알고리즘별로 장단점을 가지고 있으나 평일, 휴일 및 특송기간의 TCS 데이터를 대상으로 적용해 본 결과 많은 문제점이 도출되었다.

따라서 본 연구에서는 기존 알고리즘이 가지는 단점을 보완하면서 명확한 근거없이 정해진 평활화 계수 등의 산정이 교통류의 특성(평균 및 분산 등)에 맞게 변동하는 새로운 알고리즘을 제시하고자 한다.

1. 개발 알고리즘 과정

1단계 : 원시 데이터를 5분 단위로 수집(집계)하여 각 수집주기의 제1사분위수를 산정 후 5분 단위 집계(단, 교통량이 3대 미만인 경우 이전주기 제1사분위수 적용). 1사분위수를 기준으로 선정한 이유는 TCS 데이터가 이상치가 제거된 상태에서 대푯값을 추출하며, 이 과정에서 이상치를 제거하고도 남아있는 자료들이 운전특성 및 휴게소 이용 등으로 인해 대푯값의 왜곡으로 나타날 수가 있기 때문이며, 동일한 교통조건에서 빨리 주행하기보다는 과도한 통행시간을 나타내는 자료들이 상대적으로 많기 때문이다(도명식 외, 2004).



〈그림 7〉 제안알고리즘 과정

2단계 : 5분집계 기준값(1사분위값)을 가지고 최대 및 최소 유효범위를 산정, 즉, 이전 주기의 평균통행시간에 통행시간 파라미터를 주어 최소·최대 통행시간의 유효범위 산정. 파라미터는 거리별로 분산의 차이가 있기 때문에 일정한 파라미터를 적용하는 것이 아니라 거리에 따라 파라미터 값을 적용(〈표 3〉 참조).

$$(tt_{ABmin})^k = (tts_{AB})^{k-1}(1-\beta) \quad (9)$$

$$(tt_{ABmax})^k = (tts_{AB})^{k-1}(1+\beta) \quad (10)$$

여기서, $(tt_{ABmin})^k$: k번째 수집주기에서 AB구간의 최소 통행시간

$(tt_{ABmax})^k$: k번째 수집주기에서 AB구간의 최대 통행시간

$(tts_{AB})^{k-1}$: $k-1$번째 수집주기에서 AB구간의 유효 통행시간

β : 통행시간 파라미터(〈표 3〉 참조)

〈표 3〉에 나타난 구간(거리) 별 평균통행시간과 거리에 따른 표준편차의 차이를 고려한 파라미터는 정상교통

〈표 3〉 거리별 통행시간 파라미터 값의 범위

거리	평균통행시간(분)	표준편차(분)	파라미터
서울—수원	10	3	0.30
서울—천안	40	10	0.25
서울—대전	85	15	0.18
서울—북대구	180	20	0.12
서울—부산	300	30	0.10

류 상태를 기반으로 산정되어 유효범위를 정하는 자료이며, 교통상황별로 상이한 패턴자료를 구축한다면 상황별 유효범위를 달리 적용할 수 있을 것으로 기대된다. 단, 본 연구에서는 정상교통류 상태인 9월 4일에서 10일(8일과 9일 휴일자료 제외) 까지 평일 24시간 TCS 5분 집계데이터를 이용하여 얻은 결과이다.

3단계 : 최소·최대 통행시간의 유효범위 안에 들어가지 않는 이상치를 제거한 후 유효 통행시간의 평균 산정.

$$att_{AB}^k = \frac{\sum_{i=1}^{n} tt_{iAB}^k}{n_k} \quad (11)$$

if $att_{AB}^{k-1}(1-\beta) \le att_{AB}^k \le att_{AB}^{k-1}(1+\beta)$

then att_{AB}^k, else qt_{AB}^k

여기서, att_{AB}^k : k번째 수집주기에서의 AB구간의 유효통행시간 평균

tt_{iAB}^k : k번째 수집주기에서 i번째 차량의 AB구간의 유효통행시간

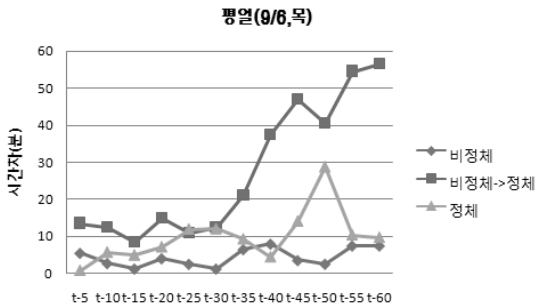
qt_{AB}^k : k번째 수집주기에서 AB구간의 제1사분 위 통행시간

n_k : k번째 수집주기에서 유효통행시간 차량수

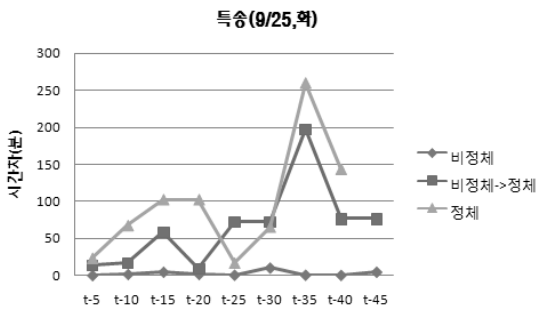
β : 파라미터(=0.18)

4단계 : 결측 보정단계로 5분 집계한 시간대에 데이터가 존재하는 경우 평균값을 적용하여, 데이터가 없는(결측) 경우에는 먼저, 1) 현재시각 기준으로 이전 30분 이내 이력데이터가 있는 경우에는 이력데이터를 사용하며, 2) 30분 이내 이력데이터가 없는 경우에는 VDS 데이터를 활용하여 통행시간을 계산하여 사용.

여기서, 결측이 발생한 경우 과거 이력데이터의 활용범위는 대상 구간의 길이, 차로 수, 유출입 및 휴게소 유무 등에 따라 차이가 나기 때문에 개별 구간별로 산정함이 바람직하지만 TCS 자료의 과거 패턴을 특성(평일, 휴일, 특송, 계절 등)별로 그룹화하여 이용할 것을 추천한다.



〈그림 8〉 평일 과거이력자료의 결측치 보정



〈그림 9〉 특송기간 과거이력자료의 결측치 보정

〈그림 8〉~〈그림 9〉는 본 연구에서 대상으로 하는 대전-서울간의 평일과 특송기간의 교통특성별(정체, 비정체→정체로 전이, 비정체)로 구분하여 현재(t=0)를 기준으로 과거자료와의 시간차를 살펴본 결과이며, 교통특성에 따라 다소 차이가 있어 확정적으로 판단하기는 어려우나 30분 이상의 과거자료를 결측보정을 위해 이용하기에는 무리가 있다고 판단된다.

5단계 : 이전 주기의 유효통행시간평균을 이용한 평활화 과정.

$$tth''_{AB_{k+1}} = \gamma_1 tth_{AB_{k-3}} + \gamma_2 tth_{AB_{k-2}} + \gamma_3 tth_{AB_{k-1}} + \gamma_4 tth_{AB_k} \quad (12)$$

여기서, $tth''_{AB_{k+1}}$: k번째 수집주기에서 AB구간의 평활화된 통행시간

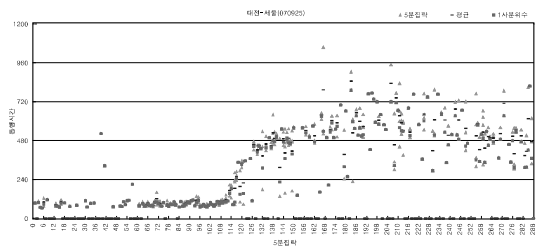
$tth''_{AB_{k-n}}$: $k-n$번째에서 AB구간의 평균유효통행시간

$\gamma_1 \sim \gamma_4$: 지수평활 계수

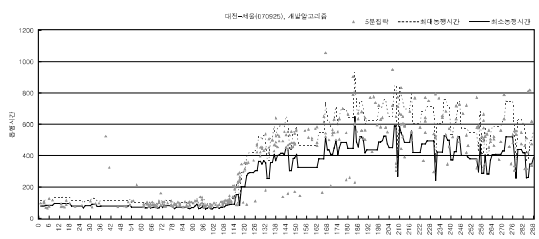
평활화는 검지기 자료의 수집오류나 노이즈에 의해 발생되는 자료의 일시적 변동을 보정하기 위한 처리과정으로 소통상태 판정과 같은 자료 활용측면에서의 불안정을 제거하고 자료의 규칙적인 연속성을 주어 자료의 신뢰성을 높여주는 처리과정이다. 일반적으로 자료의 수집주기가 짧을수록 각 수집주기마다 수집 자료는 일시적 변동을 포함하고 이로 인하여 자료 가공 시 불안정을 야기할 수 있다. 따라서 이전주기의 평활화 된 값이 아닌 이전주기의 유효통행시간평균값을 이용해 평활화 하는 방법을 제안한다.

2. 제안 알고리즘의 적용가능성 평가

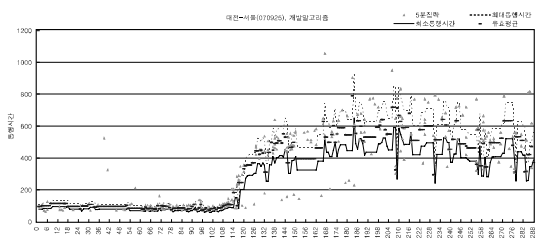
개발된 알고리즘의 적용가능성을 검토하기 위해 TCS 데이터의 분산이 큰 대전-서울 구간의 9월25일 자료를 대상으로 단계별로 알고리즘 적응 과정을 〈그림 10〉~〈그림 13〉에 나타내었다.



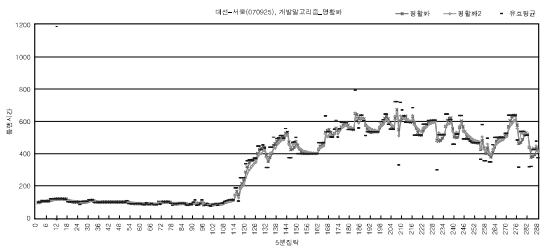
〈그림 10〉 데이터 특성



〈그림 11〉 1~2단계 과정 후 결과



〈그림 12〉 전처리 3~4단계 과정 후 결과



〈그림 13〉 전처리 5단계 과정 후 결과

(1) MAPE(Mean Absolute Percent Error)

기준값과 관측값의 차이를 기준값에 대한 비율로 나타내는 평가지표

$$MAPE = \frac{1}{N}\sum\frac{|\hat{Y}_i - Y_i|}{Y_i} \times 100$$

여기서, $\hat{Y}_i$:관측값, Y_i:기준값, N:표본수

(2) RMSE(Root Mean Square Error)

기준값과 관측값의 차이를 나타내어주는 평가지표

$$RMSE = \frac{\sqrt{\sum(\hat{Y}_i - Y_i)^2}}{N}$$

여기서, $\hat{Y}_i$:관측값, Y_i:기준값, N:표본수

(3) 변동계수

성질이 다른 집단간의 분산을 비교하는 방법으로 자료값의 크기에 따른 상대적인 비교를 위하여 편차를 집단의 평균으로 나눈 값

$$\frac{s}{\bar{x}}$$

여기서, s : 표준편차, $\bar{x}$: 평균

(4) 사분위수범위(IQR)

상위 25%에 해당하는 관측값과 하위 25%에 해당하는 관측값을 제외하고 범위를 구한 값

$$IQR = Q_{0.75} - Q_{0.25}$$

여기서, $Q_{0.75}$: 관측값을 크기순으로 나열했을 때 전체의 75%보다 큰 값 중 제일 작은 값

$Q_{0.25}$: 관측값을 크기순으로 나열했을 때 전체의 25%보다 큰 값 중 제일 작은 값

〈표 4〉 알고리즘 평가 결과

평가지표 \ 알고리즘	MAD	Trans Guide	신뢰구간	제안
MAPE	(기준)	312.13	21.69	12.97
	11.99	290.37	19.19	(기준)
RMSE	(기준)	353.49	112.64	74.17
	74.17	329.36	104.78	(기준)
변동계수	0.64	NA	0.64	0.63
사분위범위수	449.5	NA	445.75	427.5

한편, 본 연구에서 제시한 알고리즘의 객관적인 검증은 각 시간대의 대푯값을 이용하여 현재 출발시점에 있는 운전자에게 제공될 통행시간의 추정 및 예측알고리즘을 구현한 후 가능하며 이에 대한 다양한 연구도 저자들이 진행 중이며(Namkoong et al., 2008), 본 연구는 통행시간 추정 및 예측을 위한 각 시간대별 대푯값의 산출에 그 목적을 두었음을 밝힌다.

따라서 기존 방법과 본 연구에서 제안한 방법과의 성능비교를 위해 검증을 하고자 한다. 단, TCS 데이터는 OD간 통행시간의 분포가 대푯값과 차이가 있다하더라도 휴게소 이용시간 및 운전자 개인의 운전특성 등이 포함된 참값이기 때문에 기준으로 삼을 지표가 없음을 감안하여 상대적인 비교를 위해 〈표 4〉와 같이 기존의 MAD를 기준으로 한 경우와 본 연구에서 제안한 방안을 기준으로 한 경우로 분리하여 분석하였다.

평가결과에서 보는 바와 같이 분산이 큰 경우에도 개발 알고리즘을 통해 전처리를 한 결과 개선된 결과를 나타냈다. 다만, 현재 고속도로에서 사용중인 중위절대편차를 이용한 방법과는 데이터 집계 단계에서 중위값 대신 1사분위수를 사용하는 것과 대푯값으로 평균을 채택한 점에서 차이가 있으며 평가결과가 다소 개선되었음을 알 수 있다. 또한, 평활화과정에서 주로 사용하는 평활화된 이전주기값을 이용하는 방법과 제안한 평활화 방법(유효통행시간 평균이용)을 비교한 결과 제안방법이 더 좋은 결과를 나타냄을 알 수 있었다.

Ⅴ. 결론 및 향후과제

통행시간 등 교통정보의 제공에 대한 이용자들의 관심이 주 5일 근무 실시 및 유류비용의 증가, 시간가치에 대한 관심의 증대 등으로 어느 때보다 커지고 있는 실정이다.

그리고 실시간 통행시간 제공의 정도를 향상시키고 교통네트워크의 효율적인 운영을 위해서는 TCS 데이터를 기반으로 한 기초자료의 확보와 이에 따른 이상치제거 및 결측치 보정에 대한 알고리즘의 개발이 선결과제이다.

따라서 본 연구에서는 TCS자료를 이용하여 정확하고 신뢰성 있는 교통 정보 제공을 위한 DB구축을 위한 Off-line 알고리즘을 개발하였으며, 또한 거리에 따라 분산이 다르다는 점을 고려하여, 이상치제거 유효범위 산정 시 거리에 따라 상이한 파라미터값을 적용하는 방법을 제시하였고, 고속도로의 일부구간(경부선 중심)을 대상으로 평일, 휴일 및 특송기간을 대상으로 다양한 교통류 특성자료를 대상으로 개발된 알고리즘의 적용가능성을 살펴보았다. 분석 결과 기존의 알고리즘의 단점을 보완하면서 다양한 교통류 특성자료에 적합한 알고리즘을 제안하였다고 판단된다.

향후 연구로는 Hi-pass 데이터를 이용한 알고리즘 개발과 Off-line이 아닌 실시간 정보 제공을 위한 On-line 상에서도 적용할 수 있는 이상치 제거 알고리즘의 개발이 필요할 것으로 판단된다. 또한 본 연구에서는 결측보정을 이전주기값을 이용해서 했지만 좀 더 정확한 결측 보정을 위한 방법이 필요할 것으로 판단된다.

참고문헌

1. 강진기 · 손영태 · 윤여환 · 변상철(2002), "비매설식 자동차량인식장치를 이용한 구간교통정보 산출방법 연구", 한국 ITS학회 논문집,vol.1 No.1, pp.22~31.
2. 남궁성 외(2000), "ITS 기술개발연구(Ⅳ)-고속도로 통행시간 예측시스템 개발", 한국도로공사 도로연구소.
3. 도명식 · 김성현 · 배현숙 · 김종식(2004), "국도의 동질구간 선정과 이상치 제거 방법에 관한 연구", 대한교통학회지, 제22권 제7호, 대한교통학회, pp.7~16.
4. 오세창 · 김명하 · 백용현(2003), "차량검지기 교통량 데이터를 이용한 고속도로 통행시간 추정 및 예측모형 개발에 관한 연구", 대한교통학회지, 제21권 제5호, 대한교통학회, pp.83~95.
5. 원태연 · 정성원(2004), "통계조사분석", SPSS 아카데미.
6. 이지연 · 도명식 · 김성현 · 류승기(2003), "교통량 데이터의 실시간 보정로직-국도 3호선을 중심으로", 응용통계연구, 제16권, 제2호, pp.203~215.
7. 장진환 · 변상철 · 백남철 · 김성현(2005), "AVI 자료 필터링 알고리즘 개발 -일반국도를 중심으로", 대한토목학회논문집, 제25권, 제2D호, pp.1~8.
8. 최윤혁(2003), "택시GPS Probe 자료의 실시간 이상치 제거 알고리즘 개발", 아주대학교 석사학위논문.
9. Barnett, V. and Lewis, T.(1994) "Outliers in statistical data", John Wiley & Sons.
10. Dion, F., and Rakha, H.(2003) "Estimation Spacial Travel Time using Automatic Vehicle Identification Data", TRB.
11. Little, R.J.A. and Rubin, D.B.(2002) "Statistical analysis with missing data", 2nd E. Wiley Interscience.
12. Mouskos K.C. et al.(1998) "TRANSMIT System Evaluation. Final Report, Institute for Transportation, New Jersey Institute of Technology, N.J.
13. Namkoong S., Park, E., Oh, C., Do, M. and Lee, H. (2008) "A method to estimate path-travel time on expressway using toll collection system data", 2008 ITS World Congress (in review).
14. Tanaka Y, Kanayama K, Sugimura H.(1992) "Travel-time data provision system using vehicle license number recognition devices". In:Proc. of the Intelligent Vehicles 92 Symposium. Detroit, USA.
15. Whitlock M.E. and Queen C.M.(2000), "Modelling a Traffic Network with Missing Data", Journal of Forecasting, pp.561~574.
16. Yanmei Guo, Ling Qin, Tao Kong, Changqing Zheng, Haihui Shan(2007), "Extracting Travel Time Information from Automated Vehicle Identification Detectors Data", 14th World Congress on ITS, Beijing.

♧ 주 작 성 자 : 도명식
♧ 교 신 저 자 : 도명식
♧ 논문투고일 : 2008. 2. 23
♧ 논문심사일 : 2008. 5. 15 (1차)
2008. 7. 14 (2차)
2008. 7. 28 (3차)
♧ 심사판정일 : 2008. 7. 28
♧ 반론접수기한 : 2008. 12. 31
♧ 3인 익명 심사필
♧ 1인 abstract 교정필