한국ITS학회논문지
제13권, 제3호 (2014년 6월)
pp.66~77

pISSN 1738-0774

DOI http://dx.doi.org/10.12815/kits.2014.13.3.066

적응형 k-NN 기법을 이용한 UTIS 속도정보 결측값 보정처리에 관한 연구

A study on the imputation solution for missing speed data on UTIS by using adaptive k-NN algorithm

김 은 정* (Eun-jeong Kim) 배 광 수** (Gwang-Soo Bae) 안 계 형*** (Gye-Hyeong Ahn) 기 용 걸**** (Yong-Kul Ki) 안 용 주***** (Yong-Ju Ahn)

요 약

UTIS(Urban Traffic Information System)는 프로브차량을 활용하여 도시지역의 구간통행시간 정보를 직접 수집하는 방식으로 타 검지체계에 비해 상대적으로 정확한 링크 속도정보를 산출할 수 있다. 하지만, 현재 UTIS에서는 프로브차량(Probe Vehicle) 및 노변기지국(RSE)의 부족, 시스템 오류 등 다양한 요인에 의해 링크 속도정보의 수집이 누락되는 결측 구간이 발생되고 있다. 본 연구에서는 보다 정확한 여행시간 정보를 제공하기 위한 방안으로 k-NN 알고리즘을 기반으로 결측속도 정보를 효율적으로 보정할 수 있는 새로운 보정모형을 제안하였다. 제안 모형은 각 후보개체(이력 시계열 데이터)의 분포 특성에 따라 최근접이웃 개수를 탄력적으로 조정하는 적응형 k-NN 모형이다. 모형 평가 결과, 제안 모형이 결측정보를 효과적으로 보정·처리할 수 있는 동시에 ARIMA 등 타 모형에 비해 보정 오차를 크게 감소시킬 수 있는 것으로 분석되었다. 본 연구에서 제안된 결측 보정 모형은 UTIS 중앙교통정보센터에 직접 적용하여 교통정보 서비스 품질을 향상시키데 활용될 계획이다.

핵심어 : UTIS, 결측정보, 보정, 적응형 k-NN, 중앙교통정보센터

ABSTRACT

UTIS(Urban Traffic Information System) directly collects link travel time in urban area by using probe vehicles. Therefore it can estimate more accurate link travel speed compared to other traffic detection systems. However, UTIS includes some missing data caused by the lack of probe vehicles and RSEs on road network, system failures, and other factors. In this study, we suggest a new model, based on k-NN algorithm, for imputing missing data to provide more accurate travel time information. New imputation model is an adaptive k-NN which can flexibly adjust the number of nearest neighbors(NN) depending on the distribution of candidate objects. The evaluation result indicates that the new model successfully imputed missing speed data and significantly reduced the imputation error as compared with other models(ARIMA and etc). We have a plan to use the new imputation model improving traffic information service by applying UTIS Central Traffic Information Center.

Key words : UTIS, Missing Data, Imputation, Adaptive k-NN, Central Traffic Information Center

* 주저자 : 도로교통공단 교통과학연구원 책임연구원
** 공저자 및 교신저자 : 도로교통공단 교통과학연구원 선임연구원
*** 공저자 : 도로교통공단 교통과학연구원 연구위원
**** 공저자 : 도로교통공단 교통과학연구원 책임연구원
***** 공저자 : 도로교통공단 교통과학연구원 과장
† 논문접수일 : 2014년 04월 11일
† 논문심사일 : 2014년 06월 13일
† 게재확정일 : 2014년 06월 16일

Ⅰ. 서 론

UTIS(Urban Traffic Information System)는 차량내 장치(OBE: On-Board Equipment)와 노변기지국(RSE: Road side Equipment)으로 구성된 이동환경 하에서 IEEE 802.11a,e를 근간으로 하는 무선랜 기반의 첨단무선통신 기술을 활용하여 교통정보를 수집/제공하는 시스템이다. OBE가 장착된 UTIS 프로브 차량은 내장된 노드-링크 체계 및 GPS 장치를 활용하여 링크 통행시간 및 주행궤적 정보를 수집하고, 이를 RSE을 통해 각 지역교통정보센터로 전송하게 된다. 현재 UTIS에서는 운행시간이 상대적으로 긴 택시 차량 위주로 OBE를 보급하고 있지만, 방대한 도시부 도로망을 고려할 때 아직은 많은 지역에서 프로브차량이 부족한 상황이다. 또한 중요 교차로 중심으로 설치된 RSE의 부족과 시스템 오류 가능성 등도 있기 때문에 많은 도시에서 속도 데이터 수집이 누락되는 결측 구간이 발생되고 있다. 정보 미수집으로 인한 결측 구간의 존재는 교통정보 서비스의 품질에 직접적인 영향을 미칠 수 있다. 따라서 교통정보 결측 구간에 대한 보정처리는 교통정보시스템의 운영전략 상 중요한 문제 중 하나이다.

본 연구에서는 UTIS 속도정보가 미수집된 구간에 대해 효율적으로 보정·처리할 수 있는 새로운 결측보정 방법을 k-최근접이웃(k-NN) 알고리즘 기반으로 개발·평가함으로써, UTIS 교통정보 서비스의 품질을 향상시키는데 활용하고자 하였다.

Ⅱ. 관련 연구 고찰

1. 국내 연구

김정연(2006)등은 고속도로 FTMS 자료를 활용하여 각 검지기 설치지점 간 이동소요 주기를 고려하는 방법, 지점 간 연속성을 고려한 회귀분석 및 차로별 이용률을 이용한 결측 데이터 보정방안 등 총 3개 방안을 제안하였다. 이 연구에서는 이력자료 동일주기 방법과 전후지점 데이터에 의한 회귀분석 방법이 상대적으로 좋은 결측정보 추정값을 얻을 수 있는 것으로 분석되었다. 이 연구에서 제안한 결측보정 회귀식은 식 (1)과 같다. [1]

$$\begin{aligned} V_{j,l,t,d} &= \alpha_{l,v}(i.j) + \beta_{l,v}(i,j) \times V_{i,l,t,d} \\ S_{j,l,t,d} &= \alpha_{l,s}(i,j) + \beta_{l,s}(i,j) \times S_{i,l,t,d} \\ O_{j,l,t,d} &= \alpha_{l,o}(i,j) + \beta_{l,o}(i,j) \times O_{i,l,t,d} \end{aligned} \quad (1)$$

윤원식(2008) 등은 신호 처리 알고리즘을 기반으로 인접 노드의 속도정보 및 이력 데이터를 활용하여 결측 속도를 추정하는 알고리즘을 제안하였다. 결측값 추정에는 지수창함수(exponential window function)와 단순선형회귀모형을 이용하였다. 모형 구조가 비교적 간단하면서도 최근 자료만을 활용하여 결측값 추정의 효율성을 높인 방법론으로 평가할 수 있으나 평가범위가 제한적이었다는 점과 도시부 도로에서 추정력이 떨어지는 점이 한계점으로 지적될 수 있다. [2]

$$B_i = \frac{\exp^{\frac{-i^2}{w}}}{\sum_{j=0}^{w-1} \exp^{\frac{-i^2}{w}}}, \qquad C_n = \sum_{i=1}^{w} B_i S_{n-i} \quad (2)$$

여기서, B_i = 지수창함수를 이용한 가중치값
C_n = 결측 추정값
S_n = 이력자료 및 유사도로 자료값

연지윤(2009) 등은 무선통신 기반 교통정보 수집체계 하에서 발생되는 차량주행 궤적정보의 결측값을 보정하기 위한 방안을 제시하였다. 이 연구에서는 varied-window similarity measure 기법을 활용하여 선행 참조차량의 이력자료를 기반으로 결측값을 우선 보정하고, 대상차량과 결측값 발생 전후 차량과의 차두거리에 따라 가중치를 부여하는 방법이 적용되었다. [3]

하정아(2007) 등은 일반국도의 누락 교통량 데이터를 추정하기 위해 상시 교통량 조사 자료를 활용하였다. 이 연구에서는 교통량 결측자료의 보정을 통계적인 방법으로 접근하였으며, 보정대상 시간과

동일 시간의 자료를 적용할 수 있는 자기회귀분석과 보정대상 지점과 동일지점의 자료를 적용할 수 있는 계절 시계열 분석방법을 이용하여 결측 교통량을 보정하는 방안을 제시하였으며, 자기회귀모형의 추정값이 계절 시계열 모형의 추정값에 비해 오차가 적은 것으로 분석되었다. [4]

2. 해외 연구

James H. Conklin(2003) 등은 차량 검지자료 내 결측에 대한 처리를 이동평균 및 과거자료 대체를 비롯하여 EM과 같은 통계적 보정처리를 적용한 결측보정 결과를 제시하였으며, 보정 처리 과정을 통해 처리방법에 대한 수행속도 및 처리결과에 대한 추정력을 제시하였다. [5]

Chao Chen(2003) 등은 전후 지점 간의 통행패턴의 연속성을 전제로 하는 선형회귀분석을 통해 지점값을 추정하는 방법을 제안하였다. [6]

Sharma(2001) 등은 연구에서는 교통량의 결측 데이터를 추정하기 위해 ARIMA 모형과 유전자 회귀모형, 신경망 모형을 적용하였으며, 캐나다 앨버타 주의 6개 상시 교통량 조사지점의 교통 데이터를 활용하였다. 이 모형은 AADT와 DHV를 추정하기 위해서 적용되었고, 유전자 회귀모형이 결측 교통량 추정에서 가장 효과적인 것으로 제시하였으며, 추정 정확도 측면에서 ARIMA 모형이 신경망 모형보다 우수한 것으로 분석되었다. [7]

3. 시사점

속도정보와 같은 교통 데이터의 결측에 대응하여 이를 보정하기 위한 다양한 방안들은 결국 이력데이터(historical data)나 주변 유사 정보를 활용하여 추정의 정확도를 얼마나 향상시킬 수 있느냐의 문제로 귀결될 수 있다. 또한 개발된 모형의 다양한 도로환경 및 교통상황에 대한 적응성을 효과적으로 평가하기 위해서는 운영 중인 교통정보 시스템에 대한 직접 적용을 통해 평가가 이루어지는 것이 바람직하다.

국내·외의 선행 연구를 보면 대부분 학술적 차원의 모형 개발 및 모의평가 수준에 머무르고 있어 평가의 적정성과 성과의 활용성 및 확장성 측면에서 일정수준 한계가 있었음을 알 수 있다. 본 연구에서는 k-NN 알고리즘을 기반으로 UTIS 이력데이터를 활용한 새로운 적응형 결측 보정 알고리즘을 개발하고 이를 UTIS 중앙교통정보센터에 직접 적용·운영함으로써 다양한 지역 및 도로교통 환경에 대한 적응성 및 실용성도 확보토록 하였다.

Ⅲ. UTIS 속도정보 결측보정 현황 분석

1. 도시별 결측구간 발생 현황

UTIS는 OBE와 RSE으로 구성된 무선통신 기반의 이동환경 하에서 GPS와 노드-링크 체계를 활용하여 속도 중심의 교통정보를 수집한다. OBE를 장착한 프로브차량은 음영구간 주행 중에는 수집정보(위치정보, 상태정보 등)를 축적하고 있다가 RSE 통신가능 영역에 진입하는 시점에 업로드(upload)를 수행하게 되며, 통신반경 내에서는 상시 연결상태를 유지하며 수집정보를 지역센터로 전송한다.

현재 UTIS 사업이 완료된 수도권 10개 도시의 1레벨 링크를 대상으로 '13년 9월의 결측정보 발생 현황을 분석하였다. UTIS의 노드-링크 체계는 총 5레벨로 구성되어 있으며, 1레벨은 도로교통 체계를 그대로 반영하여 각 노드(교차점, 터널/교량 시종점, 운영상태 변화지점 등)와 링크를 구성한 물리적 레벨이다. 2레벨~5레벨은 교통정보 서비스를 위한 논리적 레벨로써, 하위레벨(2레벨⇒5레벨)로 진행될수록 대구간 및 중요 도로를 중심으로 링크를 구성하기 때문에 교통정보 서비스의 세밀도는 감소하게 된다. <표 1>에서 보면 서울시의 전체 1레벨 링크수가 22,075개이며, 이 중 11,965개의 링크에 결측이 발생되어 평균 결측률은 54.2%로 분석되었다.

타 도시를 보면, 광명 32.2%, 과천 8.1%, 성남 40.4% 등으로 도시별로 결측율 편차가 비교적 크게 나타나고 있음을 알 수 있다. 편도 2차로 이상 링크를 대상으로 한 경우, 평균 결측률은 약 16.5%로 분석되어 주요 도로를 대상으로는 UTIS 정보수집이

〈표 1〉 UTIS 속도정보 결측 현황
〈Table 1〉 The current situation analysis of missing UTIS data

city	# of 1st level link		#(%) of missing data link	
	all roads	over four-lane roads	all roads	over four-lane roads
seoul	22,075	12,116	11,965 (54.2%)	7,864 (64.9%)
gwang myeong	497	352	160 (32.2%)	102 (29.0%)
anyang	1,532	801	718 (46.9%)	149 (18.6%)
gwacheon	295	164	24 (8.1%)	9 (5.5%)
ansan	3,221	2,124	165 (5.1%)	103 (4.8%)
seongnam	3,145	1,670	1,271 (40.4%)	259 (15.5%)
namyangju	3,121	1,018	1,270 (40.7%)	133 (13.1%)
guri	667	316	228 (34.2%)	34 (10.8%)
hanam	695	312	32 (4.6%)	13 (4.2%)
uiwang	733	323	169 (23.1%)	33 (10.2%)
average number of missing data link			1,460 (27.1%)	792 (16.5%)

비교적 원활히 이루어지고 있는 것을 알 수 있다.

2. 기존 결측보정 방법 및 문제점

현재 중앙교통정보센터에서는 패턴데이터와 유사 시공간 데이터를 활용하여 결측된 링크 속도정보에 대한 보정을 시행하고 있다. 기존 보정 방법은 결측 링크와 유사 특성을 나타내는 링크(결측 링크의 상·하류링크와 결측링크가 포함된 상위레벨 링크)를 기본적으로 활용하게 되며, 유사 링크의 실시간 수집 데이터와 DB에 구축된 패턴 데이터를 활용하여 결측 보정값을 산출하게 된다.

기존 방법론의 결측값 보정 절차는 다음과 같이 3단계로 진행된다.

step 1 : 현 주기(Tt)에 결측이 발생된 목표 링크 탐색

step 2 : 상·하류 링크의 속도 데이터 활용 결측 보정

- 상·하류 링크 데이터가 모두 존재하는 경우
  - 상·하류 링크의 현 주기 수집 속도($S_{up}^{t_i}$, $S_{down}^{t_i}$) 및 패턴 속도($PDS_{up}^{t_i}$, $PDS_{down}^{t_i}$)를 활용하여 결측 보정값(S_{rev}) 산출

$$S_{rev} = PDS_{ms}^{t_i} \times \frac{Avg(S_{up}^{t_i} + S_{down}^{t_i})}{Avg(PDS_{up}^{t_i} + PDS_{down}^{t_i})} \quad (3)$$

- 상·하류 링크 데이터 중 하나만 존재하는 경우
  - 상류or하류 링크의 현재 속도 및 패턴 속도 활용

$$S_{rev} = PDS_{ms}^{t_i} \times \frac{S_{up}^{t_i}}{PDS_{up}^{t_i}} \text{ 또는}$$

$$S_{rev} = PDS_{ms}^{t_i} \times \frac{S_{down}^{t_i}}{PDS_{down}^{t_i}} \quad (4)$$

step 3 : 상위레벨 속도 데이터 활용 결측 보정

- 3레벨 현 주기 속도($S_{lev3}^{t_i}$)와 패턴 속도($STS_{lev3}^{t_i}$)를 활용하여 결측 보정값 산출

$$S_{rev} = PDS_{ms}^{t_i} \times \frac{S_{lev3}^{t_i}}{STS_{lev3}^{t_i}} \quad (5)$$

본 방법론은 결측 보정에 사용되는 데이터가 제한적이고 시스템 연산이 비교적 간단하여 실용적 측면에서 장점이 있는 반면, UTIS 운영 과정에서 결측 보정된 속도값의 신뢰도 및 결측 보정률이 요구 수준에 미치지 못하는 것으로 분석되어 새로운 보정 방법론의 개발 필요성이 대두되었다.

Ⅳ. UTIS 속도정보 결측보정 방법론 개발

1. 개요

새로운 UTIS 교통정보 결측 보정 방법은 k-최근접이웃(k-Nearest Neighbor) 알고리즘을 기반으로 하며, 최초 설정된 k값을 각 시계열 데이터셋의 표준편차(standard deviation)를 기반으로 임계값까지 확장시키는 적응형 모형(adjusted model)으로 개선하여 최적의 결측 보정값을 산출할 수 있도록 하였다.

2. k-NN 알고리즘

k-NN의 기본 개념은 조건부 확률을 관측자료로부터 직접 산정하는 것이다. 일반적인 k-NN 밀도함수 추정량(density estimator)은 식 (6)과 같이 표현되며, 이는 커널함수(kernel function) $K()$와 주변 관측치(neighbors)의 개수인 k로 이루어져 있다.

$$f_{GNN}(x) = \frac{1}{r_k^d(x)n}\sum_{i=1}^{n} k\left(\frac{x-x_i}{r_k(x)}\right) \tag{6}$$

k는 조건부확률의 smoothing factor이며, $K()$는 이 확률분포 양 끝단 모양을 결정하게 된다. 여기서, x는 확률변수의 임의 값이며, x_i는 i번째 관측값을, n은 관측값의 전체 개수를, r은 x와 x_i사이의 유클리디언 거리를, d는 상태공간의 차수를 의미한다. 현재 상태벡터와 과거 상태벡터와의 유클리디언 거리를 구한 후, 가까운 순으로 k개를 선택하고, 이를 k-최근접이웃이라 칭한다.

$$r_{ij} = \left(\sum_{m=1}^{d} w_m (x_{im} - x_{jm})^2\right)^{\frac{1}{2}} \tag{7}$$

일단 k개의 주변 값들이 주어지면, 각각의 값들은 거리의 함수로써 식 (8)에 의해 일종의 확률적 값을 지니게 된다.

$$K(i,j) = \frac{\frac{1}{j}}{\sum_{j=1}^{k}\frac{1}{j}} \tag{8}$$

커널함수가 결정되면, k개의 주변 값들로부터 다수의 샘플을 추출하고, 이를 통해 다음 상태에 대한 통계적 추론이 가능해진다. 비록 매개변수 k가 존재하지만 이의 민감도가 상대적으로 낮은 점과 특정 확률분포를 가정하지 않고 경험적 확률분포를 고려한다는 점에서 대표적인 비모수적 방법으로 알려져 있다. [8]

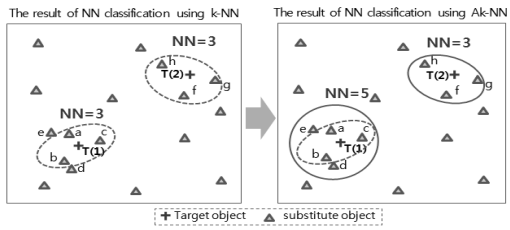
3. 결측보정을 위한 적응형 k-NN 개발

1) 최근접이웃(NN) 확장 방법

제안 알고리즘은 현재의 결측 속도값을 k-NN 알고리즘을 이용하여 추정하기 위해 특정 기간의 이력 데이터 중 유사 시계열 데이터를 '가중 유클리디언 거리(weighted euclidean distance)' 기반으로 검색한다. 본 연구의 적응형 k-NN은 기존 k-NN 알고리즘이 고정된 k값을 사용하는데서 오는 최근접이웃 선정의 경직성 및 목표 개체의 위치가 가지는 국지적 특징의 미반영 문제를 효과적으로 감소시킬 수 있다.

본 연구에서는 목표 시계열 데이터와 인접한 후보 시계열 데이터 간의 유사도를 가중 유클리디언 거리의 SD(standard deviation, σ)값으로 판단하여 최종 최근접이웃(NN) 개수를 결정하는 방법을 적용하였다. 즉, 결측보정의 대상이 되는 여러 후보 시계열 데이터들의 국지적 분산 특성 및 밀집도를 고려하여 최근접이웃(NN) 개수를 탄력적으로 조정한다. 목표 시계열 데이터를 중심으로 후보 시계열 데이터들의 밀집도가 높아 거리비율이 통계적 임계값 내에 포함되는 경우, 최근접이웃 개수를 설정된 최대값에 도달 시까지 지속적으로 확장한다.

<그림 1>에서 보면 k=3으로 설정되어 있을 경우, 기존의 k-NN 알고리즘에서는 목표개체 T(1)과 T(2) 모두의 경우에서 NN=3개의 고정된 최근접이웃을 선정하게 된다.



〈그림 1〉 적응형 k-NN을 활용한 최근접이웃 분류 결과
〈Fig. 1〉 The result of nearest neighbor classification using the adaptive k-NN algorithm

하지만 목표개체 T(1)의 경우 주변의 후보개체의 분포 밀도가 상대적으로 높은 국지적 특성을 보이며, 이와 같은 경우 해당 후보개체를 결측치 보정에 활용하는 것이 상대적으로 높은 추정 정확도를 나타낼 수 있다. 따라서 후보개체 중 e와 d를 최근접 이웃에 포함시켜 최종적으로 NN=5개의 최근접이웃을 선정하게 된다.

제안 알고리즘의 최근접이웃(NN)의 확장 규칙은 다음과 같다.(*k=n*일 경우)

❶ *p*개의 변수와 *q*개의 개체로 구성된 이력 시계열 데이터 *p*×*q* 자료행렬 중, 결측값이 발생한 목표개체에 가장 근접한 후보개체를 가중 유클리디안 거리 순서에 따라 *n*개 선정

❷ 목표개체와 각 후보개체간의 가중 유클리디안 거리를 $d_{(r,h)}$라 할 때, 목표개체로부터 *k*번째로 가까운 후보개체와의 거리를 d_k, 다음으로 인접한 후보개체와의 거리를 d_{k+1}로 함

❸ 이 두 거리의 차를 d_{k+1}-d_k, 목표개체와 후보개체 간 거리 전체의 표준편차를 σ_{dij}, *NN* 확장 임계값을 k_{max}라 하고 $d_{(r,h)}$가 정규분포를 이룬다고 가정할 때, 정규분포의 통계적 성질을 이용하여 두 개체간의 유사도($S_{k,k+1}$)를 *1sigma* 기준으로 판단(*1sigma* 기준을 적용함으로써 낮은 유사도 수준에서의 NN 확장을 방지)

❹ 두 개체가 유사 개체로 판명되면 '1'을 부여하여 *NN*에 포함시킨 후 다음 거리의 후보개체 d_{k+2}을 대상으로 유사도 $S_{(k+1,k+2)}$를 다시 판단하고 동일 방식으로 k_{max}까지 확장. 그렇지 않으면 '0'을 부여하고 *NN* 확장 종료

$$S_{(k,k+1)}=\begin{cases}1, & (d_{k+1}-d_k)-\sigma_{d_{ij}}\le 0\\ 0, & (d_{k+1}-d_k)-\sigma_{d_{ij}}>0\end{cases} \quad (9)$$

2) 적응형 k-NN을 활용한 속도 결측값 보정 절차

본 연구에서 제안한 적응형 k-NN 알고리즘을 활용한 결측값 보정절차는 다음과 같다.

step 1. k값 설정(기본값 : k=5)

step 2. 현재 주기(*Tt*)에 결측이 발생된 실시간 시계열 데이터($D^r_{m,t}$)를 탐색(1레벨 데이터 기준)

step 3. $D^r_{m,t}$의 30분전(이전 6주기)까지 실시간 시계열 데이터 셋(data set)을 구성

$$D^r_{m,t-i}=\{D^r_{m,t-6}, D^r_{m,t-5}, D^r_{m,t-4}, D^r_{m,t-3}, D^r_{m,t-2}, D^r_{m,t-1}\} \quad (10)$$

step 4. 결측 발생 링크와 동일 요일/시간대의 정상 시계열 데이터 셋을 n주전(기본값: 20) 까지 탐색

step 5. 결측 시계열 데이터 셋과 탐색된 각 이력 시계열 데이터 셋 간의 유클리디언 거리($d_{(r,h)}$)를 탐색기간에 따른 가중값(ω_P)을 고려하여 산출

$$d(r,h)=\omega_P\left\{\sum_{i=1}^{6}\left\{(D^r_{m,t-i}-D^{h_n}_{m,t-i})^2\right\}^{1/2}\right\} \quad (11)$$

여기서, $\omega_P=\begin{cases}1.00, & D^{h_1}\sim D^{h_8}\\ 1.05, & D^{h_9}\sim D^{h16}\\ 1.10, & otherwise\end{cases}$

step 6. 산출된 가중 유클리디언 거리를 크기가 작은 순으로 배열하고 최소값부터 *k*개의 이력 시계열 데이터 최근접이웃 해($NN(D^h)$)를 도출

$$NN(D^h)=\{nnD^{h_{i1}}, nnD^{h_{i2}}, nnD^{h_{i3}},\ldots, nnD^{h_k}\} \quad (12)$$

step 7. SD 기반 최근접이웃 확장 방법에 따라 최대 *kmax*=10(기본값)까지 NN 확장

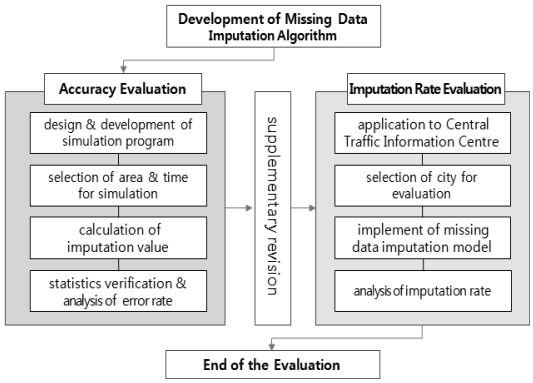
step 8. 최종 선정된 각 NN의 *t*주기 데이터를 산술평균한 값으로 결측값을 대치

$$\text{결측보정값}(RD^r_{m,t})=\frac{\sum_{i=1}^{k}(nnD^{h_i}_t)}{k} \quad (13)$$

Ⅳ. 모형 검증 및 평가

1. 개요

본 연구에서 개발한 UTIS 속도정보 결측 보정 알고리즘을 정확도와 보정율 지표로 구분하여 평가·검증하였으며, 평가절차는 <그림 2>와 같다.



〈그림 2〉 결측 보정 알고리즘 평가 절차
〈Fig. 2〉 The evaluation process of imputation algorithm

- 정확도 평가

- 결측보정 시뮬레이션 프로그램을 개발하여 실시간으로 수집된 통행속도 데이터와 제안 알고리즘에 의해 보정된 데이터를 비교·평가함. 통계적 검증과 오차율 분석을 수행

- 보정율 평가

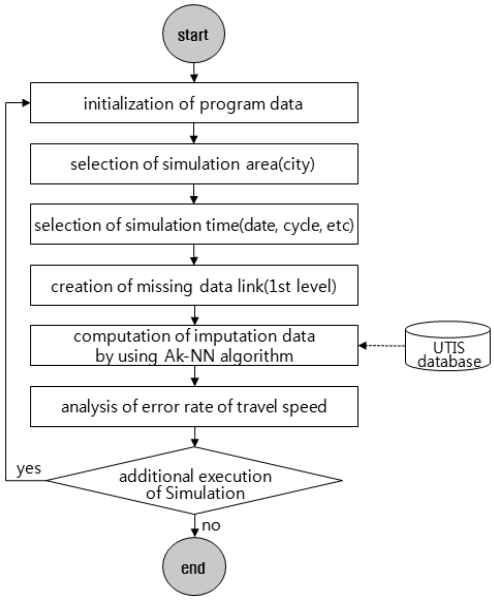
- 제안 알고리즘을 UTIS 중앙교통정보센터 시스템에 직접 시험·적용하여 교통정보 결측 링크가 보정되는 비율을 분석

2. 정확도 평가

(1) 결측 보정 시뮬레이션 프로그램 개발

결측보정 시뮬레이션 프로그램에서는 분석 대상 도시와 시간대를 우선 설정한 후 모든 1레벨 링크에 속도정보 수집이 누락되는 결측 상황을 인위적으로 조성하고, 본 연구에서 제안한 적응형 k-NN 알고리즘을 적용하여 보정 처리된 속도값을 산출하게 된다. 이를 통해 산출된 보정 속도값과 해당 주기에 시스템을 통해 실제로 수집된 실시간 속도값(기준값)을 상호 비교함으로써 제안 알고리즘의 정확도를 평가하였다.

결측보정 시뮬레이션 프로그램은 C++ Builder를 활용해 PC 기반으로 개발되었으며, 중앙교통정보센터의 교통정보 이력 데이터베이스 및 통계 데이터베이스와 직접 연계되어 UTIS가 구축된 도시의 모든 요일 및 시간대에 걸쳐 효율적인 평가가 가능토록 구성하였다.



〈그림 3〉 결측 보정 시뮬레이션 프로그램 실행 절차도
〈Fig. 3〉 The execution flow of simulation program

(2) 통계적 검증 결과

정확도 검증은 UTIS 사업이 완료된 5개 도시(서울, 성남, 용인, 광명, 남양주)를 대상으로 하였다. 제안 알고리즘에 의해 산출된 결측 보정값의 오차율 분석에 앞서 paired sample t-검정을 수행함으로써 대응된 표본(실 수집된 UTIS 속도값↔보정 속도값)으로 이루어진 두 데이터가 동일집단으로 판단할 수 있는지에 대한 통계적 근거를 제시하였다.

통계적 검증은 각 도시 단위로 수행하였으며, 적용된 유의수준(significant level, α)은 0.05이다. T-검정의 귀무가설(H_0) 및 대립가설(H_1)은 다음과 같다.

- H_0 : 결측보정된 속도값(S_r)과 실수집 속도값(S_c)의 평균은 차이가 없다(S_r=S_c)
- H_1 : $S_r \neq S_c$

〈표 2〉 동질성 검증을 위한 paired sample t-test 결과
〈Table 2〉 The result of paired sample t-test

city	df	paired sample differences statistic			t-value (2-tailed)	P-value (2-tailed)
		avg.	stdev.	avg. of se		
seoul	19,709	0.18	11.83	0.16	1.132	0.287
seongnam	3,119	-0.17	9.53	0.20	0.838	0.402
yongin	1,857	-0.06	6.07	0.14	-0.149	0.822
gwang myeong	480	-0.18	6.04	0.34	0.544	0.587
namyangju	3,103	0.18	6362	0.15	1.22	0.221

제안 알고리즘에 의해 산출된 속도값과 실수집된 속도값과의 평균값 차이 분석을 위한 paired sample t-test 결과가 <표 2>에 제시되어 있다. 분석결과를 보면, 모든 도시에서 유의확률(P-value)이 유의수준(α, 0.05)보다 큰 것으로 나타나, 두 집단의 평균값이 95% 신뢰수준에서 차이가 없는 것으로 분석되었다.

(3) 오차율 분석 결과

오차율 분석은 현재 중앙교통정보센터에서 운영중인 기존 결측보정 방법론에 의해 산출된 보정값과 제안 알고리즘에 의해 산출된 보정값과 해당 주기 실수집 데이터와의 속도 오차율을 각 도시 단위로 비교하였다.

또한 추가적으로 패턴 데이터를 직접 보정처리에 활용할 경우와도 상호 비교도 수행하였다. 평가지표로는 평균제곱근오차(RMSE: root mean square error)와 평균절대백분율오차(MAPE: mean absolute percentage error)를 활용하였다.

$$RMSE = \sqrt{\frac{\sum (T_s(t) - T_p(t))^2}{n-1}} \quad (14)$$

$$MAPE = \frac{1}{n}\sum \left| \frac{T_s(t) - T_p(t)}{T_s(t)} \right| \times 100 \quad (15)$$

여기서, $T_s(t)$: 기준값(수집값)
$T_p(t)$: 예측값(보정값)

본 연구의 제안 알고리즘 및 다른 결측 보정 방법(패턴데이터에 의한 직접 보정, 기존 방법론에 의한 보정)에 의한 오차율 분석결과가 <표 3>에 제시되어 있다.

〈표 3〉 알고리즘별 보정 오차율 분석 결과(1)
〈Table 3〉 The result of imputation error rate by algorithm(1)

city	RMSE			MAPE		
	pattern data	existing method	Ak-NN	pattern data	existing method	Ak-NN
seoul	18.8	16.8	14.9	48.5	39.1	28.8
seongnam	11.8	11.5	10.9	29	27.9	23.1
yongin	11.1	12.6	9	29	25.7	14.2
gwang myeong	10.5	10.2	7.1	22.3	21.7	8.7
namyangju	11.4	12.1	6.7	28.3	25.7	14.6
avg.	12.72	12.64	9.72	31.42	28.02	17.88
stdev.	3.43	2.49	3.34	9.95	6.58	7.98

먼저 RMSE 산출 결과를 보면, 적응형 k-NN에 의한 결측 보정값의 RMSE가 모든 도시에서 가장 낮게 산출되었으며, 5개 도시 평균값은 9.72로 기존 방법론이나 패턴데이터 직접 보정에 의한 RMSE에 비해 약 40% 이상 낮은 수준으로 분석되었다.

MAPE 분석 결과에서는 평균 MAPE가 '적응형 k-NN(17.88%)⇒기존방법론(28.02%)⇒패턴데이터 직접 보정(31.42%)' 순으로 나타났으며, 모든 도시에서 제안 알고리즘에 의한 결측 보정값의 MAPE가 가장 낮은 것으로 분석되었다.

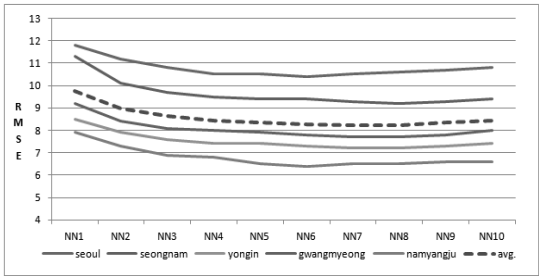
다음 단계로써 제안 알고리즘과 단기 예측에 광범위하게 활용되고 있는 ARIMA 모형과의 오차율 비교분석을 수행하였다. 5개 도시에서 2개 링크를

무작위로 선정하고 6주기(30분) 동안의 결측보정 모형에 의한 산출값과 해당 주기 실수집 속도 데이터와의 MAPE에 대한 분석을 수행하였다. ARIMA 모형에 대한 일차적 모형식별 및 진단 결과, UTIS 속도 데이터의 예측 시에는 ARIMA(1,0,1) 모형이 가장 적합한 것으로 분석되었다.

5개 도시별로 선정된 2개 링크에 대한 MAPE 산출 결과는 <표 4>와 같다. 5개 도시 평균 MAPE 산출값을 보면, 제안 알고리즘(적응형 k-NN)이 17.19%로써 가장 낮은 것으로 분석되었으며, 다음으로 ARIMA 모형(22.02%), 기존 방법론(27.56%) 순으로 나타나, 제안 알고리즘이 ARIMA 모형과의 비교에서도 상대적으로 정확한 추정력을 나타낸 것을 알 수 있다.

〈표 4〉 알고리즘별 보정 오차율 분석 결과(2)
〈Table 4〉 The result of imputation error rate by algorithm(2)

city	link ID	existing method	ARIMA (1,0,1)	Ak-NN
Seoul	1000000301	38.8	28.9	25.7
	1000002301	37.9	33	28.3
seongnam	2040000400	28.2	19.3	20.5
	2040002701	28.5	22.4	23.9
yongin	2280005200	23.2	24.2	16.2
	2280012005	25.5	22.1	14.8
gwang myeong	2130001500	19.7	17.9	8.3
	2130008200	20.5	15.4	6.9
namyangju	2220001203	26	20.7	14.5
	2220002009	27.3	16.3	12.8
avg.		27.56	22.02	17.19



〈그림 4〉 NN 확장에 따른 결측 보정값 오차율 분석 결과
〈Fig. 4〉 The result of imputation error rate by extension of NN

오차율 분석의 마지막 단계로써 적응형 k-NN 알고리즘의 최근접 이웃(NN) 개수에 변화에 따른 오차율 추이를 분석하였다. 적응형 k-NN 알고리즘은 표준편차(σ)를 기반으로 후보 시계열 데이터셋(data set)의 밀집도에 따라 최근접이웃(NN) 개수를 확장한다. 서울 등 수도권 5개 도시를 대상으로 평일 1시간 동안의 NN 확장에 따른 오차율(RMSE) 변화를 분석한 결과가 <그림 4>에 제시되어 있다.

분석결과를 보면, NN에 따른 오차율의 분포가 우하향 곡선 형태로써 오차율이 점진적으로 감소하는 것으로 나타났으며, 5개 도시의 평균 오차율이 NN=7~8에서 최소값을 나타냈다. 서울시와 남양주시는 NN=6, 그리고 성남시는 NN=8에서 최소 오차율을 나타내는 등 도시별/시간대별 교통특성에 따라 최적 NN 개수 선택이 변화되어 적응형 k-NN 알고리즘이 효율적으로 적용되었음을 보여주고 있다.

3. 보정율 평가

본 연구에서 제안한 알고리즘을 UTIS 중앙교통정보센터에 직접 적용하여 결측이 발행된 링크에 대한 보정율을 평가하였다.

알고리즘별 보정율을 평가하기 위해 중앙교통정보센터에 제안 알고리즘을 적용 후 평일 3일간의 평균 보정률을 분석하여 <표 5>에 제시하였다. 평가 도시는 UTIS 사업이 완료된 수도권의 10개 도시를 대상으로 하였다. 여기서 결측정보 보정률의 개념은 식 (16)과 같이 정의된다.

$$\text{결측정보 보정율} = \frac{\text{보정처리된 링크 개수}}{\text{결측발생된 링크 개수}} \times 100 \quad (16)$$

먼저 기존 방법론을 적용했을 경우는 안양시가 75.0%로 가장 높은 보정률을 나타냈으며, 다음으로 구리시(74.3%), 남양주시(72.5%), 의정부시(65.1%) 순으로 분석되었다. 제안 알고리즘 적용 시의 결측 보정률은 전 도시에서 기존 방법론에 비해 높게 나타났으며, 서울, 구리, 의정부 등의 도시들은 보정

〈표 5〉 알고리즘별 결측 보정률 분석 결과
〈Table 5〉 The result of imputation rate by algorithm

city	# of 1st level link	# of missing data link	existing method		Ak-NN	
			# of imputation	imputation rate(%)	# of imputation	imputation rate(%)
seoul	12,116	8,475	3,500	41.3	8,449	99.7
seongnam	1,670	220	136	61.8	209	95.0
ansan	2,124	124	56	45.0	106	85.3
anyang	801	20	15	75.0	17	85.0
guri	316	30	22	74.3	30	99.8
yongin	1,196	161	51	31.8	137	85.3
gwangju	425	120	50	41.9	114	95.0
paju	1,293	454	232	51.0	433	95.3
uijongbu	1,011	49	32	65.1	49	99.4
nam yangju	1,018	120	87	72.5	116	96.7
avg. of imputation rate			-	56.0	-	93.7
stdev. of imputation rate			-	15.7	-	6.1

률이 99%를 상회하였다.

10개 도시의 평균 결측 보정률을 비교하였을 경우, 기존 방법론이 56.0%인 반면 제안 알고리즘(적응형 k-NN)은 93.7%의 보정률을 나타냈다. 이는 제안 알고리즘이 비율적으로 약 60% 이상 많은 결측 링크를 보정처리한 결과로써, 기존 방법론이 보정하지 못한 결측 링크에 대해서도 적응형 k-NN에 의한 보정처리가 이루어졌음을 의미한다.

또한 도시별 보정률의 표준편차 분석 결과에서는 기존 방법론이 15.7, 적응형 k-NN이 6.1로 분석되어, 제안 알고리즘이 도시 규모나 교통특성, 결측비율 등에 관계없이 상대적으로 안정적인 보정율을 나타내고 있음을 알 수 있다.

Ⅴ. 결론 및 향후 연구 사항

UTIS와 같은 교통정보시스템에서 교통정보 결측구간에 대한 보정처리는 서비스 품질을 향상시키기 위한 운영전략 차원의 중요한 문제 중 하나이다. 본 연구에서는 UTIS 결측정보를 효율적으로 보정하기 위한 새로운 알고리즘을 k-NN을 기반으로 개발하여 성능을 검증하였다.

본 연구에서 제안한 방법은 각 후보개체의 밀집도에 따라 최근접이웃(NN) 개수를 탄력적으로 조정할 수 있는 적응형 결측보정 모형으로써, 이력 시계열 데이터 중 현재 패턴과 유사한 개체들을 선정하여 보정처리에 활용하게 된다. 제안 알고리즘에 대한 검증은 정확도 평가와 보정율 평가로 구분하여 수행하였으며, 평가의 효율성을 확보하기 위해 결측상황을 인위적으로 조성할 수 있는 시뮬레이션 프로그램을 개발하여 실시간 수집 데이터와의 비교를 통해 오차율을 분석하였다.

정확도 평가에서는 집단별(기준값⇔보정값) 속도 평균값의 동질성에 대한 통계적 검증과 함께, MAPE 및 RMSE를 분석지표를 활용하여 UTIS 중앙교통정보센터에서 운영 중인 기존 보정 방법론 및 ARIMA 모형 산출값과의 오차율 분석을 수행하였다. 분석 결과, 제안 알고리즘이 ARIMA 모형 산출값이나 기존 방법론에 의한 보정값보다 오차율이 약 30% 이상 감소하는 것으로 분석되었다.

보정율 평가에서는 제안 알고리즘을 UTIS 중앙교통정보센터에 적용하여 알고리즘별 결측 보정률을 평가하였다. UTIS 사업이 완료된 수도권 10개 도시에 대한 일평균 보정률에 대한 분석 결과, 기존 방법론이 평균 56.0%, 제안 알고리즘이 평균 93.7%의 보정률을 나타내어 결측정보에 대한 보정율 측면에서도 더욱 우수한 것으로 분석되었다.

본 연구에서 개발한 알고리즘을 UTIS 중앙교통정보센터에 직접 적용하여 교통정보 서비스 품질을 향상시키데 활용될 계획이다. 향후과제로는 다중결측 등 다양한 결측 패턴에 효율적으로 대응하기 위한 방법론 연구와 개발 모형에 대한 추가적인 비교 평가를 통해 신뢰도를 최종적으로 검증하는 것이 필요할 것으로 판단된다. 또한 다양한 도로환경 및 교통특성에 대한 적용성을 확대 검증함으로써 개발 모형의 완성도를 더욱 높이고 UTIS 교통정보 서비스 체계를 더욱 정교화하는 후속 작업이 필요할 것으로 판단된다.

REFERENCES

[1] J. Y. Kim, Y. I. Lee, G. S. Baek, S. Namgung, “A Study on the Imputation for Missing Data in Dual-loop Vehicle Detector System”, Journal of Korean Society of Transportation, vol. 24, no. 7, pp.27-40, Dec., 2006.

[2] W. S. Yun, H. C. Jeong, “Missing Data Estimation for Link Travel Time”, Journal of Korean Society of Transportation, vol. 26, no. 2, pp.101-107, Apr., 2008.

[3] J. H. Yeon, H. M. Kim, C. Oh, W. G, Kim, “A Comprehensive Method to Impute Vehicle Trajectory Data Collected in Wireless Traffic Surveillance Environments”, Journal of Korean Society of Transportation, vol. 27, no. 4, pp.175-181, Aug., 2009.

[4] J. A. Ha, J. H. Park, S. H. Kim, “Missing Data Imputation Using Permanent Traffic Counts on National Highways”, Journal of Korean Society of Transportation, vol. 25, no. 1, pp.121-132, Feb., 2009.

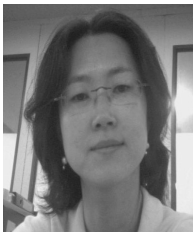
[5] J. H. Conklin, “Data Imputation Strategies for Transportation Management Systems, University of Virginia, 2003.

[6] C. Chen, J. Rice, “Detection Errors and Imputation Missing Data for Single Loop Surveillance System”, 82th TRB Annual Meeting, 2003.

[7] S. Sharma, P. Lingras, M. Zhong, “Effect of Missing Value Imputation on Traffic Parameters from Permanent Traffic Connts”, 80th TRB Annual Meeting, 2001.

[8] H. G. Lee, S. G. Kim, Y. H. Cho, K. Y. Chong, “Probabilistic Reservoir Inflow Forecast Using Non-parametric Methods”, Journal of Hydro Environment Research. May., 2008.

저자소개



김 은 정 (Kim, Eun-Jeong)

1992년 5월 ~ 현재 : 도로교통공단 교통과학연구원 책임연구원
2007년 : 서울시립대학교 박사과정 수료
1991년 7월 ~ 1992년 5월 : 한국건설기술연구원 도로연구실 연구원
1991년 : 영남대학교 대학원 교통공학 석사
e-mail : kej92@koroad.or.kr
연락처 : 02-2230-5250

배 광 수(Bae, Gwang-Soo)

1997년 3월 ~ 현재 : 도로교통공단 교통과학연구원 선임연구원
2008년 12월 : 교통기술사
2002년 7월 : 서울시립대학교 도시과학대학원 교통공학 석사
1995년 2월 : 충북대학교 공과대학 도시공학과 학사

안 계 형 (Ahn, Gye-Hyeong)

2002년 ~ 현재 : 도로교통공단 교통과학연구원 연구위원
1997년 7월 ~ 2002년 12월 : 교통개발연구원 책임연구원
1997년 5월 : University of Texas at Austin 토목공학과 박사(교통공학전공)
1986년 2월 : 서울대학교 환경대학원 도시계획학 석사(교통공학전공)

기 용 걸 (Ki, Yong-Kul)

1994년 12월 ~ 현재 : 도로교통공단 교통과학연구원 책임연구원
1991년 12월 ~1992년 12월 : 삼성전자(주) 연구원
2007년 2월 : 고려대학교 컴퓨터학과 전산학박사

안 용 주 (Ahn, Yong-Ju)

2006년 ~ 현재 : 도로교통공단 교통과학연구원 과장
2002년 3월 ~ 2006년 5월 : 포스코ICT 대리
2002년 2월 : 한양대학교 산업공학 학사