시계열 모형을 이용한 GPS 프로브 자료의 이상치 제거 알고리즘

A Time Series based Outlier Elimination Algorithm for GPS Probe Data

최기주
(아주대학교 교통공학, 교수)

장정아
(아주대학교 교통공학, 박사과정)

Key Words : GPS 프로브, 이상치 제거, 시계열 모형(ARIMA 모형), 신뢰도

목 차

Ⅰ. 서 론

첨단여행자교통정보시스템(Advanced Traveler Information System; 이하 ATIS) 분야의 주 기능은 교통정보를 수집·분석하여 다양한 매체를 통하여 제공할 수 있는 교통정보체계를 구축하는 것이다.

이러한 ATIS에서 일반적으로 교통정보를 제공하기 위해서는 가장 기초가 되는 단계는 교통정보의 수집단계이다. 교통정보를 수집하는 방법으로는 일반적으로 루프검지기, 영상검지기, 초음파 검지기등을 통한 지점검지체계와 AVL, cellular phone 및 GPS 프로브차량을 이용한 구간검지체계가 있다. 이중 구간검지체계는 수집표본수가 적절하다면 지점검지체계에 비해 공간적, 시간적으로 정확한 정보를 확보할 수 있다는 점 등으로 인해 관심이 급증하고 있는 분야이다.

그러나, 구간검지체계의 경우 GPS 프로브 차량을 이용하거나 AVL을 이용할 경우 시·공간적 데이터의 부재 및 손실 등에 인한 정보의 누락이 발생할 수 있다. 이러한 문제는 프로브 대상 차량의 확보라는 시스템의 확장을 통해 지속적으로 해결이 가능한 문제이다. 그러나 보다 근본적인 문제가 있는데 이는 프로브 차량에 얻어진 교통데이터는 도시부 도로구간에서는 샘플링에 의해 추출된 형태를 띠고 있어, 데이터 자체에 이상치적인 요소가 나타날 수 있다. 혹은 프로브 대상차량이 일반 승용차가 아닌 택시나 버스와 같은 대중교통차량일 경우 프로브 자체의 운행특성이 달라(예를 들어 택시의 경우 공차와 승차, 버스의 경우 정류장의 대기 시간 등)에 의해 이상치적인 요소가 발생할 가능성이 있다.

본 연구는 이러한 구간검지체계의 수집 및 가공체계에서 고려될 문제점을 검토해보고 그 해결책을 제시할 것이다. 2장에서는 구간검지체계에서의 GPS 프로브 차량 방법을 검토해보고, 이후 GPS 프로브 차량에 의해 구간통행시간 산출시 야기될 수 있는 문제점으로 이상치의 의미를 정립하기로 하겠다.

3장에서는 수집 교통정보의 신뢰성 제고를 위한 GPS 프로브 이상치 제거 알고리즘으로 시계열 모형을 적용하는 알고리즘을 제시할 것이다. 4장은 실제 데이터를 이용하여 알고리즘을 평가 결과를 수록하고 있다. 이후 본 연구의 한계와 결론을 5장에서 제시하기로 한다.

Ⅱ. 이론적 배경 및 개념의 정립

1. GPS 프로브를 이용한 구간 검지 체계

소통상태를 판정하여 교통운영이나 교통정보를 제공하려는 방식으로서 측정 지점에 따라 구간검지와 지점검지로 분류할 수 있다. 지점검지는 지점을 중심으로 검지체계를 설치하여 해당지점의 통행속도, 통행시간, 밀도 등에 대한 검지를 수행함을 일컫는데 반해서 구간검지는 도로구간의 특정 공간상의 검지라고 할 수 있다. 즉, 지점검지가 점 위주의 point-based 라고 하면 구간검지는 segment-based 라고 할 수 있다.

구간검지체계에서 통행시간을 수집하는 것은 다른 도로구간 또는 링크에서의 통행시간이나 평균속도를 수집하는 것을 의미한다. Circle-X 알고리즘(최기주, 신치현, 1997)과 같은 방법론으로 통행시간을 산정한다.

일반적으로 구간검지체계의 프로브 차량에 의한 시스템은

크게 5가지 유형으로 분류되며, 이들 시스템에 대하여 간단히 설명하면 다음과 같다(Turner, 1996).

- Signpost-based Automatic Vehicle Location(AVL): 프로브 차량은 구축된 signpost를 지날 때 무선통신을 통해 자료를 수집하는 시스템으로 대중교통관련부문에 주로 사용됨.
- AVL : 프로브 차량은 전자식 태그(tag)를 장착하여 노변의 송수신기와 통신을 함으로써 차량의 ID를 확인하고 자료를 수집함
- Ground-based Radio Navigation: 대중교통 혹은 상업용 차량의 관리를 위해 종종 사용되며, 자료는 프로브차량과 무선 송수신 타워 사이의 통신에 의해 자료가 수집됨
- Cellular Geo-location : 이 실험적인 방식은 휴대용 전화기의 위치추적에 의해 통행시간 자료를 수집하는 방식
- GPS : 프로브 차량은 GPS 수신기를 장착하고 GPS위성으로부터 신호를 수집하기 위해 양방향 통신을 수행함. GPS 신호로부터 위치정보를 결정하기 위해 프로브 차량의 실시간 위치 표시를 제어하는 센터에 보내지고, 통행시간 정보는 수집된 자료로부터 결정됨.

2. 교통데이터의 이상치란 무엇인가?

1) GPS 프로브 차량의 교통데이터

GPS 프로브 차량에 의하여 교통정보를 수집한다는 것은 원시적으로 시각(일반적인 최소 단위: 초)마다 차량의 위치 데이터인 원시 데이터(raw data)와 이를 특정주기(일반적인 단위: 5분주기)마다 원시 데이터를 평균화하는 1차 가공된(통합된 혹은 aggregate한) 교통 데이터로 나눌 수 있다.

2) 대표치 산정

이러한 실시간으로 수집된 원시자료에 대하여 설정된 정보의 분석주기(혹은 갱신주기) 동안 누적된 자료의 통합 과정에서 추출된 값의 대표성을 검토해 볼 필요가 있다. 일반적으로 대표하는 정보의 추출을 위해서 사용되는 통계적 방법론은 다음과 같은 산술평균의 사용이다.

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

여기서, LTT_t = 특정주기 t의 링크통행시간의 대표치

$PLTT_i$: i번째 프로브의 개별 링크통행시간

N: 특정주기 t 에서의 수집된 총 프로브 대수

대표적 중심치를 제시하는 다른 통계적 방법론으로 중앙치나 최빈치를 사용할 수도 있으나, 다른 통계치는 추정치의 바람직한 특성인 비편파성, 효율성, 일관성의 문제에서 평균을 사용하는 것보다 오히려 바람직하지 않은 결과를 초래할 수 있다. (장상희, 2001) 따라서 원시데이터를 1차 가공된 교통데이터를 산정하기 위해 위와 같은 산술평균을 사용하는 것은 바람직하다고 사료된다.

3) 이상치의 정의

이렇게 링크 통행시간의 교통데이터를 만들기 위해 산술평균으로 1차 가공을 할 경우 이상치에 대한 문제가 발생한다. 이는 산술 평균의 경우 가중치이므로, 심하게 편향된 데이터에 의해 중심의 위치가 크게 변동될 수 있는 통계치이기 때문이다. 본론을 논하기에 앞서 이러한 (GPS 프로브차량에서의) 이상치(outliner)의 개념을 살펴보면 다음과 같다.

- 이상치 정의 1: 특정 GPS 프로브(예를 들어 택시나 버스) 차량의 운행특성에 의해 일반 승용차량의 통행시간을 기준으로 하여 주행과 관련 없는 정보[1]에 의해 수집된 원시 프로브데이터
- 이상치 정의 2: 자료계열에서 관측치들의 대부분에 의해 제시된 형태를 이루지 못하는 관측치(즉 주기시간동안 특히 편향되는 링크통행시간 정보)

본 연구에서는 이러한 이상치의 개념은 후자의 개념으로 접근하려 하며 전자보다 광의적 개념으로 대표치를 평균값으로 산정할 경우 나타나는 문제점으로 볼 수 있다.

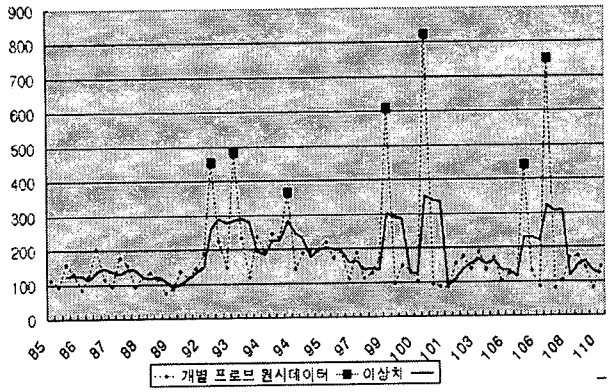
다음 [표 1]과 [그림 1]을 살펴보면[2] 이상치로 보이는 원시데이터를 포함하여 주기별 통행시간 대표치로 산출하였을 때 그 결과가 상당히 상이하게 나타나는 것을 보이는 예이다.

<표 1> 그림 1의 데이터 예제

주기	이상치 제거전		이상치제거후		제거된 이상치 개수	전후 통행시간차
	프로브대수(대)	산술평균치(초)	프로브대수(대)	산술평균치(초)		
90	1	114	1	114	0	0
91	1	106	1	106	0	0
92	3	223.33	2	147.5	1	75.83
93	6	209.66	5	163.4	1	46.27
94	6	206.33	5	178	1	28.33
95	3	181	3	181	0	0
96	2	202	2	202	0	0
97	2	142	2	142	0	0
98	2	116.5	2	116.5	0	0
99	4	240	3	122.33	1	117.67
100	3	368.67	2	105.5	1	263.17
101	3	72.67	3	72.67	0	0
102	2	145	2	145	0	0
103	2	149.5	2	149.5	0	0
105	3	123.67	3	123.67	0	0
106	5	159.8	4	104.25	1	55.55
107	1	683			1	683
108	3	101.67	3	101.67	0	0

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

2) 본 예는 2002년 12월 20일 오전 11시~1시까지 택시 GPS 프로브에 의해 수집된 데이터임



<그림 1> 원시데이터 예제

4) 이상치 제거의 필요성

위와 같이 교통자료에 이상치가 존재하면 그것은 검정통계량을 무효화시키고, 모수 추정을 왜곡시키고 결국 틀린 통계적 추론을 유도하게 된다. 즉 교통정보의 수집차원에서의 이상치 제거 알고리즘이 수반되지 않는다면 교통정보의 가공단계에도 이상치의 영향으로 수집 및 가공의 신뢰도에 영향을 미칠 수 있다.

교통수집데이터의 신뢰성에 문제가 있다는 것은 다음 두 가지의 오류가능성을 말하는 것이다.

- 이상치가 아닌데 이상치로 판별이 날 가능성
- 이상치 일 것 같은데 이상치 판별이 안 날 가능성

두 가지의 오류 가능성 모두 심각한 문제가 될 수 있지만 보다 중요한 문제는 선행적으로 이상치를 판별시켜 원시 데이터 차원에서 제거 시켜야 한다는 것이다.

따라서, 본 연구는 다음과 같은 영가설을 설정하고 문제의 해법을 모색하고 있다.

$$H_0: P\ LTT_i = \text{정상치}$$
$$H_1: P\ LTT_i \neq \text{정상치} = \text{이상치}$$

III. 시계열 모형을 이용한 이상치 제거 알고리즘

1. 시계열 모형

1) 개요

시계열 분석모형은 시간이 흐름에 따라 변하는 현상을 관찰함으로서 얻어지는 일련의 자료를 통해 모형을 수립하고 예측하는 과정을 말한다. 모형수립은 관찰된 시계열의 평균, 분산 등의 성질을 조사하고, 그 확률적 특성을 고찰하여 적합한 시계열 모형을 찾는 것이고, 예측은 모형을 토대로 미래의 값을 제시하는 것이다.

시계열 자료는 일반통계자료와는 달리 기본적으로 확률구조가 독립이 아니라 종속이다. 따라서 시계열 자료의 확률적 특성을 파악해야 적절한 모형화가 가능하다. 일반적으로 AR(p), MA(q), ARMA(p,q), ARIMA(p,n,q) 모형등이 있으며 Box-Jenckins 법에 의한 모형의 식별-모수추정-모형진단-예측과정이 널리 사용되고 있다(김원경, 1998).

2) Box-Jenkins 모형의 절차

① 모형의 식별(Model Identification) 단계

모형의 식별 단계에서 가장 중요한 것은 정상성을 확인하는 것이다. 정상성이 없는 비정상 시계열 모형이라면 정상적으로 변환 혹은 차분이 필요하며, 정상성이 확보된 시계열 모형의 경우는 관찰된 자료가 어떤 시계열 모형으로부터 추출된 표본인가를 밝혀야 한다. 이러한 모형의 식별은 표본자기상관함수와 표본 편자기 상관함수를 구하여 ARIMA(p, d,q) 모형에 적합한가를 결정한다.

② 모수의 추정(Model Estimation) 단계

모형이 식별이 된 후 다음단계로 모수를 추정하여야 한다. 적률법, 최소제곱법, 최우추정법 등으로 추정이 가능하다.

③ 모형의 진단)Model Disgnostic Check) 단계

모수의 추정 후 모형이 관찰된 시계열 자료에 얼마나 잘 부합되는가를 조사하여야 하는데 이것이 바로 모형의 진단과정이다. 일반적으로 잔차 분석과 과다 적합화 분석이 있다.

④ 모형의 예측(Model Forecasting) 단계

시계열 모형의 사용은 과거의 자료로부터 미래를 예측하는 것이다. 이론적으로는 관찰된 자료를 조건으로 하여 미래 값에 대한 조건 분포를 구하는 것과 같다.

2. 이상치 제거 알고리즘

일반적으로 예측모형에서 유용하게 사용되는 시계열 모형에서의 모형화 과정을 응용하여 이상치 제거의 기준이 될 수 있는 상한값(Upper Bound;이하 UB)와 하한값(Lower Bound; 이하 LB)값을 제시하는 알고리즘의 흐름은 다음과 같다.

① **1단계**:주기별 LTT를 이용하여 시계열 모형식을 추정한다.

- 추세를 제거, 정상 시계열을 만든다. (d 혹은 D의 결정)
- 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 그 중에 적합한 것을 고른다.

② **2단계**: 신뢰수준 α에서 UB와 LB를 결정한다.

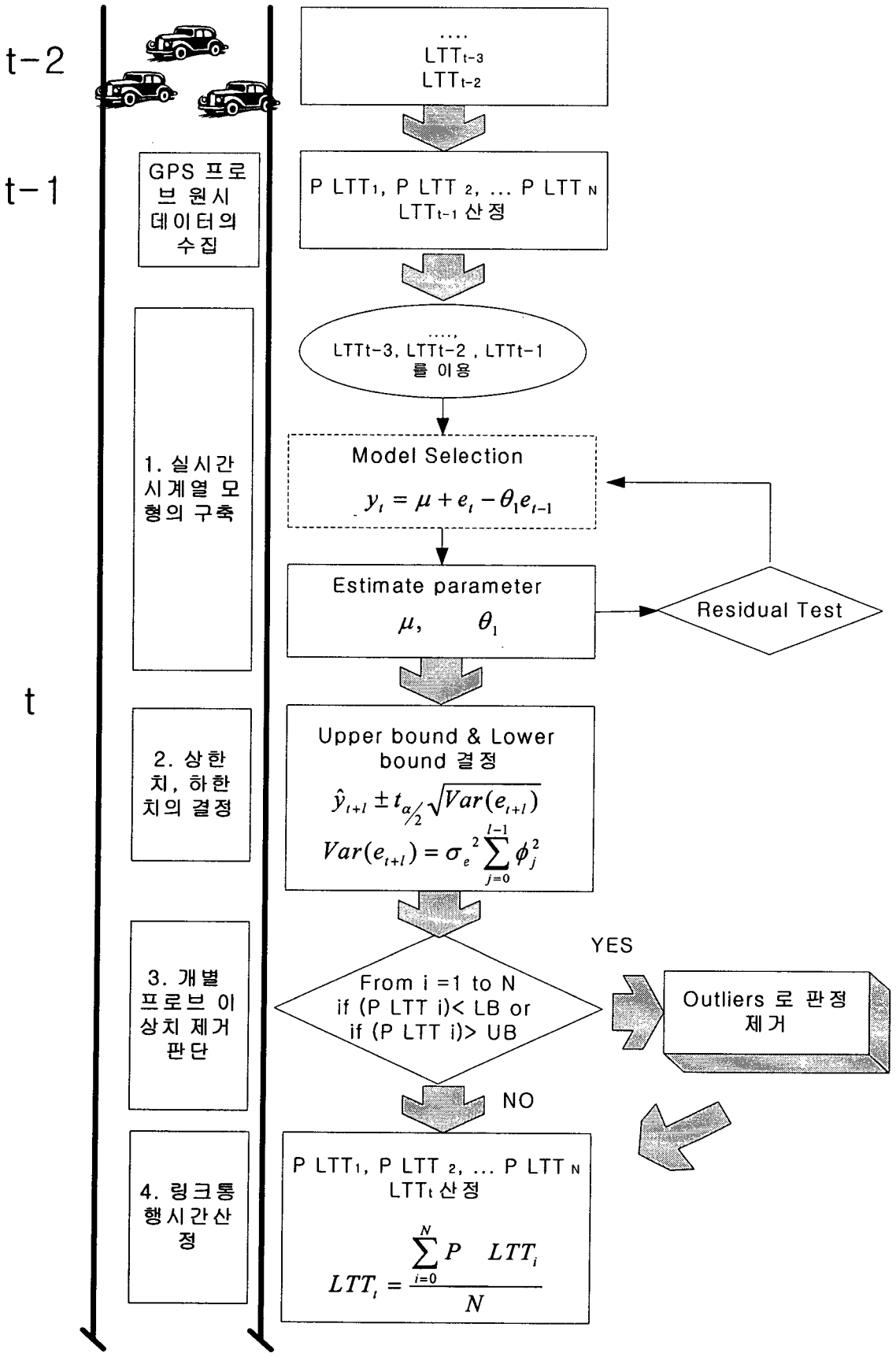
$$UB = \hat{y}_{t+l} + t_{\alpha/2}\sqrt{Var(e_{t+l})}$$

$$LB = \hat{y}_{t+l} - t_{\alpha/2}\sqrt{Var(e_{t+l})} \quad \text{단} \quad Var(e_{t+l}) = \sigma_e^2 \sum_{j=0}^{l-1} \phi_j^2$$

③ **3단계**: 현 주기(t)내에 수집된 개별프로브 링크통행시간 ($P\ LTT_i$)에 대하여 UB와 LB와 비교하여 이상치를 결정하여 제거한다.

④ **4단계**: 이후 새로이 통행시간을 계산한다.

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



<그림 5> 시계열 모형을 적용한 이상치 제거 알고리즘

IV. 모형의 적용

1. 데이터

1) 시간적 범위

2002년 12월 17일, 18일(화, 수)의 하루 24시간으로, 수집 및 분석주기는 5분단위이다.

2) 공간적 범위

다음과 같이 총 10개의 도로 구간의 GPS 프로브 데이터를 이용하였다. 여기서, GPS 원시 프로브 개별데이터는 링크통행시간(통과된)의 값을 이용하였다.

<표 2> 분석지역에 대한 설명

#	구명	도로명	기점명	종점명	구간거리
1	강남구	논현로	차병원	학동역	792
2	강남구	논현로	학동역	차병원	792
3	강남구	매봉터널	도곡업무타운	영동세브란스	709
4	강남구	매봉터널	영동세브란스	도곡업무타운	709
5	강남구	봉은사로	차병원	제일생명사거리	881
6	강남구	봉은사로	제일생명사거리	차병원	881
7	서초구	올림픽대로	한남IC	동호대교남단	1388
8	서초구	올림픽대로	반포대교남단	한남IC	2392
9	동대문	천호대로	장한평역	군자교서단	536
10	동대문	천호대로	장수삼거리	장한평역	630

3) 모형의 추정 방식

모형은 각 링크별로 17일의 240주기[3]의 데이터를 근간으로 시계열 모형을 식별하여, 초기 모수를 추정하였다. 이후 2시간 단위, 6시간, 하루 단위로 모수를 변경 추정하여 이상치 제거 알고리즘을 적용하였다.

2. 모형 적용결과

1) 시계열 모형의 식별결과

모든 링크가 다음과 같은 ARIMA(0,1,1) 즉 IMA(1,1)모형으로 식별되었다. 따라서 모형식별과정은 보통 분석가의 주관적으로 결정해야 될 때가 많으므로 추후 시스템 적용식 모두 IMA(1,1)형태로 식별하고, 일정 기간동안 업데이트 하는 방안도 좋을 것으로 사료된다.

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

2) 모수추정결과

IMA(1,1) 모형의 경우 μ, θ 값만을 추정하면 되므로 이를 추정하였을때 17일 전날 초기 추정치는 다음과 같다.

3) 하루 24시간이면, 24시간*12주기/시간 = 288주기여야 하나, 새벽 1시~5시 사이는 프로브 대수가 극히 적어 모형 추정에서 제외되었음

<표 3> 모형의 초기 24시간에 대한 모수 추정치

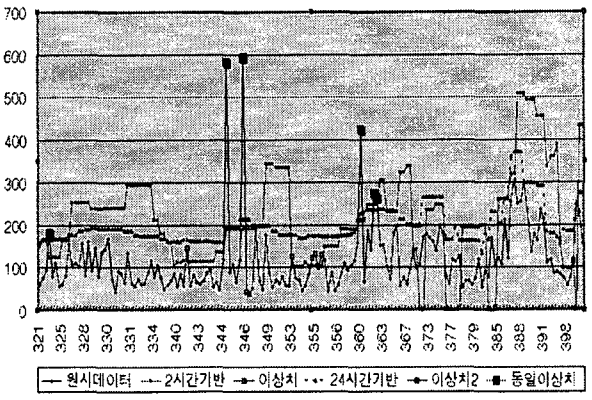
#	μ	θ
1	0.01964	0.77539
2	0.25358	0.29952
3	0.283599	0.63475
4	0.12929	0.75319
5	-0.0404	0.43985
6	-0.00846	0.71456
7	-0.09112	0.43539
8	-0.27387	0.4047
9	-0.0021	0.8112
10	-0.03826	0.54588

3) 이상치 제거 결과

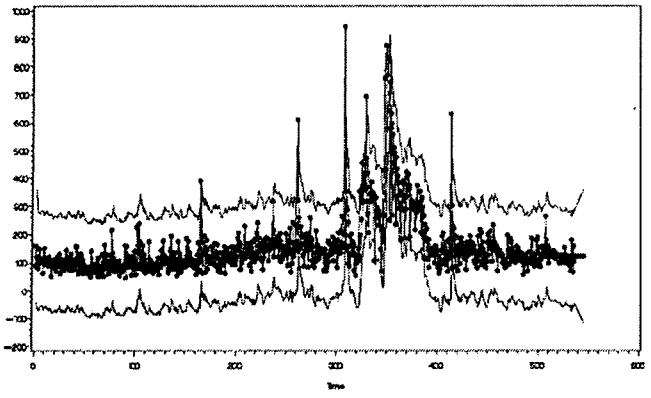
다음은 모수추정 업데이트 시간을 변동시킴에 따라 이상치 제거 개수를 나타낸 표이다. 모수 추정 업데이트 시간을 길게 함에 따라 시간에 따른 혼잡의 추세를 길게 평활화시키기 때문에 원시데이터를 이상치로 더 판단하여 제거하고 있는 경향을 보이고 있다. 이에 비해 모수 추정기간을 짧게 하여 실시간적으로 할 수록 바로 전주기대의 혼잡상황이 반영되어 이상치를 덜 제거하는 경향을 보이고 있다.

<표 4> 모수 추정기간별 이상치 제거 결과

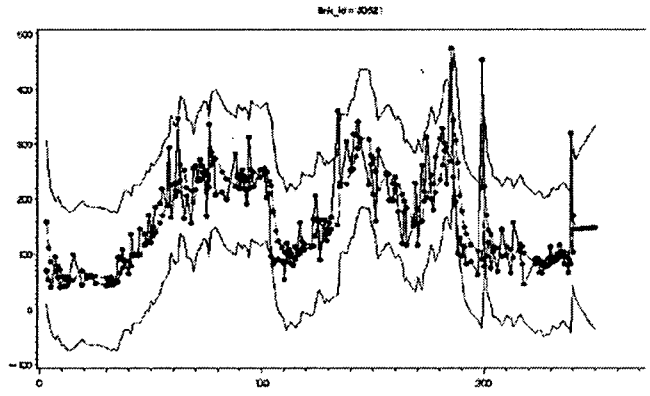
#	모수 추정 업데이트 시간			수집된 총 프로브 데이터 수
	2시간기반 추정	6시간기반 추정	24시간 기반 추정	
1	36	40	43	1399
2	22	32	63	1215
3	6	14	11	439
4	32	52	60	1256
5	45	54	58	1207
6	39	63	105	1364
7	100	112	126	1507
8	17	32	32	1230
9	4	7	9	481
10	4	8	9	255
총 합계	305	414	514	10353



<그림 8> 모수 추정기간별 이상치 제거 추세



<그림 9> 24시간 기반 링크 3의 모형예측결과



<그림 10> 24시간 기반 링크 5의 모형예측결과

V. 결론 및 향후과제

현재까지 수행된 국내외 많은 구간교통정보수집에 관한 연구 중에서 수집된 데이터의 이상치를 검지하고 제거하는 알고리즘에 대한 연구는 많지 않았다. 본 연구에서는 수집단계 차원에서의 이상치 제거의 필요성을 설명하고 이후 이상치에 대한 개념정립을 하였다. 이상치란 자료계열에서 관측치들의 대부분에 의해 제시된 형태를 이루지 못하는 관측치로서 주기별 산술평균 대표치를 산정하기 위해서는 이상치 여부를 판단하여 제거하는 알고리즘이 필수적이다.

연구에서 제안한 시계열 모형은 보통 예측력이 우수하여 추정 및 예측에도 널리 사용하는 모형이다. 본 연구에서는 시계열 모형의 신뢰구간 추정 부문을 이상치 제거 모듈과 결합하여 이상치 제거 알고리즘을 개발하였다.

그 결과 이상치의 제거에 과정에 있어 추정기간에 따라 민감하게 반응하는 결과를 볼 수 있었다.

본 연구의 한계와 향후과제에 대하여 설명하면 다음과 같다.

첫째, 적절한 시계열 모형을 찾기 위한 링크별 시간대별 혹은 요일별 추정기간, 추정에 사용되는 데이터베이스의 기간추정기간 등에 따라 다양한 시나리오가 가능하다. 실제 가장 최적의 안을 찾기 위해서는 각 시나리오들에 대한 다양한 검토가 필요하다.

둘째, 가장 최적의 이상치 제거를 위한 시계열 모형을 찾더라도 도시부 도로에서 중요한 외생 변수인 신호 변수의 요소(신호주기, 현시, 신호운영방법 등) 지속적인 혹은 일정주기별 모형의 업데이트 및 검증 작업이 수반되어야 한다.

셋째, 본 연구는 신뢰수준 α의 이상치로 판단된 개수를 비교하는 수준으로 마무리하고 있다. 여기에 중요한 문제는 실제 이상치로 판단된 프로브 개별 데이터에 대한 실제적인 확인 불가하다는 것이다. 즉 제거된 원시 프로브 값이 실제 이상 주행 등의 원인을 파악하기 힘들다는 것이다. 이러한 부문을 확인하기 위해서는 막대한 시간과 비용의 실사작업이 필요하다(개별 프로브 데이터의 운행특성이 검토되기에는 현재의 대부분의 시스템은 적용할 수 없는 부문이기도 하다)

마지막으로 본 연구에서 검토한 ARIMA 모형은 통행시간이나 통행속도의 추정 혹은 예측과정에 사용되는 것으로 본 연구에서 활용이 되게 되는 이상치 제거 모듈로 사용하기 위해서는 시스템의 적용성면에서 시스템의 실시간화, 부하 문제등의 문제성을 확인해야 된다.

본 연구에서는 도시부에서 구간검지체계에서의 수집데이터의 가공이전 전처리부문으로서 이상치의 제거에 관한 실증적 연구를 수행하고 적용가능성을 검토하였다는데 중요한 의의가 있다. 특히 프로브 데이터의 추세가 IMA(1,1)모형으로 일관적으로 추정될 수 있다는 것을 확인할 수 있었다. 또한 원시 프로브 데이터의 개별 운행 특성을 확인하지 않고, 확률적이며, 데이터 기반의 통계적 추정방법을 활용하였다는 데 쉽게 적용가능한 방법론이 될 수 있다. 그러나, 연구결과 몇 가지 한계가 도출되었으며 이를 보완하기 위한 지속적인 연구도 함께 필요하다고 사료된다.

참고문헌

1. 최기주, 신치현, GPS와 GIS를 이용한 링크통행시간 산정에 관한 연구, 대한교통학회지, 제 16권 제 2호, 177-195, 1998
2. 김원경 외, 시계열분석, 교우사, 1998
3. 장상희 외, 사회통계학, 박영사, 2001
4. 아주대학교 교통연구센터, 구간 교통정보 보정 알고리즘 개발 및 교통정보 신뢰도 평가, 2003
5. 아주대학교 교통연구센터, 구간교통정보 보정 알고리즘 개발을 위한 데이터 분석, 2002
6. Turner S, Advanced Techniques for Travel Time Data Collection, TRR No.1551, TRB, 57, 1996