▣ 論 文 ▣

국소가중다항회귀분석을 이용한 이상치제거 및 자료보정기법 개발
(GPS를 이용한 개별차량 주행속도를 중심으로)

Correction of Erroneous Individual Vehicle Speed Data Using Locally Weighted Regression (LWR)

임 희 섭
(한양대학교 교통시스템공학과 학석사과정)

오 철
(한양대학교 교통시스템공학과 교수)

박 준 형
(한양대학교 교통시스템공학과 석사과정)

이 건 우
(Univ. of California at IrvinePh. D. Candidate)

목 차

Ⅰ. 서론
Ⅱ. 국내외 관련연구 고찰
Ⅲ. GPS를 이용한 개별차량 속도자료 수집
1. 자료수집 과정
2. 자료구축 과정
Ⅳ. 방법론 적용
1. Synthetic Outlier 생성 및 주입
2. 국소가중다항회귀분석과 지수평활화
3. 파라미터 결정방법
4. 이상치 보정
5. 분석결과
Ⅴ. 결론
참고문헌

Key Words : 교통정보, 국소가중다항회귀분석, 이상치 보정, 주행속도, 필터링
Data Filtering, Locally, Weighted Regression(LWR), Outlier Correction, Travel Speed, Traffic Information

요 약

현장에서 수집되는 교통원시자료는 수집장비의 결함 및 주변환경 등에 의해 다양한 이상치가 발생한다. 원시자료의 품질은 추가 가공을 통해 생성되는 교통정보의 신뢰도에 직접적인 영향을 미치는 중요한 요인이다. 실시간으로 수집되는 교통원시자료를 1차 가공하는데 있어서 핵심은 이상치(Outlier)를 검지하고 보정하는 것이라고 할 수 있다. 본 연구에서는 GPS장비를 이용해 얻은 개별차량의 주행속도에서 발생하는 이상치를 제거하고 보정하는 기법을 제안하였다. GPS는 광범위한 교통네트워크상의 차량추적에 용이하게 사용될 수 있는 장점이 있다. 수집된 개별차량의 주행속도에서 이상치를 검지하고 보정하기 위해 국소가중다항회귀분석(LWR: Locally Weighted Regression)을 적용하였다. 또한 국소가중다항회귀분석을 수행하기 위한 파라미터 결정 알고리즘을 개발하여 적용하였다. 개발된 필터링 기법의 성능 평가를 위해 Synthetic Outlier를 생성 및 주입하여 개발된 필터링 기법을 통해 보정시키고 원시자료와 비교 분석 하였고, LWR을 이용한 기법의 상대적 성능 평가를 위해 지수평활화를 이용한 기법과 비교하였다. 평가 결과 LWR기법이 지수평활화를 이용한 기법보다 낮은 오차율을 보여 상대적으로 우수함을 검증하였다. 본 연구에서 제안한 방법론은 교통정보공학 분야의 자료처리 및 정보가공을 위한 도구로서 활용도가 클 것으로 기대된다.

Effective detection and correction of outliers of raw traffic data collected from the field is of keen interest because reliable traffic information is highly dependent on the quality of raw data. Global positioning system (GPS) based traffic surveillance systems are capable of producing individual vehicle speeds that are invaluable for various traffic management and information strategies. This study proposed a locally weighted regression (LWR) based filtering method for individual vehicle speed data. An important feature of this study was to propose a technique to generate synthetic outliers for more systematic evaluation of the proposed method. It was identified by performance evaluations that the proposed LWR-based method outperformed an exponential smoothing. The proposed method is expected to be effectively utilized for filtering out raw individual vehicle speed data.

본 연구는 국토해양부 교통체계효율화사업(2006)『U-Transportation 기반기술 개발』과제의 지원으로 수행되었음.

Ⅰ. 서론

최근 정보기술(IT)의 급속한 발전으로 인해 유비쿼터스 환경의 실현이 가속화되고 있다. 교통분야에서는 실시간으로 이용자, 교통시설, 교통수단을 네트워킹하여 보다 안전하고 편리한 인간중심의 교통시스템을 개발하고 있다. 유비쿼터스 환경에서의 교통네트워크의 구현을 위해서는 방대한 양의 자료를 수집, 처리, 가공할 수 있어야 한다.

보다 진보된 통신 및 센서 기술의 발달은 교통공학의 다양한 분야의 기술 개발에도 큰 공헌을 하고 있다. 가장 대표적인 분야가 교통정보공학(Traffic Information Engineering)라고 말할 수 있다. 교통정보공학에서는 도로교통상황을 나타내는 교통자료를 실시간으로 수집하여 이용자가 요구하는 보다 정확하고 신뢰성 있는 정보로서 가공한 후 제공하는 일련의 프로세스를 다룬다. 이 중 가장 기본이 되는 부분 중의 하나는 현장에서 수집되는 실시간 원시 교통자료(raw traffic data)를 1차 가공하는 것이다. 1차 가공의 핵심은 이상치(outlier)를 검지하고 보정하는 것으로서 다음 단계의 자료 처리 및 가공을 위한 출발점이라고 말할 수 있다.

현장의 교통원시자료는 현재 가장 널리 사용되고 있는 지점검지기인 루프검지기로부터 수집되는 집계된(aggregated) 교통량, 속도, 점유율로부터 개별차량을 추적하여 보다 미시적인 주행정보를 추출하는 단계로 발전해가고 있다. 차량의 추적을 위한 기술로는 차량영상이미지 프로세싱(VIP: Video Image Processing)과 Global Positioning Systems(GPS)를 이용한 기법 등이 적용되고 있다. 개별차량의 추적을 통해 통행시간, 통행패턴, O-D정보 등을 얻을 수 있으며, 다양한 교통상황의 변화를 보다 신뢰성있게 모니터링 할 수 있다. GPS를 이용한 차량추적 기술은 VIP 기반의 기술보다 광범위한 교통네트워크상의 차량추적에 용이하게 사용될 수 있는 장점이 있어 네트워크 교통류 제어 및 관리에 보다 효과적으로 활용될 수 있다.

본 연구의 목적은 GPS를 이용한 개별 프로브(probe) 차량의 속도를 필터링하는 기법을 개발하는 것이다. 이를 위해 국소적인 (local) 자료의 패턴변화에 능동적으로 대응할 수 있는 것으로 알려진 국소가중다항회귀분석(LWR: Locally Weighted Regression)을 개별차량 속도자료에 적용하는 새로운 기법을 제안하였다. 또한, 다양한 이상치의 패턴을 고려한 필터링기법을 개발하기 위해 인위적으로 이상치를 생성하고 주입하는 방법을 제시하였다. 아울러 자료의 필터링 및 평활화를 위해 가장 널리 사용되고 있는 지수평활화기법과 성능을 비교하였다.

본 연구에서 제시하는 개별차량의 속도 필터링 기법은 교통정보공학 분야의 다양한 자료처리 및 정보가공을 위한 도구(tool)로서 그 활용도가 클 것으로 기대된다.

본 연구의 구성은 다음과 같다. 2장에서는 국내외 교통자료처리 분야의 관련 연구를 고찰하여 research opportunity 및 본 연구의 특징을 찾는다. 3장에서는 GPS를 이용한 개별차량 속도자료 수집방법, 4장에서는 LWR에 대한 이론적 배경 및 본 연구에서 개발한 방법론을 제시한다. 아울러 방법론 적용결과를 논의한다. 마지막 5장 결론에서는 연구의 성과를 정리하여 설명하고 향후 연구내용을 제시한다.

Ⅱ. 국내외 관련 연구 고찰

본 연구의 방법론 및 분석과정 정립을 위해 LWR(Locally Weighted Regression)기법 연구, Smoothing 기법연구, 이상치 제거 및 보정기법 등 기존의 관련연구를 고찰하였다.

Toledo(2007)는 차량주행궤적 내에 존재하는 결측자료를 보정하기 위해 LWR기법을 적용하여 LWR기법 적용 시 중요 Parameter인 WindowLength와 Polynomial Order값을 변화시키면서 오차율을 최소화하는 Parameter 값을 산출하고 주행궤적을 예측하는 방법을 제시하였다. Sun(2003)은 Local Linear Regression기법을 이용하여 과거자료와 현재자료를 이용하여 가까운 미래시점의 교통자료를 예측하는 모형을 개발하였다. 그리고 구축된 모형을 Kernel smooting기법, k-nearest neighbor기법과 비교하여 Local regression기법이 타 기법보다 우수하다는 결과를 도출하였다. Cleveland (1979)는 기존의 LWR기법에서 발전된 robust한 LWR기법의 이론적배경과 원리를 제시하고 LWR에 사용되는 Parameter의 영향에 대해 산점도를 smoothing하는 과정을 통해 분석하였다.

김정연(2006)은 검지기에 의해 수집된 교통자료에 존재하는 결측자료에 대해 인접지점참조방식과 이력자료를 활용한 방법론을 적용하여 원시자료를 일정 차로 및 시간단위로 보정하는 과정을 제시하고 그 결과를 상대오차율로 비교하였다. 최기주(2006)는 단속류에서 Taxi

에 설치된 GPS자료를 이용하여 교통정보를 생성할 경우 Taxi의 실제도로주행과 관련 없는 자료를 통계적 이상치라 정의하고 이를 검지 및 제거하는 휴리스틱한 알고리즘을 개발하였다. Jun(2006)은 GPS로 수집된 이동거리, 속도, 가속도자료에 존재하는 Random error를 Kalman filtering기법을 적용하여 교통정보 산출 시 자료 내에서 error의 영향을 최소화하는 방법론을 개발하였다. Weijermars(2006)는 최대교통량과 최소교통량, 교통량보존법칙을 이용해 도시부 교차로에서 수집되는 루프검지기 자료의 이상치를 검지하는 방법을 제시하였다.

Klemiato(1998)는 다양한 수학적 이론 기반의 예측모형을 제시하고 각 모형을 입력자료에 적용한 결과로 Holt의 adaptive exponential smoothing기법이 시계열 자료에 적용하기에 적합하다는 결론을 도출하였다. Peeta(2002)는 Incidents와 Data Faults를 구분하여 이상치를 검지하였고, 이상치를 보정하기 위해 푸리에 급수를 기반으로 한 휴리스틱 알고리즘을 개발하였다.

III. GPS를 이용한 개별차량 속도자료 수집

1. 자료수집 과정

본 연구에서는 GPS 수신기와 노트북을 이용해 이동차량의 주행궤적 및 속도자료를 수집 하였다. 수집된 주행궤적(위도, 경도)과 위성사진을 이용하여 주변 환경에 따른 이상치 발생 여부를 조사하였다. 현장자료 수집을 통해 총 25개의 dataset을 수집하였다. 각 dataset은 약 3000개의 좌표 및 속도자료로 구성되어 있으며 수집간격은 1초이다. 본 연구에서는 속도자료를 이상치 보정 알고리즘의 입력자료로 이용하였다. 자료 수집결과, dataset 25개 중 이상치로 검지된 속도 자료는 총 40개이며, 한 개의 dataset에서 최대 11개의 이상치가 발생하였다. 이상치 발생지점을 지도상에서 확인한 결과 주로 고층건물이 있는 도시부, 터널, IC주변지역 등 이었으며, 이는 일시적 통신두절이 가장 큰 발생 원인인 것으로 판단된다. 〈그림 1〉은 조사 장비를 장착한 차량의 모습과 GPS자료 수집 화면이다.

본 연구를 위한 공간적인 자료 수집범위는 한양대학교 안산캠퍼스와 서울시 송파구 잠실역 사이 구간이며, 연속류와 단속류로 분류하였다.

자료수집 구간을 〈그림 2〉와 같이 지도로 표시 하였으며, 크게 3개 구간으로 구분 하였다. 잠실역에서 외곽

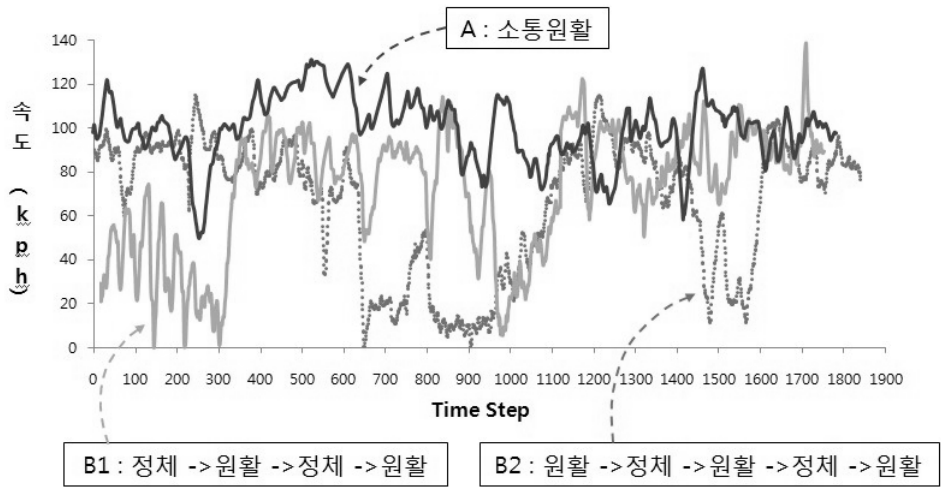
〈그림 1〉 조사장비

〈그림 2〉 자료수집경로

순환고속도로 송파 IC까지를 단속류 구간, 송파 IC부터 수인산업도로 진출부 까지를 연속류 구간, 수인산업도로 진출 후 부터 한양대학교 안산캠퍼스까지를 단속류 구간으로 구분하였고, 본 연구에서는 연속류 구간만을 대상으로 분석하였다.

2. 자료구축 과정

LWR기법을 이용한 이상치 보정 알고리즘의 성능평가를 위해 이상치를 인위적으로 주입한 자료를 구축하였다. 원시자료에 이상치를 주입할 때 고려한 사항은 주입 위치와 주입되는 이상치의 개수로 설정하였다. 그리고 교통상황에 따른 보정 알고리즘의 성능평가를 위해 수집한 속도자료들 중에서 교통상황 패턴에 따라 〈그림 3〉과 같이 크게 3가지로 구분하여 분석 하였다. 전체구간에서 정체가 없는 패턴 A자료, 정체로 시작하여 소통원활과 정체가 반복되는 패턴 B1자료, 소통원활로 시작되어 정체와 소통원활, 정체가 반복되는 B2자료로 나누어 분석 하였다.

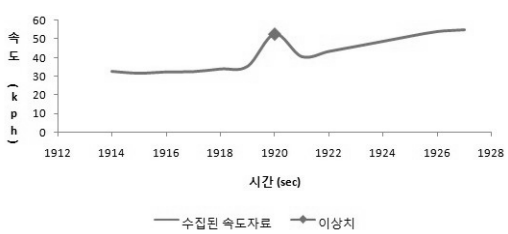


〈그림 3〉 교통상황패턴분석

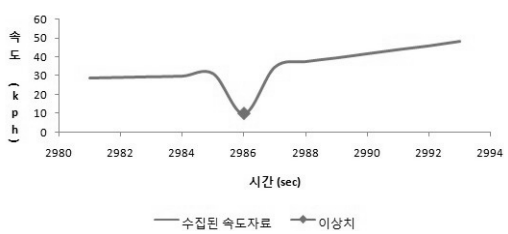
IV. 방법론 개발 및 적용

1. Synthetic Outlier 생성 및 주입

본 연구에서는 보다 체계적이고 과학적으로 이상치를 필터링하고 보정할 수 있는 방법론을 개발하기 위해 인위적 이상치를 생성 및 주입 (synthetic outlier generation and injection) 하였다. 우선 이상치를 정의하기 위해 일반 차량 의 최대가속도($3m/s^2$)와 최대감속도($-5.8m/s^2$)를 가감속도 변화 한계치의 상한값(upper bound)과 하한값(lower bound)으로 설정하였다. 실제 수집한 자료로부터 관찰된 이상치의 예를 〈그림 4〉에 제시하였다.



(a) 비정상 가속으로 인한 이상치



(b) 비정상 감속으로 인한 이상치

〈그림 4〉 이상치의 예시

본 연구에서는 이상치가 없는 원시자료에 이상치를 주입하고 개발된 필터링 알고리즘을 이용해 보정된 값과 원시자료의 값를 비교하였다. 이상치의 주입 시 이상치의 크기(magnitude)는 자료의 평활화 결과 값에 영향을 주게 되므로, 이러한 영향을 실제 상황과 유사하게 구현하기 위해 현장에서 수집된 이상치 자료 40개의 크기에 대한 확률분포함수를 추정하고, 이렇게 추정된 확률분포함수로부터 임의로 추출한 이상치를 원시자료에 주입하였다. 실제 이상치 자료의 분포는 ARENA 프로그램[9]을 이용하여 추정하였으며, lognormal 분포가 가장 적합한 것으로 분석되었다. 〈표 1〉에 분포함수 추정

〈표 1〉 분포함수 추정 결과

Distribution Summary	$f(x) = 3 + \frac{1}{6.91x\sqrt{2\pi}} e^{-(\ln(x)-2.9)^2/95.5}$	
	Square Error	0.00281
Chi-Square Test	Number of intervals	3
	Degrees of freedom	0
	Test Statistic	0.816
	Corresponding p-value	〈 0.005
Kolmogorov - Smirnov Test	Test Statistic	0.12
	Corresponding p-value	〉 0.15
Data Summary	Number of Data Points	40
	Min Data Value	3.05
	Max Data Value	16.2
	Sample Mean	5.38
	Sample Std Dev	2.94

결과를 정리하였다.

평균이 2.9, 표준편차가 6.91인 lognormal분포로부터 임의로 추출된 난수 (random number)를 개별차량의 원시속도 자료에 주입하였다. 또한 비정상적 가속에 의해 발생된 이상치(〈그림 4〉 (a))와, 비정상적 감속에 의한 이상치(〈그림 4〉 (b))의 비율은 실제 수집된 이상치에서의 비율과 같은 83%와 17%를 적용하였다. 〈그림 5〉에 분석대상자료 2%에 해당하는 개수의 이상치를 임의로 주입한 예를 도식화 하였다. 원시자료보다 높은 속도와 낮은 속도의 이상치가 일정 비율로 주입되었으며 이상치의 크기도 다양하게 주입되었음을 보여준다.

2. 국소가중다항회귀분석(LWR)과 지수평활화

1) 국소가중다항회귀분석(LWR)

LWR기법은 국소 가중회귀법에 의한 평활기법으로 가까이에 있는 점들만을 이용하여 평활화 될 각 점의 다항식을 추정하는 방법이다. 이 방법은 자료의 선형성, 주기성, 비선형성에 따라 발생되는 다양한 시계열 자료에 적용하기에 좋은 방법으로서 다항선형회귀분석과 방법이 유사하나 회귀모형을 조건에 따라 정해진 인근의 자료값만을 가지고 분석을 한다는 점이 다르다. 이 기법은 비매개변수법의 하나로 모형에 대한 가정이 없다.[10]

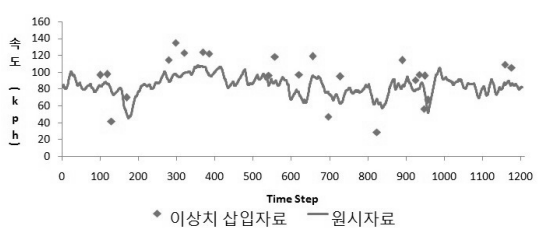
$$y(t) = f_{t_0}(t,\beta_{t_0}) + \epsilon_{t_0,t} \quad (1)$$

$f_{t_0}(t,\beta_{t_0})$: 추정된 LWR함수에 의해 시점 t_0에서 추정된 시점 t의 속도

$\epsilon_{t_0,t}$: 정규분포를 가지는 잔차

$$\min_{\beta_{t_0}} [X_{t_0} - f_{t_0}(t,\beta_{t_0})]^T W_{t_0} [X_{t_0} - f_{t_0}(t,\beta_{t_0})] \quad (2)$$

X_{t_0}: 시점 t0를 중심으로 인접한 속도자료행렬



〈그림 5〉 이상치 자료 주입 예

W_{t_0}: 인접 속도자료에 부여되는 가중치 행렬

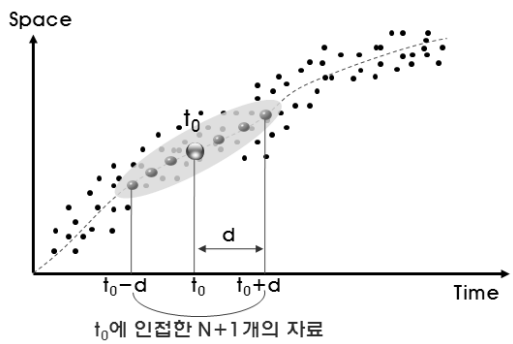
$$w(t_0,t) = (1-u(t_0,t)^3)^3 \quad (3)$$

$$u(t_0,t) = \frac{|t-t_0|}{d},\ d = \frac{N+1}{2}$$

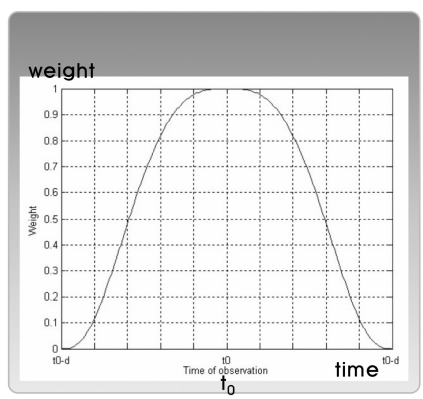
$u(t_0,t)$: 시점 t_0와 인접한 자료의 차이값을 구하는 함수

d : 시점 t_0를 기준으로 함수추정에 사용되는 인접한 점까지의 거리

추정되는 회귀함수의 형태는 식(1)과 같으며 시점 t의 속도를 예측하기 위해 〈그림 6〉와 같이 지점 t_0와 t_0에 인접한 N개의 속도자료를 사용한다. β_{t_0}는 예측지점 t의 추정된 함수인 다항식의 계수이다. 식(2)는 시점 t_0의 속도자료와 가중치를 부여한 인접한 N개의 속도자료로 최소자승법에 의해 다항 함수를 추정하는 식이다. 가중치 행렬(W)은 대칭행렬로서 본 연구에서는 〈그림 7〉



〈그림 6〉 Locally Weighted Regression 기법의 평활화 개념도



〈그림 7〉 LWR기법의 가중치 함수

와 같이 Tri-cube weight function을 사용하였다.〔7〕

LWR을 수행하기 위해서는 세가지의 주요 변수를 Parameter를 설정해야하며 이는 추정되는 다항식 함수의 형태를 설정하는 과정이다. 추정되는 함수의 차수는 Polynomial order값의 설정으로 결정되며 이에 따라 추정계수(β)의 개수도 달라진다.

함수추정 시 사용되는 시점 t_0를 기준으로 인접한 속도자료의 개수를 Window length값으로 설정할 수 있다. Window length값을 크게 설정하는 것은 함수추정 시 참조하는 자료의 수를 많게 한다는 의미로 자료를 평활화하는 강도가 증가한다.

가중치 함수는 식(3)과 같이 예측지점 t_0를 중심으로 인접한 자료에 가중치를 부여하는 함수로서 본 연구에서는 Window length의 범위 내에서 시점 t_0에 인접할수록 큰 가중치를 부여한다.〔7〕

2) 지수평활화

지수평활법이란 가장 최근 자료에 가장 큰 가중치를 두고 시간이 지남에 따라 가중치가 기하학적으로 감소되는 가중치 이동 평균 예측기법 중 하나로 지수함수에 따라 가중치를 갖는다. 지수평활법은 최근자료를 보다 의미 있는 자료로 사용 할 수 있고 과거자료의 경우 그 가중치를 체계적으로 감소시킬 수 있다는 장점이 있으며, 계산이 쉽고 계산을 하기 위한 필요자료가 적기 때문에 일반적으로 예측 시스템에서 널리 사용된다.

$$\begin{aligned} S_0 &= x_0 \\ S_t &= kx_t + (1-k)x_{t-1} \end{aligned} \tag{4}$$

식(4)는 지수평활화 기법을 수식으로 표시한 것으로 S는 평활화 된 속도 자료를 나타내며 x는 입력되는 원시자료이다. k값은 평활화 상수로 최근자료와 과거자료의 가중치를 부여하는 기준이 된다. k값이 작을수록 과거자료의 가중치가 높아져 평활화 정도가 커지게 된다.

3) LWR기법와 지수평활화기법의 차이점

LWR과 지수평활화 기법의 차이점은 다음과 같다. 첫째로 지수평활법은 해당시점에서 인접한 과거자료를 참조하여 평활화를 수행하는 방법론인데 비해, LWR기법은 해당시점을 기준으로 이전 및 이후 시점의 자료를 모두 참조하여 평활화를 수행한다.

둘째로 지수평활법에서는 과거자료에 대한 가중치를 부여할 때 단순히 k값에 따라 부여되며, LWR에서의 가중치는 적용하는 가중치 함수에 따라 다양한 가중치를 부여할 수 있다.

셋째로 지수평활법에서는 k값이 평활정도를 조정하는 parameter로서 k값이 평활정도에 절대적인 영향을 주는 반면에 LWR기법에서는 평활화 또는 예측을 하기위해 참조하는 인접자료의 개수를 설정하는 Window length와 추정하는 함수식의 차수를 결정하는 Polynomial order값에 따라 평활정도가 결정된다.

넷째로 지수평활법과는 다르게 LWR기법에서는 추정되는 결과값의 상·하한범위를 설정할 수 있어 분석 시 다양한 제한조건을 적용할 수 있다.

다섯째로 지수평활법과 LWR기법은 결과값을 산출하는 과정에서 근본적인 차이를 보이는데 LWR은 분석 대상지점을 기준으로 전후로 인접한 자료에 가중치를 부여하여 최소자승법을 통해 회귀식으로 구성된 함수를 추정하여 결과값을 산출하는 반면에 지수평활법은 인접한 과거자료로 비교적 간단한 계산과정에 의해 결과값이 산출된다.

3. 파라미터 결정방법

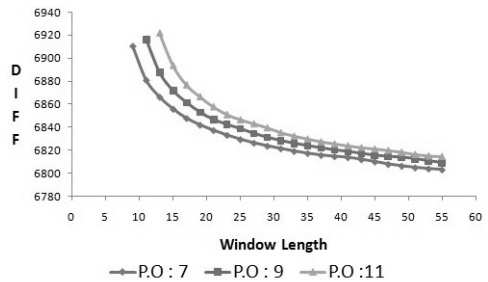
LWR기법에서는 2개의 주요 Parameter인 Polynomial Order(P.O.)값과 Window Length (W.L.)값을 자료평활화 수행 시 설정해야 한다. LWR을 이용한 보정방법과의 성능비교를 위해 동일한 입력자료로 수행한 지수평활화 기법은 평활화계수(K)에 의해 평활화 정도가 결정된다. 본 연구에서는 사전에 설정된(pre-defined) 분석주기 내에서 중앙값으로부터 각각의 자료의 차이값(difference value)의 합(DIFF, 식(5)참조)을 최소로 하는 파라미터를 이상치의 필터링을 위한 최적의 파라미터로 설정하였다. 여기에서 분석주기를 설정한 이유는 동적인 속도패턴의 변화에 필터링 기법이 최대한 능동적으로 대응할 수 있는 파라미터를 찾기 위함이며, 평균 대신 중앙값을 사용한 것은 분석주기 내에 관측된 비정상적인 이상치로 인한 영향을 최소화하기 위함이다. 본 연구에서는 30초와 60초 두 가지 분석주기를 적용하였다.

Find PO, WL, K
to minimize

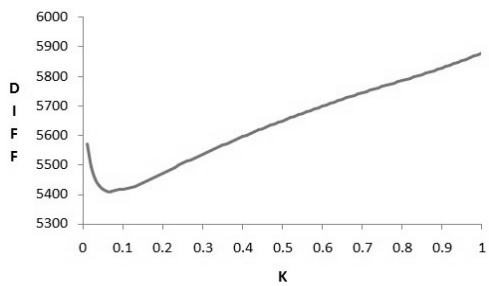
$$DIFF = \sum_{j=1}^{n}\sum_{i=1}^{m}\left|(x_{ij} - X_j^{Med})\right| \tag{5}$$

x_{ij}: j 분석주기의 i번째 속도
X_j^{Med}: j 분석주기의 중앙값

7, 9, 11 의 P.O.값과 다양한 W.L.을 적용한 한 경우에 나타나는 DIFF 값의 변화를 〈그림 8〉에 도식화하였다. 일반적으로 P.O.값이 작고 W.L.값이 클수록 평활화정도가 커진다. P.O.와 W.L.에 따른 개별차량 속도자료의 평활화 수준을 〈그림 9〉에 예로서 제시하였다.

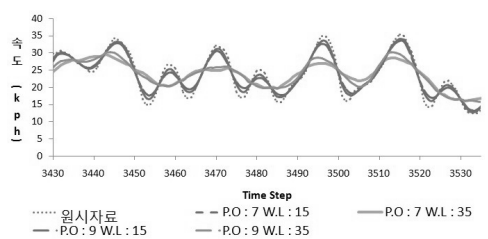


(a) 국소가중다항회귀분석

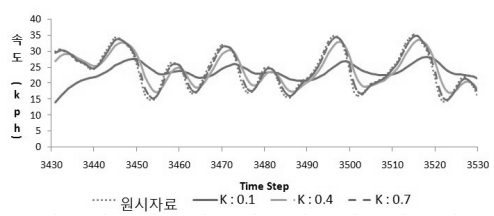


(b) 지수평활화

〈그림 8〉 파라미터 결정을 위한 분석 예시



(a) LWR 적용



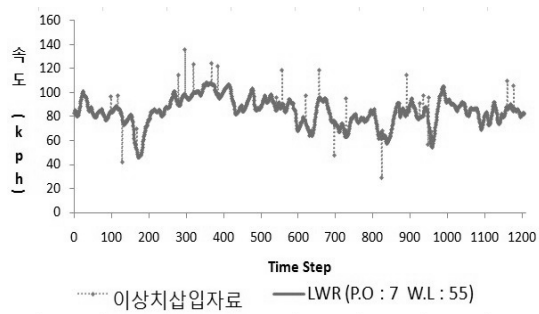
(b) 지수평활화 적용

〈그림 9〉 평활화 수행 예

4. 이상치 보정

실제 도로에서는 교통사고와 같은 돌발상황으로 인해 속도의 급격한 변화가 있을 수 있다. 돌발상황의 경우에는 이상치 발생과는 다른 주행패턴을 보인다. 이상치 발생의 경우 이상치 이전과 이후의 자료 비교 시 한계값 이상의 속도차이가 존재한다. 돌발상황의 경우에는 어느 한 속도자료가 비정상적으로 변화하는 것이 아니라 돌발상황으로 인한 시·공간적 영향권 내에 존재하는 차량들의 속도가 변화하게 된다. 본 연구에서는 이상치가 발생한 경우에 중점을 두어 분석하였다.

수집한 원시자료에 LWR기법과 지수평활화 기법을 각각 적용하여 평활화된 자료를 산출하고 이를 인위적으로 주입된 이상치 값에 대응시켜 이상치를 보정하였다. 보정 알고리즘에 의해 산출된 보정값과 참값을 가진 자료인 이상치 주입 이전의 원시자료와의 차이를 산출하여 본 연구에서 개발한 자료처리 기법의 성능을 평가하였다. 교통패턴A의 원시자료에 2%의 이상치를 주입한 후 LWR을 수행하여 이상치를 보정한 결과를 예로서 〈그림 10〉에 제시하였다.



〈그림 10〉 LWR 기법을 이용하여 이상치를 보정한 예시

5. 분석결과

본 연구에서는 모든 자료가 수집된 상태의 자료에서 이상치를 제거하고 보정하였다. 국소가중다항회귀분석을 이용한 자료처리 기법은 이미 모든 자료가 수집된 Offline자료와 실시간으로 수집되는 Online자료 모두에 적용될 수 있다. 단, Online자료에 적용 할 경우 국소가중다항회귀분석의 특성상 보정할 자료의 이후 자료를 이용하여 보정하기 때문에 일정 시간의 지연이 발생한다. 본 연구에서는 WindowLength의 최대값을 55로 설정하였기 때문에 약 1분 내외의 시간지연이 발생함을 예측

〈표 2〉 이상치 분포와 자료 패턴별 MAPE 비교

이상치 삽입기준		방법	패턴 A		패턴 B1		패턴 B2	
위치	개수		Parameter	MAPE(%)	Parameter	MAPE(%)	Parameter	MAPE(%)
Uniform	60초4개	LWR	P.O. : 7 W.L. : 25	3.763	P.O. : 7 W.L. : 39	8.585	P.O. : 7 W.L. : 39	7.660
		E.S.	K : 0.04	7.266	K : 0.05	28.001	K : 0.05	20.355
	30초2개	LWR	P.O. : 7 W.L. : 25	3.233	P.O. : 7 W.L. : 25	8.154	P.O. : 7 W.L. : 55	7.332
		E.S.	K : 0.05	5.799	K : 0.05	25.310	K : 0.06	17.898
	60초2개	LWR	P.O. : 7 W.L. : 25	3.478	P.O. : 7 W.L. : 37	7.837	P.O. : 7 W.L. : 55	8.593
		E.S.	K : 0.04	6.791	K : 0.05	26.575	K : 0.05	19.978
Random	1 %	LWR	P.O. : 7 W.L. : 55	3.647	P.O. : 7 W.L. : 13	22.007	P.O. : 7 W.L. : 55	6.991
		E.S.	K : 0.03	5.971	K : 0.07	36.800	K : 0.04	11.824
	2 %	LWR	P.O. : 7 W.L. : 55	3.590	P.O. : 7 W.L. : 13	13.460	P.O. : 7 W.L. : 55	11.599
		E.S.	K : 0.03	8.998	K : 0.05	15.066	K : 0.05	27.938
	3 %	LWR	P.O. : 7 W.L. : 55	3.467	P.O. : 7 W.L. : 21	9.752	P.O. : 7 W.L. : 39	5.805
		E.S.	K : 0.03	10.032	K : 0.05	17.669	K : 0.04	16.680

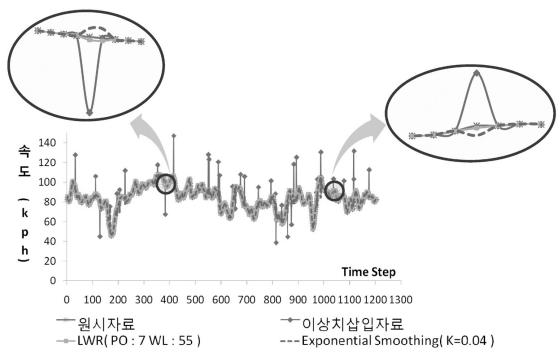
〈표 3〉 주행속도 크기에 따른 MAPE수치 비교 예

교통패턴	관측값(kph)	예측값(kph)	MAPE(%)
정체	4.92	11.73	138.16
소통원활	93.21	87.29	6.34

할 수 있다. 따라서 실시간(Online)으로 구현하기 위해서는 처리되는 자료의 궁극적인 활용방안을 고려하여 WindowLength를 조정하는 추가 연구가 수행되어야 할 것으로 판단된다.

LWR기법과 지수평활화 기법을 이용하여 자료의 이상치를 제거하고 보정하는 과정을 수행한 결과를 〈표 2〉에 정리하였다. 3가지의 교통상황패턴에 따라 원시자료에 이상치를 주입하는 방법을 위치 및 개수에 따라 다르게 적용하고 LWR과 지수평활화 기법의 Parameter값도 변화시키면서 분석을 수행하였다. 이상치의 위치에 따른 이상치 보정 결과의 변화와 개수에 따른 이상치 보정의 결과의 변화를 관찰하기 위해 6종류의 자료를 구축하였다. 이상치 삽입위치가 일정한 자료(Uniform)는 이상치의 개수를 60초마다 4개, 60초마다 2개, 30초마다 2개로 하고, 이상치 삽입 위치가 일정하지 않은 경우(Random)에는 이상치의 개수를 전체 자료의 1%, 2%, 3%로 하여 입력 자료를 구축하여 분석하였다.

본 연구에서 개발한 이상치 제거 및 보정 알고리즘의 평가를 위해 산출된 보정값과 실제 참값을 비교하여 적용된 Parameter별로 MAPE(Mean Absolute Percentage Error)를 산출하였다. 그리고 본 연구에서 제시한 LWR기법을 이용한 평활화 및 보정 기법의 성능을 상대적으로 비교하기 위해 지수평활화 기법을 적용하여 산출한 보정결과를 비교하였다.



〈그림 11〉 LWR과 지수평활화 비교 분석

모든 패턴의 자료에서 LWR기법을 적용한 결과가 지수평활화 기법을 적용한 결과보다 MAPE의 수치가 작게 산출되었다. 이 결과를 통해 본 연구에서 수집한 교통자료에서 LWR 기법을 적용한 방법이 지수평활화 기법을 적용한 방법보다 적합한 방법이라고 판단된다. 특히 패턴 A의 경우 LWR을 이용한 방법은 이상치의 위치나 개수에 관계없이 MAPE가 약 3.5%의 수치를 유지하는 반면 지수평활법의 경우에는 이상치의 개수가 늘어남에 따라 MAPE의 수치가 증가하는 결과가 산출되었다. 따

라서 지수평활법의 경우 입력자료에 포함된 이상치의 개수가 증가할수록 이상치 제거 및 보정기능이 감소되는 것을 알 수 있다.

패턴 B1의 경우에는 이상치 삽입 위치가 일정하지 않은 자료에서 MAPE의 수치가 높게 산출 되었다. 이는 이상치의 위치 선정에 있어 상대적으로 정체 구간에 많은 이상치가 분포되어서 이다. MAPE의 경우 상대적 오차를 구하기 때문에 〈표 3〉과 같이 자료의 크기가 작을 경우 절대적 오차의 크기는 작지만 비율상으론 큰 차이가 생기기 때문이다.

〈그림 11〉은 이상치 자료가 보정되는 과정을 보여주는 예로서, LWR을 적용한 방법이 지수평활화 기법을 적용한 방법보다 적합함을 알 수 있다. LWR기법을 적용한 방법의 경우 알고리즘의 결과값인 보정값과 실제 참값인 원시자료 값이 매우 유사한 값을 가지는 결과가 산출되어 본 연구에서 제시한 방법론이 높은 보정력을 가지는 것으로 판단된다.

V. 결론

본 연구에서는 GPS를 통해 수집된 개별차량 주행 속도자료의 이상치를 제거하고 이를 보정하는 방법론을 제시하였다. 국소가중다항회귀분석 기법을 이용하여 이상치 자료를 보정하고 평활화한 결과를 지수평활화 기법을 적용한 결과와 비교하였다.

소통원활상태의 교통패턴을 나타내는 원시자료에 2%의 이상치를 주입한 후 LWR기법과 지수평활화 기법을 적용하여 이상치를 보정한 결과 LWR 기법을 적용한 경우 3.590%, 지수평활화 기법을 적용한 경우 8.309%의 MAPE가 산출되었다. LWR 기법을 적용하는 방법이 지수평활화 기법보다 원시자료와 비교해 더 유사하게 보정할 수 있다는 결과가 산출되었다.

교통상황패턴에 따라 분석한 결과 속도의 변화가 크지 않은 자료에서는 높은 보정 능력을 보였지만 속도의 변화가 심할 경우에는 보정 능력이 다소 감소되는 결과가 산출되었다. 이에 대한 개선 방안으로는 속도의 변화 정도에 따라 분석 주기를 변화시켜 유동적으로 주기를 설정하는 방법이 있다. 속도 변화가 심한 경우에는 분석 주기를 짧게 하여 주기 내의 속도 분포를 유사하게 하여 분석 한다면 정확도가 높아질 것으로 예상된다.

P.O.값과 W.L값의 설정에 따라 LWR기법의 평활화 정도가 달라지므로 실제 현장에서 자료처리기법으로 활용되기 위해서는 교통상황에 맞게 Parameter값이 설정되고 자료처리과정에 적용되는 알고리즘에 대한 연구가 필요하다고 판단된다.

본 연구에서는 개발된 자료처리기법의 상대적인 성능평가를 위해 지수평활화 기법을 사용하였다. 자료를 평활화하고 보정하는 방법에는 다양한 기법들이 있으며 향후 보다 다양한 기법들과의 비교를 통해 LWR을 이용한 자료처리 방법의 타당성을 검증해야 할 것이다.

고속도로, 터널 등의 도로환경적 요인에 의해 GPS자료가 정상적으로 수신되지 않을 경우에는 이상치가 연속적으로 발생하는 패턴을 보이므로 특정 값의 보정만이 아니라 이상치가 발생한 구간의 보정과정이 필요하다. 따라서 향후 연구에서는 이상치 발생지점을 기준으로 전후의 속도 변화 패턴을 고려해 이상치 구간의 보정이 가능한 알고리즘 개발이 필요하다고 판단된다.

본 연구에서 개발한 개별차량 속도자료 처리 기법은 자료수집 시 다양한 요인에 의해 발생하는 이상치를 효과적으로 제거하고 보정함으로써 수집된 교통자료를 이용하여 보다 신뢰성있는 교통정보를 생성하는데 기여할 것으로 판단된다.

알림 : 본 논문은 대한교통학회 제59회 학술발표회(2008. 10.24)에서 발표된 내용을 수정・보완하여 작성된 것입니다.

참고문헌

1. 김정연・이영인・백승걸・남궁성(2006), “차량 검지자료 결측 보정처리에 관한 연구 (이력자료 활용방안을 중심으로)”, 대한교통학회지, 제24권 제7호, 대한교통학회, pp.27~40.
2. 최기주・최윤혁(2006), “링크통행시간 생성을 위한 이상치 제거 알고리즘 개발”, 대한토목학회지, 제26권 제5D호, pp.731~738.
3. Hongyu Sun, Henry X. Liu, Henh Xiao, Rachel R. He, and Bin Ran(2003), “Use of Local Linear Regression Model for Short-Term Traffic Forecasting” In Transportation Research Record: Journal of Transportation Research Board, No.1836, TRB, National Research Council, Washington D.C..
4. Jungwook Jun, Randall Guensler, and

Jennifer H. Ogle(2006), "Smoothing Methods to Minimize Impact of Global Positioning System Random Error on Travel Distance, Speed, and Acceleration Profile Estimates", In Transportation Research Record: Journal of Transportation Research Board, No.1972, TRB, National Research Council, Washington D.C..

5. Maciej KLEMIATO and Jan Tadeusz DUDA (1998), "Mathematical prediction of time-series in management and control", Proceedings of the Summer School on Computer Science in Management, Kraków, May pp.7~12.
6. Srinivas Peeta, Ioannis Anastassopoulos (2002), "Automatic Real-Time Detection and Correction of Erroneous Detector Data with Fourier Transforms for Online Traffic Control Architectures", Transportation Research Record, No.1811.
7. Tomer Toledo, Haris N. Koutsopoulos, Kazi I. Ahmed(2007), "Estimation of vehicle trajectories with locally weighted regression", In Transportation Research Record: Journal of Transportation Research Board, No.1999, TRB, National Research Council, Washington D.C..
8. W.A.M. Weijermars, E.C. Van Berkum (2006), "Detection of Invalid Loop Detector Data in Urban Areas", In Transportation Research Record: Journal of Transportation Research Board, No.1945, TRB, National Research Council, Washington D.C..
9. W. David Kelton, Randall P. Sadowski, Deborah A. Sadowski, "SIMULATION WITH ARENA", McGRAW-HILL INTERNATIONAL EDITIONS.
10. William S. Cleveland(1979), "Robust locally weighted regression and smoothing scatterplots", Journal of the American Statistical Association, Vol.74, No.368, pp.829~836.

♧ 주 작 성 자 : 임희섭
♧ 교 신 저 자 : 오 철
♧ 논문투고일 : 2008. 10. 24
♧ 논문심사일 : 2008. 12. 8 (1차)
2009. 2. 16 (2차)
2009. 2. 18 (3차)
♧ 심사판정일 : 2009. 2. 18
♧ 반론접수기한 : 2009. 8. 31
♧ 3인 익명 심사필
♧ 1인 abstract 교정필