大 韓 土 木 學 會 論 文 集
第26卷 第5D號 · 2006年 9月
pp. 731~738

링크통행시간 생성을 위한 이상치 제거 알고리즘 개발

A Heuristic Outlier Filtering Algorithm for Generating Link Travel Time using Taxi GPS Probes in Urban Arterial

최기주* · 최윤혁**
Choi, Keechoo · Choi, Yoon-Hyuk

Abstract

Facing congestion, people want to know traffic information about their routes, especially real-time link travel time (LTT). In this paper, as a sequel paper of the previous non-taxi based LTT generating study by Choi et al. (1998), taxi based GPS probes have been tried to produce LTT for urban arterials. Taxis in itself are good deployment mode of GPS probes although it by nature experiences boarding and alighting time noises which should be accounted. A heuristic real-time dynamic outlier filter algorithm for taxi GPS probe has been developed focusing on urban arterials. An actual traffic survey for dynamic link travel times has been conducted using license plate method for the test arterials of Seoul city transportation network. With the algorithm, it is estimated that 70% of outliers have been filtered and the relative error has been improved by 73.7%. The filtering algorithm developed here would be expected to be in use for other spatial sites with some calibration efforts. Some limitations and future research agenda have also been discussed.

Keywords: *link travel time, dynamic outlier filter algorithm, GPS probe*

요 약

교통 혼잡이 증가하면서, 경로에 대한 교통정보, 특히 실시간 구간통행시간에 대한 사람들의 관심이 증대되고 있다. 본 논문은 GPS Probe를 통해 구간통행시간을 산출했던 최기주(1998) 등의 후속 연구로써, 도시부에서 구간 통행시간을 산출하기 위해 택시를 GPS Probe로 활용하였다. 택시는 GPS Probe로 활용되기 위한 매우 좋은 수단이지만, 승객의 승하차시간 등 주행과 관계없는 불필요한 데이터가 포함되게 된다. 따라서 본 논문에서는 도시부에서 Taxi GPS를 통해 교통정보를 생성할 경우 주행과 관계없는 정보를 실시간으로 검지하여 제거하는 휴리스틱한 이상치 제거 알고리즘을 개발하였다. 평가를 위해 서울시 주요 간선축에서 번호판 조사를 실시하였으며 알고리즘을 적용한 통행시간과 비교하였다. 이상치 제거 알고리즘을 적용한 결과, 약 70%의 이상치가 제거되었으며, 실측 통행시간과의 상대 오차가 73.7%로 향상된 것으로 나타났다. 따라서 본 알고리즘을 이용할 경우 Taxi GPS를 통해 신뢰할 수 있는 실시간 교통정보를 생성할 수 있을 것으로 판단된다.

핵심용어 : 링크 통행시간, 이상치 제거, GPS Probe

1. 서 론

ITS가 도입된 이래로 교통시설의 효율적인 운영은 신뢰성 있는 교통정보를 생성하고 제공하는 것에 좌우되고 있으며, 정보화시대를 맞이하여 교통정보, 특히 구간통행시간에 대한 많은 사람들의 관심과 요구가 높아지고 있다. 교통정보를 생성하기 위한 정보수집체계는 크게 지점 검지체계와 구간검지체계로 나눌 수 있다. 이 중에서 구간검지체계는 일반 운전자들이 원하는 통행시간정보를 생성하기 위해 중요한 역할을 하고 있어, 이에 대한 많은 연구가 선행되었다. 특히 실시간 교통정보의 수집/제공을 위한 (비용-효과측면에서의) GPS의 우수성은 많은 연구결과와 사례를 통해서 입증되었으며, 최기주(1998) 등은 본 논문의 초석이 되는 GPS Probe를 통한 구간통행시간 산출에 대한 연구를 수행하였다.

그러나 구간검지체계에 대한 이러한 관심에도 불구하고, 기존 연구의 초점은 주로 고속도로 및 간선도로 등의 연속류에 대한 연구가 대부분이었으며, 도시부에서 GPS Probe를 이용한 구간 교통정보 생성에 대한 연구가 부족한 상황에서 최근 이에 대한 연구들이 수행되고 있다.(정재영 2001, 신강원 2003)

최근 교통정보 수집을 위한 GPS Probe로 일반 차량이 아닌 버스 및 택시 등의 대중교통 수단을 이용하는 방법이 시도되고 있는데, 이는 신호 교차로가 존재하는 단속류의 도시부에서 일반 차량을 통해 도시 전체의 교통정보를 수집하기 위해서는 너무나 Probe가 필요하여, 실시간으로 교통정보

*정회원 · 아주대학교 환경건설교통시스템공학부 교통공학과 교수 (E-mail : keechoo@ajou.ac.kr)
**정회원 · 아주대학교 일반대학원 건설교통공학과 박사과정 (E-mail : yoonhyuk@ajou.ac.kr)

를 수집하기가 현실적으로 불가능하기 때문이다. 이에 비해 대중교통수단인 버스 및 택시 등은 승객을 태우기 위해 도시 전체를 계속 운행(혹은 순환)하기 때문에 일반 차량보다 적은 Probe 대수로도 보다 넓은 지역을 커버할 수 있다.

또한 버스의 경우 차량에 장착된 GPS 시스템이 버스정보 안내시스템, 버스배차관제시스템 등으로 함께 사용될 수 있어 버스에 대한 승객의 서비스를 높이고 동시에 버스 회사의 이익이 증가할 수 있는 장점이 존재한다. 택시의 경우 GPS를 이용하여 승객의 호출 지점에서 가장 가까운 위치에 존재하는 택시를 연결하여 승객에게는 빠른 서비스를 제공하고 이를 통해 보다 많은 승객을 태울 수 있어 택시 회사의 이익도 함께 증가하는 시너지 효과가 발생한다.

실제로 외국에서는 버스 GPS Probe를 이용해 구간의 교통정보를 생성하는 연구가 수행되었으며(Daily 1999), 이는 버스의 경우 일정한 배차간격을 유지하여 버스 노선이 통과하는 도로에 대해 높은 유효 수집율[1]을 가지기 때문이다.

그러나, 지금까지 외국에서도 택시 GPS Probe를 이용한 교통정보 생성에 대한 연구가 수행된 경우가 없는데, 이는 외국의 경우 택시가 우리나라처럼 많지 않아 Probe로 사용되기에 적절하지 않기 때문이라고 판단된다.

이에 비해 우리나라의 경우 택시의 수가 많고 (거의) 하루 종일 운행되기 때문에 구간 교통정보 생성을 위한 GPS Probe로 사용되기에 아주 좋은 시장 여건을 가지고 있으며, 실제로 (주)로티스 및 SK(주) 엔트랙 등이 택시를 Probe를 이용해 통행시간 정보를 제공하고 있다.

그러나, 버스 및 택시 등의 Probe를 통해 산출된 통행시간은 일반 차량의 통행시간과는 달리 주행과는 승하차 시간 등이 포함되어 있어, 이를 아무런 여과 없이 링크의 대표 소통정보[2]로 사용하기에는 무리가 있다. 특히 버스의 경우는 대부분의 링크마다 한 개 이상의 정류장이 존재하여 정류장의 위치에 따라 일반 차량의 통행시간과 차이가 매우 크기도 하다.

그에 비해 택시의 경우는 링크 중간에서의 승하차시간, 승객대기시간 및 개인용무시간 등을 제외하면 일반 차량의 통행시간과 유사하다고 판단되는 바, 본 논문에서는 도시부에서 택시 GPS Probe를 이용하여 교통정보를 생성할 때 주행과 관계없는 승객을 위한 서비스 시간 때문에 일반 차량의 통행시간과 다르게 생성되는 이상치를 검지하여 제거하는 연구를 수행하였으며, 본 연구를 위해 사용된 데이터는 SK(주) 엔트랙의 것을 사용하였다.

2. GPS Probe와 이상치

2.1 GPS와 GPS Probe

GPS(Global Positioning System)는 미 국방성에서 자국의 군사 목적을 위하여 개발한 것으로 지구상 어디에서나 기후에 구애받지 않고, 표준 좌표계에서 위치, 속도, 시간측적을 가능하게 해주는 인공위성을 이용한 첨단항법시스템이다. (wooden, 1985) GPS 위성은 모두 24개로 구성되는데, 이 중 21개가 항법에 사용되며, 3개의 위성을 예비용으로 배치된다. 모든 위성은 고도 20,200km 상공에서 11시간 58분을 주기로 지구 주위를 돌고 있으며, 궤도면은 지구의 적도면과 55°의 각도를 이루고 있다. 60°씩 떨어진 총 6개의 궤도가 있으며, 한 궤도면에는 4개의 위성이 위치한다. 이와 같이 GPS 위성을 지구 궤도상에 배치하는 것은 지구상 어느 지점에서나 동시에 5개에서 8개까지의 위성을 볼 수 있게 하기 위함이다.

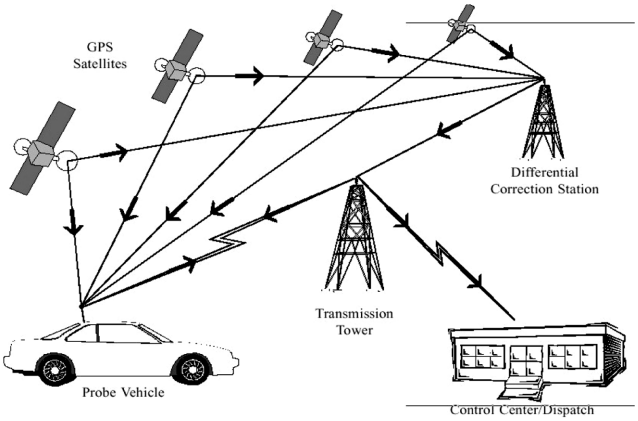

그림 1. 일반적인 GPS Probe 시스템 프로세스

GPS Probe는 Probe 차량에 GPS 수신기를 장착하고 GPS 위성으로부터 신호를 수집하기 위해 양방향 통신을 수행한다. GPS 신호로부터 Probe 차량의 실시간 위치좌표가 결정되고, 이를 통해 통행시간 정보가 산출된다.

2.2 택시 GPS Probe의 특징

전술하였듯이 버스 및 택시 등의 대중교통 GPS Probe를 이용하여 구간의 교통정보를 생성할 경우 일반 차량을 이용한 경우보다 많은 지역과 넓은 시간대에 교통정보를 수집할 수 있다.

버스의 경우는 노선이 고정되어 있어 배차간격에 따라 노선이 포함된 링크의 교통정보를 주기적으로 얻을 수 있으므로, 유효 수집율이 높은 장점이 존재하지만, 고정된 노선 이외의 지역에서는 교통정보를 수집할 수 없는 단점이 존재한다. 그러나, 버스를 통해 교통정보를 얻는다고 하더라도 버스 승강장에서의 승/하차 시간이 계속 존재하며, 버스전용차로, 회전 등의 영향이 크게 작용하므로 본 논문에서는 이를 고려하지 않기로 한다.

택시의 경우 유동 인구가 많은 특정 지역에 밀집되는 현상을 보이며, 승객을 위한 대기시간 및 개인 용무 시간이 발생하는 단점이 존재하지만, 고정된 노선이 없어 다양한 지역에서 교통정보 수집이 가능하고 거의 하루 내내 교통정보를 생성할 수 있으며, 일반 차량과 비교할 경우 네트워크 내에서의 데이터 분포측면에서 유리한 장점을 가지고 있다.

택시의 통행시간이 일반차량의 통행시간과 다른 주요 원인이 되는 대기시간 및 개인 용무 시간이 모든 택시에서 항상 발생하는 것이 아니며 일정한 시간대와 특정한 공간에 상관없이 간헐적으로 발생하여, 이와 같은 이상치를 제거하면 택

1) 유효 수집율은 통상적으로 사용되는 교통정보의 제공주기인 5분 동안에 수집되는 Probe 데이터의 수집율을 뜻한다.

2) 링크의 대표 소통정보는 링크를 주행하는 차량의 70~80%을 차지하여 링크를 대표할 수 있는, 일반 차량을 이용해 생성되는 통행시간 정보이며, 이는 또한 정보의 주요 수혜자인 일반 차량 운전자들에게 제공된다.

표 1. 일반 차량과 택시 Probe의 비교

항목	일반 차량	택시
Spatial Coverage	집 · 직장의 일부 노선	다양한 노선
Temporal Coverage	출퇴근 시간대를 비롯한 일부 시간	거의 24시간 내내

시 Probe를 이용해 산출된 통행시간을 링크의 대표 통행시간으로 사용해도 무리가 없다고 보인다. 따라서, 택시 GPS Probe는 고정된 노선이 없어 다양한 지역에서 거의 24시간 내내 교통정보 수집이 가능한 특징을 가지고 있으나, 택시 GPS를 통해 교통정보를 생성할 경우 주행과 관계없는 정보가 포함될 가능성이 있으므로 이를 제거해야 한다.

2.3 이상치의 정의 및 발생원인

택시 GPS Probe를 이용하는 목적은 운전자에게 제공하기 위한 교통정보를 생성하기 위한 것으로, 하나의(혹은 다수의) 링크(혹은 구간)를 통과하는데 걸리는 통행시간을 알고자하는 것이다. 그러나 택시의 경우 정상적으로 링크를 통과하지 않고, 승객의 승하차 등의 시간이 포함되는 경우가 있으며, 이는 정상적인 통행시간이 아닌 잘못된 통행시간, 즉 이상치가 되는 것이다.

즉, 본 연구에서 사용하는 이상치(Outlier)는 "Taxi GPS Probe를 통해서 교통정보를 수집할 경우 (일반차량의 통행시간을 기준으로 하여) 주행과 관계없는 정보가 포함된 통행시간 데이터"중에서 "그것의 존재나 누락이 추정하려는 통계적 모수에 많은 영향을 끼치는 데이터"라고 정의될 수 있으며, 그 발생 원인으로는 링크중간에서의 승객의 승 · 하차를 위한 정지, 택시 승강장에서의 승객 대기, 개인 용무, 링크 중간에서의 이면도로 통행 등이 있다.

2.4 이상치 제거의 필요성

이상치는 통계적으로 자료계열에서 관측치들의 대부분에 의해 제시된 형태를 이루지 못하는 관측치이다. 만약 자료에 이상치가 존재하면 그것은 검정 통계량을 무효화시키고, 모수 추정을 왜곡시키며, 틀린 통계적 추론을 유도한다. 평균은 산술평균을 뜻하는 것으로 자료의 합을 자료의 개수로 나눈 값이며, 이러한 평균값은 자료 1개의 영향력이 모두 동일하게 나타난다. 이와 같이 평균은 소수의 크거나 적은 이상치에 의해 크게 영향을 받기 때문에, 특히 평균을 통해 대표되는 값을 제시할 경우 이상치 제거가 매우 중요하다.

따라서 현재와 같이 택시 GPS를 통해 수집되는 데이터들을 평균하여 교통정보를 제공할 경우 이상치에 대한 효과가 매우 크게 나타날 수 있다. 특히 GPS 방식의 경우 주기내 수집되는 데이터수가 적기 때문에 1개 이상치가 갖는 영향력이 매우 커지게 된다. 게다가 Probe로 사용되는 차량이 일반 차량이 아닌 대중교통수단인 택시일 경우 승객의 서비스 시간으로 인한 이상치가 발생할 확률이 높기 때문에 이상치 검지가 필수적이라고 할 수 있다.

3. 이론적 배경

이상치를 검지하는 방법으로 가장 일반적인 것은 통계적인 방법을 이용하는 것이다. 통상적으로 이상치 검지에 사용되는 통계적인 방법들은 Box-Plot방법, 첨도 통계량(Kurtosis Statistics) 방법, Shapiro-Wilk 검정법, Dixon-Type 검정법, 그리고 Grubb's Test라고 불리는 ESD(Extreme Studentized Deviate) 검정법 등이 있다.

3.1 Box-Plot방법

Box Plot 방법은 Devore와 Peck(1986)에 개발되었으며, 통계패키지에서 이상치 탐색 방법으로 많이 사용되고 있다. 이 방법은 순서화된 자료에서 4분위수를 구하여 F_L을 제 1사분위수, F_U를 제 3사분위수라고 하며, 구간 $(F_L - k\times(F_U - F_L))$, $(F_U + k\times(F_U - F_L))$ 의 바깥쪽에 떨어지는 관측치를 이상치로 간주한다. 이 구간은 k값에 따라 변하는데, k=1.5와 k=3.0일 때가 Box Plot(Emerson and Strenio 1983, Tukey 1977) 모형의 전형적인 형태이다.

Box Plot의 주된 핵심은 중위수, 제 1사분위수(F_L), 제 3사분위수(F_U)이며, 박스는 중앙선, 중위수(F_L,F_U)를 포함한다. 범위선(fences)로 알려진 절단점은 제 1사분위수 아래와 제 3사분위수에 놓여있으며 범위선을 넘어서는 관측치를 이상치라고 판단한다.

비록 Box Plot 방법이 관측치를 분류하는 가장 효율적인 방법은 아니지만, 이 방법은 오염된 정규분포처럼 이상치가 분포의 양쪽 끝에 존재할 경우에는 효과적이다.

3.2 첨도 통계량(Kurtosis Statistics) 방법

표본 첨도는 정규성으로부터 벗어난 측도와 관측치를 검정하는데 사용되고 있으며, 이 통계량은 계산과 사용이 아주 간편하다. KS방법은 적절한 n과 a와 기각값을 비교하였는데, D'Agostino and Tietien(1971)은 표본크기가 7≤n≤50이고, a =0.01, 0.05, 0.1인 경우의 기각값을 제시하였다. N>50 보다 큰 경우 KS 백분점은 Pearson and Hartley(1966)에 나타나 있으며, 관측치 x가 이상치로 판단되면, 이 관측치를 제거하고 그 절차는 계속해서 반복한다.

이 방법의 단점은 인접한 이상치가 존재할 때 가려진다는 것이며, 이를 피하기 위해 Tain(1981)는 KS를 사용할 때 일반화 ESD 방법과 유사한 절차와 검정을 위한 적절한 기각값을 제시하였다.

3.3 Shapiro-Wilk 검정법

Shapiro와 Wilk는 이상치를 검정하는데 의미있는 특징을 가지는 정규성을 위한 검정법을 소개하였다. 그들은 $\sum_{i=1}^{n} c_i X_{(i)}$ 의 형태를 가지는 σ의 추정량에서 $N(\mu, \sigma^2)$ 인 정규분포로부터 표본의 순서통계량 $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ 을 결합하는 방법에 대해서 이론적 체계를 세웠다.

W 검정으로 알려진 정규성 검정은 $(n-1)S^2$ 로 인식하는 다른 추정량과 $b^2, D = \sum_{i=1}^{n} (X_i - \overline{X})^2$ 을 비교하며 이상치 검정 절차는 다음과 같다.

3) 자료의 성격에 따라 모평균 또는 표본평균으로 불림

1. $h=\begin{Bmatrix} n/2 & (n:\text{짝수}) \\ (n-1)/2 & (n:\text{홀수}) \end{Bmatrix}$를 정의한 다음 상수 a_{n+1-i} (i= 1, 2, …, h)를 구하고, $b=\sum_{i=1}^{n} a_{n+1-i}(X_{(n+1-i)}-X_{(i)})$를 계산한다.

2. $D=\sum_{i=1}^{n}(x_i-\bar{x})^2$을 계산한다.

3. $W=b^2/D$를 계산한다.

4. 만약 W>C이면 이상치가 존재하지 않는다고 판정하며, 이때 기각값 C는 Shapiro(1986)와 Barnett and Lewis (1984)가 계산한 값을 이용한다. 그렇지 않으면 $\bar{x}$로부터 가장 벗어난 관측치를 이상치로 간주한다. 이 관측치를 제거하고 감소된 표본에서 이와 같은 과정을 반복하여 이상치를 제거한다.

이 방법은 n값에 따라서 상수 h가 영향을 받기 때문에 이상치를 검지하는데 문제성을 가지고 있다.

3.4 Dixon-Type 검정법

Dixon Type 검정은 순서화된 표본 일부의 범위의 비를 나타내며, 이것은 유동적인 구조로 인해 의심스러운 이상치의 구체적인 패턴을 찾기 위해 설계된 많은 방법을 야기하고 있다.

1. 오른쪽에 위치한 하나의 이상치 검정법

다음 식 $r_{(11)}=\frac{x_{(n)}-x_{(n-1)}}{x_{(n)}-x_{(2)}}>\lambda_{11}$을 만족하는 $x_{(n)}$이 이상치이다. 일반적으로 $r_{j,k-1}$형태의 검정은 1보다 크지 않는 어떤 j와 k의 $x_{(n)}$-$x_{(k)}$에 의해 $x_{(n)}$-$x_{(n-1)}$을 나눈다. 표본의 왼쪽 끝에 있는 이상치에 의해서 야기되는, 가려지는 잠재적인 효과를 예방하기 위해 k=2를 선택한다. 왼쪽에 있는 한 개의 이상치는 각각의 관측치를 처음으로 음수를 택하는 것으로 동일하게 검정될 것이며 이는 를 포함하는 로 놓기 때문이다.

2. 오른쪽에 있는 두 개의 이상치를 위한 검정법

다음 식 $r_{(21)}=\frac{x_{(n)}-x_{(n-2)}}{x_{(n)}-x_{(2)}}>\lambda_{21}$을 만족하는 $x_{(n)}$, $x_{(n-1)}$이 이상치이다. 왼쪽에 있는 두 개의 이상치를 위한 검정은 각각의 관측치를 음수로 만드는 작업이다.

3. 양쪽 끝에 있는 하나의 이상치 검정

가장 작을 뿐만 아니라 가장 큰 관측치를 이상치로 판명하는 법칙을 이용하여 이상치를 판정하며 그 식은 다음과 같다.

$$r'_{(10)}=\max\left[\frac{x_{(n)}-x_{(n-1)}}{x_{(n)}-x_{(1)}},\frac{x_{(2)}-x_{(1)}}{x_{(n)}-x_{(1)}}\right]>\lambda'_{10}$$

Dixon Type 검정법은 사용하기는 매우 쉬우나 결과는 의심이 가는 이상치의 위치와 정확한 수를 올바르게 선택하는 것에 크게 의존하고, 가려지는 현상과 은닉되는 현상에 매우 민감하다는 단점이 있다.

3.5 ESD(Extreme Studentized Deviate) 검정법

ESD(Extreme Studentized Deviate)검정은 선형 모형에서 이상치를 분류하기 위해서 잘 알려진 검정법이다.(Weisbeg 1985) 표본이 하나인 경우의 이 검정법은 $t_i+\{n(n-1)\}^{1/2}\frac{x_i-\overline{x_{-1}}}{s_{-i}}$ 이다.[4] Rosner의 일반화 ESD 방법(Rosner 1983)은 가려지는 효과에 의해서 의심이 가는 것을 극복하기 위해서 ESD 검정법을 확장한 것이며, 이는 m번 추출한 후에 이상치들의 상한값이 절차에 따라 분류된다.

모든 표본으로부터 시작하여 $R_1=(\max|x_i-\overline{x_{-1}})/s$ 라 둔다. R_2는 R_1에 대응하는 경우를 제외한 후에(즉, n 1개의 표본을 가지고) R_1과 같은 방법을 이용한다. 기각값 λ_i는 다음과 같다.

$$\lambda_i=\frac{t^*\alpha,_{n-i-1}(n-j)}{[(n-i-1+t^{2*}\alpha,_{n-i-1})(n-i-1)]^{1/2}}$$

여기서, $\alpha^*=(\alpha/2)(n-i-1)$

$$M=\begin{cases} \text{만약 } R_i\le\lambda_i; i=1,2,\ldots,m \\ \max(i:R_i>\lambda_i,\ \text{그 외의 경우} \end{cases}$$

만약 M=0이면 이상치는 없는 것으로 판명하며, M≥1이면 $R_1,R_2,\ldots,R_m$에 대응하는 M인 경우가 이상치로 판명된다. 환언하면, ESD 검정법 Ts를 수정하여 m개의 이상치가 존재할 때의 검정법으로 확장할 수 있다.

3.6 통계적 이상치 검지방법 적용의 한계

그러나, 이러한 통계적인 이상치 검지 방법들은 기본적으로 모집단 및 표본 집단의 데이터가 정규분포[5]라는 전제가 바탕에 깔려있으며, 일정한 구간에서 수집되는 통행시간 데이터는 신호의 영향을 받기 때문에 적색신호에 대기하지 않은 차량군과 적색신호동안 대기한 차량군으로 분리되어 정규분포를 나타내지 않는다.

아래 그림 1은 수원시 1번 국도상의 교차로인 "수원시청사거리~권선사거리"를 오후 4:50부터 6:30까지 번호판 조사방법을 통하여 조사한 구간 통행시간을 그래프로 그린 것으로, 신호연동에 의한 구간 통행시간의 이분화가 정확하게 나타나고 있다.

또한 일반적으로 교통정보를 생성하여 제공하는 주기인 5분 동안의 통행시간 변화를 보기 위하여 한 주기의 통행시간을 분석하였다.

주기내 통행시간의 정규성을 알아보기 위하여 <그림 2>와 같이 히스토그램을 그린 결과 링크의 통행시간은 단일봉(Uni-modal)의 형태가 아닌 쌍봉(Bi-modal) 혹은 다봉(Multi-modal)의 형태를 나타내었으며, 정규성을 띄지 않았다.

이렇듯 5분 주기의 통행시간 데이터가 정규성을 띄지 않

4) 여기서 $\bar{x}$와 s는 표본 평균과 표준 편차를 나타내고, $\bar{x}_{-1}$과 s_{-i}는 i번째 데이터를 제거한 후 대응되는 양을 나타내며, 정규성 가정하에서 t_i는 자유도 t-2인 t분포를 따른다.

5) 혹은 정규분포보다 약간 편향된 오염된 정규분포

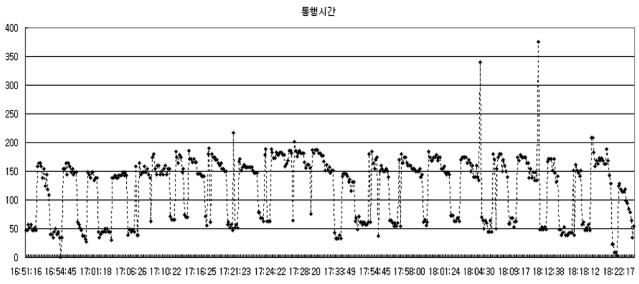

그림 2. 신호연동에 의한 통행시간의 이분화 현상

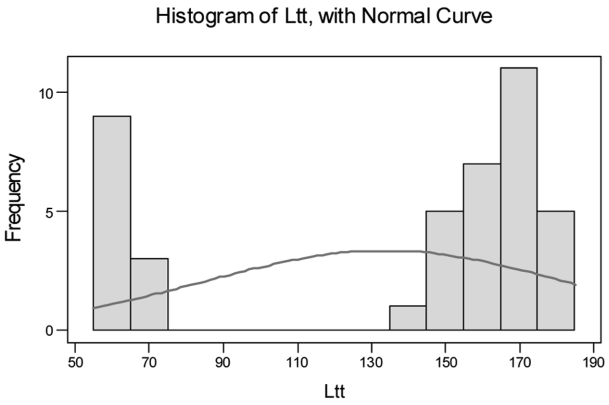

그림 3. 주기 통행시간의 정규성 분석

을 경우, 통계적인 방법을 이용하여도 이상치가 제거되지 않는다. 특히, 첨도 통계량방법, Shapiro-Wilk 검정법 등은 데이터의 정규성에 대한 가정을 전제하고 하고 있으며, Dixon-Type 검정법은 이상치로 의심되는 데이터의 위치와 개수를 정확하게 알고 있느냐에 따라 검정의 결과가 달라지는 현상이 발생한다.

또한, GPS 방식으로는 주기내 수집되는 Probe 수의 한계가 존재하여, 적은 데이터로 통계적으로 이상치를 검지하는 무리가 있으며, Box-Plot 방법, ESD 검정법의 경우 데이터수가 적은 경우 치명적인 약점이 존재한다. 따라서 본 논문에서는 통계적인 방법을 이용하지 않고 실시간으로 이상치를 검지하여 제거하는 휴리스틱한 알고리즘을 개발하고자 하였다.

4. 알고리즘 개발

4.1 알고리즘의 개요

이상치 제거 알고리즘의 기초는 이상치가 가지고 있는 특징에서 비롯되었으며, 이상치는 일반 차량의 정상적인 데이터보다 통행시간이 높은 특징을 가지고 있다. 따라서 이상치를 제거하기 위해서는 시계열적인 통행시간의 변동을 고려한 현재 주기에 나타날 수 있는 최대 통행시간 기준값이 필요하다.

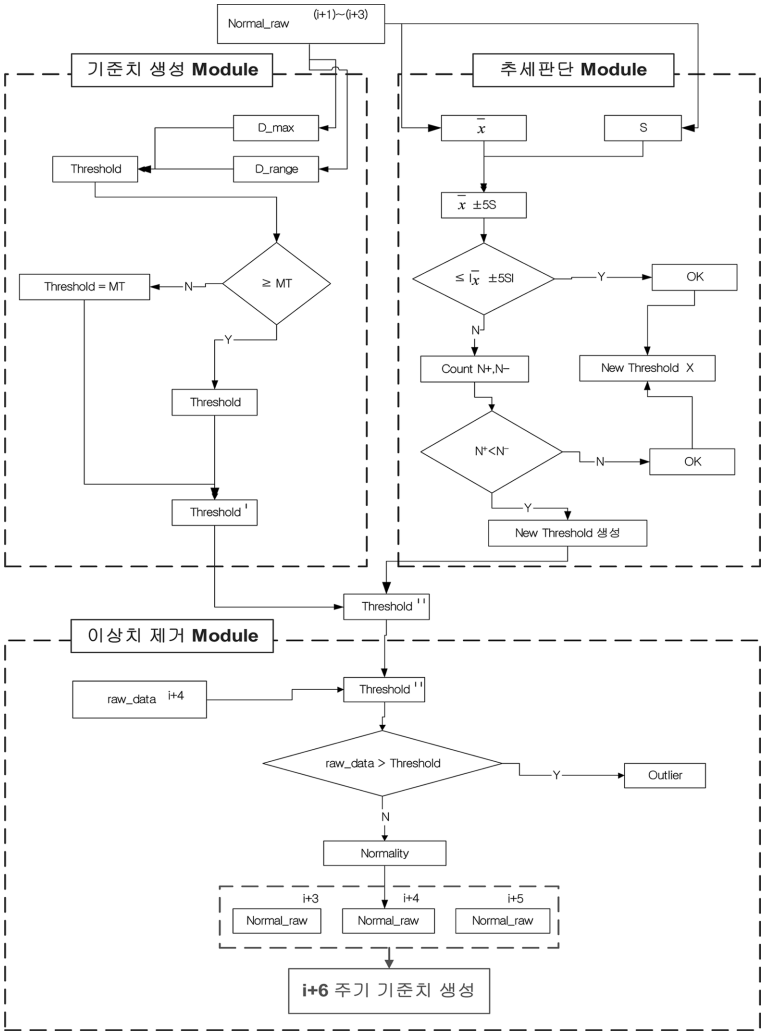

그림 4. 이상치 제거 알고리즘의 흐름도

그러나, 알고리즘은 정상적인 데이터를 이상치로 판단하는 오류를 범해서는 안된다. 왜냐하면 통행시간이 높은 데이터는 이상치일 수도 있지만, 혼잡으로 인해 통행시간이 증가하여 발생한 정상치일 수도 있기 때문이다. 특히 이상치의 특징이 높은 통행시간이므로 혼잡상황이 발생한 것과 동일한 특징을 나타내기 때문에 이를 구별하는 것이 중요하다.

4.2 알고리즘의 구성

이상치 제거 알고리즘은 혼잡의 발생을 인식할 수 있어야 하기 때문에 알고리즘의 핵심은 과거 일정시간 동안의 통행시간의 최대값과 혼잡의 발생을 고려한 오차항(Error Term)이 포함된 현재 주기의 최대 통행시간의 기준값을 생성하는 것이다. 따라서 현재 주기 최대 통행시간의 기준값은 과거 일정기간 통행시간의 최대값과 혼잡의 발생을 고려한 오차항을 더한 값으로 결정될 수 있다.

현재 주기의 최대 통행시간의 기준값을 생성하기 위한 적절한 과거 일정기간의 통행시간의 최대값을 결정하기 위해 시행착오법을 이용하여, 5분, 10분, 15분의 생성시간을 서로 비교하는 시뮬레이션을 실시하였으며, 혼잡에 민감한 알고리즘을 위해서는 이전 3주기(15분)가 기준값 생성을 위한 적절한 시간으로 결정되었다. 혼잡 및 돌발상황을 고려한 오차항(Error Term)을 결정하기 위해 이전 3주기의 최대통행시간을 기준으로 하여 통행시간의 범위(Range)와 3표준편차를 다음 주기의 최대 통행시간과 비교하였으며, 그 결과 통행시간의 범위를 이용한 방법이 혼잡 및 돌발상황을 고려하기에 적합한 것으로 나타났다.

4.3 알고리즘의 세부 모듈

이상치 제거 알고리즘은 기준값 생성모듈, 추세 판단 모듈, 이상치 제거모듈의 세부 모듈로 구성되며, 기준값 생성 모듈은 이전 3주기 데이터의 최대값과 범위를 이용하여 현재 주기의 최대 통행시간을 생성하는 모듈이다. 추세 판단 모듈은 데이터의 추세를 판단하여 혼잡 완화시 기준값을 재설정하는 모듈이며, 이상치 제거 모듈은 생성된 기준값을 이용하여 이상치를 검지하고 제거하는 모듈이며, 이들 모듈은 서로 유기적으로 연결된다.

4.4 기준값 생성 모듈

기준값 생성 모듈은 이전 3주기 데이터의 최대값과 범위를 가지고 현재 주기의 최대 통행시간 기준값을 생성하는 모듈로 수식은 다음과 같다.

○ 최대 통행시간 기준값 생성

$MAX\chi_t = D_{\_range} + D_{\_Max}\{x_i\}$,

$D_{\_range} = Max(\chi_{t-1} - \chi_{t-3})$

$MAX\chi_t$: t 주기의 최대 통행시간 기준값

$D_{\_Max}\{x_i\}$: 이전 3주기의 최대 통행시간

$D_{\_range}$: 이전 3주기 데이터의 통행시간 범위(range)

4.5 추세 판단 모듈

추세 판단 모듈은 3편차법을 이용하여 한계 이외의 데이터를 부호 개수 알고리즘으로 파악하여 혼잡 완화 추세를 판단하는 모듈로 혼잡이 완화되는 추세가 판단되면 기준값을 재설정하며, 그렇지 않은 경우는 기존의 기준값을 그대로 사용한다. 추세 판단 모듈을 통해 재설정되는 기준값은 다음과 같이 제시될 수 있다.

○ 기준값 재설정

$MAX\ \chi_t' = D_{\_range} + D_{\_Max}\{x_i\} - k \times s/\sqrt{n}\ MAX\chi_t$

: 재설정된 기준값

s : 3주기 데이터의 표준편차

n : 3주기 데이터의 개수

단, K는 추세 판단 모듈에서 설정된 값임

$k \times s/\sqrt{n}$ 은 기존의 기준값보다 좀 더 낮은 기준값으로 설정될 수 있는 한계값을 나타내며 K값은 정규분포가 아닌 데이터에서 사용되는 체비셰프 부등식에 의해 가장 적당한 K값을 5로 도출하였다. 이는 체비셰프 부등식에 의하여 96% 신뢰수준하에서 설정된 한계값이며, 혼잡의 완화라는 상태에서 기준값이 너무 낮게 떨어지는 현상을 막기 위한 통계적 방법이다.

4.6 이상치 제거 모듈

이상치 제거 모듈은 설정된 기준값과 수집된 통행시간 데이터를 비교하여 이상치로 판단된 데이터를 제거하는 모듈로 이상치를 제거한 후 정상치(일반 차량의 통행시간이라고 추정할 수 있는 정상적인 데이터)들을 이용하여 신뢰성 있는 링크별 실시간 교통정보를 생성하게 된다. 또한 제거된 이상치들을 제외한 정상치들은 다음 주기의 기준값을 생성하는 입력 데이터가 된다.

5. 알고리즘의 적용 및 평가

5.1 알고리즘의 적용 및 평가

본 연구에서 제시한 이상치 제거 알고리즘을 적용 및 평가하기 위해 2002년 12월 17, 18, 20일(화, 수, 금)에 서울시의 주요축인 강남대로, 도산대로, 양화로(신촌)의 일부 링크에서 번호판 조사를 실시하여 각 링크의 구간 통행시간을 구하였다. 실측 조사시간은 오전 7:30~10:30, 오후 2:00~5:00로 총 6시간에 걸쳐 조사하였다.

또한 통계적으로 알고리즘을 평가하기 위해 이상치가 제거되는 비율로 나타낼 수 있는 이상치 제거율(%) 분석과 이상치가 제거됨으로써 줄어드는 통행시간의 상대오차 분석을 실시하였다. 이상치 제거의 효과분석은 상대오차를 이용하여 산출하였으며, 그 식은 다음과 같다.

$$Confidence\ \ Ratio(\%) = 100(\%) - \left[\frac{1}{N}\left\{\sum\frac{|LTT_{Auto} - LTT_{Taxi}|}{LTT_{Auto}}\right\} \times 100(\%)\right]$$

즉, 실측한 통행시간과 이상치 제거전과 제거후의 상대오차 감소효과를 비교하였다.

5.2 통행시간 상대오차 분석결과

전체적으로 이상치 제거 알고리즘을 적용한 결과 표 2와

표 2. 통행시간 상대오차 분석결과

도 로 명		상대오차	
		이상치 제거전	이상치 제거후
강남대로		74.58%	74.70%
	뱅뱅사거리~서초우성APT	77.45%	77.53%
	강남역~제일생명사거리	71.70%	71.87%
양화로		67.37%	75.11%
	학동사거리~청담동사거리	70.46%	80.10%
	신사역~안세병원	68.29%	70.11%
도산대로		68.87%	71.24%
	홍대입구사거리~동교동삼거리	71.48%	73.24%
	신촌교회삼거리~동교동삼거리	66.26%	69.23%
전 체 종 합		70.94%	73.68%

같이 상대오차가 더 줄어진 것으로 나타났다. 다만 현재의 효과는 실측시간대비 각 링크의 감소효과이므로 실측시간 이외에 발생한 오류데이터에 대한 알고리즘의 제거능력은 정확히 감지될 수 없으며, 특히 실측시간이 혼잡이 발생하기 바로 전 시간이라는 것을 감안하면 이상치 제거 효과와 더불어 그래프를 통해서 각 모듈의 효과를 판단해야 할 필요성이 존재한다. 이는 상대오차가 실측대비분석임에도 불구하고 실측시간이 하루 중의 6시간이며, 이상치의 특성상 조사시간이외의 시간에도 존재할 확률이 크기 때문으로, 추가적인 이상치 제거 알고리즘의 적용 효과로 이상치의 제거율도 분석하였다.

특히 현재의 효과분석이 5분화 데이터의 비교이므로 1개 원시데이터의 이상치 제거로 인한 효과가 5분화되고, 각 링크별로, 축별로 평균화되면서 이상치 제거 효과가 감소되게 되며, 이는 실측대비 원시데이터 및 5분화 통행시간의 그래프 비교로 더욱 명확하게 나타났다. 따라서, 전체적인 이상치 제거의 효과는 비록 미비하게 나타났으나, 매주기의 실시간 정보의 효과를 고려할 때 해당 시간대에 제공하는 정보의 오류를 줄이는 것은 매우 중요하다고 사료된다.

5.3 이상치 제거율 분석결과

실측한 통행시간에 대비한 상대오차 비교를 통해 본 알고리즘이 이상치를 얼마나 효과적으로 제거할 수 있는가를 알아보았으며, 그 결과 약 70%의 이상치가 제거되는 것으로 나타났다. 따라서, 이상치 제거 알고리즘이 현재 주기에 올라온 이상치를 실시간으로 제거하여 실시간 정보의 정확도를 높일 수 있을 것으로 판단된다.

다만, 이 결과는 실측시간인 6시간 동안 원시데이터의 이상치를 제거한 결과이지만, 해당시간의 오류데이터의 제공확률을 약 70% 정도 감소할 확률은 하루 전체로 확대해서 판단하여도 무방하리라고 판단된다.

5.4 실제 이상치 제거 결과

번호판 조사를 통해 실측된 각 주기별 통행시간과 택시 probe에서 올라온 데이터, 그리고 이를 통해 생성된 기준값 및 이상치로 판정된 데이터를 그래프를 통해서 비교하였다.

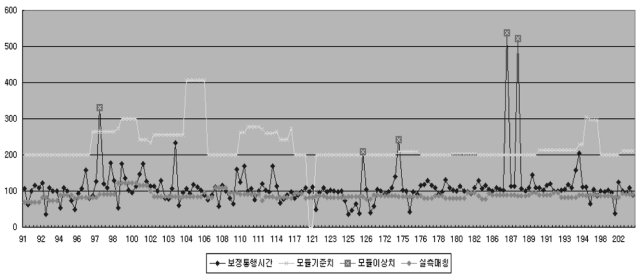

그림 5. 강남대로 "강남역~제일생명사거리" 구간

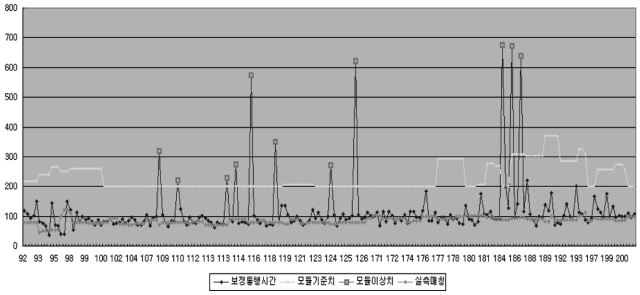

그림 6. 도산대로 "학동사거리~청담동사거리" 구간

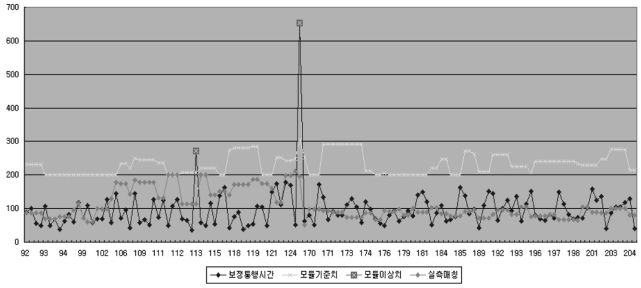

그림 7. 신촌 양화로 "신촌교회삼거리~동교동삼거리" 구간

표 3. 이상치 제거율 분석결과

도로명	이상치개수	처리개수	제거율(%)
강남대로	35	23	65.71%
도산대로	41	26	63.41%
신촌~양화로	22	18	81.82%
전체	93	62	70.32%

그림 5~7과 같이 그래프를 통해 이상치 제거 알고리즘의 효과를 평가한 결과, 조사된 실측통행시간을 이용하여 기준값과 비교했을 때 대체로 기준값이 안정적으로 실시간적인 통행시간의 변화를 잘 반영한다는 것을 알 수 있었으며, 알고리즘이 이상치의 제거에 적절한 것으로 나타났다.

6. 결론 및 향후연구과제

본 논문에서는 택시 GPS probe를 통해 신뢰할 수 있는 실시간 교통정보를 생성하기 위해 택시 GPS를 통해 교통정보가 생성될 경우 주행과 관계없는 정보가 수집된 경우, 이와 같은 이상치를 검지하고 제거하는 실시간 이상치 제거 알고리즘을 개발하였으며, 서울시 주요 축의 일부인 강남대

로, 도산대로, 양화로(신촌)의 일부 링크에 대하여 적용, 알고리즘을 평가하였다.

그 결과 전체적으로 통행시간의 상대 오차가 73.7%로 향상되었으며, 약 10%의 상대오차 감소의 효과를 보이는 링크도 존재하여 본 알고리즘을 이용하여 Taxi GPS를 통해 신뢰할 수 있는 실시간 교통정보를 생성할 수 있을 것으로 사료된다.

그러나, 이상치 제거전과 제거후의 축별 링크별 상대오차 비교와 그래프를 통한 분석을 통해 본 알고리즘의 효과를 정확히 판단하기는 무리라고 사료되는 바, 이상치로 판단되는 데이터를 얼마나 효과적으로 제거할 수 있는가를 판단하는 이상치 제거율(%) 분석을 실시하였으며, 그 결과 약 70%의 이상치가 제거되는 것으로 나타났다. 이와 같은 결과는 상대오차 비교 및 그래프를 통한 이상치 제거 효과와 더불어 이상치 제거 알고리즘의 우수성을 나타낸다고 할 수 있겠다.

다만, 현재의 이상치 제거 알고리즘은 상한값의 기준값으로 일반적인 통행시간보다 길게 올라온 이상치(Upper Outlier)는 제거할 수 있지만, GPS 매칭의 오류 및 회전으로 인해 일반적인 통행시간보다 짧게 올라온 이상치(Under Outlier)를 제거하지 못하고 있다. 따라서 이와 같은 문제를 해결하기 위한 모듈이 알고리즘에 추가되어야 할 것으로 판단된다.

또한 이상치 제거 알고리즘을 적용하기 위한 전제조건은 택시 GPS Probe를 통해 교통정보제공의 기본 단위인 5분동안 링크 통행시간 데이터가 수집되는 것이나, 이의 수집이 어려운 경우의 이를 보완할 수 있도록, 버스(BMS/BIS), 택배차량, 일반 승용차의 데이터를 이용하는 필요하며, 이에 대한 추가적인 연구가 필요할 것으로 보인다.

참고문헌

김동환(2000), **혼잡교통류에서 GPS/GIS를 활용한 링크통행시간 및 정지지체 추정기법 개발**, 석사학위논문, 아주대학교.

서울시립대학교 부설 도시과학연구원(2000), **TSD 교통정보 제공 시스템 구축 자문보고서**.

신강원(2003), **Probe 도착시간과 검지기 교통량을 이용한 링크통행시간 추정**, 석사학위논문, 아주대학교.

안상형, 이명호(1994), **현대통계학**, 학현사.

정연식(1999), **GPS Probe 및 루프검지기 자료의 융합을 통한 통행시간추정 알고리즘 개발**, 석사학위논문, 아주대학교.

정재영(2001), **GPS Probe를 이용한 정지지체 산정 및 혼잡지표 개발**, 석사학위논문, 아주대학교.

Barnett, Vic. and Lewis, Toby (1984) *Outliers in statistical data, Wiley*.

Sen, A., Thakuriah, P., Zhu, X., and Karr, A. (1997) Frequency of Probe reports and variance of travel time estimates, *American Society of Civil Engineers*.

Keechoo Choi, Chi-Hyun Shin, and Incheol Park (1998) Link travel time derivation using GPS/GIS, *Proceedings of the 1998 GIS-T Conference*, American Asso. of State Transportation & Highway Officials, Salt Lake City, UT.

(접수일: 2006.3.13/심사일: 2006.5.15/심사완료일: 2006.5.15)