▣ 論 文 ▣

NPR기반 누락 교통자료 추정기법 개발 및 적용

Development and Application of Imputation Technique
Based on NPR for Missing Traffic Data

장 현 호 (서울대학교 환경대학원 박사수료)
한 동 희 (한국도로공사 도로교통연구원 연구원)
이 태 경 (한양대학교 도시공학과 박사수료)
이 영 인 (서울대학교 환경대학원 교수)
원 제 무 (한양대학교 도시공학과 교수)

목 차

Key Words : 자료관리시스템, 이력자료, 누락자료, 비모수 회귀, 다중대체
Data Management System, Historical Data, Missing Data, NPR, Multiple Imputation

요 약

지능형 교통체계는 실시간 교통자료를 수집하고 방대한 양의 이력자료를 축적한다. 그러나 방대한 이력자료는 효율적으로 관리/이용되지 않고 있는 실정이다. ADMS와 같은 자료관리시스템이 도입되면서, 이력자료의 잠재적 활용성은 급격히 증대되고 있다. 그러나 자료관리스템의 교통자료는 다량의 누락자료를 포함하고 있다. 누락자료는 장기간에 걸쳐 빈번하게 교통자료를 이용할 수 없게 하기 때문에, 이력자료를 활용하는데 있어 주된 장애요인 중 하나이다. 따라서 누락자료 추정기법은 자료관리시스템에서 주요한 역할을 수행하게 된다. 이러한 한계를 극복하기 위하여, 본 연구에서는 자료관리스템에 탑재가 용이하며 이력자료에 포함된 누락자료를 추정하기 위한 누락자료 추정모형을 개발하였다. 개발모형은 비모수회귀식(NPR)을 기반으로 개발되었으며, 이력자료의 다양한 교통자료 패턴을 이용하고 현실적인 요구사항(변수 최소화, 연산속도, 다양한 형태의 누락자료 보정, 다중대체)을 충족하도록 설계되었다. 모형의 평가는 다양한 누락자료 형태의 상태에서 수행되었으며, 자료관리시스템에 탑재되기 위해 요구되는 정확도, 연산 수행속도에서 기존에 보고된 모형보다 우수한 성능을 보였다.

ITS (Intelligent transportation systems) collects real-time traffic data, and accumulates vest historical data. But tremendous historical data has not been managed and employed efficiently. With the introduction of data management systems like ADMS (Archived Data Management System), the potentiality of huge historical data dramatically surfs up. However, traffic data in any data management system includes missing values in nature, and one of major obstacles in applying these data has been the missing data because it makes an entire dataset useless every so often. For these reasons, imputation techniques take a key role in data management systems. To address these limitations, this paper presents a promising imputation technique which could be mounted in data management systems and robustly generates the estimations for missing values included in historical data. The developed model, based on NPR (Non-Parametric Regression) approach, employs various traffic data patterns in historical data and is designated for practical requirements such as the minimization of parameters, computational speed, the imputation of various types of missing data, and multiple imputation. The model was tested under the conditions of various missing data types. The results showed that the model outperforms reported existing approaches in the side of prediction accuracy, and meets the computational speed required to be mounted in traffic data management systems.

I. 연구의 배경 및 목적

ITS분야의 실시간 교통정보 수집기술 발전에 따라 방대한 양의 교통자료를 수집·활용할 수 있는 여건이 조성되었다. 대부분의 수집된 교통정보는 실시간 교통정보 제공 및 교통운영을 위한 목적으로 활용되고 있다. 반면, ITS의 실시간 자료수집체계를 이용하여 축적된 방대한 양의 이력자료(Historical Data)를 효율적으로 관리하고 활용하기 위한 기술수준은 낮은 실정이다. 이는 기존 ITS체계의 일차적인 목표인 실시간 정보 수집/제공을 수행한 후, 축적된 이력자료의 관리/활용에 소홀하였기 때문이다. 최근에는 이력자료를 유지·관리하기 위한 집계자료관리시스템(ADMS, Archived Data Management System)이 도입되기 시작하면서 이력자료의 활용 가능성은 매우 높아지고 있다. 그러나 방대한 양의 이력자료는 누락자료(missing data)를 포함하고 있기 때문에, 이력자료를 효과적으로 관리·활용하기 위해서는 누락자료의 보정처리가 선행되어야 하며, 이를 위해서는 누락자료 추정기법이 필수적이다. 이러한 누락자료의 대부분은 Loop 검지기의 수명 도달에 따른 교체, 도로공사 등으로 인한 파손으로 발생하게 된다. 대부분의 경우 정상적인 가동까지는 모니터링, 설계, 공사, 준공의 단계를 거침으로 상당한 기간이 요구된다.

따라서 본 연구에서는 이력자료의 효율적인 관리·활용을 위하여 비모수회귀식(NPR, Non-Parametric Regression) 기반 누락자료 추정기법을 개발하였다. NPR기법은 입력과 출력자료간의 관계를 결정하는 과정으로서 파라미터 계산 없이 새로운 관측자료를 쉽게 모형에 추가할 수 있다. 따라서 모형에 필요한 파라미터를 주기적으로 갱신할 필요가 없기 때문에 현장에서의 운영에 장점을 가지고 있다.

개발된 NPR모형은 과거자료에 포함되어 있는 다양한 패턴을 활용하고, 실제 개발된 모형이 집계자료관리시스템 등의 시스템에 탑재되는 것을 고려하여 파라미터를 최소화하고 다양한 형태의 누락자료를 처리할 수 있도록 설계되었다. 누락자료의 형태별로 시나리오를 설정하여 개발된 모형을 적용한 결과, 개발된 모형의 추정력은 기존 모형에 비하여 매우 우수하게 나타났으며, 모형의 연산 수행속도 또한 현장 적용에 무리가 없는 것으로 분석되었다.

연구의 수행과정은 기존연구 고찰, 모형의 개발 및 적용, 그리고 결론 및 향후연구로 구성된다. ①기존연구 고찰은 현행 누락자료 추정기법과 본 연구의 접근법인 NPR기법 진행되며, 보고된 누락자료 추정기법의 근본적인 문제점을 도출하도록 한다. ②모형의 개발 및 적용은 모형의 개발 시 고려사항, 누락 교통자료 추정모형 개발, 그리고 개발모형의 적용 및 평가로 구성된다.

II. 기존연구 고찰

기존연구 고찰은 국·내외 누락자료 추정기법을 통해 기존연구의 근본적인 한계를 도출하고, 본 연구에서 이용한 비모수회귀식(NPR)의 특성과 적용사례에 대한 검토를 수행하도록 한다.

1. 누락자료 추정기법

미국의 Virginia ADMS, California PeMS 등의 시스템에서는 이력자료를 이용한 통계처리, 시·공간적 추세 등을 이용한 누락자료의 보정을 수행하고 있다. 누락 교통자료 추정기법에 관한 연구는 2000년부터 ADMS의 도입과 더불어 보고되고 있다. Smith B.L. & Conklin, J.H. (2002)은 과거 차로이용률과 인접 검지기 자료를 이용하여 보정처리기법을 개발하고 적용하였다. Chen, C. 등(2003)은 단일루프검지기를 대상으로 전·후 지점 간의 선형회귀모형을 이용하여 결측자료 보정모형을 개발하였으며, Ni, D. 등(2005)은 결측자료의 형태에 따른 다양한 결측자료 보정처리 방법론을 소개하였다.

국내의 경우, 지점 검지자료를 대상으로 결측자료 추정 및 보정기법이 보고되었다. 이승재 등(2001)은 불규칙변동 분해 시계열분석 기법을 이용하여 AADT를 추정하였으며, 시계열적 변동요인으로 추세변동, 계절변동, 주기변동, 그리고 불규칙 변동을 고려하였다. 이승재 등(2002)은 판별분석과 신경망 모형을 이용하여 AADT를 추정하였으며, 신경망 모형의 적중률(%)은 평균 92.4로서 판별분석의 적중률 32.2보다 우수한 결과를 도출하였다. 강원의 등(2004)은 인공신경망을 이용하여 평일 교통량을 예측하였으며, 평균 92.5%의 정확도를 보였다. 김정연 등(2006)의 연구는 이력자료를 이용한 다양한 결측자료 보정기법을 제시하고 평가하였다. 분석결과, 이력자료를 이용한 보정기법이 가장 우수하게 나타났으나, 시공간적 종속성을 전제로 간략한 모형을 제시하였다. 이력자료 동일주기 기법의 추정오차는 교통량

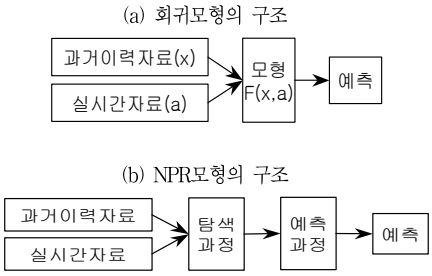
29~67%, 속도 7~10%, 시간 점유율 37~61%로 나타났다. 하정아 등(2007)은 시간대별 지점 교통량 결측자료 추정모형으로 계절 시계열 모형, 자귀회귀모형, EM 알고리즘을 비교·평가 하였다. 분석결과, 자귀회귀모형의 추정력이 우수하게 평가되었으며, 평균 추정오차는 약 15%로 나타났다. 김현석 등(2007)은 순환확률분포를 이용하여 시간대별 지점 교통량 결측자료를 추정하였으며, 결측 비율(10%~90%)에 따라 추정오차는 1.7~23.2%로 나타났다. 순환분포모형은 기존 방법보다 간단하고 이식성이 뛰어나며 결측패턴에 대해 roust한 반면, 시계열적으로 여러 개의 첨두가 존재하는 다봉형의 경우나 출/퇴근 등의 비대칭형에 대한 한계가 있다.

주요 교통변수(교통량, 속도)간 관계는 매우 동적이면서 밀접한 관련이 있다. 또한 누락자료의 특성상 교통변수 전체에 걸쳐 누락자료가 발생하게 된다. 그러나 기존연구의 고찰 결과, 보정대상은 교통량을 주로 다루고 있으며, 교통변수간 관계를 고려하여 누락된 교통변수(교통량, 속도)에 대한 다중대체법(multiple imputation)은 제시되고 않고 있다.

2. NPR기법

ITS분야에서 장래 교통상태의 예측에는 회귀모형, 시계열 기법, 인공신경망 모형, 칼만 필터링 기법 등이 주로 이용되고 있다. 이중 많이 이용되는 회귀식 모형은 설명변수를 고려하는 모수(parametric) 회귀식과 설명변수를 고려하지 않는 비모수 회귀식(NPR, Non-Parametric Regression)으로 구분된다.

(모수)회귀식은 설명변수에 의한 영향, 즉 파라미터가 장래에도 동일하게 종속변수에 영향을 미친다는 가정을 전제로 한다는 문제를 갖는다(Oswald 등, 2000). 이에 비해 NPR 기법은 현재 조건과 유사한 과거 관측치를 탐색하여 장래 상태를 추정하는데 적용이 용이하다(Smith, 2000). <그림 1>은 회귀모형과 NPR모형의 구조적 차이를 보여주고 있다. 일반적인 (모수)회귀식 모형과 달리 NPR모형은 입력과 출력자료간의 관계를 결정하는 것으로서 파라미터의 계산 없이 새로운 관측자료를 쉽게 모형에 추가할 수 있다. 또한 자료의 통계적 구조에 대한 가정을 하지 않는 (즉 distribution-free) 장점이 있다. 반면 변수간의 상관관계를 알 수 없는 체계적 접근방법이 아니며, 타 예측기법들에 비하여 연산시간 상대적으로 오래 걸린다는 단점이 있다. 즉 변수간의



<그림 1> 회귀 및 NPR 모형의 구조

상관관계를 알 수 없다(Oswald 등, 2000).

NPR기법은 과거의 패턴 즉, 경험을 바탕으로 장래의 상태를 결정하는 의사결정과정이라고 할 수 있다. NPR은 예측시간대에 상태(state)와 유사한 과거의 상태 벡터로 구성된 의사결정집단인 군집을 지속적으로 탐색하는 동적군집모형이라 할 수 있다. 이는 NPR이 예측시간대의 군집수를 정의하는 대신 현행(current) 입력상태와 유사한 과거 경우에 대한 군집을 정의하기 때문이다. NPR의 군집(neighborhood)을 정의방법은 크게 최인접 개수 k를 정하는 k-nearest neighbor 기법과 최인접 개수의 ± 범위를 설정하는 Kernel neighborhood 기법이 있다.

NPR기법을 적용하기 위해서는 방대한 양의 이력자료(즉, 과거의 정보)가 요구되며, 과거 상태와 현재 상태를 이용한 장래 상태의 예측에 있어 다양한 분야에 이용되고 있다. Robinson(1983), Mulhern & Caprara (1994)는 혼재된 무질서 상태에서 비선형 시계열 예측문제를 NPR을 기반으로 모형화하였다. Disbro & Frame(1989)와 Mulhern & Caprara (1994)의 연구는 무질서 상태에서 타 기법보다 NPR이 장점을 갖는다고 언급하였다. Karlsson & Yakawitz(1987)은 NPR을 기반으로 강우량을 예측하였으며, Mulhern & Caprara(1994)는 시장반응(market response)의 예측에 NPR을 적용하였다.

교통분야에서 NPR은 교통량 또는 통행시간 예측에 주로 적용되었다. Davis & Nihan(1991), Smith 등(2002), 그리고 Sun 등(2003)은 NPR을 단일시간대(single interval) 교통량 예측에 적용하였으며, Qi & Smith(2004)는 유고지속시간을 예측하였다. Chang, H. 등(2010)은 BMS 이력자료와 실시간 통행시간 자료를 이용하여 버스 정류장간 다중시간대 출발지 기준 경로통행시간을 예측하였으며, 우수한 결과를 도출하였

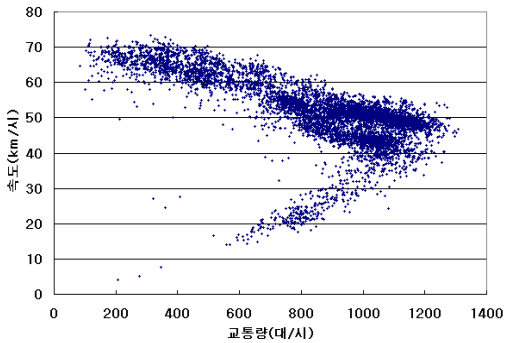
다. Oswald 등(2000)은 NPR 기법을 이용하여 고속도로 교통량을 예측하였으나, 오차가 10% 이상으로 크게 나타났으며, 요일 등 시간적 주기성에 대한 고려가 없었다. 이외에도 Smith 등(1996), Smith 등(1999)은 교통류 예측에 패턴기반 기법을 적용하였으나, 시간에 따른 교통류 상황변화의 주기성을 고려하지 않은 한계를 가지고 있다.

III. 모형의 개발 및 적용

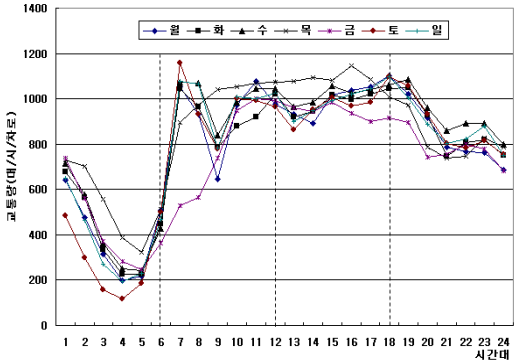
1. 모형개발의 고려사항

1) 교통자료의 특성

교통상태는 교통상황에 따라서 매우 동적이며 다양하게 나타난다. <그림 2>는 실제 교통량-속도 관계로 교통변수의 임의적 변동성을 보여주고 있다. 교통자료(교통량, 속도, 점유율)간의 관계는 다양한 정보가 혼재되어 있음으로 결정론적으로 규정하기는 참으로 어려운 일이다. 그러나 과거 이력자료의 다양한 변동에는 많은 정보를 포함하고 있다. 따라서 모형의 개발시 이력자료에 포함되어 있는 다양한 정보를 이용할 수 있는 기법을 개발하도록 한다.



<그림 2> 교통량-속도 관계



<그림 3> 교통량의 시계열 변동 및 주기성

2) 시계열 변동 및 주기성 고려

교통상태의 복잡한 다양성은 주중/주말, 각 요일별 시간대별 패턴과 시계열 변동에 내재되어 있다. 따라서 시계열 자료에 평활화 과정을 적용할 경우 시계열자료의 변동과 내재된 정보는 희석되며, 추정되는 상태의 시계열적 변동에 대한 민감도는 낮아지게 된다. <그림 3>은 동일 지점의 요일별 시간대별 교통량 시계열자료의 시계열적 변동과 시간대별 주기성의 다양성을 보여주고 있다. 교통량의 첨두 시간대는 1~3개까지 다양하게 나타나고 있다. <그림 2>와 <그림 3>을 고려하면 각 시간대의 교통상태는 매우 복잡한 다양성을 나타냄을 알 수 있다. 따라서 추정모형은 시계열적 변동과 특정 시간대별 주기성의 고려가 가능해야 한다.

3) 누락자료의 다양성에 대한 고려

누락자료는 시스템의 일시적 장애, 현장에 설치된 정보수집 장치의 일시적 장애 및 고장 등 다양한 원인으로 발생하게 됨으로 누락자료의 형태는 시·공간적으로 다양하게 나타난다. 따라서 개발되는 모형은 다양한 형태의 누락자료를 처리하고 교통변수들에 대한 다중대체가 가능하도록 개발되어야 한다.

4) 추정의 불확실성 문제

임의성을 수반하는 동적인 교통상태의 특성을 고려하면 예측상태에 대한 불확실성(uncertainty)은 예측/추정 기법의 근본적인 문제라 할 수 있다. 모형이 지나치게 정적이거나 민감하지 않은 경우 교통상태의 동적특성을 설명할 수 없으며, 지나치게 민감한 경우 교통상태의 동적특성을 고려할 수 있으나 (국부적으로) 예측오차는 증폭될 수 있다. 따라서 개발되는 모형은 요구되는 추정오차의 범위에서 교통상태를 예측하면서 안정적인 모형을 개발하도록 한다.

5) 적용의 용이성

모형의 정확도가 요구수준을 만족하면 모형의 연산수행속도와 더불어 운영상의 용이성을 고려해야 한다. 모형의 예측력이 뛰어나더라도 시스템 측면에서 연산 수

행시간이 추정정보를 활용할 수 있는 시간적 범위를 초과하거나, 운영측면에서 지속적으로 모형의 파라미터를 조정해야 한다면 실제 모형의 적용성은 현저히 떨어진다. 따라서 파라미터의 조정이 시스템적으로 간편하면서 모형의 연산속도가 빠른 모형을 개발하여야 한다.

2. 누락 교통자료 추정모형 개발

1) 누락자료 추정기법의 선정

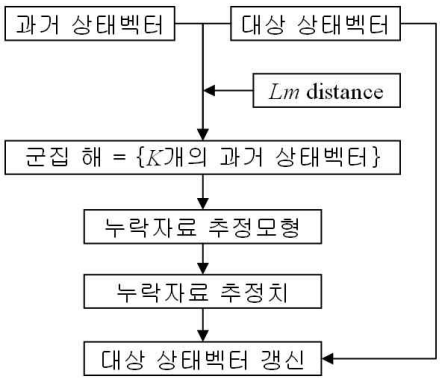
누락 교통자료 추정모형은 모형개발의 고려사항에서 언급한 5가지 요소를 충족해야 한다. Robinson (1983), Disbro & Frame(1989), Mulhern & Caprara (1994)의 연구는 혼재된 무질서 상태에서 비선형 시계열 예측에 있어 타 기법보다 NPR이 장점을 갖는다고 지적하였다. 또한 NPR의 적용이 적합한 경우는 ①과거 관측치가 평활(smooth)하지 않고 무질서한 시계열적 변동을 포함하는 경우, ②대규모 관측자료를 확보할 수 있는 경우, ③지속적으로 최근의 교통상태 자료가 확보되는 경우, ④다양한 패턴이 혼재되어 있는 경우이다. 이상을 고려할 때 NPR기반 접근법이 타 모형에 비하여 적합하다고 판단된다. 따라서 본 연구에서는 NPR을 기반으로 누락 교통자료 추정모형을 개발하였다.

2) NPR모형의 수행과정

일반적으로 NPR모형은 상태벡터(state vector), 상태벡터간 거리 메트릭(distance metric), 군집해의 자료구조, 그리고 예측모형으로 구성된다. <그림 4>는 본 연구의 누락자료 추정을 위한 NPR모형의 수행과정을 보여주고 있다. 모형의 수행과정은 거리 메트릭을 이용하여 현재(또는 대상 시간대) 상태벡터와 유사한 과거 상태벡터를 탐색하여 k개의 과거 상태벡터로 구성된 군집(neighborhood)을 구성하는 과정과 구성된 군집과 예측모형을 이용하여 장래 (또는 대상) 상태를 예측하는 과정으로 구성된다. 본 연구는 장래 예측이 아닌 누락자료 추정을 수행하게 된다. 따라서 상태벡터는 과거 상태벡터와 과거자료 중 누락된 자료가 포함된 대상 상태벡터로 구분한다.

3) 추정의 시·공간 정의

상태벡터를 구성하기 위해서는 모형의 개발에 이용되는 시·공간적 범위의 정의가 선행되어야 한다. 본 연구



<그림 4> NPR모형의 수행과정

에서 개발되는 모형은 개별 시간대에서 여러 시간대까지 개별차로의 누락 교통자료의 추정이 가능하도록 모형을 개발하고자 한다. 따라서 시간적 범위는 개별 시간대와 연속적인 $n(\geq 1)$개의 개별 시간대로 구성되는 시간대 그룹(g)으로 시간대를 정의하며, 24시간은 $m(\geq 1)$개의 시간대 그룹으로 구성된다. 그리고 공간적 범위는 차로별로 누락된 교통자료를 추정하기 위하여 차로수(l)로 정의한다. 이러한 시·공간적 정의는 시간대 길이와 시간대 그룹의 개별 시간대 개수를 조정함으로써 다양한 상태벡터의 구성을 가능하게 한다. 따라서 누락자료의 다양한 형태에 대하여 적용이 용이하다.

4) 상태벡터의 구성

기존 NPR모형의 상태벡터는 입력 상태벡터와 출력 상태벡터로 구성되며, 입력 상태벡터는 과거 상태벡터와 현행(current) 상태벡터로 구성된다. 본 연구에서는 상태평가의 시간대가 장래가 아닌 과거임으로 시간대 그룹(g)의 상태벡터는 이력(historical) 상태벡터(H_g), 대상(target) 상태벡터(T_g), 출력(output) 상태벡터(O_g)로 정의한다. 그리고 누락(missing) 정보를 표현하기 위하여 누락 상태벡터(M_g)를 추가적으로 이용한다. 그리고 각각의 상태벡터(i,j), $1 \leq i \leq n$, $1 \leq j \leq l$는 $n \times l$행렬로 구성되며 다음과 같다.

H_g^p: 교통변수(p)의 이력 상태벡터
p={교통량(q), 속도(s), 점유율(o)}
T_g^p: 교통변수(p)의 대상 상태벡터
O_g^p: 교통변수(p)의 출력 상태벡터
M_g: 누락정보 상태벡터로서 $M_g(i,j)$의 값은 누락시 0 그리고 정상시 1로 구성

5) L_m 거리

대상 상태벡터(T_g)와 유사한 이력(historical) 상태벡터(H_g)를 탐색하여 출력 상태벡터(O_g)로 구성된 군집을 구성하기 위해서는 탐색 알고리즘을 이용하게 되며, 탐색의 과정에서 일반적으로 T_g와 H_g간의 L_m 거리를 이용한다. L_m 거리는 m=1인 경우 Manhattan Distance, m=2이면 Euclidean Distance로 알려져 있다. 본 연구에서는 상태벡터간 거리가 증가할수록 상태거리가 증가하는 Euclidean거리를 이용하도록 한다. Euclidean거리는 Manhattan 거리와 같이 상태벡터 간의 절대거리가 같은 경우라도 상태벡터의 요소간 거리를 고려하기 때문에 유사한 상태간의 관계를 설명할 수 있기 때문이다.

$$L_m = (\sum_{i}^{d} |x_i - y_i|^m)^{1/m}, \quad d \geq i \tag{1}$$

본 연구에서는 모형의 구조와 연산과정을 설명하기 위하여 행렬 A와 B의 곱(×)에 대한 연산은 다음과 같이 정의하며, 이는 실제 알고리즘의 구현에 있어 매우 용이하다.

$$A \cdot B = \begin{pmatrix} a, b \\ c, d \end{pmatrix} \times \begin{pmatrix} 1, 2 \\ 3, 4 \end{pmatrix} = \begin{pmatrix} a \times 1, b \times 2 \\ c \times 3, d \times 4 \end{pmatrix} \tag{2}$$

시간대 그룹(g)의 개별 교통변수(p)에 대한 T_g^p와 H_g^p 간 Euclidean거리(U_g^p)는 $p_{\max}$을 이용하여 개별 상태벡터의 요소값을 0~1로 정규화한 후, 다음과 같이 누락 상태벡터(M_g)를 이용하여 누락자료를 제외한 정상자료간의 U_g^p을 계산한다.

$$U_g^p = \left[\sum_{i,j} \left\{ M_g(i,j) \times \frac{T_g^p(i,j) - H_g^p(i,j)}{p_{\max}} \right\}^2 \right]^{1/2} \tag{3}$$

여기서, $p_{\max}$는 교통변수(p)의 최대값

개별 교통변수(p)별 U_g^p을 산정한 후, 교통변수별 가중치(w_p)를 이용하여 교통변수간 시간대 그룹(g)의 가중된 거리(U_g)를 다음과 같이 산정한다.

$$U_g = \sum_p w_p \cdot U_g^p \tag{4}$$

여기서, $\sum_p w_p = 1.0, \quad 0 \leq w_p \leq 1.0 \quad \forall p$

6) 군집해의 자료구조

본 연구에서 군집(neighborhood)을 구성하는 이웃(neighbor) 해를 탐색하는 과정은 KNN(k-nearest neighbor) 기법을 이용하였다. 따라서 군집해의 자료구조는 k개의 이웃해로 구성되며, 교통변수(p)의 시간대 그룹(g)에 대한 자료구조는 다음과 같이 입력상태와 출력상태 벡터로 구성된다.

i	입력상태 벡터		출력상태 벡터	
	독립변수		종속변수	거리
1	$h_{g,1}^p \times M_g$	→	$h_{g,1}^p \times \lvert M_g - 1 \rvert$	$U_{g,1}$
2	$h_{g,2}^p \times M_g$	→	$h_{g,2}^p \times \lvert M_g - 1 \rvert$	$U_{g,2}$
…	…		…	…
K	$h_{g,K}^p \times M_g$	→	$h_{g,K}^p \times \lvert M_g - 1 \rvert$	$U_{g,K}$
…	…		…	…
k	$h_{g,k}^p \times M_g$	→	$h_{g,k}^p \times \lvert M_g - 1 \rvert$	$U_{g,k}$

여기서, $h_{g,K}^p \in H_g^p, \quad 1 \leq K \leq k, \quad \forall k$

7) 누락자료 추정모형

군집해가 구축되면 추정모형을 이용하여 누락자료를 추정하게 된다. 추정모형은 종속변수를 산술평균한 방법과 상태간 거리의 역수로 가중 평균한 방법을 들 수 있다. 그러나 산술평균 모형은 상태간 거리를 이용할 수 없는 단점이 있다. 이러한 단점을 극복하기 위하여 상태간 거리의 역수로 가중 평균한 방법을 많이 이용하며 산술평균보다 우수한 예측값을 보였다(Smith, B.L. 등 2002, Chang, H. 등 2010). 본 연구에서는 두 모형의 추정력을 비교한 후 적정 모형을 선정하도록 한다.

모형1: 직접평균

$$O_g^p = H_g^p \times |M_g - 1| = \frac{\sum_{i=1}^{k} h_{g,i}^p \times |M_g - 1|}{k} \tag{5}$$

모형2: $1/U_{g,K}$로 가중평균

$$O_g^p = H_g^p \times |M_g - 1| = \frac{\sum_{i=1}^{k} \left(\frac{h_{g,i}^p \times |M_g - 1|}{U_{g,i}} \right)}{\sum_{i=1}^{k} \frac{1}{U_{g,i}}} \tag{6}$$

8) k-Nearest Neighbor(KNN) 알고리즘

NPR의 군집(neighborhood)을 구성하는 방법은 최인접 개수 k를 정하는 KNN(k-nearest neighbor) 알고리즘과 상수값인 Kernel distance를 이용하여 최인접 개수의 ±범위를 설정하는 Kernel neighborhood 알고리즘이 있다. 이 두 탐색 알고리즘은 Smith, B.L. 등(2002)과 Chang, H. 등(2010)에서 적용되었다. KNN 알고리즘은 이력 상태벡터($h_g^p \times M_g$, $h_g^p \in H_g^p$)을 검색하여 대상 상태벡터($T_g^p \times M_g$)와 거리가 가장 가까운 k개의 이웃해($h_g^p \times |M_g - 1|$)를 구성하는 탐색과정으로서 본 연구에서 이용한 KNN 알고리즘은 다음과 같다.

주어진 시간대 그룹(g), $T_g^p \times M_g$, 그리고 최인접 이웃해의 개수 k에 대하여:

①군집해 자료구조의 1에서 k까지 neighbor를 초기화

②모든 이력 상태벡터(h_g^p, $h_g^p \in H_g^p$)에 대하여

ⓐ j번째 이력 상태벡터 $h_{g,j}^p \times M_g$와 대상 상태벡터 $T_g^p \times M_g$간의 $U_{g,j}$을 계산

ⓑ If $U_{g,j} < U_{\max}$ Then

(여기서, $U_{\max} = \max\{U_{g,1},..,U_{g,K},..,U_{g,k}\}$)

ⅰ) $U_{\max}$에 해당하는 neighbor를 제거

ⅱ) $U_{g,j}$의 독립변수, 종속변수, 그리고 거리를 군집에 갱신

ⅲ) 갱신된 군집에서 $U_{\max}$을 탐색

③추정모형을 이용하여 $O_g^p = H_g^p \times |M_g - 1|$을 산정

3. 개발모형의 적용 및 평가

1) 시·공간적 및 내용적 범위 설정

사례분석 대상지점은 <그림 5>와 같으며, 서울시에서 교통량과 속도 정보를 수집하기 위하여 설치·운영 중인 쌍루프 검지기이다. 수집된 자료는 방향별 차로별 시간대별로 관리·분석되고 있다. 본 연구에서는 도심방향(충정로→신문로) 자료를 이용하도록 한다. 이력자료의 시간적 범위는 2007년 3월 1일에서 5월 31일까지 시간대별 차로별 교통량과 속도이며, 5월 마지막 한 주는 개발모형의 평가에 적용하도록 한다.

회귀모형은 모수(parametric) 모형과 비모수(non-parametric) 모형으로 구분할 수 있다. 본 연구에서 제시한 모형은 비모수 회귀를 기반으로 개발되었다.

따라서 기존 모형과의 평가는 모수모형인 ARIMA모형을 이용하였다. 적용된 ARIMA모형은 시계열자료의 분석에 널리 적용되는 자기회귀모형을 기반으로 하는 ARIMA [1,0,0]을 적용하였다. ARIMA모형의 종속변수는 지속적인 누락이 발생한 차로의 교통량 (또는 속도) 시계열자료이며, 독립변수는 정상적인 차로의 교통량, 속도 시계열자료이다.

<그림 5> 사례분석 검지기 설치 지점

2) 평가 시나리오 설정

누락 교통자료는 다양한 형태를 보인다. 그러나 대부분의 경우에 있어 원인자 공사로 인한 파손과 내구연한 도달에 따른 기능장애로 인한 차로별 누락자료 발생이다. 이러한 경우 복구에 상당한 기간이 요구됨으로 상당기간(1~6개월)의 누락자료가 발생하게 된다. 따라서 본 연구에서는 <표 1>과 같이 3개 차로에서 발생할 수 있는 6개 누락자료 시나리오를 구성하고 이를 평가하도록 한다. 시나리오 1~3은 3개 차로 중 1개 차로의 교통자료가 누락된 경우이며, 시나리오 4~6은 3개 차로 중 2개 차로에서 누락자료가 발생된 경우이다. 따라서 누락자료 보정에 이용할 수 있는 대상 상태벡터(T_g^p)의 정상자료는 시나리오 1~3의 경우 2개 차로, 그리고 시나리오 4~6은 1개 차로이다.

도시부 교통량의 첨두 시간대는 1~3까지 다양하게 나타날 수 있음을 <그림 3>에서 언급하였으며, 이를 고려하여 시간간격 길이는 한 시간으로 그리고 시간대 그

<표 1> 평가 시나리오

구분	누락 차로수	1차로	2차로	3차로
시나리오1	1개 차로	×	○	○
시나리오2		○	×	○
시나리오3		○	○	×
시나리오4	2개 차로	×	×	○
시나리오5		×	○	×
시나리오6		○	×	×

주 1) ○: 정상자료, ×: 누락자료 발생

룹(g)은 6시간을 1개 그룹으로 설정하였다. 따라서 하루 24시간은 4 개 그룹으로 구분되어 추정을 수행하게 된다. 시간대 그룹(g)은 심야 시간대(1~6시), 오전 시간대(7~12), 오후 시간대(13~18시), 저녁 시간대(19~24시)로 구성된다. 그리고 추정에 이용되는 교통자료는 교통량(대/시/차로)과 속도(km/시/차로)이며, U_g의 산정시 이용되는 교통변수별 가중치는 0.5: 0.5로 설정하였다.

3) 평가지표 설정

모형의 추정력을 평가하기 위한 평가지표는 평균절대값백분율오차(MAPE; Mean Absolute Percentage Error)를 적용하였다. MAPE는 적정 k값의 산정 및 오차분석에 이용하였다. 그리고 추정 결과의 세부적인 시계열 분석은 시계열적 변동을 이용하여 분석하고, 산포도, 오차(%)의 분포와 쌍체 t검증(paired t-test)을 이용하여 통계적 검증을 수행하도록 한다.

$$MAPE(\%) = \left(\frac{1}{n} \sum_{i=1}^{N} \frac{x_i - \hat{x_i}}{x_i} \right) \times 100 \qquad (7)$$

여기서, x_i: 관측값, $\hat{x_i}$: 예측값, n: 자료의 개수

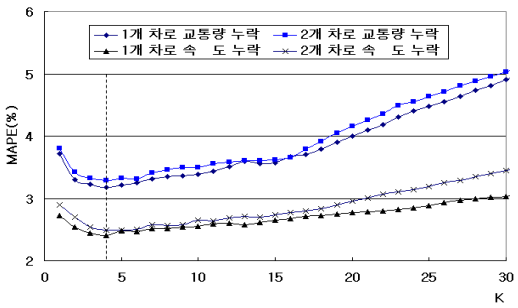
4) 적정 k값 및 오차분석 결과

<그림 6~7>은 시나리오 1~3(1개 차로 누락)인 경우와 시나리오 4~6(2개 차로 누락)에 대하여 모형별 k값에 따른 관측치와 추정치의 오차 정도를 보여주고 있다. k값이 증가할수록 오차는 감소하다 다시 증가하는 아래로 볼록한(convex) 형태의 오차곡선을 보이고 있다. 모형 1과 2에서 속도의 오차가 교통량에 비하여 낮게 나타나고 있다.

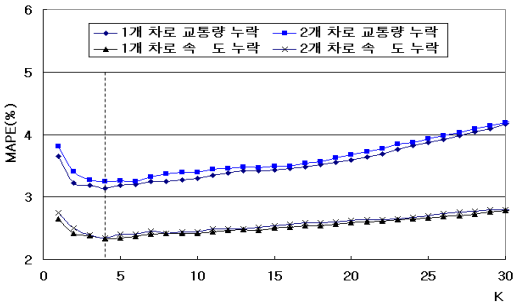
1개 차로 누락인 경우가 2개 차로 누락의 경우보다 오차가 낮게 나타나고 있으며, 이는 1개 차로 누락의 경우 2개차로 누락보다 많은 정보를 이용하였기 때문으로 판단된다. k값이 증가함에 따라 모형1은 모형2에 비하여

<표 2> 모형별 적정 k값 분석 결과

교통 변수		교통량		속도	
시나리오		1~3	4~5	1~3	4~5
누락 차로수		1	2	1	2
모형1	k값	4	4	4	4
	MAPE(%)	3.18	3.29	2.40	2.48
모형2	k값	4	4	4	4
	MAPE(%)	3.14	3.24	2.33	2.34



<그림 6> 모형1의 k값에 따른 오차



<그림 7> 모형2의 k값에 따른 오차

오차가 크게 증가하고 있으나, 최소오차가 되는 k값은 <표 2>와 같이 4로 동일하게 분석되었다.

<표 3>은 모형1, 모형2 그리고 ARIMA모형의 추정 교통변수별 시나리오별 시간그룹별 추정오차를 보여주고 있다. 추정 누락 교통량 오차의 범위는 모형1 2.18~4.63%, 모형2 2.17~4.55%, ARIMA모형 2.59~11.45%로 나타나 개발모형의 추정의 추정오차는 ARIMA

<표 3> 오차분석 결과 비교 (단위(%))

구분	교통변수	시나리오	시간 그룹(시)				
			1~6	7~12	13~18	19~24	평균
모형1	교통량	1~3	4.35	3.27	2.22	3.02	2.77
		4~6	4.63	3.36	2.18	3.03	3.04
	속도	1~3	1.75	4.43	2.26	2.15	2.32
		4~6	1.73	4.61	2.21	2.09	2.53
모형2	교통량	1~3	4.33	3.17	2.19	2.99	2.74
		4~6	4.55	3.26	2.17	3.02	3.00
	속도	1~3	1.74	4.31	2.25	2.14	2.29
		4~6	1.71	4.42	2.22	2.10	2.49
ARIMA	교통량	1~3	7.87	4.96	2.59	3.88	4.83
		4~6	11.45	5.56	3.25	4.61	6.22
	속도	1~3	1.83	2.55	1.86	1.29	1.88
		4~6	2.51	4.40	3.65	2.04	3.15

주1) 평균은 교통변수별 모든 개별 추정치의 평균임

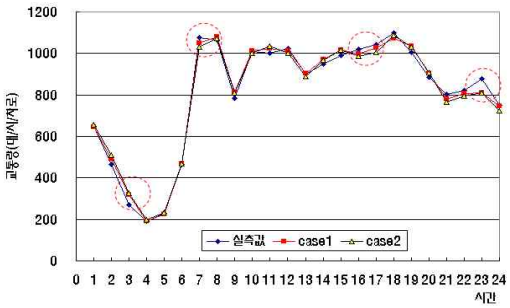
모형보다 낮게 나타났으며, 모형2의 추정오차는 모형1 보다 낮게 나타났다. 시나리오 1~3(1개 차로 누락)인 경우에 대한 시나리오 4~6(2개 차로 누락)의 교통량 추정오차는 개발모형의 경우 2% 종반에서 3% 초반대로 크게 증가하지 않고 있으나, ARIMA모형은 평균 4.83%에서 6.22%로 증가하고 있다. 또한 교통량이 상대적으로 낮은 시간그룹 1~6의 경우 개발모형의 추정오차는 5% 미만인 반면, ARIMA모형은 약 7.9%에서 11.5%로 증가하고 있다. 따라서 정상적인 시계열자료의 정보가 감소하는 경우(즉, 누락자료의 빈도가 증가하는 경우), 개발모형의 추정력이 ARIMA모형에 비하여 높으며 안정적이라고 판단된다. 또한 개발모형의 추정오차는 기존 연구(Smith 등, 2000)의 단일 시간대 교통량 예측에 NPR을 적용한 결과(예측오차 약 10%)보다 매우 낮게 나타났다.

추정 누락 속도의 오차범위는 모형1 1.73~4.61%, 모형2 1.71~4.42%, ARIMA모형 1.29~4.40%로 나타나 ARIMA모형의 추정오차 범위가 크게 나타났다. 그러나 양호한 정보의 양이 적어지는 시나리오 4~6의 경우 시나리오 1~3의 오차에 대한 오차 증가분은 ARIMA모형 1.73%, 개발모형 약 0.2% 정도로 나타나고 있다. 따라서 누락자료의 빈도가 증가하는 경우, 개발모형의 추정오차 증가량은 ARIMA모형에 비하여 낮게 나타나고 있다. 따라서 개발모형이 ARIMA모형보다 안정적이라 판단된다.

모형 2의 예측력이 모형 1에 비하여 전반적으로 우수하게 나타난 이유는 모형2의 경우 $1/U_g$을 적용하였기 때문으로 판단된다. 이것은 대상상태에 더 근접한 과거상태 일수록 예측하고자 하는 상태에 대한 더 많은 정보를 가지고 있다는 사실을 의미한다. 모든 모형에서 상대적으로 변동량이 적은 속도의 추정오차가 교통량보다 낮게 나타나고 있다. 그리고 누락자료 추정오차는 시나리오 1~3이 시나리오 4~6보다 낮게 나타났다. 이는 1개 차로 누락자료 발생의 경우가 2개 차로 누락자료 발생의 경우보다 더 많은 정보를 이용하기 때문으로 판단된다. 2개 차로 누락자료가 발생하는 경우, 개발모형의 추정오차는 교통량과 속도 모두 3%이하인 반면, ARIMA모형은 교통량 6% 이상, 속도 3%이상으로 증가하고 있다. 따라서 누락자료가 많은 경우 개발모형이 ARIMA모형보다 안정적이라고 판단된다.

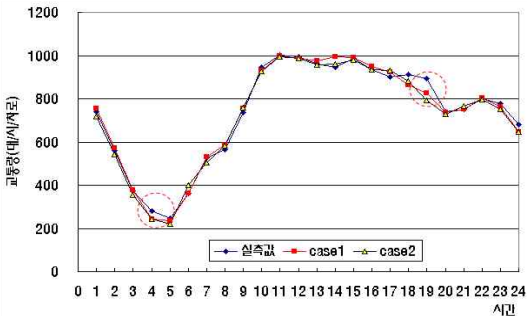
5) 시계열 변동 분석

<그림 8~15>는 개발모형과 ARIMA모형의 월요일과

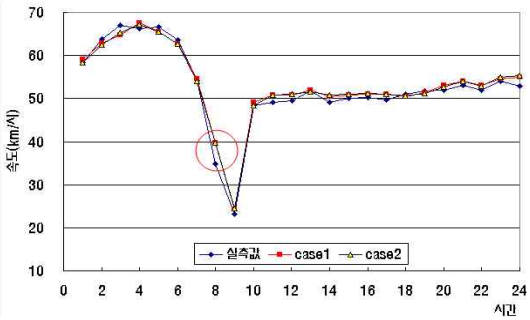


주 1) Case 1: 1차로 누락 2~3차로 정상시, 1차로 추정
주 2) Case 2: 1~2차로 누락 3차로 정상시, 1차로 추정

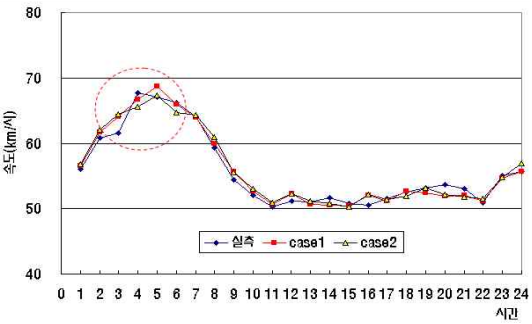
<그림 8> 개발모형의 교통량 추정결과: 월요일



<그림 9> 개발모형의 교통량 추정결과: 일요일



<그림 10> 개발모형의 속도 추정결과: 월요일

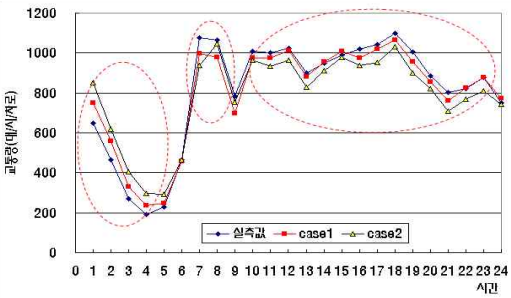


<그림 11> 개발모형의 속도 추정결과: 일요일

일요일의 시나리오 1과 4에 대한 1차로 누락자료 추정결과를 시계열 변동으로 보여주고 있다. 개발모형의 시계열 변동은 두 모형이 유사하게 나타났다. 따라서 보다 예측력이 우수한 모형2의 결과를 이용하여 분석하도록 한다. 교통변수의 상승과 하강 상태에서 개발모형의 추정오차는 타 상태에 비하여 국부적으로 높게 나타난 반면, ARIMA 모형은 개발모형에 비하여 다소 높게 나타나고 있다. 첨두/첨미에서 추정오차는 개발모형이 ARIMA모형에 비하여 낮게 나타나고 있다. <그림 8, 12>의 시간대 9~19와 같이 시계열 변동이 심한 경우, 개발모형은 낮은 오차로 시계열적 변동을 설명하고 있으나 ARIMA모형은 개발모형보다 큰 오차로 시계열적 변동을 설명하고 있다.

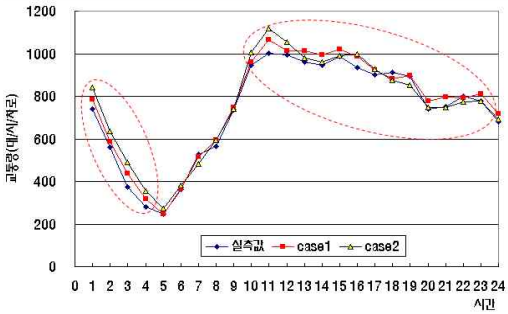
<그림 9~11, 13~15>의 시간대 11~18과 같이 시계열자료가 안정된 경우, 개발모형의 추정오차는 ARIMA 모형보다 낮게 나타나고 있다. 이는 개발모형이 ARIMA 모형보다 시계열자료에 내재된 국부적이고 임의적인 변동을 보다 잘 설명하기 때문으로 판단된다. ARIMA모형의 추정오차는 <그림 12~15>와 같이 1개 차로 정상시의 자료를 이용한 경우(case 2)에 비하여 2개 차로 정상시 자료를 이용한 경우(case 1)에 증폭되고 있고 있다. 또한 추정오차의 증가는 속도보다 교통량이 크게 나타나고 있다.

반면, 개발모형의 추정오차는 <그림 8~11>과 같이 정상자료의 양이 감소하더라도 큰 차이를 보이지 않고 있다. 따라서 누락자료의 구성비가 증가한 상태에서 누락자료의 추정 정확도는 개발모형이 ARIMA모형보다 안정적이라 판단된다.

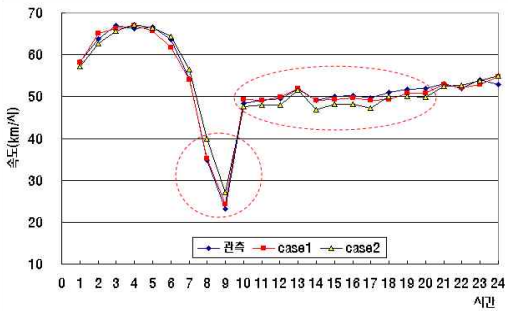


주1) Case 1: 1차로 누락 2~3차로 정상시, 1차로 추정
주2) Case 2: 1~2차로 누락 3차로 정상시, 1차로 추정

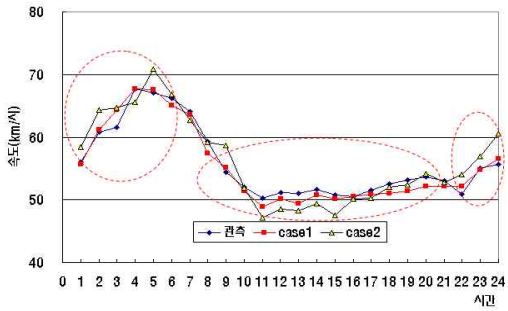
<그림 12> ARIMA모형의 교통량 추정결과: 월요일



<그림 13> ARIMA모형의 교통량 추정결과: 일요일



<그림 14> ARIMA모형의 속도 추정결과: 월요일



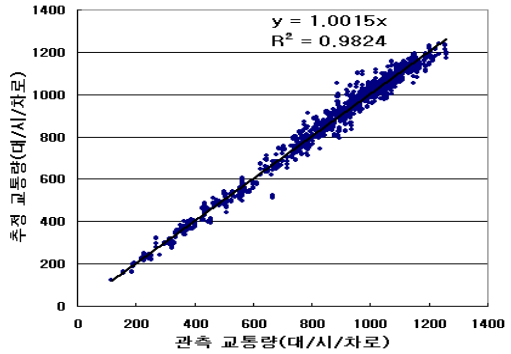
<그림 15> ARIMA모형의 속도 추정결과: 일요일

6) 오차분포 및 통계적 검증

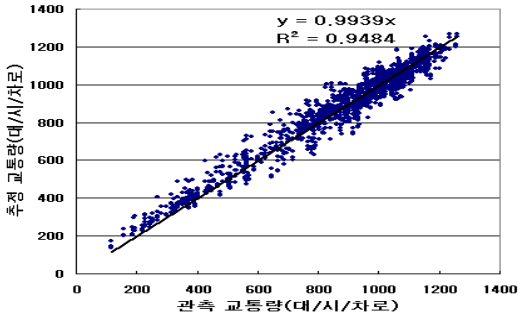
<그림 12~19>와 <그림 20~23>은 개발모형과 ARIMA 모형의 교통변수별 산포도와 추정오차(%) 분포를 각각 보여주고 있으며, 추정오차(%)는 다음과 같다.

$$\text{추정오차}(\%) = \left(\frac{\text{관측값} - \text{추정값}}{\text{관측값}}\right) \times 100 \qquad (8)$$

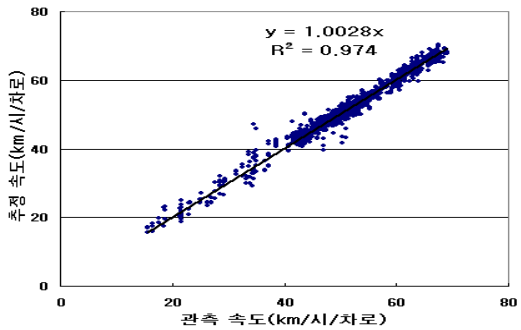
<그림 16~17>의 교통량 산포도 분석결과, 개발모형의 결정계수(R^2)는 0.982 ARIMA모형은 0.948로서 개발모형의 결정계수가 높게 나타났다. 계수값은 약 1.0으로



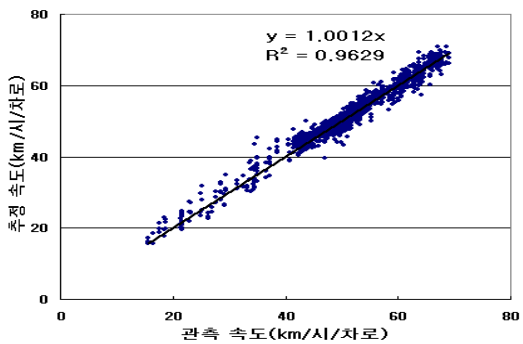
<그림 16> 개발모형의 교통량 산포도



<그림 17> ARIMA모형의 교통량 산포도



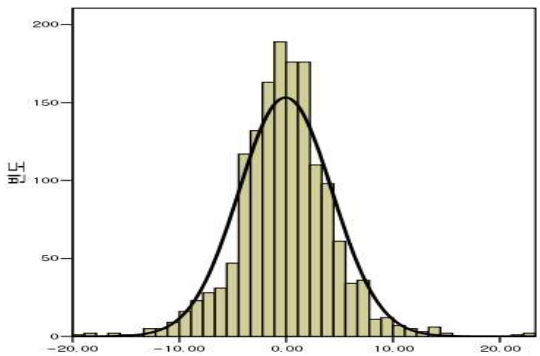
<그림 18> 개발모형의 속도 산포도



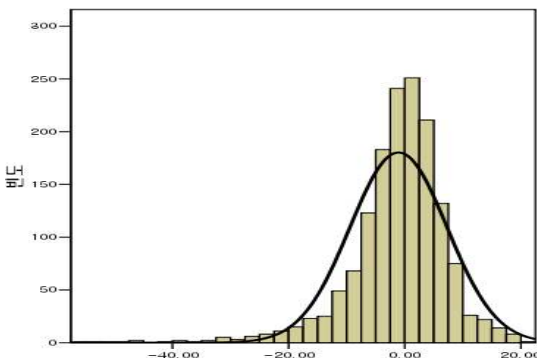
<그림 19> ARIMA모형의 속도 산포도

나타나 교통량의 정도에 따른 추정 교통량의 편향은 없는 것으로 나타났으며, 잔차의 양은 교통량이 증가함에 따라 서서히 증가하다 감소하고 있다. 그러나 ARIMA모형의 경우 관측 교통량 600대/시/차로 이하에서 추정 교통량은 관측 교통량보다 다소 높게 추정되었다. 따라서 <그림 12>와 <표 3>에서 교통량이 적은 시간대(0~6시)의 MAPE는 7.87~11.45까지 증가하였다. 교통량 추정오차 ±5%이내의 적중률은 <그림 20~21>의 추정오차 분포에서 개발모형 89.7%, ARIMA모형 72.8%로서 개발모형이 높게 나타났다. 추정오차 -10% 이하와 +10% 이상의 구성비(%)는 ARIMA모형 12.8%로서 개발모형의 2.4%에 비하여 높게 나타났다. 따라서 ARIMA모형을 적용할 경우 일부 보정 교통량에 대하여 심한 왜곡을 초래할 수 있다고 판단된다. 이는 개발모형에 비하여 ARIMA모형이 경우 국부적이고 임의적인 변동에 대한 설명력이 낮기 때문으로 판단된다.

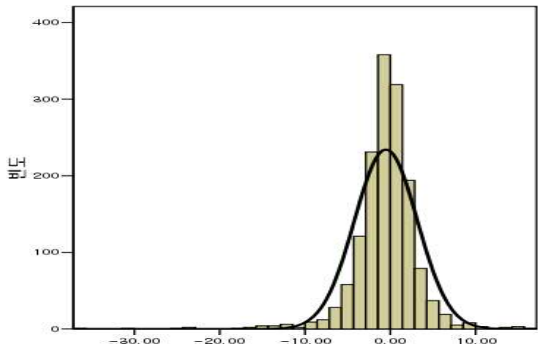
<그림 18~19>의 속도 산포도 분석결과, 개발모형의 결정계수는 0.97, ARIMA모형은 0.96으로 나타나 모형간의 차이는 없는 것으로 나타났으며, 계수값은 약



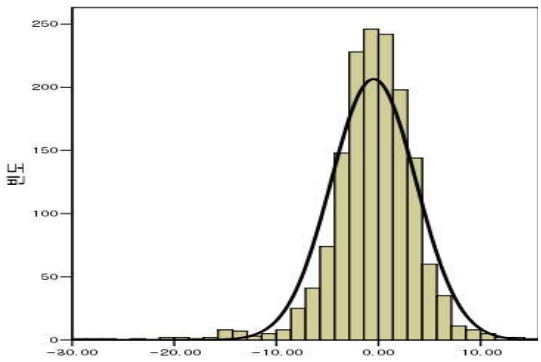
<그림 20> 개발모형의 교통량 추정오차(%)의 분포



<그림 21> ARIMA모형의 교통량 추정오차(%)의 분포



<그림 22> 개발모형의 속도 추정오차(%)의 분포



<그림 23> ARIMA모형의 속도 추정오차(%)의 분포

1.0을 동일하게 분석되었다. 속도 추정오차 ±5%이내의 적중률은 <그림 22~23>의 추정오차 분포에서 개발모형 95.4%, ARIMA모형 93.2%로 개발모형이 높게 나타났으며, 추정오차 ±5% 이내의 첨도는 개발모형이 ARIMA모형에 비하여 높게 나타나고 있다. 속도 추정오차 -10% 이하와 +10% 이상의 구성비는 개발모형 2.0%, ARIMA모형 2.3%로서 모형간 차이는 크게 나타나지 않았다.

<그림 18~19>에서 잔차의 양은 관측속도 40kph 이하에서 상대적으로 크게 나타나고 있다. 따라서 추정오차 |±10%|에서 발생하는 오차는 대부분 저속도에서 발생하고 있으며, 속도의 양이 낮기 때문에 증가된 오차로 판단된다. 따라서 두 모형의 속도 추정값은 정상자료를 왜곡하지 않을 것으로 판단된다. 그러나 ARIMA모형의 추정값은 관측 속도 40kph이하에서 다소 높게 평가되고 있다. 통계적 검증은 쌍체 t-검증(paired t-test)을 수행하였으며, 가설은 다음과 같다.

귀무가설 : H_o : $\delta=0$ 대립가설 : H_1 : $\delta>0$

<표 4> 통계적 검증 결과

구분	교통변수	t-value	p-value
개발모형	교통량	1.009	0.316
	속도	1.270	0.208
ARIMA	교통량	1.446	0.153
	속도	1.424	0.159

교통변수(교통량, 속도)별 예측결과를 관측값과 유의수준 5%에서 쌍체검증한 결과는 <표 4>와 같다. 자유도는 120이상임으로 양측검정의 $t(\infty,0.025)$=1.96이다. 검증결과, 모든 경우에 대하여 p-value는 0.025보다 크고, t-value는 1.96보다 작아 귀무가설을 기각할 수 없다. 따라서 개발모형과 ARIMA모형의 추정치는 관측치와 동일한 집단으로 분석되었다.

7) 모형의 연산 수행속도

NPR모형의 연산 수행시간은 타 모형에 비하여 많은 시간이 소요된다고 알려져 있다. 연산 수행속도의 문제는 NPR 뿐만 아니라 이력자료를 이용하는 모든 누락자료 예측기법의 문제점이며, 이는 탐색해야할 이력자료의 양과 지원시스템에 의해 결정되기 때문이다. 따라서 시스템 탑재를 통한 현장 적용을 평가하기 위해서는 모형의 연산 수행속도를 고려해야 한다. 개발모형은 1주일(7일), 하루 4개의 시간그룹, 6개의 누락자료 시나리오, 1~50까지의 k값 시나리오로 총 8,400 (=7×4×6×50)번의 반복연산을 수행하였으며, 일반 데스크 탑 컴퓨터에서 약 3.8초가 소요되었다. 따라서 본 연구에서 개발된 모형이 자료관리시스템 등에 탑재될 경우 연산저하로 인한 문제는 없을 것으로 판단된다. 이는 모형의 개발 시 시간대 그룹(g)을 이용하여 시간대별 교통자료의 패턴을 인식함과 동시에 연산횟수를 줄였기 때문이다. 이상의 성능평가 결과를 토대로 모형별 장/단점을 정리하면 <표 5>와 같다.

Ⅴ. 결론 및 향후연구

1. 결론

이력자료를 효율적으로 운영·관리하기 위해서는 집계자료 관리시스템(ADMS)의 도입과 더불어 누락자료를 신뢰성을 담보하면서 효율적으로 추정할 수 있는 누

<표 5> 모형별 장/단점 비교

구분	장점	단점
개발모형	• 모형이 간단, 적용이 용이 • 보정성능이 우수함 • 다양한 누락자료의 형태에 적용이 용이함 • 통계적 구조에 대한 가정을 하지 않음 • 비선형적 관계를 이용함 • 파라미터의 주기적 갱신이 필요하지 않음 • 다중대체가 용이함 • 대량의 교통자료 보정에 적용이 용이 • 임의 변동에 대한 보정력 높음. • stochastic 상태에서 설명력이 높음 • 시스템 탑재가 용이함 • 모형의 확장이 용이	• 변수간 상관관계를 알 수 있는 체계적 접근방법이 아님 • heuristic한 접근방법임 • 다량의 이력 교통자료와 패턴이 필요함 • 이력자료가 방대할 경우, 연산 수행속도가 타 모형에 비하여 느림. 이는 이력자료 기반의 보정모형의 공통적 단점임
ARIMA	• 계열내 관측값의 상관관계를 이용함 • 통계적으로 변수간의 상관관계 파악 가능 • 보정성능이 우수함 • 파라미터의 주기적 갱신이 필요하지 않음 • 상대적으로 다량의 이력자료가 필요하지 않음	• 모형 복잡, 적용이 어려움 • 선형분석 기법을 적용 • 다양한 누락자료의 형태에 따라 순차적인 반복연산 필요 • 다중대체가 어려움 • 대량의 교통자료 보정기법으로서는 한계가 있음 • 임의적인 변동에 대한 보정력 낮음 • stochastic 상태에서 설명력이 낮음 • 모형의 확장이 어려움

락자료 추정기법이 필수적이다. 이러한 인과관계를 고려하여 본 연구에서는 이력자료를 이용한 비모수회귀식(NPR)기반의 누락자료 추정모형을 개발하였으며, 결론은 다음과 같이 요약할 수 있다.

첫째, 개발된 모형은 자료관리시스템에 탑재를 목표로 개발되었다. 이를 위하여 이용 가능한 자료, 모형의 연산 수행속도, 모형의 적용시 운영측면, 그리고 교통자료의 특성을 고려하여 모형을 개발하였다.

둘째, 이력자료에는 다양한 형태의 누락자료가 포함된다. 따라서 본 연구에서는 누락자료 정보행렬(시간대 개수×차로수)과 교통변수별 행렬을 적용함으로서 다중대체가 가능하도록 하였으며, 개별 시간대에서 여러 시간대까지 다양한 형태의 행렬을 이용한 행렬단위의 누락자료 추정을 가능하게 하였다. 따라서 일시적인 누락자료와 장기 누락자료에 대한 적용이 가능하다. 또한 인접차로의 정상자료를 이용함으로서 실제 교통상태를 반영하도록 하였다.

셋째, 기존의 결측자료 추정기법은 지점별 시간 또는 일 교통량의 추정을 목적하였으며, 차로별 누락자료 추정모형을 개발하지 않았다. 그러나 본 연구에서는 인접차로의 정상자료를 이용하여 차로별 시간대별로 누락자료 추정을 수행함으로서 보다 미시적으로 누락자료를 추정하였으며, 그 결과 또한 기존 연구에 비하여 우수하게 나타났다.

넷째, 상류부 그리고/또는 하류부 검지기를 이용한 누락자료 추정기법과는 달리 독립적으로 설치된 검지기에서도 적용이 가능한 기법을 개발함으로서 모형의 적용시 공간적 전이성을 고려하였다. 마지막으로 블록 단위 즉, 여러 시간대×차로수의 행렬을 이용한 모형을 개발함으로써 기존에 NPR모형이 가지고 있는 연산 수행속도의 문제를 해결함으로써 현장 적용이 가능하도록 하였다.

2. 향후연구

본 연구의 사례대상 지역은 서울 도심의 단속류 구간으로서 반복적인 교통상태가 고속도로보다 강하게 나타난다. 따라서 더 많은 이력자료 구축을 통한 적용 연구가 필요하다. 또한 연속류를 대상으로 모형의 개발 및 적용 연구가 추가적으로 필요하며, 시간대 길이(time length)를 1시간에서 30분, 15분, 5분으로 짧게 하여 보다 시계열적 변동이 큰 경우에 대한 모형의 적용 및 평가가 수행되어야 할 것이다.

참고문헌

1. 이승재 · 백남철 · 권희정 · 최대순 · 도명식(2001), “불규칙변동 분해 시계열분석 기법을 사용한 AADT 추정”, 대한교통학회지, 제19권 제6호, 대한교통학회, pp.65~73.
2. 이승재 · 백남철 · 권희정(2002), “단기조사 교통량을 이용한 AADT 추정연구”, 대한교통학회지, 제20권 제6호, 대한교통학회, pp.59~68.
3. 강원의 · 백남철 · 윤해경(2004), “AIC(AKaike's Information Criterion)을 이용한 교통량 예측 모형”, 대한교통학회지, 제22권 제7호, 대한교통학회, pp.155~159.
4. 김정연 · 이영인 · 백승걸 · 남궁성(2006), “차량 검지자료 결측 보정처리에 관한 연구(이력자료 활용방안을 중심으로)”, 대한교통학회지, 제24권 제7호, 대한교통학회, pp.27~40.
5. 김현석 · 남두희 · 임강원 · 이영인(2007), “순환확률분포를 이용한 교통량 결측자료 보정 모형”, 대학교통학회지, 제25권 제4호, 대한교통학회, pp.109~121.
6. 하정아 · 박재화 · 김성현(2007), “일반국도 상시 교

통량자료를 이용한 교통량 결측자료 추정", 대학교통학회지, 제25권 제1호, 대한교통학회, pp.121~132.

7. Altman, N.S. (1992), "An introduction to kernel and nearest-neighbor nonparametric regression", The American Statistician, Vol. 46, No. 3, pp.175~185.
8. Chang, H. et al. (2010), "Dynamic multi- interval bus travel time prediction using bus transit data", Transportmetrica, Vol. 6. No. 1, pp.19~38.
9. Chen, C., Kwon, J., Rice, J., Skabardonis, A., and Varaiya, P. (2003), "Detecting errors and imputing missing data for single loop surveillance systems", TRR Vol. 1855, pp.160~167.
10. Ni, D., Leonard, J.D., Guin, A., Feng, C. (2005), "Multiple imputation scheme for overcoming the missing values and variability issues in ITS data", JTE, Vol. 131, Issue 12, pp.931~938.
11. Davis, G. and Nihan, N. (1991), "Nonparametric regression and short-term freeway traffic fore- casting", JTE, Vol. 117, No. 2, pp.178~188.
12. Disbro, J.E., Frame, M. (1989), "Traffic flow theory and chaotic behavior", New York State Department of Transportation Report FHWA/NY/SR-98/91, New York.
13. Karlsson, M. and S. Yakowitz (1987), "Rainfall-runoff forecasting methods, old and new", Stochastic Hydrology and Hydraulics, Vol. 1, No. 4, pp.303~318.
14. Mulhern, F. J. and R. J. Caprara (1994), "A nearest neighbor model for forecasting market response", International Journal of Forecasting, Vol. 10, No. 2, pp.191~207.
15. Oswald R. K., Scherer W.T. and Smith B.L.(2000), "Traffic flow forecasting using approximate Nearest Neighbor Nonparametric regression", Research project report for U.S. DOT University transportation center.
16. Qi, Y., and Smith, B.L. (2004), "Identifying nearest-neighbors in a large-scale incident data archive", TRR, 1879, pp.89~98.
17. Robinson, P. (1983), "Nonparametric estimators for time series", Journal of Time Series Analysis, Vol. 4, No 3, pp.185~207.
18. Smith, B.L. (1995), "Forecasting freeway traffic flow for intelligent transportation system applications", Doctorial dissertation. Department of civil engineering, University of Virginia, Charlottesville.
19. Smith B.L., and Conklin J.H. (2002), "The use of local lane distribution patterns to estimate missing data values from traffic monitoring systems", TRR, Vol. 1811, pp.50~56.
20. Smith, B.L., B.M. Williams, and R.K. Oswald (2002), "Comparison of parametric and nonparametric models for traffic flow forecasting", TR Part C, Vol. 10. pp.303~321.
21. Sun, H. et al. (2003), "Use of local linear regression for short-term traffic forecasting", TRR, 1936, pp.143~150.

♧ 주 작 성 자 : 장현호
♧ 교 신 저 자 : 한동희
♧ 논문투고일 : 2008. 10. 24
♧ 논문심사일 : 2008. 12. 2 (1차)
2009. 7. 27 (2차)
2009. 10. 9 (3차)
2010. 5. 6 (4차)
2010. 6. 3 (5차)
♧ 심사판정일 : 2010. 6. 3
♧ 반론접수기한 : 2010. 10. 31
♧ 3인 익명 심사필
♧ 1인 abstract 교정필