순환확률분포를 이용한 교통량 결측자료 보정 모형에 관한 연구

A Study on Modelling the Missing Data Imputation for Traffic Volume using Circular Probability Distribution

김 현 석
(한국건설기술연구원 선임연구원)

남 두 희
(한성대학교 정보시스템학과 교수)

임 강 원
(서울대학교 환경대학원 교수)

이 영 인
(서울대학교 환경대학원 교수)

목 차

I. 서론

1. 연구의 배경 및 필요성

도로의 교통특성 자료를 기반으로 구축된 도로용량편람의 공학적 검증 및 업데이트와 교통량 통계자료 수집·분석, 최근 국내에서 도로 이용의 효율성 제고를 목표로 하여 확대 구축중인 지능형 교통시스템(ITS)의 운영은 기본적으로 신뢰성있는 교통자료의 수집을 필요로 한다.

그러나 미국 TTI는 텍사스주에서 운영중인 교통자료 수집장비의 자료 결측률(missing rate)을 약 16~93%로 보고하고 있으며, Chandra and Al-Deek(2004)은 주간고속도로 I-4에 설치·운영중인 루프검지기의 자료 결측률을 약 15%로 보고하고 있다. 또한, 미국의 미네소타 교통부에서 운영중인 교통량 상시조사장비중에 약 40%의 장비가 결측 자료를 포함하고 있다는 보고를 하였다[24].

국내의 경우도, ITS사업의 차량검지기나 주로 교통량 통계연보 발간을 목적으로 하는 운영하는 교통량 상시조사장비의 자료 결측률도 공식적으로 보고된 바는 없으나, 국외의 사례에 준하는 것으로 추정하고 있다.

대부분 교통자료 수집장비에서의 자료 결측(data missing)은 장비 자체적인 결함에 주로 기인하나, 그 밖에도 도로 조건의 변화에 의해서도 발생하며, 수집장비에 대한 유지관리 수준을 최대한 높인다고 해도 결측률 0%를 확보하는 것은 현실적으로 불가능하다.

자료 결측의 심각성은 현실적으로 거의 대부분의 조사(survey) 과정에서 발생한다라는 점이다. 비단 교통 분야뿐만 아니라 인문사회 분야나 기상학, 생물학, 지구과학 등 거의 모든 분야의 조사과정에서 발생하며, 인력식이든 기계식이든 조사방식에 관계없이 발생한다. 한편, 최근에 실시간 교통정보 제공을 위해 확대 구축되고 있는 ITS사업이나 BIS사업과 교통량 통계자료 수집을 위한 교통량조사 사업에서도 차량검지기가 증설되고 있는 추세이므로, 자료 결측 문제는 앞으로도 계속 심화되어 이슈화될 것으로 예상된다.

이와 같이, 교통자료 수집장비에서의 자료 결측의 발생은 현실적으로 불가피한 현상으로 볼 수 있으며, 이와 같은 자료 수집과정에서 발생하는 결측을 신뢰성있게 추정하여 보정하였던 선행연구의 대부분은 교통량 자료의 결측값 보정시 통계적 검증없이 시간적인 임의의 종속성만 고려함으로

서, 보정 성능이 떨어지는 단점을 노출하고 있다. 이들 연구에서 적용했던 기법들 또한, 교통량 자료가 가지고 있는 가장 큰 특징인 주기적 순환성(periodic circularity)이 제대로 반영되지 못함으로서, 적용상 한계를 노출하고 있다.

2. 연구의 목적 및 내용

본 연구는 교통자료 수집장비에서의 자료결측 문제를 새롭게 제안하는 순환분포모형의 적용 및 평가를 통해 실질적으로 해소할 수 있는 이론적인 준거틀을 제공하는데 그 목적이 있으며, 구체적으로는 다음과 같다.

첫째, 과거 프로파일 이용법이나 선형 보간법 등 기존의 ad-hoc 또는 heuristic 접근법과 모형기반 및 최근의 알고리즘 기반의 체계적인 접근법 등 관련 선행 연구들을 분석하여 그 한계점과 이를 극복할 수 있는 대안을 모색한다. 둘째, 선행연구의 한계를 극복할 수 있는 대안으로서, 순환성이 강한 교통량 자료의 결측 보정을 위한 순환분포모형을 교통 특성별(도시부 및 지방부)로 개발하여 기존의 보정 모형과의 비교 평가를 통하여 그 성능 및 모형의 적용성을 확인한다. 셋째, 결측률과 결측 양상(missing mechanism)에 따른 비교분석을 통하여 자료 결측의 심각성과 결측 조건별 순환분포모형의 보정 효과 등에 대해 분석한다.

II. 자료 결측 현황 및 선행연구 고찰

1. 자료 결측 현황

자료 결측은 현실적으로 교통 분야뿐만 아니라 거의 대부분의 조사(survey) 과정에서 발생한다. 인문사회 분야나 기상학 및 생물학, 지구과학 등 거의 모든 분야의 조사 과정에서 발생하며, 인력식이든 기계식이든 조사 방식에 관계없이 발생한다. 또한, 자료 결측은 결과적으로 자료의 수집 비용을 증가시키며, 수집 정보의 신뢰성을 저하시킨다.

다양한 사회 현상의 메커니즘을 파악하기 위해 대부분 인력식으로 자료를 조사 수집하는 사회과학 분야의 경우, 조사의 대부분은 전수 조사(census)가 아닌 표본 조사 방식으로 시행되고 있다. 통계학과 확률 이론에 근거한 표본추출이론(random sampling theory)은 표본에서 추출한 값이 모집단의 추정치로서 대표성을 유지하는 것을 중요한 원칙으로 삼고 있다.

그러나, 표본조사 자료의 왜곡이나 무응답(non-response) 여부에 따라 모집단의 확률적 속성이 유지되기 어려운 경우가 빈번히 발생한다. 무응답이나 왜곡된 사례수가 많을 경우 이를 기초로 한 통계적 추론이나 의사 결정에 오류가 포함될 가능성이 커진다. 자동화된 차량검지기를 정보 수집단(information terminals)으로 이용하는 교통관리 시스템이나 모니터링 시스템에서 발생하는 자료의 결측은 국내외적으로 심각한 수준이다.

자료의 결측은 대부분 교통정보 수집 장비가 정상적으로 작동하지 않는 경우가 전제되기 때문에 정상적 가동의 정도를 나타내는 가동률(operating ratio)과는 반대의 의미로 해석할 수 있다.

국내에서 현재 운영중인 교통관리 및 모니터링 시스템의 2005년도 기준 교통정보 수집 시스템의 월별 가동률 현황을 살펴보면 다음과 같다.

- 교통량조사 사업 : 연중 최저 81.4 ~ 연중 최고 92.3%, 연평균 86.8%
- 국도ITS 운영관리 : 연중 최저 84.6 ~ 연중 최고 96.3%, 연평균 92.3%

ITS 서비스의 일환으로 대중교통 서비스의 질적인 향상을 목표로 하는 버스정보시스템(BIS)의 경우에는 버스의 위치 정보 등을 GPS와 무선통신을 활용하여 수집하고 있다. 이와 같은 버스정보시스템의 경우 교통량조사나 국도ITS 사업과 정보 수집 방법이나 정보의 단위, 내역 등에는 다소 차이가 있으나, 안양시에서 운영중인 버스정보시스템의 2006년 3월에서 5월까지 월별 버스정보 수집율(좁은 의미의 가동율)을 살펴보면 다음과 같다.

- 안양시 BIS사업 : 최저 92 ~ 최고 94%, 월평균 93%

따라서, 자료의 미수집율 또는 결측률(%)은 100−자료 수집율(%)로 정의될 수 있다. 본 연구에서는 교통량 상시조사지점별로 연간 수집되어야 할 시간 교통량 자료의 총 수량(24시간×365일=8,760개 시간 교통량)을 알고 있는 경우이므로, 다음과 같이 결측률을 정의하였다. 이러한 개념정의를 바탕으로 결측률별 순환분포모형의 보정 성능을 평가하였다. 즉, 결측률(%)은

$$\frac{\text{일정기간 결측된 자료의 수량}}{\text{일정기간 수집되어야 할 자료의 수량}} \times 100$$

2. 결측 자료 보정 방법론

완전한 자료가 요구되는 연구에 있어서는 이용가능한(available) 자료를 최대한 많이 확보해야 하는데, 이런 관점에서 결측값이 있는 자료를 삭제하는 방식인 Deletion Method는 결측의 처리에 있어서 적절하지 못하므로, 결측값을 삭제함으로써 발생하는 표본자료의 부족을 극복할 수 있는 Imputation 방식이 널리 사용되고 있다. 이러한 Imputation 방식을 구분하는 기준이나 시각도 연구자에 따라서는 다소 차이가 있다.

Little and Rubin(1987)은 ad-hoc 방식의 보정 기법과 모형기반(model-based)의 보정 기법으로 구분하면서, 특히 이식성이나 검증 가능성 등을 전제로 하는 융통성과 대규모 자료 분석을 위한 가용성(availability), 그리고 다른 통계적인 추론(statistical inference)이 가능하다는 점을 지적하면서 모형기반 보정 기법 사용을 적극적으로 권유하였다[20].

Pigott(2001)은 미국 인구조사국에서 실시한 2000년도 미국총인구조사에서 실제 인구보다 과소조사된 결과에 대해 어떻게 처리할 것인가를 두고 의회(U.S. Congress)와 연방 대법원(U.S. Supreme Court)에서 심한 논란(debate)에 휩싸인 것을 예로 들면서, 주요 원인으로 지목된 결측(무응답) 조사자료 즉, 불완전 자료(incomplete data)의 문제를 해소할 수 있는 실제적 대안은 신뢰성 있는 보정임을 주장하였다. 또한, 그는 Deletion 방식이나 과거 프로파일 이용법 등 ad-hoc 방식이 사용상 편의성은 있으나, 신뢰성 측면에서 한계가 명확하기 때문에 advanced 모형기반의 보정 기법을 사용할 것을 제안하기도 하였다.

Conklin and Scherer(2003)는 그들의 연구를 통해 보정 기법을 heuristic 기법과 통계적 기법으로 구분하여 각 기법의 대표적 방법이나 모형을 소개하고 보정 능력을 평가하였다. 이중 heuristic 기법으로는 Listwise Deletion과 Pairwise Deletion 방법, 과거 프로파일 이용법, 그리고 인근 검지기자료 이용법을, 통계적 기법으로는 EM과 DA 모형을 선정하여 이들 기법의 장·단점에 대한 심도있는 분석을 수행하였다. heuristic 기법은 이해가 쉽고 신속하게 적용할 수 있다는 점을 가장 큰 장점으로 들었으며, 통계적 기법은 이해 및 컴퓨팅(computing)이 부담스러우나, 보정 신뢰도가 가장 큰 장점임을 강조하면서 현재의 컴퓨팅 기술이나 기법의 발달로 인해 결측 자료에 대한 최선의 처리 방법은 advanced 통계적 기법을 이용한 보정임을 주장하였다.

<표 1> 결측자료 보정 기법 비교

구 분	기 법	개 념
ad-hoc 또는 heuristic	과거 프로파일	과거 동 시점의 프로파일 자료를 이용
	계수 이용	각종 계수를(HF, DF, MF) 이용
	보간법	전·후 가까운 관측값을 이용하여 보정
	평균대체법	관측값들의 평균으로 결측값을 보정
	Hot-Deck	결측값을 관측값(donor)에서 채택하는 기법
	Cohen	관측값의 분포를 양분하여 보정에 이용
	대체율법	보조변수를 보정에 이용
모형기반	회귀대체법	결측값을 회귀모형식으로부터 도출된 예측값으로 보정
	ARIMA	계열내 관측값의 상관관계를 이용하여 보정
알고리즘 기반	EM	관측값과 매개변수를 이용하여 반복적으로 결측값을 보정
	MI	여러차례의 시뮬레이션을 이용하여 결측값 추정
	DA	결측값과 모수들을 반복적으로 시뮬레이션하여 보정

III. 순환분포모형 개발

1. 이론적 배경

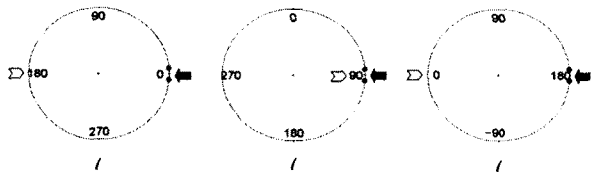
순환분포모형이란 관측 자료의 주기적인 순환성(periodic circularity)을 반영하여 통계적인 분포로 나타낸 모형을 의미하며, 생물학에서는 동물들의 이동방향, 지질학에서는 지구자기장(magnetic field)의 방향, 기상학에서는 바람의 방향 등 주로 각도나 방향으로 표현될 수 있는 분야에서 주로 활용된다. 또한, 의학에서 특정 질병에 따른 월별 사망률이나 시간의 경과에 따른 경제 시계열 자료 등도 시간의 주기를 방향의 자료로 전환하여 분석할 수 있다.

일반적으로 p-차원의 방향자료(directional data)는 일반성을 유지하고 원점을 중심으로 한, 크기가 1인 p-차원의 초평면(hypersphere)상의 점으로 나타낼 수 있다. 이 경우, $p=2$인 경우의 방향자료를 순환자료(circular data) 또는 각자료(angular data)라 하고, $p=3$인 경우의 방향자료를 구형자료(spherical data)라고 한다. 또한, 순환자료에서 방향성을 무시할 경우 즉, $\theta(0 \le \theta < 180^\circ)$와 $\theta+180^\circ$를 같게 취급하는 경우의 자료를 축자료(axial data) 또는 귀속자료(orientation data)라고

한다.

이와 같은 방향자료를 기존의 선형분석 기법을 이용할 경우 나타나는 문제점을 다음의 <그림 1>에 예시하였다. 이중 (a)의 경우, 동쪽(E)을 기준 방향(zero direction)으로 반시계 방향으로 측정된 5° 와 355° 자료의 평균방향은 0° 즉, 동쪽(E)이 분명하지만, 반면 산술평균은 (5° +355°)/2 = 180° 즉, 서쪽(W)을 나타낸다. 이러한 현상은 2개의 자료가 기준 방향에 대해 가까운 값임에도 불구하고 두 값의 차이가 350° (355° -5°)로 크게 나타나기 때문이며, 이는 자료가 순환성을 가지고 있기 때문이다. 선형분석이 가지는 또 다른 문제점은 동일한 자료에 대해 기준 방향이 바뀔 경우에 산술평균도 매번 변한다라는 점이다. (b)의 경우를 예로 들면, 북쪽(N)을 기준 방향으로 시계(clockwise) 방향으로 보면 상기의 자료는 85° 와 95° 의 값을 가진다. 이 경우는 산술평균이 90° 로 다시 동쪽(E)을 나타내게 된다. 마지막으로 (c)의 경우 다시 서쪽(W)을 기준 방향으로 시계방향으로 180° 까지, 반시계방향으로 -180° 까지 측정하면 상기 자료는 175° 와 -175° 가 되어 산술평균은 0° 가 되고 따라서 평균방향은 다시 서쪽(W)이 된다.

Σ : 산술평균 ⬅ : 평균방향



(a) (b) (c)

<그림 1> 선형분석(산술평균)과 순환분석(평균방향)의 차이

이와 같이 방향성을 가진 자료에 대한 분석에 일반 선형분석 기법을 사용할 경우 방향성(시간적 반복성)을 제대로 반영할 수 없기 때문에 통계적인 오류의 가능성이 있으며, 방향자료의 분석을 위한 통계적 도구를 사용하여야 한다. 연속형의 순환확률분포(circular probability distribution)의 순환확률변수 θ의 확률밀도함수 $f(\theta)$는 다음의 성질을 만족한다.

$$f(\theta) \geq 0, \quad \int_0^{2\pi} f(\theta)\, d\theta = 1$$

$$f(\theta) = f(\theta + k \cdot 2\pi), \quad k = 0, \pm 1, \pm 2, \cdots$$

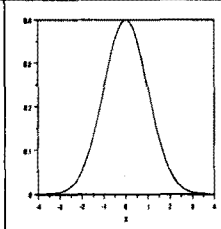
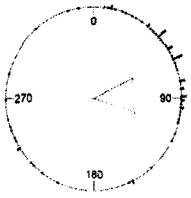
여기서, 성질 (c)는 순환분포 $f(\cdot)$의 주기성(periodicity)을 나타내는 것으로 선형분포모형과의 차이를 보여주고 있다. 위의 성질을 만족시키는 순환확률분포는 이미 알려진 선형분포들로부터 다양한 수학적 방식(wrapping, characterizing, offset, stereographic projection)을 통해 유도할 수 있다.

유도 방법 중에 Wrapping의 원리는 다음과 같다. 임의의 실수 공간에서 선형 확률변수를 X라 하고, 이의 밀도함수를 $f(x)$라 하면, 이때 선형확률변수 X는 다음의 Modulo 변환 $\theta = X \pmod{2\pi}$을 통해 순환확률변수로 변환된다. 이때, 순환확률변수 θ의 확률밀도함수는 다음과 같이 주어진다.

$$g(\theta) = \sum_{k=-\infty}^{\infty} f(\theta + 2\pi k), \quad 0 \leq \theta < 2\pi$$

이와 같이 Wrapping의 과정을 통하여 유도된 순환분포를 Wrapped 순환분포라고 한다. 겹친 정규(Wrapped Normal) 및 겹친코쉬(Wrapped Cauchy) 분포가 대표적인 Wrapped 순환분포이다.

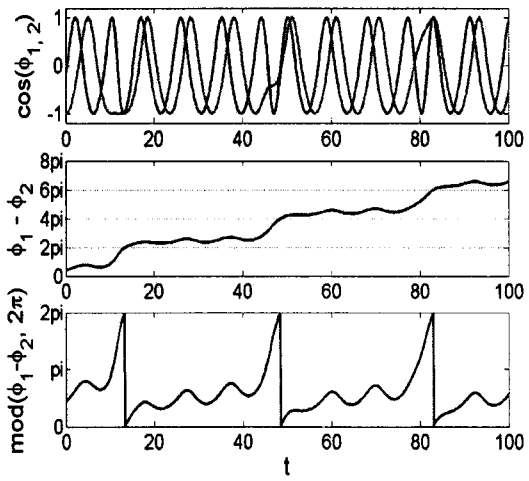
<표 2> 선형확률분포 및 순환확률분포 비교

구분	선형 확률 분포	순환 확률 분포
개념	· 연속형 2차원 선형 자료를 표현하는 확률분포함수 · 2차원 평면(plane) 상에 표현	· 연속형 2차원 방향 자료를 표현하는 확률분포함수 · 원점(0,0)을 중심으로 크기가 1인 초평면 상에 표현
형태	 · 선형분포를 따르는 연속형 확률변수 X의 일반적 범위는 $-\infty \leq X \leq \infty$	 · 순환분포를 따르는 연속형 확률변수 Θ의 범위는 $0 \leq \Theta < 2\pi$
모수	· 평균(μ) · 분산(σ^2)	· 평균방향(μ) · 순환분산(V_0) · 집중모수(κ)
종류	· 정규분포 · 코시분포 · 지수분포 · 이중지수분포	· von Mises 분포 · 겹친정규분포 · 겹친코시분포

2. 순환분포모형 적용 사례 분석

Ravindran(2002)은 우선 Embedding 접근법 및 Intrinsic 접근법, Wrapping 접근법 등 순환자료를 모델링하기 위한 통계적인 접근법에 대하여 분석하였으며, Jander(1957)의 개미 이동방향 자료와 미국 위스콘신주 밀워키 기상대에서 1975년 4월 18일부터 6월 29일까지 4일 간격으로 수집한 풍향(wind direction) 및 오존 농도 자료를 대상으로 겹친정규, 겹친코시, 겹친이중지수 순환분포모형을 개발하였다. 이들 순환확률분포의 모수(parameter) 추정은 DA 알고리즘을 사용하였다. 구축된 모형의 적합도(goodness-of-fit)를 GGC(Gelfand and Ghosh Criteria)를 이용하여 검증하였으며, 가장 낮은 값을 나타낸 겹친코시분포가 잘 적합된 것으로 보고하였다[27]. 또한, 전술한 밀워키 기상대의 풍향 및 오존 농도 자료를 기반으로 하여 순환회귀모형(circular regression model)을 개발하였는데, DA 알고리즘으로 모형의 모수를 추정하였다. 겹친정규분포가 가장 낮은 GGC를 나타냈으며, 그 다음이 겹친이중지수, 겹친코시 분포의 순으로 나타난 것으로 보고하였다.

순환분포모형과 관련한 국내의 연구로 오영남(2006)이 있다. 신호간의 위상차(phase difference) 방향자료를 대상으로 순환분포모형을 개발하였다. 그는 X축을 시간, Y축을 진폭(amplitude)이라고 했을 경우에 동일한 주파수(frequency)를 가지는 두 신호간의 위상차가 반복적인 주기성을 가지는 Sine 곡선(sinusoidal curve)의 형태로 나타나는데 착안하여 이를 순환확률분포를 이용하여 모형으로 개발하였다.



<그림 2> 위상차 $\phi_1 - \phi_2$ 및 $\phi_1 - \phi_2 (\mathrm{mod}\, 2\pi)$의 시간에 따른 변화

이와 같이 도출된 위상차 자료에 대하여 대표적 단봉형 대칭 순환분포모형인 von Mises 분포와 겹친정규 분포, 겹친코시 분포에 대하여 EM 알고리즘을 이용하여 각 모형의 모수에 대한 최대우도 추정치(Maximum Likelihood Estimator, MLE)를 도출하였는데, 겹친코시 분포가 가장 잘 적합된 것으로 보고하였다.

3. 자료의 수집 및 분석

본 연구의 분석 대상인 교통량(traffic volume)은 도로의 한 지점 또는 그 단면을 단위시간 동안 통과한 차량의 수를 의미한다. 도로를 통과하는 단위 시간당의 교통량은 도로 시설물의 효용 척도(MOE of utilities)로 사용되며, 다른 지점과 상대적 비교를 통해 각 도로 구간의 역할을 추정 또는 평가할 수 있는 지표로도 사용된다. 또한, 교통량 자료는 도로 계획 및 설계와 도로 운영 등에 폭넓게 이용되며, 교통 계획과 관리 계획 수립과 관련된 여러 분야에서 활용빈도가 높은 중요한 자료 중의 하나이다.

현재 일반국도의 교통량 변동을 시계열적으로 파악하기 위해 약 440지점에 교통량 상시조사장비(Permanent Traffic Count, PTC)가 설치 운영되고 있다. 1년 365일 상시 조사된 교통량 자료로 해당 지점의 시간대별, 일별, 월별, 계절별 등 시계열적 특성을 파악할 수 있다.

순환분포모형의 여러 교통 특성에 대한 설명력(보정 능력)을 높이기 위해서는 가급적 많은 그룹을 대상으로 모형을 구축해야 하나, 모형의 가용성을 감안하면 무한정으로 그룹수를 증가시킬 수는 없다. 따라서, Conklin, et al.(2003)의 주장대로 모형의 설명력과 가용성간 적절한 타협점을 모색하였다.

우선, 대상 자료인 시간교통량(hourly volume) 및 일교통량(daily volume) 자료를 활용하면서, 4~9개의 그룹수 제약조건을 만족시킬 수 있는 그룹핑 방안을 모색하였다. 그 결과 첫 번째, 1년 8,784개의 시간 교통량 자료를 이용하여 산출할 수 있는 교통 특성 지표인 설계시간계수(K)를 통하여 상시조사지점을 도시부, 지방부, 관광부 3가지로 구분하였다.

두 번째, 일 교통량을 이용하여 주중(월~금)과 주말(토, 일)의 교통 특성이 상이한 점에 착안하여 주중 5일간의 일 교통량의 평균과 주말 2일간의 일 교통량 평균을 구한 후, 이를 비교하여 3가지 카테고리로 구분하였다. 시간 교통량 특성을 기준

으로 상시조사지점을 3가지로 나누고, 이를 다시 일 교통량 특성을 기준으로 3가지로 구분하면 3 × 3 = 9개 그룹으로 그룹핑이 된다. 다음의 <표 3>는 주중(월~금)과 주말(토, 일) 요일의 평균 일 교통량을 그 차이가 5%가 되도록 하여 그룹핑한 결과를 나타낸 것이다.

<표 3> 주중 및 주말 요일의 평균 일 교통량 5% 차이 기준의 그룹핑 결과

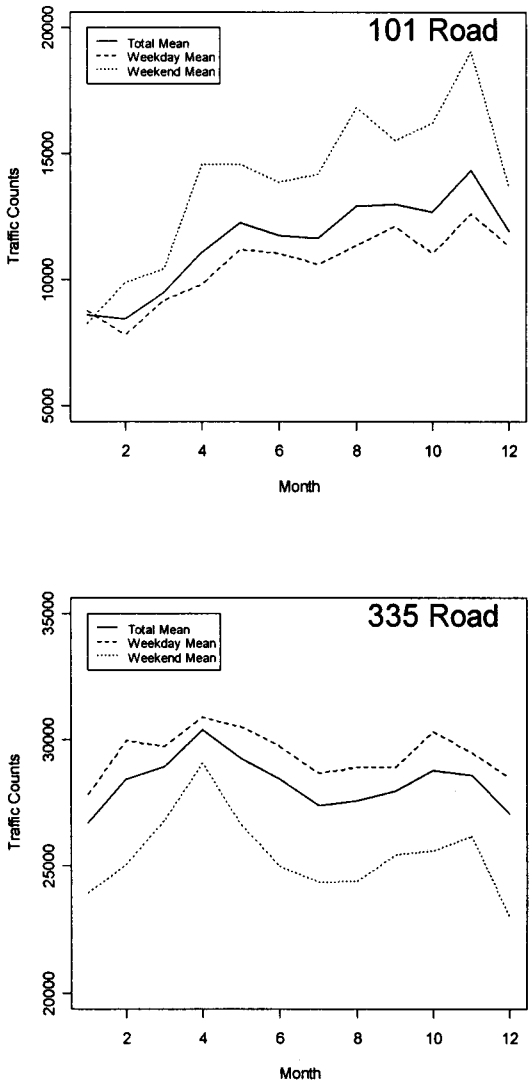
구분	주중 ≒ 주말 평균일교통량	주중 > 주말 평균일교통량	주중 < 주말 평균일교통량
도시부	그룹 1 (36지점)	그룹 2 (78지점)	그룹 3 (46지점)
지방부	그룹 4 (21지점)	그룹 5 (14지점)	그룹 6 (132지점)
관광부	그룹 7 (0지점)	그룹 8 (0지점)	그룹 9 (26지점)

상기와 같이 전체 교통량 상시조사지점을 9개의 그룹으로 그룹핑하여 이중 그룹 2(도시부)와 그룹 6(지방부)을 대상으로 모형 개발을 위한 대표지점을 선정하였다. <표 5>는 대표지점 내역을 나타낸 것이다.

<표 5> 도시부 및 지방부 대표지점 내역

구분	주중vs.주말 평균 일교통량	그룹	지점번호	주소	도로구간	호선	2004년 AADT
도시부	주중>주말	그룹2	335	경기 용인 모현 왕산	삼계리 ~ 오포면	45	28,297
지방부	주중<주말	그룹6	101	경기 평택 현덕 인광	현덕면 ~ 안중면	39	11,537

그룹핑 분석을 통해 335 도로는 도시부 도로인 그룹 2에 속하며, 101 도로는 지방부 도로인 그룹 6에 속하는 것으로 분류되었다. 지방부 도로인 101 도로는 주중의 평균 일교통량이 주말의 평균 일교통량보다 적은 양상을 보여주고 있으나, 도시부 도로인 335 도로는 주중 평균 일교통량이 주말 평균 일교통량보다 많은 양상을 보여주고 있다. 또한, 지방부 도로인 101 도로의 경우는 월요일부터 금요일까지 거의 비슷한 패턴을 보여주고 있으나, 토요일과 일요일은 다소 다른 양상을 나타내는 것으로 분석되었다. 그러나 도시부 도로인 335 도로는 월요일부터 토요일까지는 비슷한 패턴을 보여주지만, 일요일에만 다소 교통량이 적은 것으로 나타나고 있다. 이러한 현상은 시기적으로 2004년도가 아직 주 5일제 근무가 전면적으로 확산되지 않은 시기였기 때문인 것으로 판단된다. 특히, 일요일 교통량의 감소는 도시부 도로의 전형적인 특성을 나타내는 것으로 볼 수 있다. <그림 3>은 지방부 도로인 101 도로와 도시부 도로인 335 도로의 일 교통량의 연간 변화를 나타낸 것이다. 그림에서도 볼 수 있듯이 101 도로는 지방부 도로의 일반적인 특성을 나타내고 있으며, 335 도로는 도시부 도로의 일반적 특성을 나타낸다고 할 수 있다.



<그림 3> 101도로(지방부)와 335도로(도시부)의 일교통량 연간 변화

4. 순환분포모형 개발

순환분포모형(circular probability distribution model)을 개발하기 위해 R(R-2.2.1) 통계 컴퓨팅

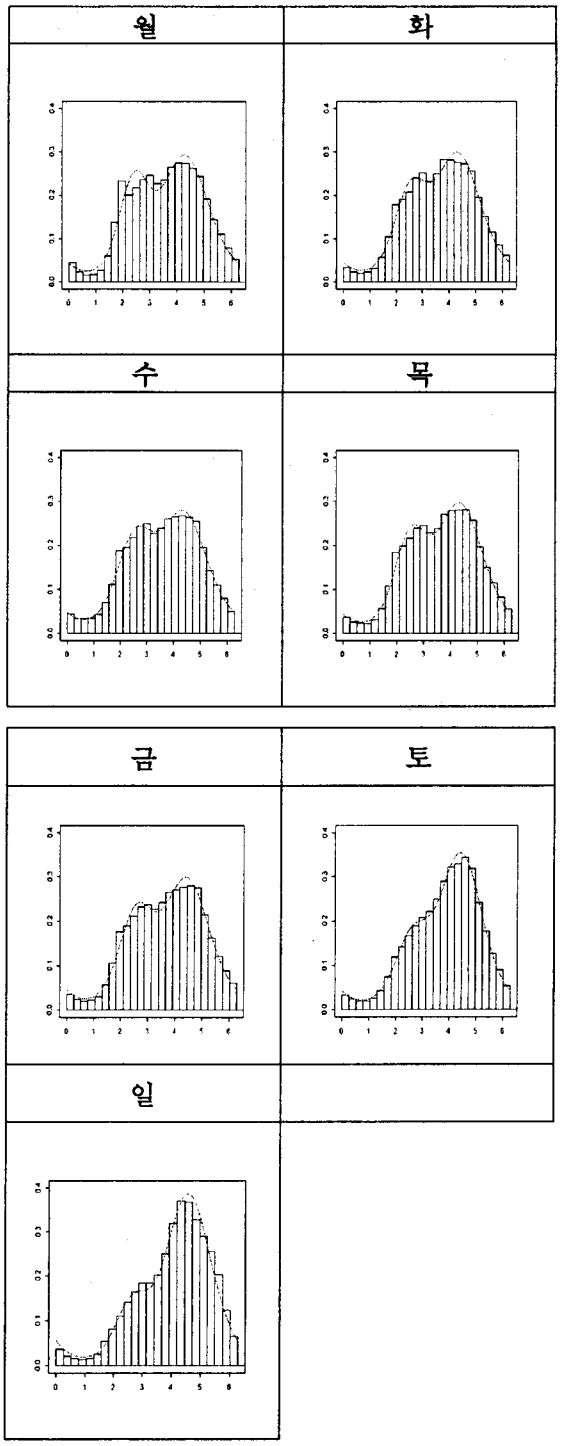
(statistical computing) 전문 언어가 사용되었다. 혼합형 분포모형인 경우에는 직접 산출할 수 있는 라이브러리가 아직 제공되지 않아 선형분포에서 순환분포로 전환하는 Wrapping에 관한 수학적인 이론을 바탕으로 이를 프로그래밍하였다.

본 연구에서는 도시부와 지방부 도로에 대해 각각 7개 요일별로 순환분포모형을 개발하였다. 요일별 순환분포모형은 각 요일의 시간교통량에 대해 순환성을 가지고 있는데, 이러한 순환성은 반드시 시간적인 연속성을 전제로 하지 않으므로 각 요일별 순환분포모형은 서로 독립적으로 분포한다(i.i.d. : independent and identically distributed). 도시부와 지방부 도로에 대한 요일별 순환분포모형은 모두 쌍봉형(bimodal)의 혼합 von Mises 분포를 이용하여 최적 적합되었다. 현재 하나의 봉우리(peak)을 가지는 단봉형(unimodal)의 대칭 순환분포(von Mises분포, 겹친정규분포, 겹친코시분포 등)에 대한 모수의 추정은 최대경사법이나 Newton-Raphson Method, EM 알고리즘을 이용하면 가능하나, 2개 이상의 봉우리를 가지는 다봉형(multimodal) 순환분포의 경우에는 분포가 가지는 확률적인 특성으로 인해 혼합 von Mises 분포만이 EM 알고리즘의 적용이 가능하며, 여타 혼합 겹친정규분포나 혼합 겹친코시분포에 대한 효율적인 모수 추정 방법은 아직 이론적인 연구가 진행 중에 있다.

모수 추정 결과에서 혼합률 모수는 혼합 von Mises 분포에서 각 봉우리에 대한 가중 확률(weighted probability)의 크기를 표현한 것이며, 각 혼합률 모수의 합은 언제나 1로 나타난다. <표 5>은 101 도로(지방부)의 시간교통량 자료를 기반으로 구축한 요일별 순환분포모형을 나타낸 것이며, <그림 4>는 이를 가로축을 라디안으로 하는 히스토그램으로 나타낸 것이다.

<표 5> 101 도로에 대한 요일별 순환분포모형

요일	$\hat{\alpha}_1 vM(\hat{\mu}_1, \hat{\kappa}_1) + \hat{\alpha}_2 vM(\hat{\mu}_2, \hat{\kappa}_2)$
월	$.29 \cdot vM(2.40,3.43) + 0.71 \cdot vM(4.30,1.38$
화	$.30 \cdot vM(2.52,2.76) + 0.70 \cdot vM(4.34,1.41$
수	$.34 \cdot vM(2.53,2.33) + 0.66 \cdot vM(4.38,1.34$
목	$.30 \cdot vM(2.50,2.96) + 0.70 \cdot vM(4.35,1.41$
금	$.33 \cdot vM(2.56,2.63) + 0.68 \cdot vM(4.44,1.50$
토	$.25 \cdot vM(2.61,4.44) + 0.75 \cdot vM(4.44,1.68$
일	$.18 \cdot vM(2.62,3.27) + 0.82 \cdot vM(4.57,1.69$



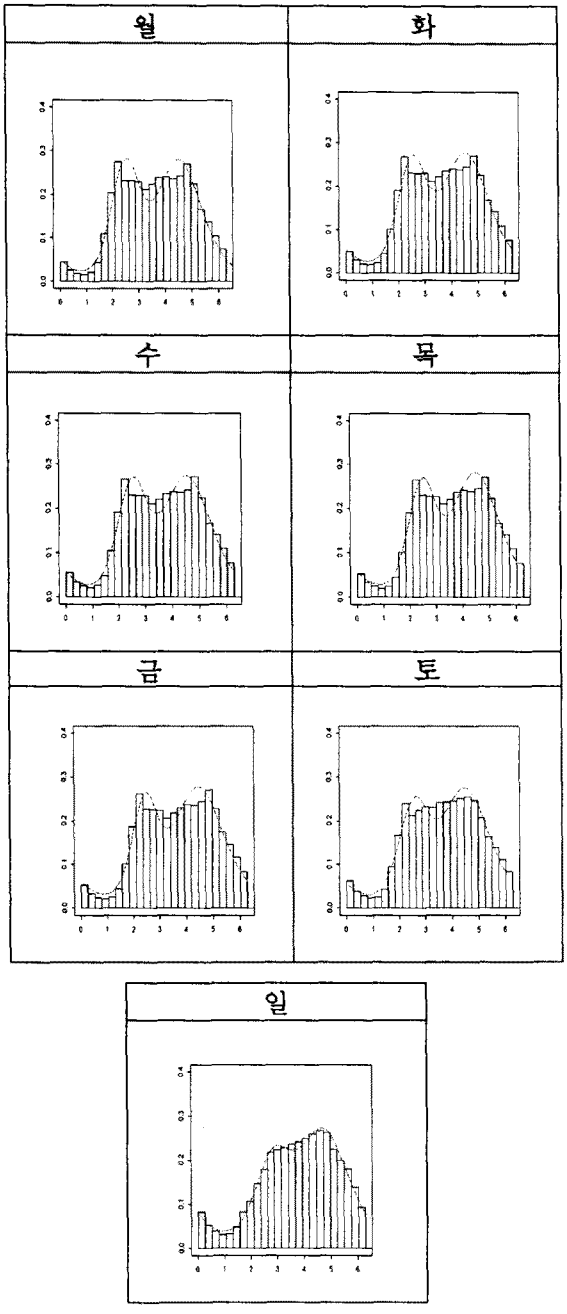
<그림 4> 101 도로의 요일별 순환분포모형

또한, 전형적인 도시부 도로의 특성을 가지는 그룹 2의 대표 지점인 335 도로의 경우는 일요일을 제외한 요일들이 서로 비슷한 시간교통량의 패턴을 보여주고 있다. 특히, 도시부 도로의 특성인 출근 시간대와 퇴근 시간대에 교통량의 집중 현상이 뚜렷이 나타나고 있다. 다음의 <표 6>은 335 도로에 대한 시간교통량 자료를 기반으로 개발한 요일별 순환분포모형의 함수식을 나타낸 것이다.

<표 6> 335 도로에 대한 요일별 순환분포모형

요일	$\hat{\alpha}_1 vM(\hat{\mu}_1, \hat{\kappa}_1) + \hat{\alpha}_2 vM(\hat{\mu}_2, \hat{\kappa}_2)$
월	$0.34 \cdot vM(2.49,3.56) + 0.66 \cdot vM(4.52,1.45)$
화	$0.33 \cdot vM(2.50,3.53) + 0.67 \cdot vM(4.52,1.37)$
수	$0.33 \cdot vM(2.49,3.43) + 0.67 \cdot vM(4.53,1.34)$
목	$0.29 \cdot vM(2.44,3.99) + 0.71 \cdot vM(4.47,1.30)$
금	$0.26 \cdot vM(2.42,4.49) + 0.74 \cdot vM(4.44,1.19)$
토	$0.30 \cdot vM(2.53,3.27) + 0.70 \cdot vM(4.49,1.28)$
일	$0.30 \cdot vM(2.82,2.35) + 0.70 \cdot vM(4.72,1.21)$

또한, 다음 <그림 5>는 335 도로의 시간교통량 자료를 기반으로 구축한 요일별 순환분포모형을 히스토그램으로 표현한 것이다.



<그림 5> 335 도로의 요일별 순환분포모형

IV. 모형의 평가

1. 평가 방법

그룹 2(도시부)의 335 도로 및 그룹 6(지방부) 101 도로의 2004년과 2005년 시간교통량 자료를 기준 자료(reference data)로 하여 결측률(10~90%, Δ=10) 및 단일 결측(결측률에 따라 무작위로 결측 생성) 및 연속 결측(처음 10%만 무작위 결측시키고 20%부터는 연속하는 2개를 결측, 30%는 3개, 40%는 4개 등으로 해당 결측률을 만족할 때까지 결측을 생성)에 따라 시간교통량 자료를 임의 결측시킨다. 각 모형을 이용하여 상기와 같이 결측시킨 자료를 보정하고, 이를 기준 자료(완전 자료)와의 oMAPE, MAPE 및 oRMSE, RMSE을 도출하여 성능을 평가하였다. 비교 대상 모형중 가변수 회귀모형을 월 효과, 요일 효과, 시간 효과 등을 주 효과(main effects)로 하고 이들 주 효과의 2차 교호 효과(interaction effects)를 포함시킨 가변수 회귀모형 1과 월 효과, 요일 효과, 시간 효과 등 주 효과만을 포함시킨 가변수 회귀모형 2로 나누어 평가하였다. 또한, 도시부 및 지방부 도로의 2005년 자료를 이용하여 모형의 예측력도 평가하였다. 그리고, 계절 ARIMA 모형의 경우는 시계열의 사전 조정(차분)과 모형 식별 등 과정을 거쳐 과도 적합(over fitting)된 형태로 모형의 모수가 도출되었으나, 모형 진단(diagnostics)에서 여러 차례의 잔차간 독립성 가정을 위반하는 등 최적합 모형 도출에 어려움이 있어 모형의 비교 평가에서는 제외하였다. 또한, 시나리오에 따른 모형의 비교 평가와는 별도로 순환분포모형의 비용 효과성을 검증하기 위하여 도시부 및 지방부 도로의 전체 시간교통량 자료(2004년 8,784개 및 2005년 8,760개)를 토대로 이를 전체 자료량 대비 1/2, 1/3, 1/4, 1/8, 1/16, 1/32, 1/53 등의 비율로 줄여가면서 혼합 von Mises 분포의 모수를 도출하였으며, 기반 자료량에 따른 결측 보정 신뢰도 변화를 확인하였다. 이를 통하여 Gold, et al.(2001)의 주장처럼 순환분포모형이 시간교통량 자료 결측 보정시 비용 효과적인 방법이라는 사실을 확인할 수 있었다.

2. 시나리오에 의한 비교 평가

평가 결과를 살펴보면, 가변수 회귀모형 1과 2, 순환분포모형 모두 지방부인 101 도로보다 도시부인 335 도로의 oMAPE와 oRMSE가 상대적으로

작은 것으로 나타났다. 이는 지방부 도로보다 도시부 도로의 교통 양상이 다소 안정적(stable)이라는 사실에 기인하는 것으로 판단된다. 구체적으로, 101 도로에서의 시간교통량에 대한 결측보정 성능은 순환분포모형을 이용한 보정 방법이 복잡한 모형인 가변수 회귀모형 1과 가변수 회귀모형 2보다 더 좋은 성능을 보였으나, 335 도로에서는 가변수 회귀모형 1이 순환분포모형보다 보정 성능이 더 우수한 것으로 나타났으며, 가변수 회귀모형 2는 순환분포모형보다 저조한 성능을 나타냈다.

결측 양상에 따른 결과를 살펴보면, 결측 비율이 같을 경우에 단일 결측보다 연속 결측의 oMAPE와 oRMSE가 낮은 것으로 나타났다. 이러한 결과는 교통량 양상이 반복적 주기성이 강하기 때문에 나타나는 현상으로 분석되며, 이로 인해 시간적으로 이웃한 자료간에 상관성이 높아 자료를 단일 결측시킬 때보다 연속 결측시킬 때 보정 성능이 좋아지는 것으로 분석된다. 또한, 결측 비율에 따른 결측 양상별 oMAPE 변화 추이를 보면 단일 결측인 경우 Ni, et al.(2005)의 연구에서처럼 선형적으로 단조 증가하는 현상을 나타내지만 연속 결측인 경우에는 결측 비율이 커질수록 그 증가율이 점점 감소하는 양상을 보였다. 이는 overall 지표(oMAPE 및 oRMSE)가 결측된 자료의 수량(n)에 대한 평균 오차를 도출하는 것이 아니라 전체 자료의 수량(N)에 대한 평균 오차를 도출한다는 점도 하나의 원인으로 분석된다. 그리고, oMAPE 10%는 FHWA의 Traffic Monitoring Guide에서 제안하고 있는 교통통계 자료의 신뢰도 수준이 90%임을 감안하면, 정책적 시사점을 가진다고 할 수 있다[12].

또한, 2004년 시간교통량을 기반으로 구축한 순환분포모형의 예측 능력을 검증하기 위해 같은 지점의 2005년 시간교통량 자료를 이용한 결측 보정을 수행하였으며, 이를 oMAPE, oRMSE, MAPE, RMSE 지표를 통해 평가하였다. 평가대상으로는 가변수 회귀모형 2를 선정하였으며, 2004년 대비 2005년의 교통량 성장률(growth factor)을 1로 가정하였다. 이식성과 장기예측 측면에서도 순환분포모형이 가변수 회귀모형 2보다 좋은 성능을 보이는 것으로 나타났다. 단일(무작위) 결측의 경우 101 도로에서는 가변수 회귀모형 2가 결측비율 30%에서 oMAPE가 약 11% 정도로 나타났으나, 순환분포모형은 결측비율 40%에서 oMAPE가 9% 정도로 나타났다. 또한, 335 도로의 경우 가변수 회귀모형 2는 결측비율 50%에서 oMAPE가 9～10% 정도로 나타났으나, 순환분포모형은 결측비율 60%에서 oMAPE가 10～11% 정도인 것으로 나타났다. 연속 결측의 경우에는 101 도로에서는 가변수 회귀모형 2가 결측비율 20%에서 oMAPE가 10% 정도를 나타냈으며, 순환분포모형은 결측비율 50%에서 oMAPE가 10% 정도를 나타냈다. 335 도로에서는 가변수 회귀모형 2가 결측비율 70%에서 oMAPE가 10～11%로 나타났으나, 순환분포모형은 결측비율 80%에서 oMAPE가 10～11% 정도로 나타났다. 이러한 결과를 통해 순환분포모형을 이용한 결측 보정은 기존 방법보다 이식성이 뛰어나며 장기 보정(예측)에도 적합하다는 사실을 확인하였다. 다음의 <표 7> 및 <표 8>는 도로별 단일 결측시 각 모형의 결측율별 보정효과(oMAPE/ oRMSE)를 비교한 것이다.

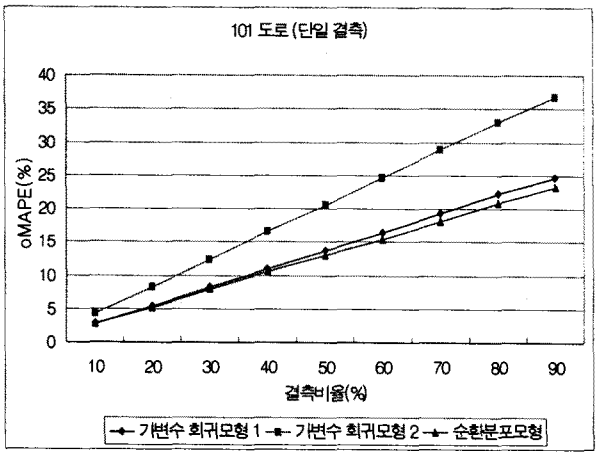
<표 7> 101 도로 결측 보정효과(단일 결측)

결측비율(%)	101 도로					
	가변수 회귀1		가변수 회귀2		순환분포모형	
	oMAPE (%)	oRMSE	oMAPE (%)	oRMSE	oMAPE (%)	oRMSE
10	2.779	50.475	4.291	63.110	2.699	55.681
20	5.272	67.206	8.164	84.453	5.146	73.307
30	8.162	83.794	12.232	105.019	7.778	90.895
40	11.020	95.069	16.531	119.946	10.557	103.377
50	13.693	104.106	20.430	130.097	13.009	112.433
60	16.397	112.114	24.594	146.706	15.365	122.160
70	19.256	120.773	28.894	157.736	18.145	131.711
80	22.145	130.732	32.961	169.639	20.777	142.253
90	24.639	138.057	36.834	178.681	23.224	150.031

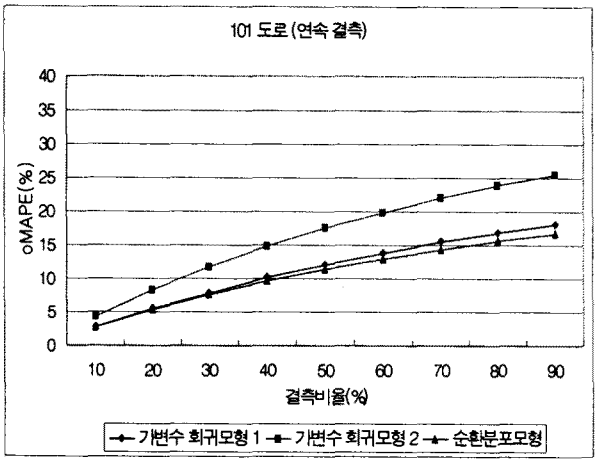
<표 8> 335 도로 결측 보정효과(단일 결측)

결측비율(%)	335 도로					
	가변수 회귀1		가변수 회귀2		순환분포모형	
	oMAPE (%)	oRMSE	oMAPE (%)	oRMSE	oMAPE (%)	oRMSE
10	0.895	36.821	1.725	52.111	1.701	58.146
20	1.784	49.579	3.563	74.019	3.511	82.362
30	2.662	59.690	5.334	92.578	5.168	99.260
40	3.631	70.694	7.329	108.878	7.080	117.180
50	4.509	79.110	9.012	119.398	8.846	131.097
60	5.406	88.321	10.625	130.184	10.564	149.804
70	6.375	96.101	12.621	142.237	12.475	162.235
80	7.253	101.687	14.391	152.720	14.133	171.433
90	8.141	106.970	16.229	161.514	15.943	181.085

다음 <그림 6>과 <그림 7>은 101 도로의 단일 결측 및 연속 결측시 모형별 보정 효과를 그래프로 나타낸 것이다.



<그림 6> 단일 결측시 보정효과(101도로)



<그림 7> 연속 결측시 보정효과(101도로)

V. 결론 및 향후 연구

본 연구를 통해 Batschelet(1981) 이후 교통학 분야에 새롭게 도입되는 순환분포모형의 이론적 배경, 특징등에 관해 개관하였다. 교통 특성별 순환분포모형의 개발을 위해 353 지점의 교통량 자료를 대상으로 9개의 그룹으로 그룹핑하여 이중 도시부와 지방부에서 임의 선정된 대표 지점의 시간 교통량 자료를 기반으로 모형을 개발하였다.

모형의 모수인 평균방향, 집중도, 혼합율 모수는 최우 추정기법인 EM 알고리즘을 이용하여 추정하였으며, 개발 모형에 대한 χ^2 적합도 검정을 통해 통계적 유의성도 검정하였다. 순환분포모형의 결측 조건별 보정 성능을 분석한 결과, 연속형의 확률분포모형인 순환분포모형이 관측값의 추세를 이용하는 가변수 회귀모형에 비해 대체로 우수한 것으로 나타났다. 또한, 순환분포모형은 비교적 적은 수의 모수를 이용, 시간교통량 현상을 잘 표현할 수 있을 뿐만 아니라, 기반자료량의 증감이나 다양한 결측양상에 대하여 robust한 확률분포의 특성도 보여 순환분포모형을 이용한 교통량 자료 결측 보정 방법이 매우 비용-효과적이라는 사실을 확인하였다. 따라서, 본 연구를 통해 분석 및 확인된 여러 연구 결과들은 향후 차량검지기 증설에 따른 자료결측 문제의 심화 및 이슈화 등을 전제로 할 때, 매우 희망적인 것으로 판단된다.

참고 문헌

Batschelet, E.(1981), Circular Statistics in Biology, Academic Press, London.

Box, G. and Jenkins, J.(1970), Time Series Analysis : Forecasting and Control, Holden Day, San Francisco.

Dempster, A.P., Laird, N.M., Rubin, D.B.(1977), Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm, Journal of Royal Statistical Society, B39 pp.1~38.

Little, R.A., Rubin, D.B.(1987), Statistical Analysis with Missing Data, John Wiley & Sons, New York.

Ni, D., Leonard, J.D., Guin, A., Feng, C.(2005), Multiple Imputation Scheme for Overcoming the Missing Values and Variability Issues in ITS Data, Journal of Transportation Engineering, vol.131, pp.931~938.

Pigott, T.D.(2001), A Review of Methods for Missing Data, Educational Research and Evaluation vol.4 pp.353~383.

오영남(2006), 방향자료에 대한 분석과 모형화, 석사학위논문, 충북대학교.

장진환, 백남철(2005), 교통량 결측 자료 대체기법 연구, 2005년도 학술발표회 논문집, 대한토목학회.