한국ITS학회논문지
제8권, 제5호 (2009년 10월)
pp. 1~11

통행시간 추정을 위한 TCS 데이터의 전처리 모형 개발

A Development of Preprocessing Models of Toll Collection System Data for Travel Time Estimation

이 현 석* (Hyun-Seok Lee)　　남 궁 성** (Seong J. Namkoong)

요 약

TCS (Toll Collection System) 데이터는 원시 데이터 자체로서도 구간의 교통상황을 어느 정도 반영할 수 있는 교통특성을 내포하고 있다. 그러나 TCS 데이터에는 이상치가 포함되어 있어 이러한 데이터는 해당 구간의 통행시간을 대표한다고 볼 수 없으므로 만약 이러한 이상치들이 포함되어 있음에도 불구하고 제거하지 않고 집락을 한다면 이상치들로 인해 통행시간은 크게 왜곡 될 가능성이 있다.

특히 장거리 구간일수록 통행시간의 분산이 증가하여 동일구간 동일시간대라도 다양한 통행시간이 분포하고 있다. 구간이 길어질수록 통행시간의 변동이 심하여 적절한 통행시간 대푯값을 구하기가 어렵다. 따라서 TCS 자료를 이용하여 통행시간의 대푯값을 산정하기 위해서는 통행시간의 변동 특성을 파악하는 것이 중요하다. 본 연구에서는 TCS 데이터의 전처리 기법을 개선하되 구간의 길이와 교통상황에 따른 통행시간의 변동을 고려하여 TCS 원시데이터로부터 시·공간적 통행패턴을 파악할 수 있는 의미 있는 통행시간을 추출하고자 한다.

Abstract

TCS Data imply characteristics of traffic conditions. However, there are outliers in TCS data, which can not represent the travel time of the pertinent section, if these outliers are not eliminated, travel time may be distorted owing to these outliers. Various travel time can be distributed under the same section and time because the variation of the travel time is increase as the section distance is increase, which make difficult to calculate the representative of travel time. Accordingly, it is important to grasp travel time characteristics in order to compute the representative of travel time using TCS Data.

In this study, after analyzing the variation ratio of the travel time according to the link distance and the level of congestion, the outlier elimination model and the smoothing model for TCS data were proposed. The results show that the proposed model can be utilized for estimating a reliable travel time for a long-distance path in which there are a variation of travel times from the same departure time, the intervals are large and the change in the representative travel time is irregular for a short period.

Key words: Link travel time, outlier removal, smoothing, spatial detection, toll collection system

* 주저자 : 한국도로공사 도로교통연구원 교통연구실 선임연구원
** 공저자 : 한국도로공사 도로교통연구원 교통연구실 수석연구원
† 논문접수일 : 2009년 7월 14일
† 논문심사일 : 2009년 10월 8일
† 게재확정일 : 2009년 10월 12일

Ⅰ. 서 론

1. 연구의 배경 및 목적

최근에 국도를 중심으로 DSRC, AVI, GPS 등의 구간정보를 활용한 통행시간 연구가 활발히 진행되고 있으나, 국내 고속도로의 경우 아직까지 기·종점 간의 통행시간 산정시 VDS에 의한 지점정보를 이용하고 있다. 지점검지기에 의한 통행시간 산출시 몇 가지 문제점이 있는데, 우선 검지기 속도의 시·공간적 집계에 있어서 교통량 및 거리에 대한 영향을 반영하여 조화평균을 구하고 있으나 검지기의 설치간격이 조밀하지 않은 구간에서는 혼잡상태 및 교통류 전이상태에서 정상상태와는 달리 구간속도를 제대로 산출하지 못하고 있다 [1]. 또한 차량이 해당구간을 이동하면서 경험하게 될 통행시간은 현재 시각의 검지기별 통행속도가 아닌 차량이 각 검지기를 통과할 때의 통행시간이나, 현재는 고속도로 전구간의 VDS를 대상으로 현 시점에서의 지점속도를 동시에 수집하여 계산한 순간 통행시간으로서 경로상의 통행시간이라기 보다는 단순히 현 시점에서의 도로상황을 나타내는 값이라 할 수 있다.

따라서 교통상황의 변화를 반영할 수 있는 통행시간을 산출하기 위해서는 구간정보를 이용해야 하는데 현 상황에서 고속도로 통행료 수납시스템 (TCS : Toll Collection System) 자료는 추가적인 설치비용 없이 수집할 수 있는 가장 신뢰성 있는 구간정보로서 통행시간 산정시 운전자가 경로 내에서 실제로 경험한 교통상황을 반영할 수 가 있다. 한국도로공사의 전국 262개소의 영업소에서 수집하는 TCS 자료는 일평균 약 320만 건이며 고속도로 이용차량의 출발/도착 영업소간 통행시간을 정보를 수집할 수 있는 교통자료이다. TCS 자료는 개별 운전자가 실제 주행하면서 기·종점 간에 경험한 교통상황 등 동적특성이 반영된 참값이라는 점에서 통행시간 산정시 매우 유용한 자료임에도 그 동안 통행시간 정보의 산출에는 이용되지 못하고 영업 정산용으로만 활용되어 왔다.

신뢰할 수 있는 교통정보를 제공하기 위해서는 관측되는 데이터의 정도뿐 아니라 관측 데이터로부터 의미 있는 교통정보를 추출하기 위한 전처리 과정이 매우 중요하다. 이는 모든 교통정보가 관측 데이터로부터 가공되거나 알고리즘을 이용하여 처리된 후 제공되기 때문이다. 본 연구에서는 TCS 원시데이터를 적절히 가공·정제하여 의미 있는 교통정보를 추출하기 위한 전처리 기법을 제시하고자 한다.

2. 연구의 내용 및 방법

TCS 데이터는 원시 데이터 자체로서도 구간의 교통상황을 어느 정도 반영할 수 있는 교통특성을 내포하고 있다. 그러나 TCS 데이터에는 휴게소에서 장시간 체류했거나 차량고장에 의한 길어깨 정차 등의 이유로 동일한 단위시각에서 출발했어도 다른 차량보다 월등히 늦게 도착한 차량들의 통행시간과 과속, 빈번한 차로변경, 길어깨 주행 및 버스 전용차로 주행 등으로 다른 차량에 비해 월등히 빨리 도착한 차량들의 통행시간이 포함되어 있다. 이러한 데이터는 해당 구간의 통행시간을 대표한다고 볼 수 없으므로 만약 이러한 이상치들이 포함되어 있음에도 불구하고 제거하지 않고 집락을 한다면 이상치들로 인해 통행시간은 크게 왜곡 될 가능성이 있다 [2].

특히 장거리 구간일수록 통행시간의 분산이 증가하여 동일구간 동일시간대라도 다양한 통행시간이 분포하고 있다 [3]. 구간이 길어질수록 통행시간의 변동이 심하여 특히 서울~대전 구간의 경우 출발시간대 별로 적절한 통행시간 대푯값을 구하기가 어렵다. 결국 통행시간의 변동이 심하면 TCS 자료를 통해서 구간의 통행패턴을 파악하기가 힘들다. TCS 자료를 이용하여 통행시간의 대푯값을 산정하기 위해서는 통행시간의 변동 특성을 파악하는 것이 중요하다.

동일한 값이더라도 통행시간의 변동은 구간의 길이에 따라 달리 적용되어야 한다. 예를 들면 10분이라는 통행시간의 변동은 서울~기흥 구간에서는 교통상황의 변화를 나타낼 수 있지만 서울~대전 구간에서는 교통상황의 변화를 나타낼 수 없다. 또한, 정체에서 비정체, 비정체에서 정체로 전이되는 시간대에서의 급작스런 통행시간의 변동을 반영하기 위해서는 동일구간이라 하더라도 혼잡상태에 따라 통행시

간의 변동을 달리 적용해야 한다 [4].

이상치가 제거된 집계시간별 대푯값이라 하더라도 하루의 통행시간 패턴을 놓고 볼 때는 비정상적인 값을 나타낼 수 있다. 즉 어떤 구간의 하루 중 통행시간은 그 구간의 정체상황에 의해서 서로 연관을 가지며 그 연관에 의하여 어떤 일정한 패턴을 보이게 되므로 이러한 일정한 패턴을 벗어나는 잡음에 해당하는 대푯값들은 TCS 자료가 보유하고 있는 의미 있는 패턴을 왜곡시키지 않기 위해서 평활화의 과정을 거쳐야 한다.

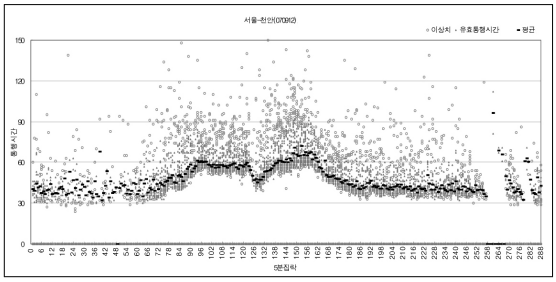
본 연구에서는 TCS 데이터의 전처리 기법을 개선하되 구간의 길이와 교통상황에 따른 통행시간의 변동을 고려하여 TCS 원시데이터로부터 시·공간적 통행패턴을 파악할 수 있는 의미 있는 통행시간을 추출하고자 한다.

Ⅱ. 기존문헌 고찰

기존의 전처리 알고리즘은 대부분 산술평균을 사용하여 통행시간 대푯값을 구하였다. 관측데이터의 표준편차를 구하여 데이터의 상·하한 유효범위를 제한하기는 하지만 데이터의 변동 폭이 큰 경우 정상치에서 크게 벗어나는 몇 개의 이상치가 표준편차를 증가시켜 평균값을 왜곡시킬 수 있다. 또한 대부분의 연구에서 통계적인 방법을 이용하고 있으나 Box-Plot 법, Shapiro-Wilk 검정법 등 통계적인 방법은 편향되거나 중심에서 많이 벗어난 자료를 이상치로 취급하여 버리는 과정에서 분포의 정규성을 임의로 가정하지만 구간 소요시간 등의 교통 데이터는 정규분포 대신 우 편향된 분포를 가지는 등 기존의 통계적인 방법을 그대로 적용하기에는 많은 문제가 있다.

1. 통계적 방법

일반적으로 통계적 방법을 이용하여 이상치를 제거하는 순서는 다음과 같다. 먼저 개별차량의 구간 통행시간을 산정한 후 상한 값과 하한 값을 초과한 값을 제거하는 데, 이때 강진기 외(2002)가 사용한 상한 값은 구간의 설계속도의 2배를 초과하는 구간



<그림 1> 통계적 방법에 의한 이상치 제거
<Fig. 1> Outlier elimination by static method

통행시간이며 하한 값은 해당구간을 10km/h 통행할 때의 통행시간이다 [5]. 단, 이러한 하한 값을 보이는 구간 통행시간 값들의 개수가 전체 구간 통행시간 값들의 50% 이상을 초과할 경우는 신뢰구간의 68%를 초과한 값을 이상치로 간주하여 제거한다. <그림 1>은 이상치 제거 기법을 통해 제거되고 남은 유효데이터와 그 유효데이터의 평균을 나타낸 것이다. 그림에서 보는 바와 같이 상대적으로 큰 폭의 유효데이터가 남게 되어 전체적으로 평균값이 크게 산정되는 단점이 있으며 특히 관측차량대수가 적거나 정체시에 소수의 이상치에 의해서 평균값이 크게 왜곡되는 경향이 있다.

2. TransGuide 알고리즘

TransGuide는 미국의 샌 안토니오에서 운영 중인 고속도로 교통관리시스템으로써 연속적인 AVI 리더기 사이의 링크통행시간은 해당 수집주기 내에서 사용자가 정한 범위를 초과하는 통행시간 값들을 자동적으로 제거하는 이동평균 알고리즘에 의해 추정된다.

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

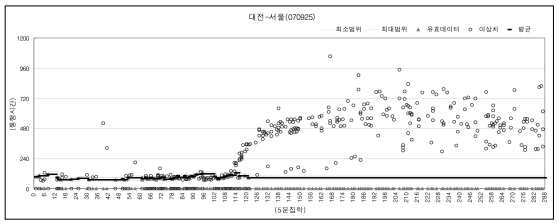
Stt_{ABi} = t동안 AB구간을 주행한 유효차량 수
t_{Ai} = 개별차량 i의 A지점 통과시각
t_{Bi} = 개별차량 i의 B지점 통과시각
t= 통행시간 수집주기

t_w = 이동평균 창
l_{th} = 통행시간 파라미터(0.2)
tt_{ABi} = t시간대의 유효차량의 평균통행시간
$tt_{ABi}^{'}$ = $t-1$시간대의 유효차량의 평균통행시간

TransGuide 알고리즘에서 주요한 파라미터는 t_w와 l_{th}이다. t_w는 현주기의 평균통행시간을 추정할 때 고려되어야 할 수집주기를 의미하며 보통 2분으로 설정하여 현재시간 t와 (t-t_w) 시간대 사이에 B지점에 도착한 차량들만 유효한 차량으로 간주한다. l_{th}는 이전 수집주기와 현재 수집주기간의 유효 통행시간 차이를 의미하는데 0.2로 설정하여 전 단계의 통행시간 추정치와 20% 이상 차이가 나면 이상치로 판단한다.

TransGuide 알고리즘은 현재 수집주기의 정상치를 판정하기 위해 바로 이전 수집주기의 평균통행시간만을 이용해 최대·최소값으로 유효 범위를 정한 후 유효범위에 속하는 유효 데이터의 평균을 구하는 방법으로 이전 주기에서 수집된 유효 통행시간의 데이터 수와 값이 다음 주기의 데이터에 영향을 미치게 된다.

최초 수집한 데이터가 편향된 크기의 값을 가지는 경우에는 참값이 제거되고 이상치가 남게 될 가능성이 있으며 집락을 한 데이터의 개수가 적은 경우 이상치의 영향을 크게 받아 대표값이 왜곡될 가능성이 크다. 또한 특송기간과 같이 통행시간이 급격하게 증가하여 이전 시간대에 비해 20% 이상 차이가 날 경우 이를 유효범위로 인식하지 못하고 이상치로 판단하게 된다. 즉, l_{th}가 일정한 값으로 설정되어 혼잡의 상태나 구간거리에 따른 통행시간의 변동을 고려할 수 가 없다.



<그림 2> TransGuide 알고리즘에 의한 이상치 제거
<Fig. 2> Outlier elimination by transguide algorithm

3. Transmit 알고리즘

Transmit은 뉴욕과 뉴저지에서 운영 중인 교통관리시스템으로써 AVI자료를 이용한 통행시간 추정방법은 기본적으로 TransGuide와 비슷하지만 현재 통행시간을 추정하기 위해 이동평균을 사용하는 대신 15분 수집주기 동안의 자료의 평활화 기법을 이용한다. 즉 이상치를 제거하지 않고 평활화 된 이전주기 자료를 사용한다.

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

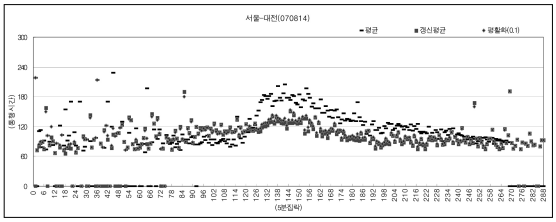
tt_{ABk} = k시간대의 AB구간의 평균통행시간
t_{Ai} = A지점에서 검지시각
t_{Bi} = B지점에서 검지시각
n_k = k시간대에서 관측된 차량의 대수

해당 수집주기의 통행시간을 위와 같이 평균한 후 갱신된 평균통행시간을 구하기 위해 과거의 동요일, 동시간대 자료를 이용해 평활화하고 이전 수집주기 동안 평활화 된 통행시간 값을 이용해 현재 수집주기의 평활화 된 평균통행시간을 구한다.

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

$tth_{AB\,k}$ = k 시간대의 과거 평활화 된 값
$tth''_{AB\,k}$ = k 시간대의 평활화 된 값
$tth''_{AB\,k-1}$ = $k-1$시간대의 평활화 된 값
α = 평활화 계수(0.1)

Transmit 알고리즘의 평활화 계수의 산정근거는 불확실하여 유고(incidents)가 발생하지 않았을 경우에는 갱신된 평균통행시간 값에 10%의 가중치를 부여하고 만약 유고가 발생했을 경우에는 0%의 가중치를 부여한다. 따라서 이러한 형태의 평활화는 유고시 통행시간 자료는 이동평균에 포함되지 않으며 이력 데이터베이스에는 전형적인 반복적 통행시간만을 포함한다. <그림 3>과 같이 정상치와 차이가 큰 이상치가 많이 발생하는 분석구간에서는 단순히 평



<그림 3> Tansmit 알고리즘에 의한 이상치 제거
<Fig. 3> Outlier elimination by transmit algorithm

활화만으로 통행시간을 산출하는데 한계가 있어 비정상적인 연속류의 특성을 가진 구간에 적용하기에는 어렵다.

Ⅲ. 이상치 제거 모형

1. 이상치 제거 기본식

통행시간을 5분 간격으로 집계하여 대표값을 산정하였을 경우 통행시간의 시계열적 변화는 불규칙적이며 동일집계간격 내에서 변동이 큰 이상치를 내포하고 있다. 일반적으로 동일집계간격 내에서의 이상치를 제거하기 위하여 통행시간의 분포를 정규분포로 가정한 후 평균값에서 표준편차의 몇 배수 이상 차이가 나는 관측치($\mu \pm n\sigma$)를 이상치로 간주하는 방법을 사용한다. 그러나 표준편차는 평균으로부터의 거리를 제곱하므로 평균값과의 거리가 큰 이상치가 오히려 전체 관측치의 표준편차에 더 큰 영향을 미칠 수 있다.

집계간격 내 통행시간 대표값 산정시 중위절대편차를 이용한 이상치 제거모형을 적용하면 다음과 같다.

$$MAD = median|x_i - x_{med}|$$

MAD : 중위절대편차
x_i : 집계간격 내에서 관측된 i번째 통행시간
x_{med} : 집계간격 내에서 관측된 통행시간의 중앙값

MAD를 정규분포의 표준편차로 근사화 시키면 다음과 같다.

$$\hat{\sigma} = K \cdot MAD$$

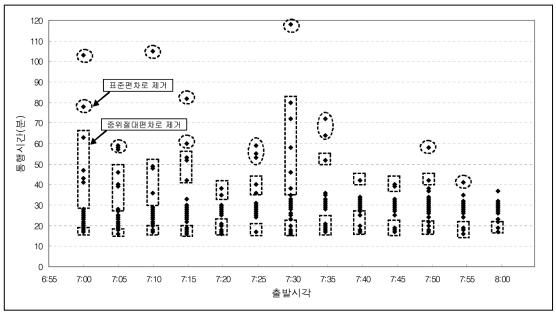
K는 MAD를 정규분포의 표준편차와 동일하게 만드는 조정계수이고, 대칭분포에서 MAD는 1사분위수와 2사분위수 사이의 거리이므로 다음의 확률식이 성립한다.

$$\frac{1}{2} = P(|X-\mu| \le MAD) = P\left(\left|\frac{X-\mu}{\sigma}\right| \le \frac{MAD}{\sigma}\right)$$
$$= P\left(|Z| \le \frac{MAD}{\sigma}\right)$$

$\sigma = 1$인 정규분포에서 $\frac{MAD}{\sigma} = \Phi^{-1}(3/4) \approx 0.6745$ 이므로 $\hat{\sigma} = 1.4826 \cdot MAD$, $z_i^{MAD} = \frac{|x_i - x_{med}|}{1.4826 \cdot MAD}$

결국 개별 통행시간 관측값의 MAD를 표준정규분포를 따르는 확률변수, z_i^{MAD}로 근사화한 후 이를 미리 설정된 제거변수와 비교하여 $z_i^{MAD} > z_{cut}$이면 이상치로 판단한다. 이때 표준정규분포에서 z=3일 때 확률이 99%이므로 일반적으로 z_{cut}=3으로 설정한다.

2009년 2월 2일 서울영업소에서 7:00~8:00에 출발하여 안성에 도착한 모든 차량들에 대하여 출발시각 기준 5분단위로 집계하여 표준편차와 중위절대편차를 이용하여 이상치를 제거하면 <그림 4>와 같다. 표준편차를 이용하는 경우 7:00 출발 차량들 중에서 103분의 통행시간을 갖는 관측치가 전체 관측치의 편차를 왜곡시켜 63분과 78분의 통행시간을 갖는 관측치가 이상치로 제거되지 못해 통행시간의 대표값은 25.2분이 된다. 반면에 중위절대편차를 이용하는 경우 26분 이상과 16분 이하의 통행시간을 갖는 관측치들이 모두 이상치로 제거되어 통행시간의 대표값은 21.0분이 된다. 즉 중위절대편차를 이용할 경우



<그림 4> 이상치 제거 방법 비교
<Fig. 4> Comparison of outlier elimination

동일집계 간격 내에서 군집주행의 행태를 벗어나는 이상치들을 더 효과적으로 제거할 수 있다.

2. 제거변수, z_{cut}의 결정

통행거리가 길어질수록 동일 시간대에 출발한 차량들의 통행시간 변동은 일반적으로 증가한다. 동일 집계간격 내에서의 통행시간 변동을 통행거리에 따라 비교하기 위해서는 통행시간의 편차를 통행시간 대표값으로 나눈 통행시간 변동비를 이용한다.

$$CV_t = \sigma_t / \mu_t$$

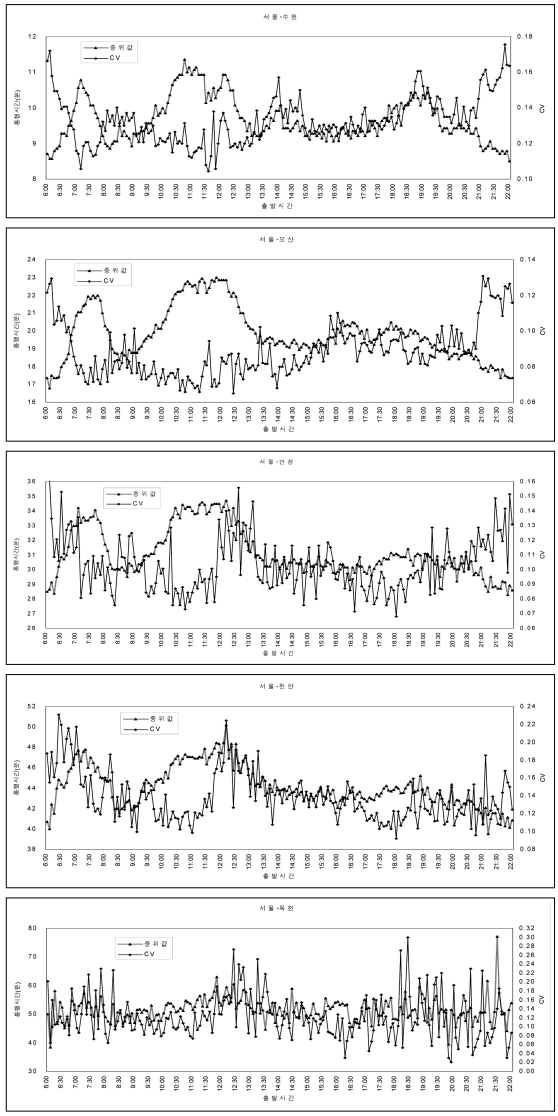
CV_t : t시간대에 출발한 차량들의 통행시간 변동비
μ_t : t시간대에 출발한 차량들의 통행시간 대표값
σ_t : t시간대에 출발한 차량들의 통행시간 편차,

$$= \sqrt{\frac{\sum_{i=1}^{n}\left(T_t^i - \mu_t\right)^2}{n-1}}$$

T_t^i : t시간대에 출발한 i번째 차량의 통행시간
n : t시간대에 출발한 총 차량대수

구간거리에 따라서 혼잡 및 비혼잡시의 통행시간 변동비를 분석하기 위하여 5분단위의 출발시각대별로 평균하여 통행시간 중위값과 변동비(CV)를 나타내면 <그림 5>와 같다. 서울을 기점으로 오전 7시 전후로 1시간의 오전 첨두시간대를 거쳐 다시 10시~12시 사이에 혼잡시간대를 형성하고 있음을 알 수 있다. 통행시간이 증가할수록 통행시간의 변동비는 감소함을 보이며 혼잡시간대에는 변동비가 0.07~0.11, 비혼잡 시간대에는 0.14~0.20의 분포를 보인다.

특히 서울~수원, 서울~오산 구간의 경우 대체적으로 통행시간과 변동비의 음의 관계를 잘 보여주고 있으나 구간 거리가 길어지는 서울~안성, 서울~천안의 경우 통행시간에 대해 변동비가 다소 크게 진동하기 시작하며, 서울~목천의 경우 변동비는 출발시간대나 통행시간과는 상관없이 불규칙적으로 큰 값을 나타낸다. 구간거리에 따른 변동비의 변화를 고려했을 때 TCS를 이용한 통행시간 산정시 단일 구간거리는 최대 70km가 넘지 않아야 한다.



<그림 5> 시간대별 변동비의 변화
<Fig. 5> Variation ratio change by departure time

중위절대편차를 이용할 경우 기본적으로 z_{cut}=3으로 설정하였다. 그러나 통행시간 변동비가 증가할수록 통행시간의 변동도 크므로 변동비의 크기에 따라 z_{cut}을 조절할 필요가 있다. 본 연구에서는 통행시간 추정시 20%까지의 변동은 인정한다는 전제하에 변동비(CV), 즉 σ/μ=0.1 까지는 z_{cut}=3을 기본으로 설정하고 σ/μ가 증가함에 따라 z_{cut}을 점차적으로 감소시키고자 한다. 따라서

<표 1> CV에 따른 z_{cut}
<Table 1> z_{cut} according to CV

CV	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.20
z_{cut}	3.00	2.73	2.50	2.31	2.14	2.00	1.88	1.76	1.67	1.58	1.50

변동비에 따른 z_{cut}은 다음과 같다.

$CV \times z_{cut} = 0.1 \times 3$에서 $z_{cut} = \frac{0.1 \times 3}{CV}$

Ⅳ. 평활화 모형

1. 평활화 기본 모형

중위절대편차를 이용한 이상치 제거는 5분 집계간격 내에서 비정상적인 통행시간을 갖는 개별차량을 제거하는 것이다. 개별차량의 이상치가 제거되었더라도 집계간격 내 차량수가 적을 경우에는 5분 간격의 대표값이 짧은 시간대에 급격한 변동을 보일 수 있다. 또한 중위절대편차를 이용한 방법은 차량군이 두개의 무리로 군집주행을 할 경우 이상치를 제대로 판별할 수가 없다.

2009년 1월 23일 오산~천안의 8시 20분 출발차량은 총 4대로서 각각 20분, 22분, 35분, 39분의 통행시간을 보인다. 만약 통행시간 자료가 20분, 22분, 35분이라면 35분 통행시간의 z_i^{MAD}는 4.4가 되어 이상치로 판정이 되어 통행시간 대표값은 21분이 되지만, 통행시간 자료가 20분, 22분, 35분, 39분인 경우에는 35분, 39분의 z_i^{MAD}는 각각 0.6, 0.9가 되어 이상치로 판정이 되지 않아 통행시간 대표값은 29분이 된다. 8시 15분 출발차량의 통행시간 대표값이 22분이고 8시 25분 출발차량의 통행시간 대표값이 24분임을 감안하면 8시 20분에 출발한 35분, 39분의 통행시간은 이상치로 보는 것이 타당하다고 판단되나 중위절대편차를 이용한 집계간격 내에서의 이상치 제거 방법으로는 이를 판별할 수가 없다. 따라서 집계간격 내에서의 이상치 제거 이후에는 전·후시간대의 통행시간 변동을 고려한 평활화 과정이 필요하다.

TCS 자료에 의한 출발시각 기준의 5분 집계 통행시간은 일반적으로 시계열적 변동을 나타낸다. 특정 시점에서의 통행시간 변동량은 직전 시간대 의 통행시간 대표값과의 차이를 나타낸다.

평활화의 기본모형은 다음과 같다.

$$\hat{t_n} = \hat{t}_{n-1} + k(t_n - \hat{t}_{n-1})$$

$\hat{t_n}$: n 시간대의 평활화 된 통행시간 대표값
$\hat{t}_{n-1}$: n-1 시간대의 평활화 된 통행시간 대표값
t_n : n 시간대에 관측된 통행시간 대표값
k : 평활상수 (0~1), $k = e^{|t_n - \hat{t}_{n-1}|}$

즉 n 시간대의 평활화 된 통행시간 대표값은 직전 시간대의 평활화 된 값을 기본으로 하여 n 시간대의 통행시간 변동량을 더해주는데 변동량이 합리적인 수준이면 변동량을 최대한 인정하여 통행시간 관측값에 최대한 가깝게 해주고 변동량이 크면 통행시간 관측값이 이상치일 가능성이 있다고 판단하여 직전 시간대의 평활화 된 값에 가깝게 취한다. 여기서 평활화상수, k는 구간거리에 따라 달리 적용해야 한다. 통행시간 변동량의 분포는 구간거리에 따라 다르기 때문이다.

기본적으로 k는 0~1의 값을 갖는다. k=0일 때는 $\hat{t_n} = \hat{t}_{n-1}$이 되어 n 시간대의 통행시간 대표값은 이상치로 판단되어 이전 시간대의 평활화 된 통행시간 값을 사용한다. k=1일 때는 $\hat{t_n} = t_n$이 되어 n 시간대의 통행시간 대표값을 100% 인정한다. 0<k<1일 때는 $\hat{t_n} = \hat{t}_{n-1} + k(t_n - \hat{t}_{n-1})$이 되어 n 시간대의 변동량을 k 값에 따라 달리 적용한다. 즉 통행시간의 순변동량이 클수록 현재 시간대의 관측값이 이상치일 가능성이 크므로 변동량을 조금만 반영하기 위하여 k값을 작게 하여 평활화가 크게 되게 한다.

2. 평활상수, k의 결정

1) 거리조절계수, r의 도입

통행시간 변동량의 분포는 구간거리에 따라 달라진다. 출발시각이 5분 차이가 날 때 목적지까지의 통

행시간이 20분 차이가 나는 것은 서울~수원(17,2km)과 서울~천안(69.4km)에서 그 의미가 다르다. 즉 동일한 변동량이라도 구간거리가 긴 서울~천안보다는 구간거리가 짧은 서울~수원에서 이상치로 판정될 가능성이 더 크다. 따라서 평활화의 정도를 나타내는 k값은 구간거리에 따라 달리 적용되어야 한다.

구간거리에 따라 동일 변동량의 영향을 차별화하기 위하여 평활화상수, k의 결정모형에 거리조절계수, r을 도입하여 다음과 같이 변환한다.

$$k = e^{\frac{1}{r}\left|t_n - \hat{t}_{n-1}\right|}$$

r값은 로지스틱 함수를 도입하여 하한 값은 1로, 상한 값은 3으로 설정하여 구간거리에 따라 r=1~3의 값을 갖도록 다음과 같이 나타낸다.

$$r = \frac{r_{\max} - r_{\min}}{1 + \frac{r_{\max} - r_{\min}}{r_{\min}} e^{-m(d-l)}} + 1 = \frac{2}{1 + 2\,e^{-m(d-l)}} + 1$$

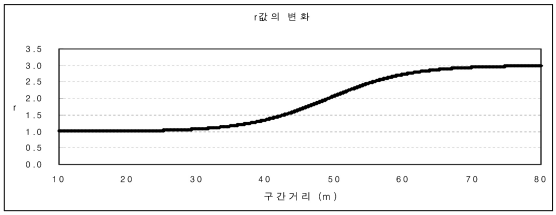
r값의 변화를 나타내는 로지스틱곡선의 형상을 결정하기 위하여 구간거리에 따른 통행시간 변동의 분포를 구해보면 <표 2>와 같다.

<표 2> 거리별 통행시간 변동 누적 분포
<Table 2> Accumulation distribution of travel time variation by length

구간	서울~수원 (17.2km)		서울~오산 (31.4km)		서울~안성 (49.5km)		서울~천안 (69.4km)	
변동 (분)	차량수 (대)	분포 (%)	차량수 (대)	분포 (%)	차량수 (대)	분포 (%)	차량수 (대)	분포 (%)
0	1,708	63.5	1,194	44.4	754	28.1	487	18.1
1	950	35.3	1,249	46.5	1,139	42.4	848	31.6
2	26	1.0	218	8.1	517	19.2	570	21.2
3	3	0.1	20	0.7	175	6.5	294	10.9
4	1	0.0	4	0.1	58	2.2	183	6.8
5			1	0.0	22	0.8	121	4.5
6			2	0.1	6	0.2	63	2.4
7					4	0.1	32	1.2
8					6	0.2	30	1.1
9					4	0.1	20	0.8
10					1	0.0	9	0.3
계	2,688	100.0	2,688	100.0	2,688	100.0	2,684	100.0

<표 3> r 값의 적용
<Table 3> Application of r-value

구간	수원 (17.2km)	기흥 (22.3km)	오산 (31.4km)	안성 (49.5km)	천안 (69.4km)
r	1.0	1.0	1.1	2.0	3.0



<그림 6> 거리조절계수의 변화 (m=0.17, l=45)
<Fig. 6> Change of distance adjustment coefficient

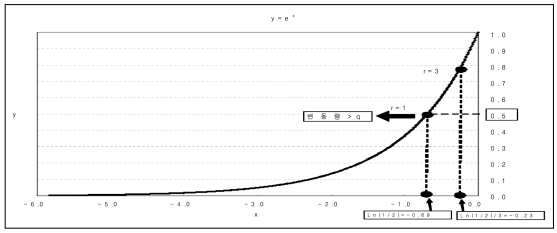
통행시간 변동의 90% 누적분포를 고려할 때 서울~수원, 서울~기흥, 서울~오산은 통행시간 변동이 1분, 서울~안성은 2분, 서울~천안은 4분 정도 임을 알 수 있다. 1달 동안의 평균값이 아닌 하루 동안의 통행시간 변동은 더 큰 값이 나올 수도 있으나 평균 변동분포로서 구간거리에 따른 변동의 변화비를 비교하여 m=0.17, l=45로 하여 r을 다음과 같이 설정한다.

$$r = \frac{2}{1 + 2\,e^{-0.17(d-45)}} + 1$$

이를 본 연구에 적용한 구간별 r값은 <표 3>과 같다. 결국 구간길이가 짧아질수록 r값이 작아지고 k값도 작아져서 평활화의 정도가 심해짐을 알 수 있다.

2) 허용변동량의 결정

평활화상수, k를 구할 때 거리조절계수, r의 도입으로 구간거리에 대한 영향을 반영하였으므로 다음으로는 기준이 되는 허용 변동량을 결정해야 한다. <그림 7>은 $y = e^x$의 지수함수를 나타낸 것으로 y축은 평활화상수, k를 x축은 허용변동량을 나타낸다. x=ln(0.5)=-0.69일 때 k=0.5를 기준으로 하여 변동량, $\left|t_n - \hat{t}_{n-1}\right|$이 허용변동량, q 보다 크면 x값을 좌측으로 이동시켜 k값을 작게 한다. 즉 통행시간이 1/2이 되는 ln(0.5)을 기준으로 통행시간 변동에 대한 영향



<그림 7> 허용 변동량의 도입
<Fig. 7> Introduction of allowable variation

을 나타내는 항, $|t_n - \hat{t}_{n-1}|/q$과 구간거리에 대한 영향을 나타내는 항, $1/r$을 도입하여 최종적으로 평활화 상수, k를 구하는 모형은 다음과 같이 구한다.

$$k = e^{\frac{\ln(0.5)}{r}\frac{|t_n - \hat{t}_{n-1}|}{q}}$$

여기서 ln(0.5)는 $y = e^x$에서 y=0.5인 x의 값을 나타낸다. 즉 x=$\frac{\ln(0.5)}{r}\frac{|t_n - \hat{t}_{n-1}|}{q}$=-0.69=ln(0.5)이면 k=0.5이므로 변동량의 1/2만 취한다는 의미로서, 결국 구간의 길이에 따라 변동량의 1/2만 취하는 기준점을 달리 적용하는 것이다.

통행시간 변동분포를 고려하여 q=10분으로 하면 서울-수원 구간은 r=1이며 통행시간 변동량에 따라 다음과 같이 k를 적용한다.

i) 변동량, $|t_n - \hat{t}_{n-1}|$ =10분이면, $\frac{|t_n - \hat{t}_{n-1}|}{q}$=1 이므로 k=0.5이고 통행시간의 변동량을 1/2만 반영

ii) 변동량, $|t_n - \hat{t}_{n-1}|$ > 10분이면 $\frac{|t_n - \hat{t}_{n-1}|}{q}$ >1 이므로 $\ln(0.5)\frac{|t_n - \hat{t}_{n-1}|}{q}$는 ln(0.5)보다 커져 그래프의 좌측으로 이동하여 k값이 작아지고 지고, 결국 통행시간의 변동을 조금만 반영

iii) 변동량, $|t_n - \hat{t}_{n-1}|$ < 10분, $\frac{|t_n - \hat{t}_{n-1}|}{q}$ < 이므로 $\ln(0.5)\frac{|t_n - \hat{t}_{n-1}|}{q}$는 ln(0.5)보다 작아지므로 그래프의 우측으로 이동하여 k값이 커지고, 결국 통행시간의 변동을 많이 반영

<표 4> 구간거리와 변동량에 따른 k값
<Table 4> k-value by section length and variation

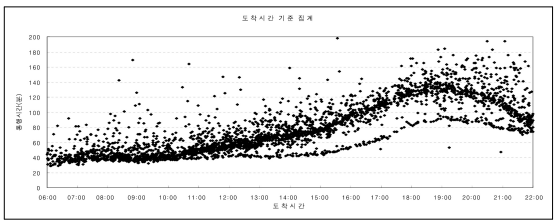
$\lvert t_n - \hat{t}_n \rvert$	$\frac{\lvert t_n - \hat{t}_{n-1} \rvert}{q}$	$\frac{\ln(0.5)}{r}\frac{\lvert t_n - \hat{t}_{n-1} \rvert}{q}$			k		
		r=1	r=2	r=3	r=1	r=2	r=3
5	0.5	-0.35	-0.17	-0.12	0.71	0.84	0.89
10	1.0	-0.69	-0.35	-0.23	0.50	0.71	0.79
15	1.5	-1.04	-0.52	-0.35	0.35	0.59	0.71
20	2.0	-1.39	-0.69	-0.46	0.25	0.50	0.63
25	2.5	-1.73	-0.87	-0.58	0.18	0.42	0.56
30	3.0	-2.08	-1.04	-0.69	0.12	0.35	0.50
35	3.5	-2.43	-1.21	-0.81	0.09	0.30	0.45

<표 4>는 q=10분으로 하였을 때 구간거리와 변동량에 따른 k값의 변화를 나타낸다.

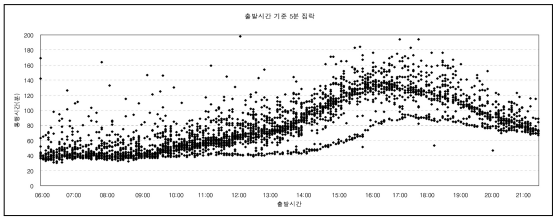
Ⅴ. 모형의 적용

본 연구에서 개발한 TCS 데이터 전처리 모형을 2009년 1월 23일 서울~천안의 데이터를 이용하여 적용해보면 <그림 8~11>과 같다. 이날은 설 연휴 전날로서 경부선 부산방향의 교통량이 오후 늦게까지 꾸준히 증가하여 본 연구에서 제시한 방법이 통행시간의 증감 패턴을 잘 반영할 수 있는지를 확인할 수가 있다.

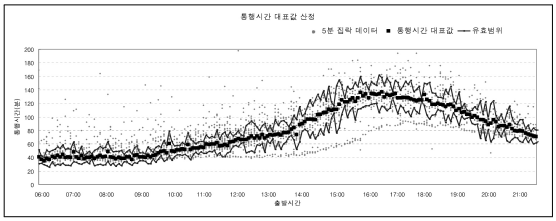
<그림 8>의 도착시각 기준의 TCS 원시데이터를 <그림 9>와 같이 출발시각 기준의 5분 집계간격으로 데이터를 정렬하면 통행시간의 증가 패턴은 어느 정도 파악 할 수 있으나 출발 시간대별로 통행시간의 분포가 다양함을 알 수 있다. 따라서 통행시간의 변동에 따라 z_{cut}을 달리하여 통행시간의 유효범위를 설정하여 이상치를 제거하면 <그림 10>과 같다. z_{cut}



<그림 8> TCS 원시데이터
<Fig. 8> TCS raw data



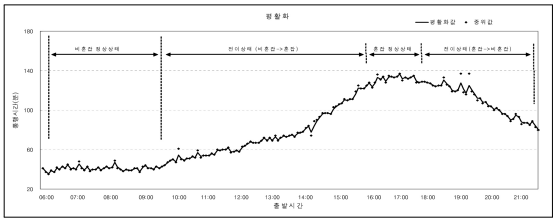
<그림 9> 출발시각 기준 5분 집락
<Fig. 9> 5-Minute cluster by departure time



<그림 10> 이상치 제거
<Fig. 10> Outlier elimination

을 일률적으로 3.0으로 적용하는 것보다 통행시간의 변동에 따라 z_{cut}을 달리 적용함으로써 이상치를 효과적으로 걸러낼 수 있음을 알 수 있다. 이를 다시 평활화를 통해 극단치를 제거하면 <그림 11>과 같이 최종적으로 통행시간의 대표값을 구할 수 있다.

설 연휴 전날의 교통패턴은 시간의 흐름에 따라 비혼잡 정상상태, 정체로의 전이상태, 혼잡 정상상태, 비혼잡으로의 전이상태 등의 양상을 모두 보이고 있으며 본 연구에서 제시한 전처리 기법을 통해 TCS 원시데이터로부터 통행패턴을 적절히 추정할 수 있음을 알 수 있다. 특히 통행시간의 급격히 증가하는 시간대에서도 전처리를 통한 통행시간의 대표값이 이를 잘 반영함을 알 수 있다.



<그림 11> 평활화
<Fig. 11> Smoothing

Ⅵ. 결 론

TCS 자료가 통행시간의 정보로서 가치를 갖으려면 출발시각 기준으로의 변환, 이상치 제거, 집락, 대표값 선정, 평활화 등의 전처리 과정을 거쳐야 한다.

본 연구에서는 TCS 대표값 설정시 동일 집계간격내의 편차를 줄이고 이상치에 의한 영향을 최소화 하도록 중위절대편차의 제거변수를 조절하였다. 구간거리와 지·정체 상황에 따른 통행시간 변동비의 특성을 파악하여 제거변수를 시·공간적으로 차등 적용한 결과, 통행시간 대표값의 상·하한 신뢰 폭을 좁힐 수 있었다. 또한 구간거리에 따른 거리조절계수와 통행시간의 허용 변동량을 도입하여 통행시간 평활화 모형을 제안하였다. 제안모형의 적용 결과 특히 통행시간이 급작스럽게 증가하는 전이 시간대에서도 추정된 통행시간이 참값에 좀 더 가까운 패턴을 보였다. 본 연구에서는 TCS 데이터의 전처리 기법을 제시하였으나, 향후 DSRC, AVI 데이터 등의 특성도 반영할 수 있는 통합적인 전처리 기법이 필요하며, 특히 통행시간 예측분야에 활용하기 위해서는 실시간 전처리 기법으로의 확장이 필요하다.

참고문헌

[1] 도로교통연구원, *고속도로 차량검지기 자료 조사·분석 및 활용기법 개발*, 2007. 12.

[2] 도로교통연구원, *고속도로 통행시간 신뢰도 제고방안 연구*, 2008. 12.

[3] H. Nishiuchi, K. Nakamura, S. Bajwa, E. Chung, and M. Kuwahara, “Evaluation of travel time and OD variation on the Tokyo metropolitan expressway using ETC data,” Proc. 22nd ARRB Conference (CD), Canberra, Australia.

[4] D. Francois and R. Hesham, “Estimating spatial travel time using automatic vehicle identification data,” 82nd Annual Meeting Preprint CD-ROM, Transportation Research Board, January 12-16, Washington, D.C

[5] 강진기, 손영태, 윤여환, 변상철, “비매설식 자동차량 인식장치를 이용한 구간교통정보 산출 방법 연구,” *한국ITS학회논문지*, 제1권, 제1호, pp. 22-32, 2002. 12.

저자소개

이 현 석 (Lee, Hyun-Seok)

2009년 : 서울대학교 건설환경공학부 박사 (교통공학전공)
2002년 6월 ~ 현재 : 한국도로공사 도로교통연구원 교통연구실 선임연구원

남 궁 성 (Namkoong, Seong J.)

1996년 : 한양대학교 도시공학과 박사 (교통전공)
1996년 10월 ~ 현재 : 한국도로공사 도로교통연구원 교통연구실 수석연구원
2008년 1월 ~ 현재 : 명절특별수송대책기간 한국도로공사 교통예보관
2006년 12월 ~ 현재 : 첨단교통정보기술지원센터 (Center for OASIS) 운영
2002년 9월 ~ 2004년 10월 : University of Virginia, Smart Travel Leb. Research Scientist