한국ITS학회논문지
제7권, 제6호 (2008년 12월)
pp.21~30

사분위편차 및 관리도 모형에 의한 GPS 수집기반 구간통행속도 데이터 이상치 제거방안 연구

The Quartile Deviation and the Control Chart Model of Improvement Confidence for Link Travel Speed from GPS Probe Data

한 원 섭* 김 동 효** 현 철 승*** 이 호 원*** 오 영 태**** 이 철 기*****
(Won-Sub Han) (Dong-Hyo Kim) (Cheol-Seung Hyun) (Ho-Won Lee) (Yong-Tae Oh) (Choul-Ki Lee)

요 약

GPS를 탑재한 프로브차량에 의해 수집되는 교통정보(구간통행속도)는 차량검지기와 같이 특정링크에 대하여 연속적으로 교통정보를 수집하는 방식이 아니다. 따라서 단속교통류 구간에서 신호시간 등의 영향으로 수집되는 정보의 안정성과 대표값 등에 문제점이 있다. 본 연구는 GPS를 장착한 프로브차량에 의해 수집된 구간통행속도 데이터를 필터링하는 방법을 개발하는데 목적이 있다. 시간간격별로 수집되는 데이터에 대하여 사분위편차와 관리도에 의해 이상치를 제거하였다. 사분위편차를 적용한 결과는 제거율이 0~3.7% 이고, 개별 관리도에 의한 결과는 제거율이 0.3~7.2% 이었다. 두 방법 모두 교통소통이 원활한 새벽시간 대에 이상치 제거율이 낮고, 낮 시간대에 이상치 제거율이 높은 것으로 나타났다. 문제점으로 지적되는 것은 모형에 충실할 경우 Low Bound에서 이상치의 제거기준이 낮게 설정되는 것이다. 따라서 시스템 운영과정에서 경험적인 사항이 반영되어져야 할 것으로 검토된다.

Abstract

The travel speed collected by the prove-car equipped with the GPS has the problems, which are the data's stability and finding out the representative travel speed, by the influence of the traffic signal and etc. at the interrupted traffic. This study was conducted to develop the method of filtering the outlier data from the data collected by the prove-car. The method to remove the outlier data from the serial data which were collected by the prove-car was adapted to each of the quartile deviation statistics model and the management graphic statistics model. The rate of removing the outlier data by the quartile deviation method was 0 ~ 3.7% while the rate by the management graphic statistic methods was 0.3~7.2%. Both methods show the low removal rate at the dawn time when the traffic is inactivity, on the other hand the remove rate is high during the daytime. However, both methods have the problem such that the threshold level for removing the outlier data was established at the low bound in the case as good as the statistics model. Therefore, it is required for the experience calibration.

Key words: GPS, prove-car, quartile deviation, control chart

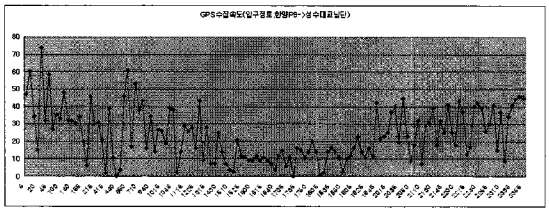
* 주저자 : 도로교통공단 교통과학연구원, 수석연구원
** 공저자 : 도로교통공단 교통과학연구원, 연구위원
*** 공저자 : 도로교통공단 교통과학연구원, 선임연구원
**** 공저자 : 아주대학교 건설교통공학부 교수
***** 공저자 : 아주대학교 ITS대학원 교수
† 논문접수일 : 2008년 11월 7일
† 논문심사일 : 2008년 12월 10일(1차), 2008년 12월 18일(2차)
† 게재확정일 : 2008년 12월 22일

I. 서 론

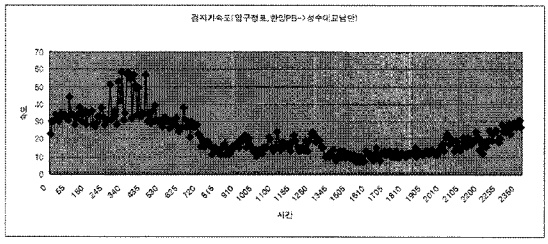
1. 연구 배경 및 목적

현재 교통정보수집장치로 이용되고 있는 신호검지기, 무인단속검지기 및 교통량검지기 등 도로에 매설된 검지기는 지점속도 수집방식으로 구간통행속도를 수집하는 데에는 한계가 있다.

GPS 단말기에 의한 교통정보의 수집은 차량검지기와 같이 특정링크에 대하여 시간대별로 연속적으로 교통정보를 수집하는 체계가 아니고, 도로망 내를 운행하는 프로브차량(Probe-Car)에 의해 산발적으로 수집되는 데이터이다. 따라서 단속교통류 구간에서는 신호시간(적색신호 대기, 녹색신호 통과 및 신호변환 도착 등)의 영향으로 수집되는 정보의 안정성과 대표값 등에 문제점이 있다(<그림 1> 및 <그림 2> 참조) [1]. 또한 교통정보수집을 목적으로 GPS를 장착하는 차량이 택시, 트럭이나 버스 등 특정 목적을 갖고 운행되는 차량을 이용할 경우가 대부분으로 차량의 주행특성에 따라 수집되는 자료는 실제 교통상황과는 다른 데이터(이상치)가 수집될 수 있다. 따라서 본 연구에서는 GPS를 탑재한 프로브차량에 의해 수집되는 구간통행속도정보가 신뢰성을 확보하기 위하여 이상치를 제거하는 방안(Filtering & Smoothing)을 모색하는데 목적이 있다.



<그림 1> GPS에 의한 구간통행속도 수집특성
(압구정로 한양PB→성수대교남단)
<Fig. 1> Link travel speed data from GPS data



<그림 2> 신호검지기에 의한 통행속도 수집특성
(압구정로 한양PB→성수대교남단)
<Fig. 2> Link speed data from traffic signal detector data

2. 연구 범위 및 내용

GPS를 탑재한 프로브차량에 의해 수집되는 구간통행속도정보의 신뢰성을 확보하기 위해서는 수집되는 데이터를 분석하여 교통상황을 대표하는 정보로 가공하는 과정이 요구된다. 따라서 본 연구는 GPS와 수치지도(링크-노드)를 이용하여 구간통행속도정보를 수집하여 데이터 이상치를 제거하는 방법을 연구범위로 정하였다. 현재 GPS 프로브차량에 의해 운영되고 있는 서울지방경찰청의 교통정보 자료를 이용하여 이상치 제거방안을 마련하여 교통정보의 신뢰성을 높일 수 있는 방안을 제시한다.

첫째, GPS 프로브차량에 의하여 수집된 교통정보에 대하여 탐색적 자료 분석(EDA : Exploratory Data Analysis)을 수행한다. 탐색적 자료 분석은 분석을 수행하기 전 단계에 주로 수행되며, 데이터 특성을 파악하기 위하여 기초통계량 및 히스토그램 등을 작성하는 방법이다.

둘째, 사분의 편차 및 관리도에 의하여 GPS 수집 데이터의 이상치를 제거하는 방법을 개발하고, 두 방법별 특성을 비교한다.

II. 관련 이론 및 연구

1. GPS 수집 데이터의 이상치 정의

현재 도시부의 교통관리를 목적으로 운영되고 있는 교통정보시스템의 교통정보를 수집/분석/제공하는 주기를 5분으로 설정하고 있다. 5분 단위의 대표치로 산정하기 위하여 링크 통행시간의 교통데이터를 산술평균하기 때문에 1차 가공 시 이상치 처리에 문제가 발생할 수 있다. 이는 산술평균은 심하

게 편향된 데이터에 의해 중심의 위치가 크게 변동될 수 있는 통계치이기 때문이다 [2][5].

일정시간 수집되는 평균데이터를 기준으로 이전과 현재 데이터의 차가 클수록 데이터의 안정성이 떨어진다는 가정 하에 GPS 구간통행속도 데이터의 안정성을 검지기 통행속도 수집데이터를 다음 식과 같이 이전 데이터와 현재 데이터 차의 평균값을 전체 평균속도와의 비율로 분석하였다.

$$VR = (\sum(Vi+1 - Vi) / n) / Vavg$$

여기서, VR = 수집데이터 진동계수

Vi = (i)번째 수집데이터

Vi+1 = (i+1)번째 수집데이터

Vavg = 분석시간동안 수집데이터 평균속도

n = 분석자료 수

위 식을 적용하여 3개 도로 구간에서 신호검지기와 GPS 데이터 간의 안정성을 분석한 결과 <표 1>과 같이 GPS 데이터의 진동이 신호검지기자료에 비하여 2배 이상 커, 평균데이터의 안정성을 약화시킬 수 있음을 보여주고 있다.

이러한 GPS 프로브차량에서 발생할 수 있는 이상치에 대한 개념 정의를 하면 다음과 같다 [3].

○ 이상치 정의 1 : 특정 GPS 프로브차량의 운행특성을 일반 승용차량 통행기준으로 적용함으로써 주행과 관련 없는 정보[1]에 의해 수집된 원시 프로브 데이터

○ 이상치 정의 2 : 자료계열에서 대부분 관측치에 의해 제시된 형태를 이루지 못하는 일부 관측치(즉, 주기시간동안 특히 편향되는 링크 통행시간 정보)

<표 1> 신호검지기와 GPS 데이터 간의 안정성 비교
<Table 1> Comparison between the traffic signal detector and GPS data

도로명	기종점	진동	GPS 진동
도산대로	영동삼단지→도산공원	2/27=0.07	10/31=0.3
강남대로	영동시장보→제일생명	4/21=0.19	8/20=0.4
압구정로	한양PB→성수대교남단	4/21=0.19	8/20=0.4

1) 택시 프로브의 경우 승하차를 위한 링크 중간에서의 정지 시간, 호객행위를 위한 공차 대기, 이면도로 통행, 개인용무 등 다양한 요인으로 기인될 수 있음.

이영인(2002)은 연속류와 단속류로 구분하여 신뢰성 있는 구간통행속도를 산출하는 적정 표본수에 대한 연구를 수행하였다 [4]. 심상우(2006)는 수집주기내 통과하지 못하는 GPS데이터를 MAPE와 MAE를 사용하여 구간통행속도 산출기법을 제안하였으며, 정확도는 1.98%, 4.75초로 실측치와 큰 차이를 보이지 않았다고 분석하였다 [5].

2. ARIA모형에 의한 GPS데이터 이상치 제거[3]

1) 구간검지체계의 GPS 데이터 이상치 제거

최기주 · 장정아의 '시계열기반의 GPS 프로브 자료의 이상치 제거 알고리즘 개발(2004.12.)'에서는 구간검지체계의 GPS 프로브 원시데이터의 1차 가공을 위한 평균화과정에서 발생할 수 있는 문제점으로 이상치 문제에 대하여 검토하고 있다. 이상치의 미 제거는 구간검지기의 수집 및 가공 신뢰도를 저하시킬 수 있으므로 이상치에 대한 부분을 수집주기별로 판단하여 제거해주여야 하는 것으로 하였다. 연구에서는 이상치 제거 알고리즘으로 ARIMA 모형을 적용시켜 실시간적 신뢰구간 추정과정들 제시하였다.

일반적으로 예측모형에서 유용하게 사용되는 시계열모형에서의 모형화 과정을 응용하여 이상치 제거의 기준이 될 수 있는 상한값(Upper Bound: 이하 UB)과 하한값(Lower Bound: 이하 LB)값을 제시하고 현 주기(t)내에 수집된 개별 프로브 링크 통행시간($LTT_{(프로브)}$)에 대하여 UB와 LB와 비교하여 이상치를 결정하여 제거한다. 제시하는 알고리즘의 단계에 대한 설명은 아래와 같다.

① 1단계: 주기별 LTT(Link Traver Times, 링크통행시간)를 이용하여 시계열 모형식을 추정한다.

· 추세를 제거, 정상 시계열을 만든다. (d 혹은 D의 결정 하에 정상화시킨다)

· AR과 MA의 차수를 결정한다.(p, q 및 P, Q의 값을 결정한다)

- 차수가 결정된 ARIMA 모형을 시계열자료를 이용해서 계수들을 추정한다.
- 앞의 추정 결과가 만족스러운지 판단한다.
- 여러 개의 비교적 만족스러운 모형이 추정되면 그중에 적합한 것을 고른다.

② 2단계: 신뢰수준에서 UB와 LB를 결정한다.

$$UB = \hat{y}_{t+l} + t_{a/2}\sqrt{Var(e_{t+l})}$$

$$LB = \hat{y}_{t+l} - t_{a/2}\sqrt{Var(e_{t+l})}$$

단 $Var(e_{t+l}) = \sigma_e^2 \sum_{j=0}^{l-1} \phi_j^2$

③ 3단계: 현 주기(t)내에 수집된 개별프로브 링크 통행시간($LTT_{i프로브}$)에 대하여 UB와 LB와 비교하여 이상치를 결정하여 제거한다.

④ 4단계: 이후 새로이 링크통행시간을 계산한다.(이상치 제거 후 링크 통행시간)

$$LTT_t = \frac{\sum_{i=0}^{N} LTT_{i프로브}}{N}$$

2) 모형 적용결과

① 시계열 모형의 식별결과

모든 링크가 다음과 같은 ARIMA(0,1,1) 즉 IMA(1,1)모형으로 식별되었다. 따라서 모형 식별과정은 보통 분석가의 주관적으로 결정해야 될 때가 많으므로 추후 시스템 적용 시, 모두 IMA(1,1)형태로 식별하고, 일정 기간 동안 업데이트 하는 방안도 좋을 것으로 사료 된다.

$X_{t+1} - X_t = \mu + e_t - \theta \cdot e_{t-l}$

X_t: t 주기의 링크통행시간

μ : IMA(1,1) 모형의 모수 추정치(평균)

e_t : t 주기의 잔차

θ : IMA(1,1) 모형의 모수 추정치(기울기)

② 모수추정

IMA(1,1) 모형의 경우 μ, θ 값만을 추정하면 되므로 이를 추정하였을 때 17일 전날 초기 추정치는 <표 2>와 같다.

<표 2> 모형의 초기24시간에 대한 모수 추정치 3]
<Table 2> Evaluation Values of IME Model during 24 hours

링크번호	μ	Θ
1	0.01964	0.77539
2	0.25358	0.29952
3	0.283599	0.63475
4	0.12929	0.75319
5	-0.0404	0.43985
6	-0.00846	0.71456
7	-0.09112	0.43539
8	-0.27387	0.4047
9	-0.0021	0.8112
10	-0.03826	0.54588

③ 이상치 제거 결과

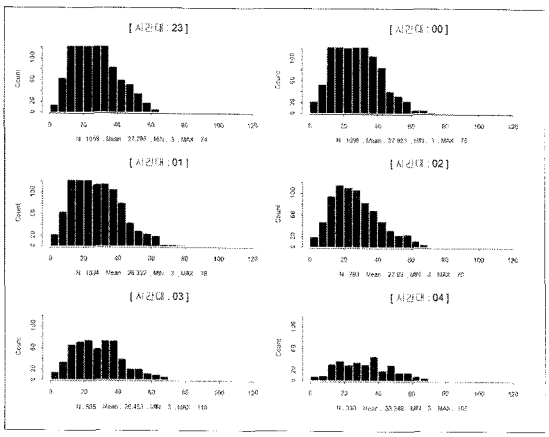
모수추정 업데이트 시간을 변동시킴에 따라 이상치 제거 개수는 2시간 기반 추정의 경우 평균 2.9%의 프로브 개수가 이상치로 판단되었고, 6시간 기반의 경우 4%, 24시간 기반의 경우 평균 5%가 이상치로 제거되다.

모수 추정 업데이트 시간을 길게 함에 따라 시간에 따른 혼잡의 추세를 길게 평활화 시키기 때문에 원시데이터를 이상치로 더 판단하여 제거하고 있는 경향을 보이고 있다. 이에 비해 모수 추정기간을 짧게 하여 실시간적으로 할수록 바로 전주기대의 혼잡상황이 반영되어 이상치를 덜 제거하는 경향을 보이는 것으로 제시하고 있다.

3. 분위편차와 관리도 정의

1) 사분위편차

관측값을 순서대로 정렬 한 후 25% 위치한 값을 1사분위수(Q1), 75% 위치한 값을 3사분위수(Q3)라 하는데, 3사분위수와 1사분위수의 차이를 사분위편차(IQR)라 한다. 사분위편차를 이용한 이상치 제거 방법은 다음과 같이 1사분위수에서 사분위편차의 1.5배만큼을 뺀 값보다 작거나 3사분위수에서 사분위편차의 1.5배만큼을 더한 값보다 큰 값을 이상치로 판단한다.



<그림 3> 링크(신사역→쌍방울앞) 속도데이터 히스토그램(23~04시)
<Fig. 3> Histogram of link speed data

2) 관리도

관리도란 데이터가 시간이 지남에 따라 데이터의 중심 및 산포가 어떻게 변화하는가를 그래프로 나타낸 것이다. 개개의 측정값(x) 관리도는 부분군을 추출할 때 한 번에 여러 개의 제품들을 추출할 수 없는 경우에 사용된다. x관리도에서는 한 개의 측정값이 얻어지면 곧 관리도에 점으로 기록되므로, 각각의 측정으로부터 공정의 안정상태의 판정 및 조치까지 시간적인 지연이 없는 것이 특징이다.

교통 데이터의 경우, 정보의 수집 특성상 한 번에 하나의 제품밖에 얻을 수 없는 경우에 개개의 측정값과 이동범위 관리도를 이용할 수 있다.

Ⅲ. 사분위편차 및 관리도에 의한 이상치 제거

1. 기준값 선정 및 탐색적 자료 분석

1) 기준값 선정

GPS에 의해 수집된 데이터를 5분 단위로 교통정보를 가공 및 제공하기 위해서는 자료의 집단화로 인한 대표치 선정이 필요하다. 일반적으로 대표하는 정보의 추출을 위하여 사용되는 통계적 방법론은 다음과 같은 산술평균을 사용 한다 [1].

$$LTT_t = \frac{\sum_{i=0}^{N} PLTT_i}{N}$$

여기에서, LTT_t = 특정주기 t의 링크통행시간 대표치

$PLTT_i$ = i번째 프로브의 개별링크통행시간

N = 특정주기 t에서 수집된 총 프로브 대수

2) GPS 데이터의 탐색적 자료 분석

분석 대상 데이터는 현재 서울시에서 운영되고 있는 GPS 수집데이터를 대상으로 하였다. 이중에 GPS 프로브 자료의 수집 데이터 건수가 많은 링크를 대상으로 하였다. 시간대별 수집 데이터 건수가 일정수준 이상으로 수집되는 도로의 링크를 선정하였다.

<표 3> 강남대로(신사역→쌍방울앞) 링크속도에 대한 기초통계량
<Table 3> Statistics for speed data of Kangnam road

시간대	수집건수	평균속도	표준편차	최소값	1사분위수	중앙값	3사분위수	최대값
00	1,098	27.923	13.3	3	17	27	36	75
01	1,034	28.332	14.598	3	16	27	38	76
02	793	27.83	14.077	3	17	26	37	70
03	585	29.453	15.573	3	17	29	39	118
04	330	33.248	16.445	3	19.25	33	44.75	105
05	298	34.977	16.283	3	22	33.5	48	75
06	335	35.104	16.47	4	22	34	47	78
07	494	31.945	14.397	3	21	31	43	71
08	759	24.962	14.912	3	12	23	36	69
09	763	26.085	13.883	3	15.5	24	36	71
10	814	22.345	13.689	3	11	20	32	78
11	913	20.522	13.417	3	10	18	29	68
12	640	24.295	15.05	3	12	22	35	68
13	636	22.057	13.407	3	11	19	31	77
14	828	16.702	12.251	3	7	12	23	61
15	885	16.95	12.483	3	7	13	23	67
16	966	16.795	12.629	3	7	13	23	71
17	835	16.97	12.979	3	7	13	23	67
18	744	18.14	12.549	3	8	14	24	68
19	767	20.621	13.758	3	10	17	31	69
20	746	23.669	13.552	3	12	22	34	70
21	814	25.455	12.953	3	15	25	35	63
22	992	25.052	13.375	3	14	23.5	34	67
23	1,058	27.295	13.075	3	16.25	26	35	74

<표 3>과 같이 강남대로의 신사역→쌍방울교차로 구간의 링크에 대하여 2007년 7월 20일 00시~8월 21일 13시 약 1개월 간 수집된 데이터를 대상으로 하였다. 1개월 동안 시간대별로 700건 이상의 자료가 수집되고, 총 데이터 건수는 18,127건이다. 데이터의 특성은 최소값이 3km/h이고 최대값이 70km/h정도 이며, 중앙값은 10~30km/h 내에 분포하였다. 시간대별 속도변화는 새벽 4~7시대가 30km/h 이상이고, 오전 혼잡시간대는 20km/h 정도이고, 오후 혼잡시간대에는 10km 정도인 것으로 나타났다.

분석 대상 데이터에 대한 속도 분포 특성을 파악하기 위하여 각 시간대별 히스토그램분석을 수행하였다. 속도의 분포 특성은 다음과 같다(<그림 3> 참조).

오후 시간대의 경우 30km/h 이내로 수집된 데이터가 약 75% 이상 차지하며, 04시~07시간대의 경우 수집 데이터의 개수가 500개 이하로 타 시간대에 비해 데이터 수집량이 적다. 새벽시간대, 오전시간대, 오후시간대, 저녁시간대 별로 속도 데이터의 분포가 거의 유사하다.

<표 4> 사분위편차에 의한 이상치 제거
<Table 4> Outlier of IQR

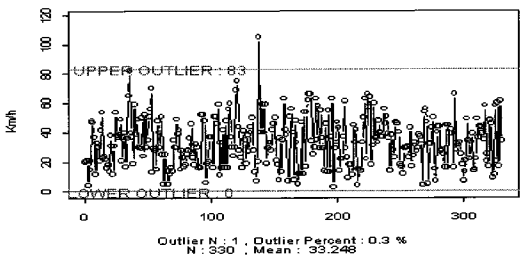
시간대	수집건수	평균속도	상한값	Outlier 개수	Outlier 비율(%)
00	1098	27.923	64.5	8	0.73
01	1034	28.332	71	4	0.39
02	793	27.83	67	2	0.25
03	585	29.453	72	5	0.85
04	330	33.248	83	1	0.3
05	298	34.977	87	0	0
06	335	35.104	84.5	0	0
07	494	31.945	76	0	0
08	759	24.962	72	0	0
09	763	26.085	66.75	2	0.26
10	814	22.345	63.5	3	0.37
11	913	20.522	57.5	18	1.97
12	640	24.295	69.5	0	0
13	636	22.057	61	5	0.79
14	828	16.702	47	19	2.29
15	885	16.95	47	26	2.94
16	966	16.795	47	36	3.73
17	835	16.97	47	28	3.35
18	744	18.14	48	21	2.82
19	767	20.621	62.5	4	0.52
20	746	23.669	67	2	0.27
21	814	25.455	65	0	0
22	992	25.052	64	2	0.2
23	1058	27.295	63.125	4	0.38

2. 사분위편차 이상치 제거

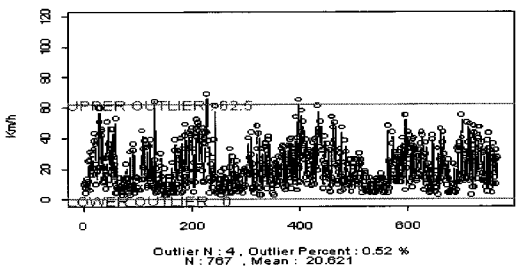
이상치를 제거하기 위하여 사분위편차를 이용한다. 관측값을 순서대로 정렬 한 후 25% 위치한 값을 1사분위수(Q1), 75% 위치한 값을 3사분위수(Q3)라 하는데, 3사분위수와 1사분위수의 차이를 사분위편차(IQR)라 한다. 사분위편차를 이용한 이상치 제거방법은 다음과 같이 1사분위수에서 사분위편차의 1.5배만큼을 뺀 값보다 작거나 3사분위수에서 사분위편차의 1.5배만큼을 더한 값보다 큰 값을 이상치로 판단한다.

○ 사분위편차를 이용한 이상치 제거방법에서의 하한 기준값 : Q1 - 1.5*IQR

○ 사분위편차를 이용한 이상치 제거방법에서의



(a) 04시간대



(b) 19시간대

<그림 4> 사분위편차에 이상치 제거 결과
<Fig. 4> Outlier of IQR

상한 기준값 : Q3 + 1.5*IQR

하한 기준값이 0 미만인 경우 이상치 제거 대상이 속도 값이므로 0으로 값을 변경한다. 분위편차를 이용하여 GPS 데이터의 이상치를 제거한 결과는 <표 4>와 <그림 4]>와 같다. 이상치 제거율이 0~3.7% 범위이고, 교통소통이 원활한 새벽시간 대에 이상치 제거율이 낮고, 낮 시간대에 이상치 제거율이 높은 것으로 나타났다.

3. 관리도에 의한 이상치 제거

개개의 측정값을 이용하여 관리도를 작성하는 경우에는 부분군의 크기를 $n=1$로 하고, x관리도와 R_s(인접한 두 측정값의 차) 관리도를 흔히 같이 사용한다. $x-R_s$ 관리도에서는 평균을 관리하기 위해서는 개개의 측정값을 사용하고 분산을 관리하기 위해서는 인접한 두 데이터간의 범위를 이용하였다. 만약 개개의 데이터 $x_1, x_2, \ldots, x_n$이 $N(\mu, \sigma^2)$의 수집과정에서 얻어졌다면

$$UCL = \mu + 3\sigma$$
$$UCL = \mu - 3\sigma \quad (1)$$

μ와 σ의 추정 값은 다음 식(2)로 구할 수 있다.

$$\hat{\mu} = \bar{x} = \sum x_i / n,$$
$$\hat{\sigma} = \frac{\overline{R}}{d_2} \quad (2)$$

그러나 개개의 측정값 관리도에서는 범위를 계산할 수 없으므로, 이 경우에는 인접한 두 측정치간의 차이인 이동범위(Moving range) R_s을 사용하였다. 이동범위의 수는 $k-1$개 있으므로, 이동범위의 평균 $\overline{R_s}$는

$$\overline{R_s} = \frac{\sum R_s}{k-1} \quad (3)$$

에 의하여 계산되고, 범위를 구하는 데 사용된 데이터의 수는 두 개이므로 $d_2 = 1.128$이다. 따라서 실제로 사용되는 x관리도의 중심선과 관리한계선은 다음과 같다.

$$CL = \bar{x}$$

$$UCL = \bar{x} + 3\left(\frac{\overline{R_s}}{d_2}\right) = \bar{x} + \frac{3\overline{R_s}}{1.128} = \bar{x} + 2.66\overline{R_s}$$

$$LCL = \bar{x} - 3\left(\frac{\overline{R_s}}{d_2}\right) = \bar{x} - \frac{3\overline{R_s}}{1.128} = \bar{x} - 2.66\overline{R_s} \quad (4)$$

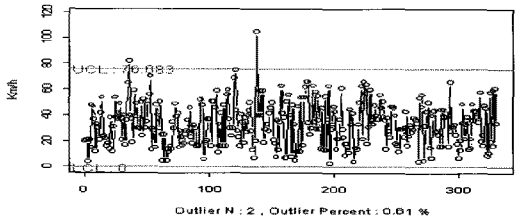
R_s관리도는 R관리도의 UCL과 LCL을 그대로 사용할 수 있다.

$$CL = d_2\hat{\sigma} = d2\left(\frac{\overline{R}}{d2}\right) = \overline{R}$$

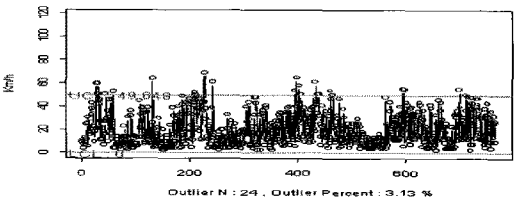
$$UCL = (d_2 + 3d_3)\hat{\sigma} = (d_2 + 3d_3)\frac{\overline{R}}{d_2} = \left(1 + 3\frac{d_3}{d_2}\right)\overline{R} = D_4\overline{R}$$

<표 5> 관리도에 의한 이상치 제거
<Table 5> Outlier of control chart

시간대	수집 건수	평균 속도	관리도에 의한 이상치 제거		
			상한값	Outlier 개수	Outlier 비율(%)
00	1098	27.923	63.412	11	1
01	1034	28.332	65.034	9	0.87
02	793	27.83	62.504	10	1.26
03	585	29.453	65.614	12	2.05
04	330	33.248	76.083	2	0.61
05	298	34.977	74.429	2	0.67
06	335	35.104	74.861	1	0.3
07	494	31.945	67.906	2	0.4
08	759	24.962	56.826	22	2.9
09	763	26.085	61.056	11	1.44
10	814	22.345	52.577	25	3.07
11	913	20.522	47.475	34	3.72
12	640	24.295	56.365	18	2.81
13	636	22.057	51.912	15	2.36
14	828	16.702	39.78	49	5.92
15	885	16.95	40.842	64	7.23
16	966	16.795	40.721	66	6.83
17	835	16.97	40.486	58	6.95
18	744	18.14	44.693	33	4.44
19	767	20.621	49.648	24	3.13
20	746	23.669	55.046	14	1.88
21	814	25.455	58.782	5	0.61
22	992	25.052	58.958	17	1.71
23	1058	27.295	60.493	8	0.76



(a) 04시간대



(b) 19시간대

<그림 5> 관리도에 의한 이상치 제거
<Fig. 5> Outlier of control chart

$$LCL = (d_2 - 3d_3)\hat{\sigma} = (d_2 - 3d_3)\frac{\overline{R}}{d_2} = \left(1 - 3\frac{d_3}{d_2}\right)\overline{R} = D_3\overline{R} \quad (5)$$

그러나 이 경우는 이동범위를 구하는 데 두 개의 데이터만 사용되었으므로 $D_4 = 3.27$이고, D_3는 값이 없으므로 생각할 필요가 없다.

$$CL = \overline{R_s}$$

$$UCL = D_4\overline{R_s} = 3.72\overline{R_s}$$

$$LCL = D_3\overline{R_s} = 0.00\overline{R_s} = 0.0 \quad (6)$$

위와 같은 과정에 의해 구해진 개별 관리도에 의한 GPS 데이터의 이상치를 제거한 결과는 <표 5>, <그림 5>와 같다. 이상치 제거율이 0.3~7.2% 범위이고, 교통소통이 원활한 새벽시간 대에 이상치 제거율이 낮고, 낮 시간대에 이상치 제거율이 높은 것으로 나타났다.

Ⅳ. 결과 분석

1. 결과 분석

1사분위수, 3사분위수에 의하여 계산된 사분위편차를 이용한 이상치 제거방법과 개개의 측정값과 이동범위($x - R_s$) 개별 관리도 방법에 의한 이상치 제거방법을 이용하여 각 시간대별로 수집된 속도데이터에 대하여 이상치를 제거하였다. 두 가지 방법에 의하여 제거된 이상치의 개수를 비교해 보면 <표 6>과 같이 사분위편차를 이용 이상치 제거율이 0~3.7% 범위이고, 개별 관리도에 의한 이상치를 제거율이 0.3~7.2% 범위로 관리도 방법에 의하여 제거된 이상치의 개수가 많음을 알 수가 있다. 두 방법 모두 교통소통이 원활한 새벽시간 대에 이상치 제거율이 낮고, 낮 시간대에 이상치 제거율이 높은 것으로 나타났다.

이는 분석대상 링크의 경우 관리도 방법을 이용하여 계산된 한계선이 사분위편차를 이용하여 계산된 관리한계선보다 값이 낮은 데에서 비롯된 것이

<표 6> 사분위편차 및 관리도 방법에 의한 이상치 제거 개수 및 비율
<Table 6> Ration of IQR and control chart Outlier

시간대	수집 건수	평균 속도	사분위편차에 의한 이상치 제거			관리도에 의한 이상치 제거		
			상한 값	Outlier 개수	Outlier 비율(%)	상한값	Outlier 개수	Outlier 비율 (%)
00	1098	27.92	64.5	8	0.73	63.41	11	1
01	1034	28.33	71	4	0.39	65.03	9	0.87
02	793	27.83	67	2	0.25	62.50	10	1.26
03	585	29.45	72	5	0.85	65.61	12	2.05
04	330	33.25	83	1	0.3	76.08	2	0.61
05	298	34.98	87	0	0	74.43	2	0.67
06	335	35.10	84.5	0	0	74.86	1	0.3
07	494	31.95	76	0	0	67.91	2	0.4
08	759	24.96	72	0	0	56.83	22	2.9
09	763	26.09	66.75	2	0.26	61.06	11	1.44
10	814	22.35	63.5	3	0.37	52.58	25	3.07
11	913	20.52	57.5	18	1.97	47.48	34	3.72
12	640	24.30	69.5	0	0	56.37	18	2.81
13	636	22.06	61	5	0.79	51.91	15	2.36
14	828	16.70	47	19	2.29	39.78	49	5.92
15	885	16.95	47	26	2.94	40.84	64	7.23
16	966	16.80	47	36	3.73	40.72	66	6.83
17	835	16.97	47	28	3.35	40.49	58	6.95
18	744	18.14	48	21	2.82	44.69	33	4.44
19	767	20.62	62.5	4	0.52	49.65	24	3.13
20	746	23.67	67	2	0.27	55.05	14	1.88
21	814	25.46	65	0	0	58.78	5	0.61
22	992	25.05	64	2	0.2	58.96	17	1.71
23	1058	27.30	63.13	4	0.38	60.49	8	0.76

다. 따라서 이상치 제거 모형은 시스템 운영과정에서 경험적인 사항이 반영되어져야 할 것으로 사료된다.

2. 기존 연구와 결과 비교

선행 연구인 '시계열기반의 GPS 프로브 자료의 이상치 제거 알고리즘으로 ARIMA 모형을 적용하였을 때, 모수추정 업데이트 시간을 변동시킴에 따라 이상치 제거 개수는 2시간 기반 추정의 경우 평균 2.9%의 프로브 개수가 이상치로 판단되었고, 6시간 기반의 경우 4%, 24시간 기반의 경우 평균 5%가 이상치로 제거되었다. 모수 추정 업데이트 시간을 길게 함에 따라 시간에 따른 혼잡의 추세를 길게 평활화 시키기 때문에 원시데이터를 이상치로 더 판단하여 제거하고 있는 경향을 보이고 있다. 이에 비해 모수 추정기간을 짧게 하여 실시간적으로 할수록 바로 전 주기대의 혼잡상황이 반영되어 이상치를 덜 제거하는 경향을 보이는 것으로 제시하고 있다.

따라서 본 연구 적용한 사분위편차를 이용 이상치 제거율은 3.7%로 시계열기반의 ARIMA 모형(5%)을 적용하였을 때 에 이상치 제거율이 높다. 반면에 개별관리도에 의한 이상치 제거율은 최대 7.2%로 ARIMA 모형에 비하여 이상치 제거율이 높은 것으로 나타났다.

V. 결 론

약 1개월간 수집된 특정 링크의 GPS 데이터에 대하여 사분위편차와 관리도 방법에 의한 이상치 제거를 분석하였다. 시간대별로 연속적으로 수집되는 데이터에 대하여 연속적으로 방법별 시간대별 이상치 제거 기준에 따라 이상치를 제거하였다. 사분위편차를 이용하여 GPS 데이터의 이상치를 제거한 결과는 제거율이 0~3.7% 범위이고, 교통소통이 원활한 새벽시간 대에 이상치 제거율이 낮고, 낮 시간대에 이상치 제거율이 높은 것으로 나타났다. 또, 개별 관리도에 의한 GPS 데이터의 이상치를 제거한 결과는 제거율이 0.3~7.2% 범위이고, 교통소통이 원활한 새벽시간 대에 이상치 제거율이 낮고, 낮 시간대에 이상치 제거율이 높았다.

문제점으로 지적되는 것은 모형에 충실할 경우 Low Bound에서 이상치의 제거기준이 낮게 설정되는 것이다. 따라서 이상치 제거 모형은 시스템 운영과정에서 경험적인 사항이 반영되어져야 할 것으로 판단되면, 또한 교통사고 발생과 같은 유고 시, GPS 이상치 제거 방안도 추후 세밀한 연구가 진행되어야할 것으로 판단된다.

참고문헌

[1] 도로교통안전관리공단, *서울지방경찰청 종합교통정보센터 수집장치별 신뢰성 분석 연구*, 2004. 8.

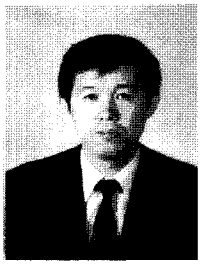
[2] 이영우, 임채문, "GPS 수집자료를 이용한 링크 통행시간 분포 특성 분석," *대한교통학회지*, 제22권, 제5호, pp. 7~17, 2004. 10.

[3] 최기주, 장정아, "시계열기반의 GPS 프루브 자료의 이상치 제거 알고리즘 개발," *대한교통학회지*, 제22권, 제6호, pp. 67~75, 2004. 12.

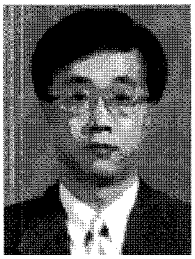
[4] 이영인, 이정희, "구간통행속도 제공을 위한 최소 표본수 결정에 관한 연구," *대한교통학회지*, 제20권, 제3호, pp. 55~67, 2002. 6.

[5] 심상우, 최기주, "혼잡상황에서 링크 미통과 GPS 프로브 데이터를 활용한 링크통행시간 추정기법 개발," *대한교통학회지*, 제24권, 제5호, pp. 7~18, 2006. 8.

저자소개

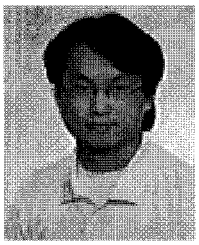


한 원 섭 (Han, Won-Sub)
1984년 3월 ~ 현재 : 도로교통공단 수석연구원
1987년 2월 : 연세대학교 전자공학과 석사 졸업
1985년 3월~1987년 2월 : 연세대학교 산업대학원 전자공학과 졸업(공학석사)
1978년 3월~1980년 2월 : 숭실대학교 전자공학과 졸업(공학학사)



김 동 효 (Kim, Dong-Hyo)
2002년 12월~현재 : 도로교통안전관리공단 연구위원
1999년 3월~2003년 3월 : 명지대학교 겸임교수
1995년 9월~2002년 12월 : 교통개발연구원 책임연구원
1989년 9월~1996년 12월 : Northwestern University 토목공학과 졸업(박사)
1986년 3월~1989년 8월 : 교통개발연구원 연구원
1983년 12월~1986년 2월 : 한국과학기술연구원 연구원
1982년 3월~1984년 2월 : 서울대학교 토목공학과 졸업(석사)
1976년 3월~1982년 2월 : 서울대학교 조경학부 졸업(학사)

현 철 승 (Hyun, Cheol-Seung)
1995년 6월~현재 : 도로교통공단 선임연구원
2003년 2월 : 성균관대학교 기계설계과 박사 졸업
1994년 2월 : 성균관대학교 기계설계과 석사 졸업
1992년 2월 : 성균관대학교 기계설계과 학사 졸업



이 호 원 (Lee, Ho-won)
2005년 아주대학교 박사졸업 예정(건설교통공학부)
1995년 6월~현재 : 도로교통안전관리공단 선임연구원
1994년 7월~1994년 12월 : 교통개발연구원 위촉연구원
1993년 3월~1995년 2월 : 아주대학교 일반대학원 석사(교통공학 전공)
1985년 3월~1993년 2월 : 아주대학교 공과대학 학사(산업공학 전공)

오 영 태 (Oh, Yong-Tae)
2005년~현재 : 대한교통학회 제 1부회장
2004년~2005 : 아주대학교 교무처장
2004년~현재 : 건교부 중앙도시교통정책 심의위원
1993년~현재 : 아주대학교 환경건설교통공학부 교수

이 철 기 (Lee, Choul-Ki)
1991년 : 아주대학교 대학원(석사)
1998년 : 아주대학교 대학원(교통공학박사)
2000년 : 미국 Texas A&M University TTI(Texas Transportation Institute) Visiting Scholar 과정
2004년 : 서울지방경찰청 교통개선 기획실장 및 COSMOS 추진 기획단장
현 재 : 아주대학교 교통연구센터 부센터장
현 재 : 아주대학교 ITS 대학원 교수