비모수 회귀분석을 이용한 실시간 통행시간예측에 관한 연구

Travel Time Prediction Using Nonparametric Regression Analysis

박 희 원*, 박 창 호**, 전 경 수***, 이 성 모****

(* 서울대학교 도시공학과 석사과정)

(** 서울대학교 도시공학과 교수)

(*** 서울대학교 도시공학과 교수)

(**** 서울대학교 도시공학과)

목 차

I. 서론
1. 연구의 배경 및 목적
II. 기존 연구의 고찰
1. 과거 프로파일 접근법
2. 시계열 모형
3. 신경망 기술을 이용한 방법
4. 비모수 회귀분석
5. 동적 통행배정 모형을 이용한 방법
6. Traffic Simulation 모형을 이용한 방법
III. 예측모형의 설계
1. 비모수 회귀분석
2. 최근린 분석
3. 알고리즘 구성
4. 데이터의 잡음제거
5. 통행시간 추정
IV. 평가 및 결론
참고문헌

I.서론

1. 연구의 배경 및 목적

각종 ITS(Intelligent Transport System) 관련연구에서 신뢰성 높은 통행시간 예측은 효과적인 교통관리나 신속한 여행자 정보제공과 같은 분야에서 매우 중요한 역할을 한다. 이를 위해 현재까지 다양한 예측 기법들이 개발되어져 왔으나 단순한 기법만으로는 미래의 통행시간을 예측하는 것은 기술적으로 많은 어려움이 따르며, 이로 인해 여러 가지 예측기법들의 장점을 취합해 종합적으로 활용하는 기술이 필요하였다.

따라서 본 연구개발은 신뢰할 수 있는 차량 통행시간의 예측을 위해 기존 통행시간 예측 기법들을 면밀히 비교·분석하여 동적이고 종합적인 예측모형을 개발하고 이를 통해 예측된 단기의 미래 통행시간을 운전자들에게 제공함으로써 출발 직전의 여행자들에게는 보다 빠른 경로의 선택을 가능하게 하고, 혼잡한 도로망 위의 운전자들에게는 대체 경로의 선택을 가능하게 하는, 통행시간 예측 시스템을 개발하는데 그 목적이 있다.

II. 기존 연구의 고찰

1. 과거 프로파일 접근법

과거 프로파일 접근법은 통행량 및 통행시간에 대한 과거 프로파일이 얻어질 수 있다는 가정을 전제로 한다. 이 접근법은 장래의 통행량 예측을 위해 단순히 과거 통행량의 평균값만을 이용하며 통행류의 주기적 특성만을 고려한다. 장점은 수행이 쉽다는 점과 빠른 실행속도에 있다. 단점은 정적 접근이라는 점이다. 예를 들어 교통사고 발생시 이에 대해 반응할 방법이 없다(Smith and Demetsky, 1997).

이 접근법은 AUTOGUIDE, LISB 등의 통행자 정보 시스템 뿐만 아니라 도시교통통제 시스템(UTCS) 등 다양한 분야에서 검증되었다.

2 시계열 모형

시계열자료는 시간 순서대로 배열된 통계 관측치의 집합이다. 관측된 통행량 계열은 실제값과 잡음 두 부분으로 구성된다. 따라서 외부의 영향에 기인하는 잡음을 소거하는 것이 시계열 모형의 핵심이다. 잡음을 소거하기 위해 자기회기(AR)와 이동평균(MA) 모형이 고안되었으며

그러한 기초 모형들을 이용하면 일반적인 이산 고정(Discrete Stationary) 시계열은 ARMA 모형으로 정의 될 수 있다.

박스-젠킨스(Box-Jenkins) 모형은 ARMA에 근거한다. 이 모형은 ARIMA(Autoregressive Integrated Moving Average) 모형이라 불리며 이전 모형에서 이용되던 고정된 초기패턴을 요구하지 않는다.

ARIMA 모형은 UTCS 및 고속도로의 통행량 예측에 주로 적용되어 왔다(Smith and Demetsky, 1997). 또한 최근의 시계열법인 칼만 필터링(Kalman Filtering)은 시계열 분석에서 추정오차를 최소화하는데 적용되어 왔으며 (Ben-Akiva et al., 1995), VARMA(Vector Autoregressive Moving Average) 모형은 도로망을 구간화하고 구간 도로망간의 상관관계를 고려하는 다변량 시계열분석에 응용되었다.

3. 신경망 기술을 이용한 방법

교통류 패턴에서 현재 통행시간을 추정하는 등의 복잡한 문제를 푸는 데에는 학습능력을 가진 이 접근법이 적합할 수 있으며 (Palacharla and Nelson, 1995), 최근에 교통류 모형, 교통신호 제어, 교통 계획 등의 응용 분야에서 각광을 받아왔다(Smith and Demetsky, 1997). 그러나 분석하고자 하는 교통망이나 데이터의 구조가 복잡해질수록 더욱 복잡한 구조의 신경망을 필요로 하고, 이는 학습에 필요한 연산시간을 기하급수적으로 늘어나게 하는 단점이 있다.

국내에서와 마찬가지로 역전파 기법 및 다층학습(Multi-Layer Learning)기법 등을 이용한 신경망 기술의 이용이 많다.

국외에서도 신경망은 이용자가 알 수 없는 블랙박스에서의 예측이 단점으로 지적되어왔으며 이를 보완하기 위하여 신경망의 체계적인 구축이나 학습방법에 많은 연구가 이루어져 왔다.

현재까지 시도되어온 신경망의 트레이닝 과정은 매우 복잡한 연산을 거치며, 이를 실시간 데이터에 응용하기에는 많은 어려움이 따른다.

4. 비모수 회귀분석

비모수 회귀는 입력치의 상황이 예측시기의 시스템 상황과 유사한 과거사례 집단을 파악하는 동적 군집모형(Dynamic Clustering Model) 혹은 패턴인식(Pattern Recognition) 문제로 간주된다.

이 방법은 예측시기 이전에 여러 집단을 고려하지 않고 현재 입력 상황과 유사한 과거 사례(혹은 근접치)의 집단만을 고려하기 때문에 동적인 특성을 지닌다.

근접치를 찾아내는 과정의 복잡성 때문에 통행시간 예측에는 많이 쓰여지지는 않았으나 탐색방법의 최적화를 통해 비교적 짧은 시간 내에 예측을 수행할 수 있다.

탐색방법의 최적화를 위해 고난도의 데이터베이스 기술과 컴퓨터 프로그래밍 기법을 필요로 한다.

5. 동적 통행배정 (Dynamic Traffic Assignment) 모형을 이용한 방법

동적 통행배정 모형들은 시간에 따라 달라지는 교통수요의 변화를 모형이 규정하는 시간 범위 안에서 복잡한 수학식의 연산을 통하여 구하기 위해 개발되었다.

동적 통행배정 모형의 응용에 있어 교통수요의 추정은 매우 중요하고 시간대에 따라 변화하는 교통수요를 최대한 반영하는 동적 O-D(Origin-Destination)표를 작성하는 것이 가장 큰 문제점으로 지적되었다.

현재까지 실험된 동적 통행배정모형으로는 정적인 이용자 평형(User Equilibrium)이나 시스템 최적(System Optimal) 모형들에 기초하여 개발된 것이 주류를 이루고 있으나 일부 학자들에 의해 인공지능(Artificial Intelligence)의 유전자 이론(Genetic Algorithm) 등을 이용한 새로운 모형들도 시도되고 있다. 정적인 모형과는 달리 동적인 모형은 특히 첨두 통행시간동안 일어날 수 있는 동적인 상황들(예를 들면 혼잡이나 정체)의 실시간 교통통제를 효과적으로 할 수 있도록 가변적인 통행량과 통행시간을 가정한다. 기술적 모형(Descriptive Model)과 규범적 모형(Normative Model)의 두 가지 유형이 있다. 기술적 모형은 주어진 교통상황에서 사용자가 어떻게 행동하는가를 최적화를 통해 설명하고자 하며, 규범적 모형은 시스템 전체의 차원에서 최적화를 위해 어떻게 시스템이 행동하

는가를 설명하고자 한다.

6. Traffic Simulation 모형을 이용한 방법

시뮬레이션 모형을 이용한 방법에는 여러 가지 유형들이 있으며 대표적인 것으로는 이산시간(Discrete Time)·이산사건(Discrete Event) 모형, 미시적·준거시적·거시적 모형 및 확정적(Deterministic)·확률적(Stochastic) 모형 등이 있다.

대부분의 시뮬레이션 모형은 이산시간 모형을 기초로 하며 이는 시간을 알려진 일련의 시간단위로 분절화하여 시뮬레이션한다. 또한 대부분의 모형은 확률적 모형 및 미시적 모형의 범주에 포함된다.

장단기의 교통상황 변화를 비교적 짧은 시간 내에 예측하는 것이 시뮬레이션 모형들의 주된 개발 목적이나, 고려하여야 하는 시뮬레이션의 세부모형들이 복잡해지고 교통 네트워크가 커질수록 시뮬레이션에 필요한 연산시간은 기하급수적으로 늘어나는 단점이 있다.

III. 예측모형의 설계

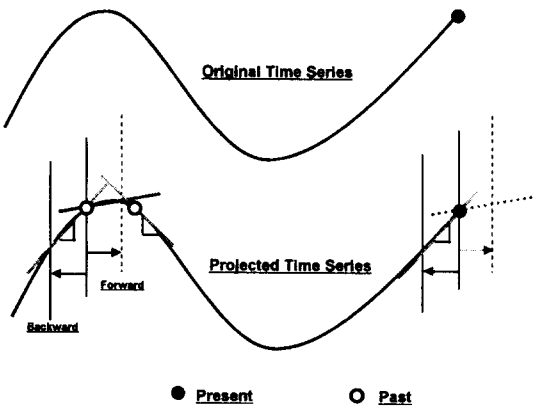
1. 비모수 회귀분석

연속형 반응함수 Y와 이를 설명해주는 한 개의 설명변수 X의 관계가 회귀함수 f 에 의해 주어지는

$$Y_i = f(X_i) + \varepsilon_i$$

의 회귀모형으로 설정된다고 하자. 오차항 ε_i 는 평균이 0이고 분산이 σ^2 이라고 가정한다. 지금까지 다루어 왔던 고전적 선형회귀모형, 일반화 선형모형, 비선형회귀모형 등에서는 회귀함수 의 형태가 사전에 주어졌었다. 다시 말해서, 단순선형회귀에서는 $f(x) = \beta_0 + \beta_1 x$로, 우리의 관심은 β_0와 β_1을 추정하고, 또한 $\beta_0 + \beta_1 x$라고 가정한 것이 타당한가를 알아보기 위해 잔차를 검토하거나 적합도 검정을 하였다. 일반화 선형모형도 마찬가지다. $g(\mu) = \beta_0 + \beta_1 x$ (단, g는 연계함수이고, $\mu = E(Y)$)라는 모형을 미리 설정하고 회귀모형을 적합시켰다. 즉, 연계함수 g가 주어지면 회귀함수 f는 자동적으로 정해진다. 비선형 회귀모형도 설명변수와 회귀계수가 비선형으로 결합된 형태의 함수 f가 미리 주어진다. 예를 들면, $f(x) = \beta_0 + \beta_1 e^{-\beta_2 x}$이 비선형 회귀모형이다. 이처럼, 회귀함수 f의 형태가 미리 주어져 있는 상황에서, 회귀계수의 추정에 관심이 있는 경우를 모수 회귀모형(parametric regression model)이라 부른다.

한편, 회귀함수 f 에 대한 형태를 미리 설정하지 않고 단지 함수 f 의 어떤 요건을 만족시키는 함수군에 속한다고 가정하는 것이 비모수 회귀모형(nonparametric regression model)이다. 함수 f 가 만족시켜야 되는 요건은, 예를 들면 두 번 미분 가능하고 f^2이 적분가능하다 등이 될



<그림 1> 비모수 회귀분석의 개념

수 있다. 따라서 함수 f 는 무수히 많은 원소를 갖는 함수군에 속한다. <그림 1>은 비모수 회귀분석의 개념을 나타낸 것이다.

2. 최근린 분석

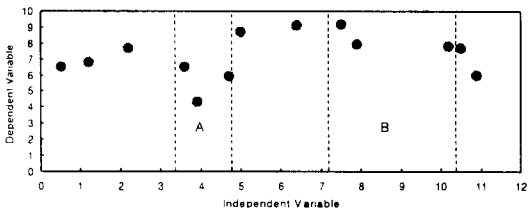
비모수 회귀분석에서 근접치를 찾기 위해 최근린 분석(k-Nearest Neighbor) 알고리즘이 많이 쓰이며 이를 수식으로 표현하면 아래와 같다 (Hardle, 1990).

$$W_{ki}(x) = \begin{cases} n/k, & \text{if } i \in J_x; \\ 0 & \text{otherwise.} \end{cases}$$

$$\hat{m}_k(x) = n^{-1} \sum_{i=1}^{n} W_{ki}(x)\ Y_i$$

여기서 근접치 $\hat{m}_k(x)$는 가중치 $[\ W_{ki}(x)\ ]_{i=1}^{n}$ 에 입력치를 곱한 값의 합을 평균하여 구한다. 이때 J_x 는 x에 대한 최근린 관측치들의 집합을 의미한다.

<그림 2>는 최근린 분석을 이용한 비모수 회귀분석의 예를 보여주고 있다. 이변수함수 $f(x)=y$에서 $x_A=4$이고 $x_B=9$인 경우의 y 값을 구한다고 가정하면, 독립변수의 값이 13 이전에 관측된 x, y 값들이 그림에 나와 있다. 이때 $k=3$이라 정하면 x_A와 x_B에 가장 가까운 각각 3개씩의 관측치들이 포함되며, y 추정치는 인근 관측치들의 평균을 구함으로써 얻어진다. 본 예에서 x_A의 경우 세 인근 관측치의 y값 6.5, 4.3, 5.9의 평균값인 5.6이 그리고 x_B는 8.2가 추정값으로 얻어진다.



<그림 2> 최근린 예

3. 알고리즘 구성

Predict: $V(t+15)$

Given: $V(t)$

$V_{hist}(t)$

$V_{hist}(t+15)$

본 연구에서는 예측의 주기를 15분으로 정한다. 왜냐하면, 이보다 짧은 주기의 통행의 변동은 불안정하기 때문이다.(McShane and Roess 1990). 그리고, Highway Capacity Manual(1994)는 운영분석에서 15분 통행을 권장하고 있다.

[알고리즘]

0. 최근린 리스트(NB)를 초기화. $(1,2,\cdots,k)$

1. 데이터 베이스의 각 원소 c에 대하여
   a. $Distance(X(t),X_c)$
   b. $Distance(X(t),X_c) < MaxDIST(NB)$이면,
      i. NB에서 최대값을 제거
      ii. NB에 c를 추가
2. $V(t+D)$를 추정

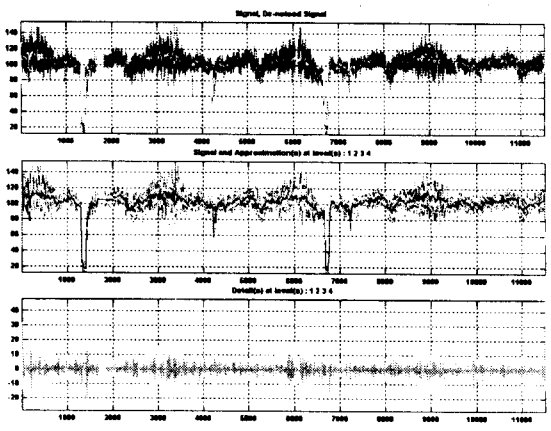
$$V(t+D)=\frac{\sum_{q\in NB} Vf_q}{k}$$

여기에서, $MaxDIST(NB)$는 $X(t)$와 가장 큰 차이를 보이는 값을 말한다.

4. 데이터의 잡음제거

통행속도의 시계열 데이터가 지니는 잡음을 제거하기 위하여 웨이블렛 변환을 이용하였다. 이때 기저 웨이블렛은 Daubechies 웨이블렛을 이용하였고, <그림 3>에서 보는바와 같이 이산 웨이블렛 분해 기법으로 시계열 데이터를 4단계로 분해하고, 일반적으로 잡음의 성격을 띄는 가장 빈도수가 높은 고주파 신호부분을 제거하여 입력 시계열 데이터의 추세를 그대로 유지하면서 세부적으로는 선형의 특성을 보이는 시계열로 전처리하였다.

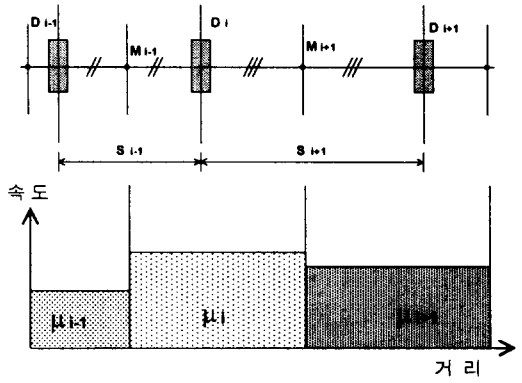
특히 웨이블렛 변환을 이용하면 이동평균을 취한 것보다 시계열의 추세는 그대로 유지하면서 세부적으로는 선형의 형태를 띄게 되어 선형 국지적(Local) 회귀분석의 일종인 비모수 회귀분석에 더욱 유리한 것으로 판명되었다.



<그림 3> 웨이블렛 변환을 이용한 시계열데이터 잡음제거

5. 통행시간 추정

통행시간은 예측된 지점별 통행속도를 이용하여 <그림 4>와 같이 추정한다. 예를 들면 15분 예측시 D_{i-1} 지점과 D_i 지점까지의 중점인 M_{i-1}까지의 속도는 D_{i-1}지점에서의 예측속도($\mu i-1$)가 되고 M_{i-1}에서 그 다음 두 지점간의 중점인 M_{i+1}까지의 속도는 D_i 지점의 예측속도(μ_i)가 된다. 그러나 D_{i-1} 지점에서 D_{i+1}지점까지 두 구간 이상을 예측할 시에는 먼저 D_{i-1}에서 D_i까지의 통행시간(S_{i-1})을 구하고 D_i 지점의 속도를 예측할 때에는 15분 후의 예측이 아닌 D_{i-1}에서 D_i까지의 통행시간(S_{i-1})을 더한 $15+S_{i-1}$의 예측을 수행한다.



<그림 4> 예측 속도를 이용한 구간별 통행시간의 추정

[Sisiopiku, Rouphail and Tarko, 1994]

IV. 평가 및 결론

1. 평가

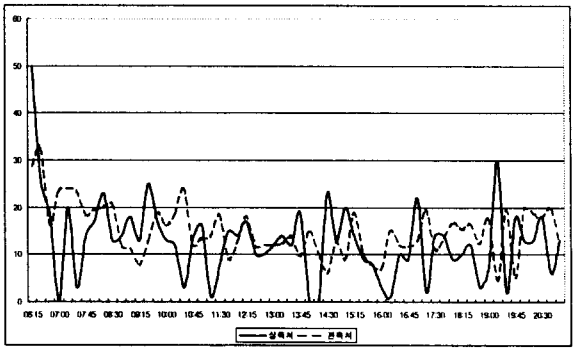
예측모형을 프로그램 한 후 컴퓨터 상에서 연산한 결과를 비교 분석하였다. 주안점은 예측의 정확성을·평가하는데 있으며, 예측치와 실측치를 비교하여 RMSE (Root Mean Square Error)와 RMPE (Root Mean Percent Error) 및 상관계수를 구해 평가한다. RMSE와 RMPE 및 상관계수를 구하는 식은 아래와 같다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i)^2}$$

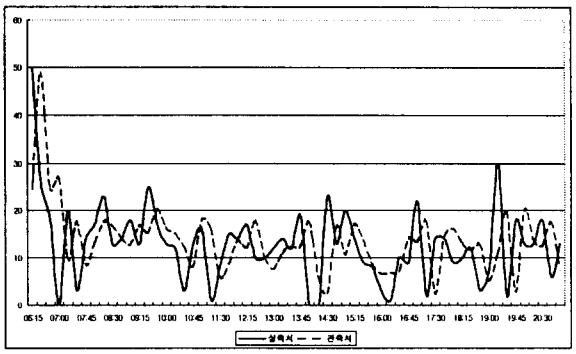
$$RMPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{x}_i - x_i}{x_i}\right)^2}$$

$$\rho = \frac{\sum_{i=1}^{n}(\hat{x}_i - \hat{x}_{mean})(x_i - x_{mean})}{n\hat{\sigma}\sigma}$$

여기서 x_i는 실측치, $\hat{x}_i$는 예측치, x_{mean}은 실측치 데이터 x_i의 평균, $\hat{x}_{mean}$은 예측치 데이터 $\hat{x}_i$의 평균이다. 상관계수 ρ는 추가로 실측치와 예측치 각각의 표준편차 σ 및 $\hat{\sigma}$를 구해서 얻는다. 이때 RMSE와 RMPE는 예측 결과와 실측치간의 편차를 이해하는데 쓰이고, 상관계수는 예측결과와 실측치간의 상관관계를 이해하는데 쓰인다. 즉 편차가 작을수록, 상관계수가 1에 가까울수록 예측의 결과가 정확하다고 할 수 있다.



<그림 5> 실측치와 예측치(웨이블렛 이용안함)



<그림 6> 실측치와 예측치 (웨이블렛 이용)

결과를 살펴보면, 15분 예측에서 실측치를 그대로 적용한 경우와 웨이블렛 변환을 적용한 후에 예측을 수행한 경우에 예측값의 차이가 있음을 알 수 있다.

RMSE값을 살펴보면, <그림 4>, <그림 5>의 경우 모두 9 부근에 나타난다. 이것은 국지적으로는 웨이블렛을 적용한 예측이 더 낫더라도 대상시간 전체에서는 예측치와 실측치간의 편차의 차이가 없다는 것을 의미한다.

그러나, 일반적으로 웨이블렛 변환을 이용하면 시계열의 추세는 그대로 유지하면서 세부적으로는 선형의 형태를 띄게 되어 비모수 회귀분석을 이용한 예측에 정확도를 향상시킬 수 있다.

이러한 차이가 나는 주된 이유는 데이터의 수집시간 간격이 15분으로 너무 크기 때문이다. 이것은 통행의 국지적 추세를 반영하기에는 미흡하다고 볼 수 있다.

2. 결론 및 향후 연구과제

본 연구개발은 과거 데이터베이스와 실시간 데이터를 이용하여 통행시간 예측모형을 개발하여, 예측된 단기간의 통행시간을 이용자들에게 제공하여 보다 빠른 경로 및 대체 경로의 선택을 가능하게 할 수 있는 토대를 마련하는데 있다.

향후에는 예측된 통행시간을 경로안내에 활용할 수 있는 방안에 대한 연구가 필요하다. 실시간 예측 정보를 바탕으로 주행 중 경로가 갱신될 수 있는 알고리즘의 개발이 요구된다.

참고문헌

1. Jinsoo You, Tschangho John Kim(1999), "A GIS-based Travel Time Forecasting Model: Prototype model development and A Preliminary Result", Pacific Regional Science Conference
2. Härdle, W(1990)., "Applied Nonparametric Regression", Cambridge University Press
3. Smith, B. L. and Demetsky, M. J. (1997), "Traffic Flow Forecasting: Comparison of Modeling Approaches", Journal of Transportation Engineering, vol. 123, no 4, p. 261-266
4. Misiti, M., Misiti, Y., Oppenheim, G., and Poggi, J.-M. (1997), "Wavelet Toolbox for Use with Matlab", User's guide, Version 1, The Math Works, Inc.
5. Ben-Akiva, M., Cascetta, E. and Gunn, H. (1995), "An On-Line Dynamic Traffic Prediction Model for an Inter-Urban Motorway Network." Urban Traffic Networks: Dynamic Flow Modeling and Control, Gartner, N. H. and Improta, G. (eds.), New York: Springer-Verlag, p. 83-122.
6. Sisiopiku, V. P., Rouphail, N. M. and Tarko, A.(1994), "Estimating Travel Times on Freeway Segments", Advance Working Paper Series, Number 32.