구간검지 교통자료 이상치 제거 방법론 고찰

A Review on Section Data Cleaning Methods

장진환

Ⅰ. 서론

1993년 대전 엑스포를 계기로 국내에 도입되기 시작한 실시간 교통정보시스템은 1997년 국도 ITS 시범 구축, 2002년 월드컵 성공적 개최를 위한 첨단교통모델도시 사업 등을 거쳐 2012년 말 현재 고속도로 100%, 국도 약 20%, 약 30여 지자체 도시부 간선도로에 구축・운영 중에 있다. 실시간 교통정보시스템의 주요한 목적은 전방 도로 상황을 실시간으로 운전자에게 전달하여 우회도로 이용, 출발시간 조정 등을 통해 도로 이용효율을 극대화하는데 있다. 이러한 목적을 달성하기 위해서는 운전자가 제공되는 교통정보를 신뢰할 수 있어야 하는데 이를 위해서는 제공되는 실시간 교통정보의 정확도가 중요하다. 정확한 교통정보 수집을 위해 기존에는 주로 루프 검지기, 영상 검지기 등 지점 검지기를 주로 활용했지만, 최근 들어 실(實) 통행시간을 수집할 수 있는 AVI[1], DSRC[2], UTIS[3] 등 구간 검지기가 확대 도입되고 있다.

그러나 구간 검지기를 활용하여 신뢰성 있는 교통정보를 제공하기 위해서는 적정 표본 수 확보, 이상치[4] 제거, 시간 처짐 현상 극복 등 몇 가지 이슈들을 해결해야한다. 이러한 이슈 중 적정 표본 수 및 시간 처짐 문제는 집락간격 확대, 설치간격 축소 등을 통해 일정부분 해결이 가능하다. 그러나 주・정차, 교차도로 유출・입, 갓길 주행 등에 의해 발생하는 이상치는 상기의 방법으로 해결이 불가능하다. 특히 유출・입 지점 수가 많고 도로변 개발 정도가 큰 도시부 도로, 국도 등의 경우에는 이상치 문제가 더욱 더 중요하게 부각된다. 그림 1은 국도 3호선 갈현IC-갈마터널 구간에서 DSRC를 이용하여 수집한 개별 차량의 통행시간 분포를 나타낸 것인데 많은 이상치가 관측됨을 알

장진환 : 한국건설기술연구원 첨단교통연구실, jhjang@kict.re.kr, 031-910-0684, 031-910-0338

1) 차량번호판을 인식하여 구간 통행시간을 수집하는 장비
2) 하이패스 단말기를 이용하여 구간 통행시간을 수집하는 장비
3) UTIS 차량 단말기를 이용하여 구간 통행시간을 수집하는 장비
4) 주어진 도로, 교통, 환경 조건하에서 해당 구간을 통행하는 소요되는 기대 통행시간(expectation)에서 현저하게(markedly) 벗어난(deviate) 관측치(observation)

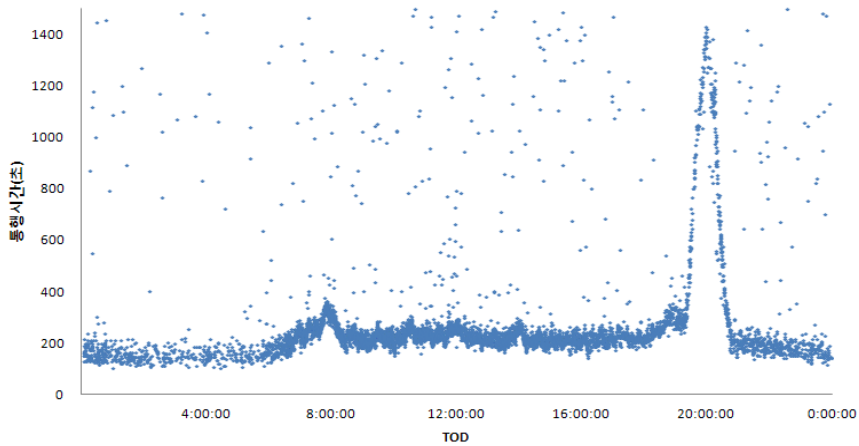

그림 1. 국도 3호선(갈현 IC-갈마터널) 구간 검지 원시 자료

수 있다. 자칫 이러한 이상치가 교통정보 생성에 포함될 경우 제공되는 정보의 신뢰성 저하는 자명해 보인다.

많은 자연과학 분야에서 일반적으로 사용되는 이상치 제거 방법은 평균, 표준편차 산출 후 일정 신뢰구간(예. 95%, 99%) 이외 범위를 제거하거나 회귀식 도출 후 회귀식 신뢰구간 이외의 범위 값을 제거하는 것이다. 그러나 그림 1에서 보듯이 구간검지 자료 이상치는 유효치보다 과다하게 큰 이상치가 관측되거나 해당 수집주기 내에 유효치보다 이상치가 더 많이 관측 되는 등 일반적인 자연과학 분야에서 접하는 이상치와는 그 특성이 다소 상이하다. 따라서 이러한 구간검지 자료 특성에 적합한 이상치 제거 방법론 개발을 위해 국내·외 연구자들의 많은 노력이 있어 왔다. 그러나 현재까지도 모든 상황에 적합한 최적의 방법론은 개발되지 않았고 단지 특정한 상황별 목적별로 적정한 방법론만이 있을 뿐이다. 따라서 교통정보센터 실무자(연구자)들은 이러한 방법론 특성을 면밀히 파악하여 해당 센터에 적합한 알고리즘을 적용(또는 개발)하는 것이 필요하다.

이에 본 고에서는 교통정보 센터에서 적정 알고리즘 도입(또는 개발)을 통한 구간검지 실시간 교통정보 신뢰성 향상을 위하여 기존의 국내·외 구간검지 교통자료의 이상치 제거(필터링) 방법론을 고찰한 후, 실제 교통정보시스템 운영의 관점[5]에서 각 알고리즘별 장·단점 등을 면밀히 분석하고자 한다.

Ⅱ. 필터링 방법론

1. 강진기 방법

1) 수식

$$T_{is} = |T_i - \tilde{T}| \le \sigma \qquad (1)$$

여기서,
T_{is} : 유효 통행시간
T_i : 개별 통행시간,
$\tilde{T}$: 극단치를 제외한 개별 통행시간 평균
σ : 개별 통행시간 표준편차

$$T_s = \frac{\sum T_{is}}{n_s} \quad V_s = \frac{S_s}{T_s} \qquad (2)$$

여기서,
S_s : 구간 S의 거리

5) 시스템 장애(장비, 통신, 센터 등) 및 장애복구 상황, 다양한 데이터 유형(유효치보다 이상치가 더 많이 관측) 등 고려

n_s : T_{is} 개수(유효 통행시간 수)
T_s : 구간 S의 평균 통행시간
V_s : 구간 S의 평균 통행속도

2) 설명

강진기는 2002년 일반국도에 설치한 비매설식 AVI 교통정보 신뢰도 평가를 위해 상기 이상치 제거 방법론을 적용했다. 우선 극단치(설계속도 2배 이상 또는 10km/h 이하 속도)를 제거한 후 산술평균 및 표준편차를 구한다. 그 다음, $\mu \pm \sigma$(식(1)) 이내의 통행시간 값만 유효치로 간주하여 평균 통행시간(속도)를 산출한다(식(2)).

3) 장점

강진기 방법론은 수식이 단순하고 이해가 용이하여 실제 교통정보시스템(S/W) 구현이 용이하다.

4) 단점

강진기 방법론은 통행시간 분포가 정규분포를 따른다는 기본 전제가 필요하다. 그러나 일반적으로 통행시간 분포는 Log-normal 분포를 따른다고 알려져 있기 때문에 방법론 적용에 대한 이론적 근거가 명확하지 않은 단점이 있다. 또한 일부 이상치 값이 과도하게 클 경우 표준편차도 따라서 커지기 때문에 유효치 범위가 불필요하게 커질 수 있고, 이에 따라 이상치가 유효치로 판정되는 경우를 초래할 수 있다.

2. TransGuide 알고리즘

1) 수식

$$Stt_{ABk} = \{t_{Bi} - t_{Ai} | t - t_w < t_{Bi} \le t\} \cap \{tt'_{ABk}(1 - l_{th}) \le t_{Bi} - t_{Ai} \le tt'_{ABk}(1 + l_{th})\} \quad (3)$$

$$tt_{ABk} = \frac{\sum_{i=1}^{Stt_{ABk}} (t_{Bi} - t_{Ai})}{Stt_{ABk}} \quad (4)$$

여기서,
Stt_{ABk} : AB 구간, k 수집주기 유효 표본집단
t_{Ai} : A지점에서 관측된 i 차량 관측시각
t_{Bi} : B지점에서 관측된 i 차량 관측시각
t_w : 집락간격
I_{th} : 통행시간 유효범위 파라미터
tt_{ABk} : AB 구간, k 수집주기 유효표본 평균
tt'_{ABk} : AB 구간, $k-1$ 수집주기 유효표본 평균

2) 설명

TransGuide 알고리즘은 Southwest Research Institute(SwRI)에서 개발하여 미국의 샌 안토니오 FTMS에서 실제 운영 중이다. 이는 해당 수집주기 내에서 사용자가 정한 범위를 초과하는 통행시간 값들을 자동적으로 제거한 후(식(3)) 나머지(유효치)를 평균하여 해당 수집주기 평균 통행시간을 산출한다(식(4)). TransGuide 알고리즘에서 주요한 파라미터는 t_w와 l_{th}인데, t_w는 집락간격을 의미하고, l_{th}는 유효 통행시간 범위를 의미한다.

3) 장점

TransGuide 알고리즘 또한 수식이 단순하고 이해가 용이하여 실제 교통정보시스템(S/W) 구현이 용이하다. 또한 이는 통행시간 분포의 정규성 가정을 포함하지 않음에 따라 이론적 근거 부족 문제도 발생하지 않는다. 특히 유출입 지점이 많지 않아 이상치가 과다하게 발생되지 않는 연속류 도로에 적합하다고 알려져 있다.

4) 단점

TransGuide 알고리즘은 통행시간 유효범위가 l_{th} 파라미터에 의해 결정되기 때문에 l_{th} 값을 크게 할 경우 이상치가 유효치로 판정될 가능성이 커지고, 반대로 l_{th} 값을 작게 할 경우 통행시간 패턴이 급격히 변하는 경우 유효치를 이상치로 판정하는 오류를 범할 수 있다. 또한 이전주기 통행시간

값을 기반으로 현재 수집주기 통행시간 유효범위를 설정하기 때문에 시스템 구축, 장애 복구 등에 따른 데이터 수집 재개 시 초기 수집주기 내에 있는 이상치 제거가 불가능하다. 따라서 초기 수집주기에 과도한 이상치가 존재할 경우 이후의 통행시간 유효범위에 오류가 발생하여 결국 시스템 불능(통행시간 과다 오차) 상태를 초래할 수 있다. 이는 특히 시스템 일시적 장애(장비고장, 통신 두절 등) 발생시 약점으로 작용될 수 있다.

이러한 문제점을 예방하기 위해 장애 복구 시 장애 발생 직전 자료를 기반으로 초기 수집주기 자료의 유효범위를 설정한다 하더라도 또 다른 문제가 발생할 수 있다. 예를 들어 그림 1 자료에서 12시에 장애가 발생하고 20시에 장애가 복구 되었다고 가정하자. 이 경우 장애 전・후의 통행시간 차이가 커서(비첨두, 첨두) 자칫 장애 복구 후 수집 자료를 전체를 이상치로(유효치임에도 불구하고) 판정할 수 있다.

3. 도로공사 방법론

1) 이상치 제거 방법

구분	이상치 제거 방법
표준편차가 통행시간의 평균값 대비 5% 미만일 경우	데이터 집단의 상위 3%, 하위 2% 샘플 제거 (제거율 5%)
표준편차가 통행시간의 평균값 대비 10% 미만일 경우	데이터 집단의 상위 5%, 하위 5% 샘플 제거 (제거율 10%)
표준편차가 통행시간의 평균값 대비 15% 미만일 경우	데이터 집단의 상위 8%, 하위 7% 샘플 제거 (제거율 15%)
표준편차가 평균값 대비 15% 이상일 경우	표준편차 범위 밖의 샘플 제거

2) 설명

도로공사 알고리즘은 고속도로 DSRC 교통정보시스템에서 이상치를 제거하는 방법론으로써, 이는 해당 수집주기 내의 통행시간 값들의 변동계수(Coefficient of Variation) 값에 따라 상・하위 값들을 제거한다.

3) 장점

도로공사 방법론 또한 수식이 단순하고 이해가 용이하여 실제 교통정보시스템(S/W) 구현이 용이할 뿐만 아니라 표본의 변동성이 커질수록 이상치가 존재할 가능성이 커진다는 기본 가정 또한 합리적인 것으로 판단된다.

4) 단점

도로공사 방법론은 수집자료의 특성에 관계없이 일정 부분의 이상치를 제거하기 때문에 유효치도 이상치로 간주하여 필터링 되는 경우가 발생한다. 또한 해당 수집주기 내에 유효치는 없고 이상치만 존재할 경우[6] 이상치 전체 제거가 불가능할 뿐만 아니라 샘플 수가 적을 경우(6 이하) 이상치 제거 수가 0.5(6*0.08=0.48) 미만임에 따라 반올림 하여도(1 미만) 이상치 필터링이 불가능하다. 만약 소수점 이하를 올림처리 한다면 샘플수가 많은 수집주기에서 이상치 과다 제거 문제가 발생할 수 있다.

4. Dion 알고리즘

1) 수식

$$T_{AB}(k) = \begin{cases} \exp((1-\alpha)\ln(T_{AB}(k-1)) + \alpha\ln(\tilde{T}_{AB}(k))), & n_{v,k} > 0 \\ T_{AB}(k-1) & n_{v,k} = 0 \end{cases} \quad (5)$$

$$\sigma^2_{AB}(k) = \begin{cases} (1-\alpha)\sigma^2_{AB}(k-1) + \alpha\tilde{\sigma}^2_{AB}(k), & n_{v,k} > 1 \\ \sigma^2_{AB}(k-1) & n_{v,k} = 0, 1 \end{cases} \quad (6)$$

$$S_{AB}(k) \equiv \{h | t - T_w < t_{B,h} \le t\} \cap \{m | T^{\min}_{AB}(k) < t_{B,m} - t_{A,m} < T^{\max}_{AB}(k)\} \quad (7)$$

$$T^{\min}_{AB}(k) = \exp(\ln T_{AB}(k) - n_\sigma(k)\sigma_{AB}(k)) \quad (8)$$

$$T^{\max}_{AB}(k) = \exp(\ln T_{AB}(k) + n_\sigma(k)\sigma_{AB}(k)) \quad (9)$$

6) 국도 DSRC 자료를 분석한 결과, 교통량이 극히 적은 새벽시간대에 해당 수집주기(5분) 내에 이상치만 존재하는 경우도 발생한다.

$$\tilde{T}_{AB}(k)=\sum_i (t_{B,i}-t_{A,i})/n_v(k),\ i\in S_{AB}(k-1) \quad (10)$$

$$\tilde{\sigma}^2_{AB}(k)= \quad (11)$$
$$\begin{cases} 0, & n_v(k)=0 \\ (\ln(t_{Bi}-t_{A,i})-\ln T_{AB}(k))^2, & n_v(k)=1 \\ \sum_i (\ln(t_{Bi}-t_{A,i})-\ln T_{AB}(k))^2/n_v(k), & n_v(k)>1 \end{cases}$$

$$\alpha=1-(1-\beta)^{n_v(k)} \quad (12)$$

$$n_\sigma(k)=\lambda+\lambda[1-(1-\beta_\sigma)^{n_0(k)}] \quad (13)$$

여기서,

$T_{AB}(k)$, $\sigma^{2}{}_{AB}(k)$: AB 구간, k 수집주기 평활화된 평균 및 분산

$\tilde{T}_{AB}(k)$, $\tilde{\sigma}^2{}_{AB}(k)$: AB 구간, k 수집주기 통행시간 평균 및 분산

$S_{AB}(k)$: AB 구간, k 수집주기 유효 샘플 집단

$T_{AB}^{\min}(k)$, $T_{AB}^{\max}(k)$: AB 구간, k 수집주기 최소, 최대 통행시간

$\alpha, \beta, \lambda, \beta_\sigma$: 파라미터

$n_v(k)$: k 수집주기 유효 표본수

$n_0(k)$: 유효 표본수가 0인 이전 수집주기 개수

2) 설명

Dion 알고리즘은 TransGuide 알고리즘 단점을 보완하기 위해 개발된 알고리즘으로써, 통행시간 분포가 Log-normal 분포를 따른다는 가정 하에 개발되었다. 이는 TransGuide 알고리즘의 정적(static) 파리미터(l_{th}) 단점을 보완하기 위해 평활화된 통행시간 값(평균, 표준편차)을 기반으로 하여 다음 주기의 유효 통행시간 범위를 가변적으로 설정한다(식(5)-(11)). 이 경우, 유효 샘플수($n_v(k)$)에 따라 평활화 가중치를 적용하는 파라미터를 개발하였고(식(12)), 샘플 수가 적은 상황에서 발생하는 급격한 통행시간 패턴 변화 시에도 알고리즘 성능을 유지하기 위해 유효 샘플 수가 없는 수집주기 개수($n_0(k)$)를 반영하여 유효 통행시간 범위의 확장이 가능한 파라미터(식(13))도 추가하였다.

3) 장점

Dion 알고리즘은 타 알고리즘에 비해 이론적 근거가 명확하고 이상치 필터링 성능이 비교적 우수하다. 특히 새벽시간대 또는 비첨두 돌발상황 발생 시 등 샘플 수가 적은 상황에서 발생하는 급격한 통행시간 패턴 변화 시에도 이상치 제거 성능이 우수하다. 또한 통행시간 평활화 방법도 유효 샘플 수에 따른 가중치를 부여하기 때문에 합리적인 것으로 판단된다.

4) 단점

Dion 알고리즘은 수식이 복잡하고 이해가 어려울 뿐만 아니라 수집 자료에 적합한 파라미터 값을 설정해야 하는 어려움 등이 있어 실제 시스템 적용성이 낮다는 단점이 있다. 또한 TransGuide 알고리즘과 마찬가지로 이전주기 통행시간 값을 기반으로 현재 수집주기 통행시간 유효범위를 설정하기 때문에 시스템 구축, 장애 복구 등에 따른 데이터 수집 재개 시 TransGuide 알고리즘과 동일한 문제점을 야기할 수 있다.

5. 장진환 알고리즘

1) 수식

$$(tts_{AB})_k=\alpha(tt_{AB})_{k-1}+(1-\alpha)(tts_{AB})_{k-1} \quad (14)$$

$$(Stt_{AB})_k=\{t_{Bi}-t_{Ai}|t_k-t_{k-1}<t_{Bi}\le t_k\} \cap\{(tt_{ABmin})_k\le t_{Bi}-t_{Ai}\le(tt_{ABmax})_k\} \quad (15)$$

$$(tt_{ABmin})_k=(tts_{AB})_{k-1}(1-l_{th}) \quad (16)$$

$$(tt_{ABmin})_k=(tts_{AB})_{k-1}(1+l_{th}) \quad (17)$$

$$\alpha=1-(1-\beta)^{n_{vk-1}} \quad (18)$$

$$(tt_{AB})_k=\frac{\sum_{i=1}^{n_{vk}}(t_{Bi}-t_{Ai})}{n_{vk}} \quad (19)$$

여기서,

$(tts_{AB})_k$: AB 구간, k 수집주기 평활화된 통행시간
$(tt_{AB})_k$: AB 구간, k 수집주기 유효 샘플 평균 통행시간
$(Stt_{AB})_k$: AB 구간, k 수집주기 유효샘플 집단
α, β, l_{th} : 파라미터
n_{vk} : k 수집주기 유효 샘플 수
$(tt_{ABmin})_k$, $(tt_{ABmax})_k$: AB 구간, k 수집 주기 최소, 최대 통행시간

2) 설명

장진환 알고리즘은 Dion 알고리즘의 기본가정(통행시간 Log-normal 분포)이 충족되지 않는 일반국도 AVI 자료 필터링을 위해 개발된 알고리즘으로써, 기본적으로 이전 수집주기 자료를 기반으로 다음 수집주기 유효치의 상·하한 값을 설정한다는 점에서 TransGuide 알고리즘과 유사하지만, 이전 주기 값을 그대로 적용하는 대신 가중치(식(18))를 적용한 평활화된 값을 적용한다는 점에서 차이가 있다(식(14)-(17)). 이러한 평활화 값 적용은 단속류 교통류 특성으로 인해 통행시간이 급격하게 변화하는 일반국도에 적합한 것으로 분석되었다.

3) 장점

장진환 알고리즘은 통행시간 분포에 대한 어떠한 가정도 필요하지 않을 뿐만 아니라 Dion 및 TransGuide 알고리즘의 장점과 일반국도 AVI 교통자료 특성을 면밀한 분석을 통하여 개발하였기 때문에 특히 일반국도 구간검지 자료 필터링에 적합하다.

4) 단점

장진환 알고리즘 역시 Dion 알고리즘과 같이 수식이 다소 복잡하여 실제 시스템 적용성이 낮다는 단점이 있고 Dion 알고리즘 등과 마찬가지로 이전주기 통행시간 값을 기반으로 현재 수집주기 통행시간 유효범위를 설정하기 때문에 Dion 알고리즘과 동일한 단점이 있다.

6. Ma 알고리즘

1) 수식

$$T_{AB}(k)=\begin{cases}\exp(median(TT\ln_{AB}(k))), & n_k \geq N_{\min}\\ T_{AB}(k-1) & else\end{cases} \quad (20)$$

$$\mu^2_{AB}(k)=\begin{cases}\sum_j \kappa(k,j)^2/(n_k-1), & n_k \geq N_{\min}\\ \mu^2_{AB}(k-1), & else\end{cases} \quad (21)$$

$$\kappa(k,j)=\ln(\hat{T}_{AB}(k,j)) - median(TT\ln_{AB}(k)), j\in S_{AB}(k) \quad (22)$$

$$S_{AB}(k)\equiv\left\{m|T^{\min}_{AB}(k)\leq t_{B,m}-t_{A,m}<T^{\max}_{AB}(k)\right\} \quad (23)$$

$$T^{\min}_{AB}(k)=\exp(\ln T_{AB}(k)-\lambda\mu_{AB}(k)) \quad (24)$$

$$T^{\min}_{AB}(k)=\exp(\ln T_{AB}(k)+\lambda\mu_{AB}(k)) \quad (25)$$

여기서,

$T_{AB}(k)$, $\mu^2_{AB}(k)$: AB 구간, k 수집주기 중앙값 및 중앙값 대비 분산도
$\hat{T}_{AB}(k,j)$, $TT\ln_{AB}(k)$: AB 구간, k 수집주기 통행시간 및 통행시간 집합
$S_{AB}(k)$: AB 구간, k 수집주기 유효 샘플 집단
$T^{\min}_{AB}(k)$, $T^{\max}_{AB}(k)$: AB 구간, k 수집주기 최소, 최대 통행시간
λ : 파라미터(신뢰구간)

2) 설명

Ma 알고리즘은 Dion 알고리즘의 현장 적용성 어려움 문제를 해결하기 위해 개발한 알고리즘으로써 Dion 알고리즘과 달리 위치 척도를 평균 대신 중앙값을 사용했고(식(20)-(22)), 과거 주기 값에 기반하여 현재 수집주기 유효 값 범위를 설정하는 대신 도로공사 방법론과 같이 현재 수집주기 값을 이용하여 현재 수집주기의 유효 값 범위를 설정했다(식(23)-(25)).

3) 장점

Dion 알고리즘에 비해 수식이 간단하고 이해가 용이하여 현실 적용성이 다소 개선되었고, 현재 수집주기 유효 값 범위 설정 시 과거 자료를 이용하지 않음에 따라 시스템 구축, 장애 복구 등에 따른 데이터 수집 재개 시 TransGuide, Dion 등 알고리즘에서 발생하는 초기 수집주기 이상치 제거 한계 문제점을 해결할 수 있다.

4) 단점

Ma 알고리즘 역시 도로공사 알고리즘과 같이 해당 수집주기 내 이상치만 존재할 경우 이상치를 유효치로 판정할 수 있다는 단점이 있고, 표본 수가 적을 경우(6 미만) 적용상 한계점이 있다고 알려져 있다고 알려져 있다.

7. Boxel 방법론

1) 이상치 판별 방법론

단계	내용
1	OBU 보급률 추정
2	구간 매칭대수 수집 (OBU 장착차량)
3	전체 교통량 추정 (OBU 보급률, 매칭대수 이용)
4	구간 통행속도 측정 (OBU 장착차량)
5	Proxy 밀도 추정 (전체 교통량, 통행속도 이용)
6	속도-밀도 관계식(Greenshield) 도출 (Least median of squares 방법 활용)
7	속도-밀도 관계식 95% 신뢰구간 설정
8	신뢰구간(95%) 벗어난 자료를 이상치로 판정

2) 설명

Boxel 방법론은 Bluetooth 검지기로부터 수집된 구간검지 자료의 필터링을 위해 개발되었다. 이는 Bluetooth 보급률과 구간통행속도를 이용하여 Proxy 밀도를 추정한 후 속도-밀도 관계식을 이용하여 이상치를 검출한다. 이 경우 속도-밀도 관계식은 중앙값을 이용한 최소자승법을 적용함에 따라 극단적인 값을 갖는 이상치의 영향을 최소화하고자 하였다.

3) 장점

Boxel은 널이 알려진 교통류 이론을 이용하여 이상치 제거 방법론을 개발했다는 점에서 이론적 근거가 명확할 뿐만 아니라 동일 수집주기 내의 자료를 이용하여 시스템 장애 복구 등에 따른 Dion 알고리즘 등 단점 극복이 가능하다. 아울러 이상치 영향을 최소화하기 위하여 최소자승법 적용 시 평균이 아닌 중앙값을 사용했다는 점에서 우수하다.

4) 단점

Boxel 방법론 역시 해당 수집주기 내 이상치만 존재하는 경우 이상치를 유효치로 판정할 수 있다는 단점이 있고, 연속류 교통류 모형을 이용함에 따라 단속류에는 적용성이 저하된다는 단점이 있다. 또한 OBU 보급률에 따라 Proxy 밀도 추정식이 달라지기 때문에 OBU 보급률에 대한 주기적인 관측·갱신이 요구되는데 이는 현실적으로 많은 어려움이 따른다.

8. Clark 방법론

1) 수식

$$QD = \frac{Q_3 - Q_1}{1.34898} \tag{26}$$

$$M_e \pm F_1(n) t_{0.975,n^*} QD \tag{27}$$

$$F_1(n) = \sqrt{\frac{1}{\frac{2}{\pi} + \frac{1}{n}(\frac{6}{\pi} - 1)}}, \quad n \text{ 짝수} \tag{28}$$

$$F_1(n) = \sqrt{\frac{1}{\frac{2}{\pi} + \frac{1}{n}(\frac{4}{\pi} - 1)}}, \quad n \text{ 홀수} \tag{29}$$

$$n^* = \frac{1}{[F_2(n)]^2} - 1 \quad (30)$$

여기서,

QD : 사분위편차(quartile deviation)
Q_3, Q_1 : 삼(일)사분위 수
1.34898 : 변환 계수
M_e : 해당 수집주기 중앙값
$F_1(n)$: 중앙값 사용에 따른 보정계수
$t_{0.975,\, n^*}$: 자유도(n^*) t(95% 신뢰구간) 통계량
$F_2(n)$: 계수(참고문헌 Kendal(1973) 참조)

2) 설명

Clark 방법론은 일정 신뢰구간을 벗어난 자료를 이상치로 판정한다는 점에서 강진기 방법론과 유사하지만 평균값 대신 중앙값, 표준편차 대신 사분위편차를 사용했다는 점에서 다르다(식(26)-(27)). 또한 중앙값 사용에 따른 보정계수(식(28)-(30))를 적용하여 t 통계량 사용에 따른 이론적 한계를 극복하였다.

3) 장점

강진기 방법론과 마찬가지로 Clark 방법론 역시 이해가 쉽고 실제 적용성이 용이하다는 장점이 있으며, 평균 대신 중앙값을 사용함에 따라 극단적 이상치에 따른 영향을 상대적으로 덜 받는다는 장점이 있다.

4) 단점

그러나 만약 Q_3 또는 Q_1이 이상치이고 그 값이 과도하게 크거나 작을 경우 사분위편차도 따라서 커지기 때문에 유효치 범위가 불필요하게 커질 수 있고, 이에 따라 이상치가 유효치로 판정되는 경우를 초래할 수 있다. 실제로 DSRC 수집자료를 살펴보면 특히 새벽 등 교통량이 적은 시간대에 유효치보다 이상치가 더 많이 존재하여 통행시간 오차가 크게 발생하는 경우가 발생한다.[7)]

9. Haghani 방법론

1) 방법론

단계	내용
1	이동평균과 히스토그램 분석에 의한 이상치 제거(극단적 이상치 제거)
2	수집주기별 평균±1.5×표준편차 초과 값 제거
3	샘플 수 3이하 수집주기(5분) 자료 제거
4	변동계수(표준편차/평균) 1이상 수집주기 자료 제거

2) 설명

Haghani 방법론은 Bluetooth 검지기를 이용하여 수집한 통행시간(속도) 자료를 고속도로 실시간 통행시간 정보 평가를 위한 기준 값으로 사용하기 위해 상기와 같은 이상치 제거 방법론을 개발했다. 따라서 Haghani 방법론 중 이동평균과 히스토그램 분석에 의한 이상치 제거 방법은 24시간 자료가 수집된 후 적용 가능함에 따라 실시간 교통정보시스템에 적용하기에는 무리가 있다.

3) 장점

1단계 방법론을 제외하고는 이해가 용이하여 실제 적용성이 우수하다고 할 수 있다. 또한 기존 방법론과 상이하게 변동계수를 이상치 제거 방법론으로 활용했다는데 의의가 있다. 일반적으로 이상치들은 유효치들에 비해 변동성이 크기 때문에 이러한 변동계수 적용방법은 수집주기 내에 이상치만 존재할 경우에도 이상치 전체 제거가 가능하다.

4) 단점

전술 했듯이 방법론 중 일부는 실시간 적용상

7) DSRC, UTIS의 경우 AVI와 달리 방향별 자료수집이 어렵기 때문에 특히 유출입이 많은 단속류 도로에서는 많은 이상치가 발생할 가능성이 높음

한계가 있고, 평균 및 표준편차를 사용함에 따라 강진기 방법과 동일한 오류를 범할 수 있다. 또한 샘플 수가 일정수준 이하일 경우 해당수집주기 자료 전체를 필터링함에 따라 교통량이 적은 지방부 도로 등에 적용시 한계점이 있다.

10. Ferguson 검정

1) 수식

$$\sqrt{b_1} = \sqrt{n}\sum_{i=1}^{n}(x_i-\bar{x})^3/[\sum(x_i-\bar{x})^2]^{3/2},\ \text{단측검정}\ (31)$$

$$b_2 = n\sum_{i=1}^{n}(x_i-\bar{x})^4/[\sum(x_i-\bar{x})^2]^2,\ \text{양측검정}\quad (32)$$

2) 설명

대부분 통계적 이상치 제거 방법은 표본집단 내에 하나 또는 두 개의 이상치를 제거하는 기법인데 반해 Ferguson-검정은 반복적으로 통계량($\sqrt{b_1}$, b_2)을 계산하여 표본집단 내 다수의 이상치를 제거하는 방법론으로 사용된다. 즉, 식 31 또는 32를 이용하여 계산된 $\sqrt{b_1}$ 또는 b_2가 표 1의 값을 초과할 경우 평균값으로부터 최대 이격된 관측치가 이상치로 판정되는데, 이 과정은 더 이상 이상치로 판정되는 관측치가 없을 때까지 반복된다.

표 1. Ferguson-검정 통계량

통계량	유의 수준	표본수(n)					
		5	10	15	20	25	50
$\sqrt{b_1}$	1%	1.34	1.31	1.20	1.11	1.06	0.79
	5%	1.05	0.92	0.84	0.79	0.71	0.53
b_2	1%	3.11	4.83	5.08	5.23	5.00	4.88
	5%	2.89	3.85	4.07	4.15	4.00	3.99

3) 장점

동일한 표본집단 내에 다수의 이상치가 존재하고 다수의 이상치 중 평균값과 크게 차이나는 (anomalous) 값(극단치)이 존재할 경우 극단치는 평균과 편차를 비정상적으로 이동(shift)시키고 이로 인해 증가된 신뢰구간 범위는 자칫 다른 이상치들을 유효치로 판정되게 한다. 그러나 Ferguson-검정은 이상치가 존재하지 않을 때까지 반복적으로 통계량을 계산하여 이상치를 제거하기 때문에 타 통계적 방법론에 비해 구간검지 자료 이상치 제거 방법론으로 적합하다고 할 수 있다. 또한 계산과정이 단순하여 실제 시스템 적용성이 우수하다고 할 수 있다.

4) 단점

Ferguson 검정 역시 도로공사 알고리즘과 같이 해당 수집주기 내 이상치만 존재할 경우 이상치를 유효치로 판정할 수 있다는 단점이 있고, 이상치보다 유효치가 적을 경우(또는 표본수가 적은 경우) 이상치를 유효치로 판정할 수 있다는 한계가 있다. 실제로 $\sqrt{b_1}$ 통계량은 이상치가 50%까지 존재할 때만 우수한 성능을 나타내고, b_2 통계량은 21%까지 존재할 때만 유효하다고 알려져 있다(ASTM, 2008).

11. 방법론 요약

이상에서 살펴본 바와 같이 현재까지 구간검지 자료 필터링 방법론은 주로 이전 수집주기 유효치(평균 등)를 이용하여 현재 수집주기의 유효치 범위를 정하거나 현재 수집주기 자료에 기반한 신뢰구간을 이용하여 유효치 범위를 정하는 방식으로 개발되었다. 이전 수집주기 값에 기반한 현재 수집주기 필터링 방법들은 주로 시스템 장애 복구 등에 따른 초기 수집주기 이상치 제거 또는 비첨두시 장애발생 후 첨두시 장애 복구 등에 따른 이상치 제거에 한계점을 내포하고 있고, 현재 수집주기 자료 기반 신뢰구간 방법론 역시 극단치 또는 해당 수집주기 내에 유효치보다 이상치가 많을 경우에 한계점을 드러낸다. 표 2는 각 방법론별 개요, 장단점을 정리한 것이다.

표 2. 구간검지 자료 필터링 방법론 요약

방법론	개요	장점	단점
강진기 방법론	극단치 제거 후 신뢰구간($\mu\pm\sigma$) 벗어난 관측치 제거	이해가 용이하고 수식이 간단	지나치게 큰 이상치(극단치)에 의한 신뢰구간 확대
TransGuide 알고리즘	이전주기 평균값±20%를 현재주기 유효범위로 설정	이해가 용이하고 수식이 간단	장애복구 등 발생에 따른 초기수집주기 이상치 제거 불가
도로공사 방법론	변동계수(σ/μ)에 따른 이상치 비율(5, 10, 15%) 등 적용	이해가 용이하고 수식이 간단	이상치 과다 제거, 적은 샘플 수에서 이상치 제거 한계, 이상치만 존재할 경우 한계
Dion 알고리즘	통행시간 분포가 Log-normal 분포를 따른다는 가정하에 평활화 평균, 표준편차를 이용한 신뢰구간법 적용	이론적 근거 명확, 적은 샘플 수에서 급격히 변하는 통행시간 패턴에서 우수	수식 복잡, 장애복구 등 발생에 따른 초기수집주기 이상치 제거 한계
장진환 알고리즘	이전주기 평활화된 평균값±30%를 현재주기 유효범위로 설정	국도 통행시간 패턴 적용성 우수	장애복구 등 발생에 따른 초기수집주기 이상치 제거 한계
Ma 알고리즘	현재주기 중앙값 및 표준편차에 의한 신뢰구간법 적용	장애복구 등 발생에 따른 초기수집주기 이상치 제거 가능	수집주기내 이상치만 존재할 경우 제거 불가, 표본수 6이하 적용 한계
Boxel 방법론	교통류 모형(속도-밀도)식의 신뢰구간에 따른 이상치 제거	장애복구 등 발생에 따른 초기수집주기 이상치 제거 가능	단속류 적용 한계, OBU 보급률에 따른 교통류 모형 지속적 업데이트 필요
Clark 방법론	중앙값과 사분위 편차를 이용한 신뢰구간법 적용	극단치 영향 최소화, 장애복구 등 발생에 따른 초기수집주기 이상치 제거 가능	1사 또는 3사분위수가 이상치일 경우(또는 이상치만 존재할 경우) 이상치 제거 한계
Haghani 방법론	평균과 표준편차를 이용한 신뢰구간법, 변동계수 등 적용	변동계수를 사용함에 따라 이상치만 존재할 경우에도 필터링 가능	극단치에 의한 신뢰구간 확대
Ferguson 검정	단측, 양측검정별 통계량의 반복적 계산을 통해 이상치 제거	극단치 영향 최소화, 장애복구 등 발생에 따른 초기수집주기 이상치 제거 가능	이상치만 존재할 경우 또는 이상치보다 유효치가 적은 경우 한계

Ⅲ. 필터링 방법론 개발방향

전 장에서 고찰한 바와 같이 아쉽게도 현재까지 구간검지 자료 필터링 방법 중 모든 구간검지 데이터 유형에 적합한 방법론은 없는 실정이다. 따라서 향후에는 실제 구간검지기 운영상 나타날 수 있는 구간검지 데이터 유형에 적합한 보편적인 방법론 개발이 필요할 것으로 사료되는 바, 본 장에서는 이에 대한 방향을 제시하고자 한다.

표 3은 구간검지 시스템(현장장비, 통신, 센터 시스템) 운영상 나타날 수 있는 데이터(이상치) 유형을 정리하였고 각각의 이상치에 대한 적정 처리방안을 제시하였다.

실제 ITS에서는 많은 장애가 발생하고 이에 대한 장애복구가 이루어진다. 장애 유형에 따라 몇 시간 이내 복구가 이루어지기도 하지만 몇 주의 기간이 소요되기도 한다. 이러한 경우, 과거 수집주기에 기반한 현재 수집주기 이상치 제거 방안은 교통류 패턴 변화, 첨두/비첨두 등의 영향으로 적절하지 않은 것으로 판단되는 바 현재 수집주기 자료에 기반한 이상치 제거가 합당해 보인다.

한편, 교통량이 적은 지방부 국도의 경우 특히 새벽시간대에는 수집되는 샘플이 극히 적거나(1 또는 2) 해당 수집주기 내에 샘플이 존재하지 않는 경우도 발생하고 더욱이 수집되는 자료가 모두 이상치인 경우도 종종 발생한다. 이러한 경우에는 현재 수집주기 자료를 이용하여 이상치를 제거하기가 사실상 불가능하므로 과거 수집주기에 기반

표 3. 구간검지 데이터 유형별 필터링 방법론 개발방안

유형	설명	적정 처리방안
Ⅰ	시스템 장애복구 등에 따른 초기 수집주기 내 이상치 존재	해당 수집주기 내의 자료를 이용한 이상치 필터링
Ⅱ	비첨두시 시스템 장애 후 첨두시 시스템 복구(또는 그 반대 상황)	
Ⅲ	적은 교통량 상태에서 급격한 통행시간 변화 (주로 새벽 시간대에서 오전피크 시로 전이시간)	해당 수집주기 내의 자료를 이용한 이상치 필터링 또는 샘플수/유효샘플이 없는 수집주기 개수 등 반영 가능한 파라미터 적용
Ⅳ	정상범위에서 과다하게 벗어난(또는 타 이상치에 비해 상대적 변동성이 큰) 이상치 존재	과거 수집주기 유효치에 기반한 현재 수집주기 유효치 범위 설정 또는 중심위치 척도를 중앙값으로 사용
Ⅴ	해당 수집주기 내에 이상치만 존재(주로 샘플 수가 3 미만인 경우 발생)	과거 수집주기 유효치에 기반한 현재 수집주기 유효치 범위 설정 또는 수집자료 변동성 등에 따른 수집주기 전체 자료를 이상치로 판정
Ⅵ	해당 수집주기 내에 유효치보다 이상치가 많이 존재	

한 현재 수집주기 유효치 범위를 설정하여 이상치를 필터링 하는 방안이 적절해 보인다.

마지막으로 UTIS, DSRC 등과 같이 방향별 데이터 수집이 불가능한 구간검지 체계에서는 해당 수집주기 내에 수집자료 많은(10개 이상) 경우라 하더라도 유효치보다 이상치가 더 많이 존재하는 경우가 드물긴 하지만 발생한다. 이러한 경우, 현재 수집주기 자료를 이용한 신뢰구간 등을 이용하여 이상치를 필터링 할 경우 자칫 이상치를 유효치로 판정하고 유효치를 이상치로 판정하는 경우가 발생할 수 있다. 그러나 다행히도 이상치는 유효치에 비해 변동성이 큰 경우가 대부분이기 때문에 이러한 변동성 지표를 이상치 판정로직이 추가하는 것이 필요할 것으로 사료된다.

Ⅳ. 결론

본 고에서는 최근 들어 무선통신 기반 차세대 ITS 도입, 하이패스 단말기 보급률 확대 등으로 점차 구축 영역이 확대되고 있는 ITS 구간검지 교통정보시스템에서 교통정보 신뢰도 확보를 위한 구간검지 자료 필터링 방법론에 대해 고찰하였다.

구간검지 방법론은 크게 두 가지 형태로 구분되는데 첫째는 과거자료를 기반으로 한 현재 수집주기 유효치 범위를 설정한 후 유효범위 밖의 데이터를 필터링하는 방법이고, 두 번째는 현재 수집주기 자료를 이용하여 신뢰구간을 도출하거나 신뢰구간을 나타내는 통계량을 구한 후 일정 신뢰구간 범위 이외의 자료를 필터링 하는 방법이다. 그러나 각 방법론별로 장・단점이 존재할 뿐 현재까지 모든 구간검지 데이터 유형에 적합한 방법론은 없는 실정이다. 따라서 본 고에서는 실제 ITS 운영시 나타날 수 있는 구간검지 데이터를 유형별로 정리한 후 각 유형에 적합한 필터링 개발방안도 병행하여 제시하였다.

본 고는 기 구축・운영 중인 구간검지 교통정보 시스템의 성능개선, 향후 구축 예정 시스템의 교통정보 신뢰성 확보, 더 나아가 실용성이 우수하면서 보편적으로 활용 가능한 구간검지 필터링 방법론 개발에 긴요하게 활용될 수 있을 것으로 기대된다.

참고문헌

강진기, 손영태, 윤여환, 변상철 (2002), 비매설식 자동차량인식장치를 이용한 구간교통정보 산출 방법 연구, 한국ITS학회논문집, 제1권 제1호, 한국ITS학회.

장진환, 변상철, 백남철, 김성현 (2005), AVI 자료 필터링 알고리즘 개발(일반국도를 중심으로), 대한토목학회논문집, 제25권 제2D호, 대한토목학회.

한국도로공사 (2008), DSRC를 활용한 도로교

통정보 검지시스템 실용화 기술개발.

ASTM International (2008), Standard Practice for Dealing with Outlying Observations, Designation: E 178-08.

Boxel D. V., Schneider IV W. H., Bakula C. (2011), An Innovative Real-Time Methodology for Detecting Travel Time Outliers on Interstate Highways and Urban Arterials, TRB 2011 Annual Meeting CD-ROM.

Clark S. D., Grant-Muller S., Chen H. (2002), Cleaning of Matched License Plate Data, Transportation Research Record No.1804.

Dion F., Rakha H. (2006), Estimating Dynamic Roadway Travel Times Using Automatic Vehicle Identification Data, Transportation Research Part B, Elsevier.

Haghani A., Hamedi M., Sadabadi K. F., Young S., Tarnoff P. (2010), Data Collection of Freeway Travel Time Ground Truth with Bluetooth Sensors, Transportation Research Record No. 2160.

Kendal M. G., Stuart A. (1973), The Advanced Theory of Statistics, Volume 1: Distribution Theory, Edward Arnold, London.

Ma X., Koutsopoulos H. (2010), Estimation of the Automatic Vehicle Identification Based Spatial Travel Time Information Collected in Stockholm, IET Intelligent Transport Systems Vol.4, Issue 4.

Southwest Research Institute (1998), Automatic Vehicle Identification Model Deployment Initiative-System Design Document, Report prepared for TransGuide, Texas Department of Transportation.