한국ITS학회논문지
제3권, 제1호 (2004. 3 · 6)
pp.45~52

실시간 누락 교통자료의 대체기법에 관한 연구

Study on Imputation Methods of Missing Real-Time Traffic Data

장 진 환* (Jin-hwan, Jang) 류 승 기** (Seung-ki, Ryu) 문 학 룡*** (Hak-yong, Moon) 변 상 철**** (Sang-cheal, Byun)

요 약

현재 여러 지자체에서 혼잡한 도시교통의 이동성 및 안전성을 향상시키기 위해 첨단교통관리체계(ITS)를 구축 · 운영중인데 이러한 시스템에서 수집하는 교통상황에 대한 실시간 자료가 노면상황, 악천후, 통신 및 장비자체의 결함 등으로 인해 수많은 자료가 결측된다. 이러한 결측 자료로 인해 통행시간 예측 및 각종 연구가 불가능한 경우가 발생하며 또한 도로의 계획과 기하구조 설계시 기본 자료가 되는 AADT 및 DHV 등의 교통 파라메터들이 과소 또는 과대 추정될 수 있어서 심각한 손해를 끼칠수 있다. 따라서 본 연구에서는 부득이하게 누락되는 교통량 자료에 대해 전 · 후기간 평균, 회귀모형, EM, 시계열 모형들을 활용한 대체기법들을 살펴보았고, 그 결과 시계열 모형을 이용한 대체의 경우 MAPE, 불균등계수, RMSE 가 각각 5.0%, 0.030, 110으로 가장 좋은 결과를 보였고 나머지 대체기법들은 평가지표에 따라 조금씩 다른 결과를 보였으나 대체로 만족할 만한 수준의 결과를 낳았다.

Abstract

There are many cities installing ITS(Intelligent Transportation Systems) and running TMC(Traffic Management Center) to improve mobility and safety of roadway transportation by providing roadway information to drivers. There are many devices in ITS which collect real-time traffic data. We can obtain many valuable traffic data from the devices. But it's impossible to avoid missing traffic data for many reasons such as roadway condition, adversary weather, communication shutdown and problems of the devices itself. We couldn't do any secondary process such as travel time forecasting and other transportation related research due to the missing data. If we use the traffic data to produce AADT and DHV, essential data in roadway planning and design, We might get skewed data that could make big loss. Therefore, the study have explored some imputation techniques such as heuristic methods, regression model, EM algorithm and time-series analysis for the missing traffic volume data using some evaluating indices such as MAPE, RMSE, and Inequality coefficient. We could get the best result from time-series model generating 5.0%, 0.03 and 110 as MAPE, Inequality coefficient and RMSE, respectively. Other techniques produce a little different results, but the results were very encouraging.

Key Words : 대체, ITS, 회귀분석, EM, 시계열, MAPE, RMSE, 불균등 계수

* 회 원 : 한국건설기술연구원
** 비회원 : 한국건설기술연구원 선임연구원
*** 회 원 : 한국건설기술연구원 선임연구원
**** 비회원 : 한국건설기술연구원
† 논문접수일 : 2004년 2월 18일

Ⅰ. 연구의 배경 및 목적

현재 여러 지자체에서 혼잡한 도시교통의 이동성 및 안전성을 향상시키기 위해 첨단교통관리체계(ITS)를 구축·운영 중이다. 또한 한국건설기술연구원에서는 건설교통부의 위탁을 받아 매년 도로교통량 통계연보 및 전국 일반국도의 교통정보를 제공하기 위해 일반국도 상에 약 400여대의 상시 교통량 조사장비를 설치·운영 하고 있다.

이러한 시스템은 교통상황에 대한 실시간 자료를 수집해 통행자들에게 유용한 정보를 제공하고 수집된 교통량 자료를 도로의 계획 및 설계시 기본자료가 되는 도로교통량 통계연보에 사용되기 때문에 교통상황에 대한 모든 자료를 수집할 수 있어야 함에도 불구하고 노면상황, 악천후, 통신 및 장비자체의 결함 등으로 인해 수많은 교통량 자료가 결측된다.

이러한 결측 자료로 인해 통행시간 예측 및 각종 연구가 불가능한 경우가 발생하며 또한 도로의 계획과 기하구조 설계시 기본 자료가 되는 AADT 및 DHV 등의 교통 파라메터들이 과소 또는 과대 추정될 수 있어서 심각한 손해를 끼칠수 있다.

따라서 본 연구에서는 부득이하게 누락되는 교통량 자료에 대해 전·후기간 평균, 회귀모형, EM, 시계열 모형들을 활용한 대체 기법들을 살펴보고 적절한 사후 평가를 통해 최적의 대체[1]기법을 제시하고자 한다.

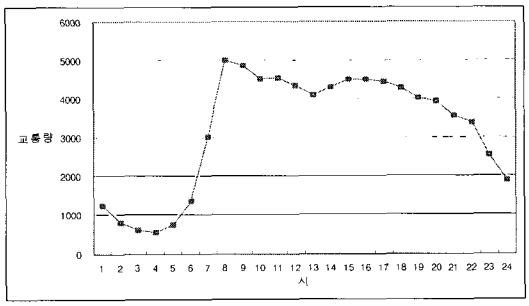
교통량 결측 자료에 대해 미국의 AASHTO의 "Guidelines for Traffic Data Program"에서는 대체(Imputation)가 부적절 하다고 했지만,[5] 이는 분석자가 과거의 자료를 적절히 이용하지 못했을 경우를 가정하고 있다.[6]

실시간으로 수집되는 이러한 교통자료에 대한 유효성 검증 및 누락 자료에 대한 대체에 관해서 Brian L. Smith 등은 미국 텍사스 교통관리센터에서 수집되는 교통량 자료를 이용해 경험적 방법 및 EM(Expectation-Maximization), DA(Data Augmentation)을 이용해 누락자료를 대체했고,[6] Satish Sharma 등은 캐나다의 알버타주에 위치한 상시 교통량 조사장비의 자료를 이용해 계수방법, 회귀, 시계열, 유전자 알고리즘을 이용한 신경망분석을 이용해 누락자료를 대체했으며[7], Z. Wall 등은 장비로부터 불량으로 전송되는 실시간 교통량 자료에 대해 유효성 검증 및 보정 알고리즘을 개발했다.[8] 또한 Chan Chen 등은 캘리포니아에서 운영 중인 상시 교통량 조사 장비로부터 전송되는 교통량 자료에 대해 불량 자료 검지 알로리즘과 누락자료를 대체하는 기법들을 제시했다.[9]

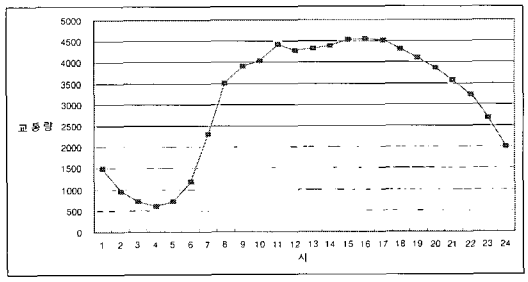
1) 대체(Imputation) : 누락된 자료를 추정된 자료로써 채우는 것.

Ⅱ. 분석 지점 현황 및 분석 시나리오

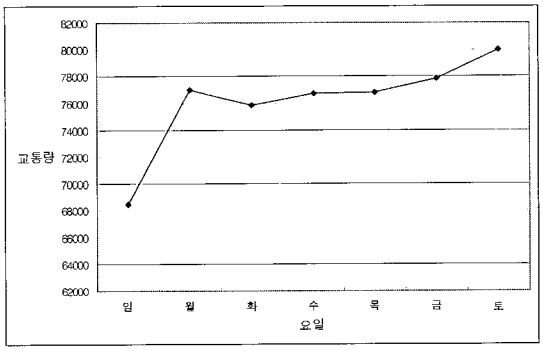
본 연구에서는 현재 한국건설기술연구원의 TMS (Traffic Monitoring System)에서 운영 중인 상시 교통량 조사지점의 자료를 이용했다. 분석지점은 경기도 광주시 중대리에 위치해 있고, 일반국도 3호선 상에 위치해 있으며, 분석지점의 2002년 AADT는 76,111대/일, K_{30} 값은 0.07로써 도시부 성격을 가지고 있으며,[1] 시간대별, 일별, 월별 교통량 패턴을 살펴보면 <그림 1>, <그림 2>, <그림 3>, <그림 4>와 같다.[2]



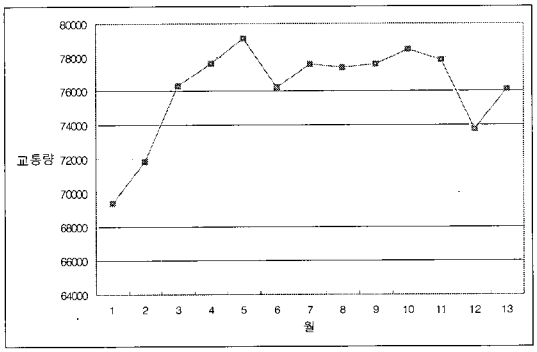
〈그림 1〉 주중 시간교통량 패턴



〈그림 2〉 주말 시간교통량 패턴



〈그림 3〉 요일 교통량 패턴



〈그림 4〉 월 교통량 패턴

분석지점의 교통량 패턴은 주중 일교통량의 경우 오전 첨두시에는 첨두 패턴을 나타내지만, 오후 첨두의 경우 평활된 패턴을 나타냄으로 인해 오후 첨두시에는 지체현상을 보일 것으로 판단된다. 한편, 주말 교통량에 비해 상대적으로 주중 교통량이 많으며 연중 교통량의 차이가 심하지 않음으로 보아 전형적인 교외부 도로의 형태를 나타내는 것으로 판단된다.

분석에 사용된 교통량 자료는 상기 분석지점의 2003년 자료 중 누락된 자료가 없는 2003년 3월 4일부터 6월 26일까지의 자료를 이용했다. 본 분석에서는 평일 자료 중 수요일 자료를 이용했는데, 상기기간의 수요일 자료를 무작위 추출(random sampling)에 의해 추출된 3월 19일과 4월 16일의 24시간 교통량 자료를 고의로 누락시킨 후 다양한 대체기법을 활용해 고의로 누락된 자료의 대체 값과 실제 수집자료 값을 비교해 최적의 대체기법을 제시했다.

Ⅲ. 평가 지표

1. 평균절대백분비오차(Mean Absolute Percent Error)

평균절대백분비오차(MAPE)은 변동계수 그룹별 교통량 조사횟수에 따른 AADT 추정치가 실제 AADT로부터 실제 AADT에 비해 평균적으로 얼마나 떨어져 있는가를 나타내는 지표로써, 수식으로 표현하면 식 (1)와 같다.

$$MAPE = \frac{\sum |PE_t|}{n} \qquad (1)$$

$$PE_t = \frac{e_t}{V_t} \cdot 100\%$$

여기서, e_t = t시간대의 오차, t = 1~24

2. 평균제곱오차제곱근(Root Mean Square Error)

평균제곱오차제곱근(RMSE)은 오차가 큰 곳에 가중치를 주어 평균하는 지표로써 이는 모형의 개발이나 분석에 널리 사용되어진다. 수식은 식 (2)와 같다.

$$RMSE = \sqrt{\frac{\sum e_t^2}{n}} \qquad (2)$$

3. 불균등계수(Inequality Coefficient)

불균등계수(U)는 타일(H. Theil, 1966)에 의해서 개발된 방법으로 계량경제모형에서 구한 예측값의 정확성에 대한 척도로서 식 (3)과 같다.[3]

$$U = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}[Y_i - X_i]^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}Y_i^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2}} \qquad (3)$$

여기서, Y_i = t시간대의 실제값

X_i = t시간대의 대체값

Ⅳ. 이론적 고찰

1. 경험적 방법(Heuristic Techniques)

1) 전 · 후일 평균

누락된 날짜의 전·후일 교통량을 산술평균 하는 방법으로서 식 (4)와 같다. 이는 주중 교통량이 누락되었을 경우 사용 가능한 방법이지만, 주말에는 주중과 교통량 패턴이 상이함으로 인해 적절하지 않을 것으로 판단된다.

$$v_t^d = (v_t^{d-1} + v_t^{d+1})/2 \quad (4)$$

여기서, v: 교통량, d: 보정일, t: 시간대

2) 전 · 후주 평균

누락된 날짜의 전·후주 교통량을 산술평균하는 방법으로 식 (5)와 같다. 일반적으로 교통량 패턴은 주기성을 가지기 때문에 이 방법은 모든 요일에 대해 보편적으로 사용될 수 있을 것으로 보인다.

$$v_t^d = (v_t^{d-7} + v_t^{d+7})/2 \quad (5)$$

2. 통계적 방법(Statistical Techniques)

1) 회귀 모형(Regression Model)

회귀분석(regression analysis)은 알려진 변수와 알려지지 않은 변수 사이의 관계를 수리적인 식, 측 예측방정식으로 발전시킨 것으로서,[4] 결측값 대체가 필요한 변수를 종속변수로 하고 일련의 연관 요인들을 설명변수로 하는 회귀 모형을 만들어 결측값 대체에 활용하는 방법이다.

본 연구에서는 무작위 추출된 수요일 교통량 자료에 대해 수요일과 가장 상관관계가 ($r=0.97$) 높은 화요일 교통량을 독립변수로 하는 단순선형 회귀모형을 이용했다. 단순회귀식은 식 (6)과 같다.

$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad (6)$$

여기서, $\varepsilon_i = N(0, \sigma^2)$이고 서로 독립

α, β는 미지의 모수

X_i : i번째 주어진 고정된 X의 값

2) EM(Expectation Maximization)

EM 알고리즘(Expectation-Maximization Algorithm)은 E(=Expectation) 단계와 M(=Maximization) 단계로 구성된다. 총자료 X를 관측부분 X_{obs}와 결측부분 X_{mis}로 구분하여 표기할 때, E 단계에서는 관측 X_{obs}에 조건화하여 결측 X_{mis}에 대한 추정값을 구하여 X_{mis}를 대체한다.(파라메타 θ의 잠정 추정값 θ_0을 사용하여) 그리고 M 단계에서는 X_{obs}와 X_{mis}를 모두 써서 θ에 대한 우도(likelihood), 즉 $L(\theta X_{obs}, X_{mis})$를 최대화한다. 여기서 구한 θ값을 θ_0으로 놓고 다시 E 단계로 돌아간다. 이와 같이 E 단계와 M 단계를 반복함으로써 X_{obs}에서의 파라메타 θ에 대한 가능도 $L(\theta X_{obs})$를 최대로 하는 θ값을 찾는다.[10] 일반적으로 동일한 변수 세트가 사용된다면 회귀대체와 EM대체간의 차이는 크지 않다. 그 이유는 회귀대체가 EM 알고리즘에서 E 단계에 해당되기 때문이다.

본 연구에서는 EM 대체를 위해 통계적 분석 프로그램인 SPSS를 사용했고, 자료는 회귀분석에서 사용한 자료와 동일한 화요일 자료를 이용했으며 또한 EM 대체를 위해서는 특정 확률모형을 전제해야 하는데 본 연구에서는 수요일과 더불어 화요일의 분포를 정규분포라고 가정했다.

3) 계절 시계열모형(Seasonal ARIMA Model)

교통량과 같은 시계열자료는 일정한 시간 간격을 두고 동일한 현상이 반복되는 경향을 가지고 있다. 이와 같이 반복적인 현상이 계속 일어나는 시간 간격을 계절주기라고 하며, 이러한 특징을 지닌 시계열을 계절시계열(seasonal time-series) 이라고 한다. 이러한 계절시계열을 분석하는 방법으로는 윈터스의 지수평활법, 분해법에 의한 시계열분석 등이 있으나, 이와 같은 방법은 시계열이 4가지 성분

인 추세, 계절, 순환, 불규칙성분 등으로 서로 독립적으로 구성되어 있다는 가정하에서 이용하는 방법이다. 그러나 실제로 관측되는 시계열은 그 변동이 확률적이며, 계절변동 역시 확률적으로 다른 변동들과 연관되어 있기 때문에 쉽게 분해되지 않는다.[3]

따라서 본 연구에서는 확률 시계열모형인 ARIMA 모형을 계절 시계열 모형에 확장한, 일반적으로 널리 사용되는 승법계절통합혼합 $[ARIMA(p,d,q)]$ $(P,D,Q)_s$모형을 이용했으며 모형구축에 사용된 자료는 누락된 수요일 자료를 제외한 나머지 자료를 이용해 분석했다. 계절주기가 s이고 d차 차분과 s차 계절차분한 승법계절통합혼합모형의 일반식은 식 (7)과 같다.

$$\phi(B)\Phi(B)(1-B)^{d}(1-B^{s})^{D}Z_{t}=\theta(B)\Theta(B)a_{t} \quad (7)$$

여기서, $\phi, \Phi, \theta, \Theta$: 모수, B : 후진연산자, Z_t : t 시간대의 시계열 값

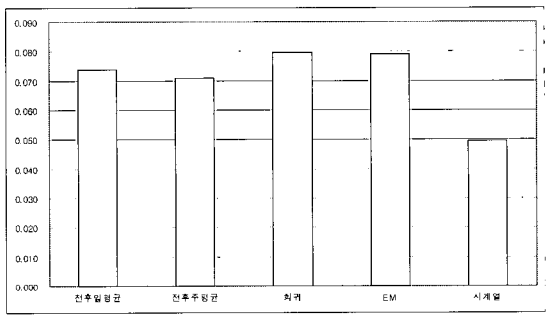
V. 분석 결과

누락된 자료의 대체시 일반적으로 널리 사용되는 통계적 기법인 회귀분석과 EM 알고리즘, 시계열 분석을 통해 교통관리센터에서 실시간으로 수집되는 교통량 자료의 누락자료를 대체했고, 아울러 교통량 패턴은 반복적으로 발생함으로 인해 경험적으로 얻을 수 있는 전·후일과 전·후주 자료의 산술평균에 의한 대체도 병행하여 각종 정량적 평가지표를 이용한 분석결과는 <표 1>과 같았고, 평가지표에 따라 그림으로 나타내면 <그림 5>~<그림 7>과 같은 결과가 나왔다.

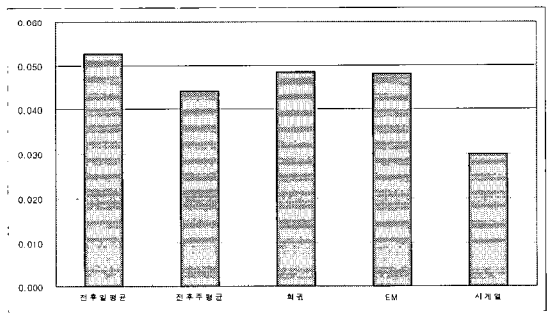
<표 1>과 <그림 5>~<그림 7>에서 볼 수 있듯이 여러 대체 기법을 통해 누락된 교통량 자료를 대체한 결과는 비교적 만족할 만한 수준이었다. 대체기법간에 다소 차이는 있었지만, 일반적으로 시계열에 의한 대체가 MAPE 5.0%, 불균등 계수 0.030, RMSE 110으로 가장 좋은 결과를 낳았다. RMSE가 110으로 나옴으로 인해 시계열에 의한 대체시 평균적으로 110vph의 교통량의 오차가 발생하는 것으로 분석되었다.

〈표 1〉 분석결과

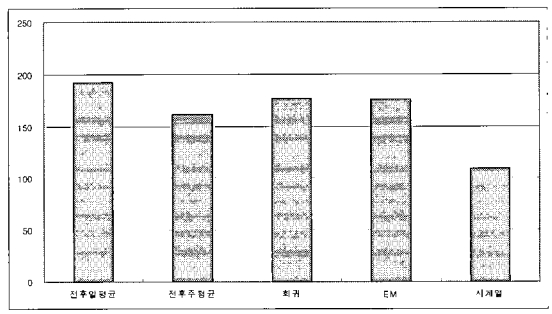
평가지표		MAPE	불균등계수	RMSE
대체기법	전 · 후일평균	7.4%	0.053	192
	전 · 후주평균	7.1%	0.044	161
	회귀분석	8.0%	0.049	176
	EM	7.9%	0.048	176
	시계열분석	5.0%	0.030	110



〈그림 5〉 MAPE 분석결과



〈그림 6〉 불균등계수 분석결과



〈그림 7〉 RMSE 분석결과

본 분석에서 사용한 계절 시계열 모형은 승법계절통합혼합 모형[$ARIMA(1,0,0)(1,1,0)_{24}$]이었고, 식별된 모형식은 식 (8)과 같고,

$$Z_t = B^{24} + \Phi B^{24} - \Phi B^{48} + \phi B - \phi B^{25} - \phi\Phi B^{25} + \phi\Phi B^{49} + a_t \quad (8)$$

추정해야 할 모수는 상수항 a_t와 ϕ, Φ 이다. 본 시계열모형의 모수를 추정하기 위한 자료는 2003년 3월 4일부터 6월 26일까지의 수요일 자료 중 고의적으로 누락시킨 3월16일과 4월19일 자료를 제외한 나머지 수요일 자료를 이용했으며, 모수를 추정한 결과는 <그림 8>과 같았다.

Model Description:
Variable: ZT
Regressors: NONE

Non-seasonal differencing: 0
Seasonal differencing: 1
Length of Seasonal Cycle: 24

Parameters:

AR1 _____ < value originating from estimation >
SAR1 _____ < value originating from estimation >
CONSTANT _____ < value originating from estimation >

95.00 percent confidence intervals will be generated.

FINAL PARAMETERS:

Number of residuals 336
Standard error 119.30061
Log likelihood -2085.568
AIC 4177.136
SBC 4188.5873

Analysis of Variance:

	DF	Adj. Sum of Squares	Residual Variance
Residuals	333	4843976.7	14232.634

Variables in the Model:

	B	SEB	T-RATIO	APPROX. PROB.
AR1	.4391165	.0487130	9.014364	.00000000
SAR1	-.5064919	.0476281	-10.634310	.00000000
CONSTANT	2.6223706	7.8751959	.332991	.73935050

〈그림 8〉 시계열분석 모수 추정 결과

따라서 추정된 모형 식은 식 (9)와 같다.

$$Z_t = Z_{t-24} - 0.5065Z_{t-24} + 0.5065Z_{t-48} + 0.4119Z_t - 0.4119Z_{t-25} + (0.4119)(0.5065)Z_{t-25} - (0.4119)(0.5065)Z_{t-49} + a_t \quad (9)$$

Ⅵ. 결론 및 향후과제

본 연구에서는 현재 여러 지자체에서 구축·운영되고 있는 교통관리센터(TMC)에서 수집되는 실시간 교통자료 중 여러요인으로 인해 부득이하게 누락되는 교통량 자료에 대해 여러 대체기법들을 이용해 임의누락 시킨 자료를 대체해서 실제 수집된 실 교통량 자료와의 비교·평가를 통해 누락된는 교통량 자료의 대체시 발생하는 오차를 분석했다.

분석 결과, 모든 대체기법에 대해 최대오차가 8%를 초과하지 않는 것으로 나타났으며, 이는 도로의 계획, 설계 및 여러 교통관련 연구에 사용될 수 있는 유용한 교통량 자료를 다양한 대체기법을 이용한 교통량 자료의 대체가 유효함을 입증할 수 있을 것으로 판단된다. 그 중에서 시계열 분석에 의한 대체가 가장 좋은 결과를 낳았고, 나머지 분석방법은 평가지표에 대해 다소 상이한 결과가 나왔다. 평가지표별로 살펴본 결과 일반적으로 불균등 계수가 평균절대백분비오차(MAPE)보다 낮은 오차율을 나타냈다.

경험적 방법인 전·후기간의 산술평균을 이용한 대체방법도 상당히 높은 수준의 정확도를 유지할 수 있었으나, 이는 과거자료를 이용한 모형화를 통한 통계적 분석방법에 비해 전·후 기간의 자료가 모두 존재하는 경우에만 사용할 수 있는 단점이 있다.

본 분석에서는 실시간 자료가 누락 없이 모두 수집된 일반국도 3호선 상에 위치한 단일지점에 대해 교통량이 가장 안정적이라고 판단되는 수요일의 교통량 자료만을 이용했으나, 향후 다양한 교통특성을 보이는 여러 지점에 대해 평일 및 주말의 다양한 요일에 대한 분석이 뒤따라야 할 것으로 보인다.

참 고 문 헌

[1] 도로의 구조·시설 기준에 관한 규칙 해설 및 지침, 건설교통부, 2000. 3

[2] 2002 도로교통량 통계연보, 건설교통부, 2003. 2

[3] 예측방법의 이해, 이덕기, SPSS 아카데미, 2001

[4] 경영 및 경제를 위한 통계학, 박우동 외, 세영사, 1996

[5] AASHTO Guidelines for Traffic Data Programs, American Association of State Highway and Trasportation Officials, 1992

[6] Exploring Imputation Techniques for Missing Data in Transportation Management Systems, Brian L. Smith et al. TRB 2003 Annual Meeting, 2003

[7] Effect of Missing Value Imputation on Traffice Prameters Estimations from Permanent Traffic Count, Satish Sharma et al. TRB 2003 Annual Meeting, 2003

[8] An Algorithm for the Detection and Correction of Errors in Archived Traffic Data, Z. Wall et al. TRB 2003 Annual Meeting, 2003

[9] Detecting Errors and Imputing Missing Data for Single Loop Surveillance Systems, Chao Chen et al. TRB 2003 Annual Meeting, 2003

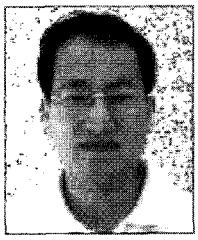
[10] Maximum likelihood from incomplete data via the the EM algorithm, Dempster, A.P. et al. Journal of the Royal Statistical Society, 1977

[11] The Fuzzy-Neural Network Traffic Prediction Framework with Wavelet Decomposition, Heng Xiao et al. TRB 2003 Annual Meeting, 2003

[12] Short-Term Traffic Forecasting Using the Local Linear Regression Model, Hongyu Sun et al. TRB 2003 Annual Meeting, 2003

〈저자소개〉

장 진 환(Jin-hwan, Jang)
2001년 : 경주대학교 학사
2004년 : 서울시립대학교 교통공학 석사
2001년~현재 : 한국건설기술연구원



류 승 기(Seung-ki, Ryu)
1990년 : 충북대학교 학사
1999년 : 충북대학교 전기공학 박사
1994년~현재 : 한국건설기술연구원

문 학 룡(Hak-yong, Moon)
1990년 : 숭실대학교 학사
2001년 : 숭실대학교 전기공학 박사
1997년~현재 : 한국건설기술연구원

변 상 철(Sang-cheal, Byun)
1994년 : 충북대학교 학사
2003년 : 서울시립대학교 교통공학 박사수료
1996년~현재 : 한국건설기술연구원