

■ 論 文 ■

고속도로 통행료수납자료를 이용한 통행시간 예측모형 개발

Development of The Freeway Operating Time Prediction Model
Using Toll Collection System Data

강 정 규

(한국도로공사 도로연구소 연구위원)

남 궁 성

(한국도로공사 도로연구소 수석연구원)

목 차

- I. 서론
 - 1. 연구 배경 및 목적
 - 2. 연구 범위 및 방법
- II. TCS 데이터의 특성과 예측
- III. 요구기능과 예측모형의 선정
- IV. TCS 데이터의 시계열 패턴분석
 - 1. 분석자료
 - 2. 시계열 분석 방법
 - 3. 시계열 패턴분석
 - 4. 분석결과 종합
- V. 고속도로 통행시간 예측모형 개발
 - 1. Modular 신경망 예측모형의 개요
 - 2. 개발모형의 기본구조
 - 3. 모형의 최적구조 설계
 - 4. 학습자료의 구성
 - 5. 모형의 수행결과 및 평가
- VI. 모형의 평가 및 분석
 - 1. 누적확률분포를 이용한 오차분석
 - 2. 실제치-예측치 변동 분석
- VII. 결론
- 참고문헌

Key Words : 고속도로, 통행시간, 예측, 시계열분석, 신경망

요 약

경인고속도로가 개통된 1969년 이후로 우리나라 고속도로는 국토의 대동맥으로서 꾸준한 확장과 신설을 거듭해 왔으며, 조만간 고속도로 3,000km시대를 맞이하게 될 전망이다. 이에 고속도로는 건설과 확장 위주에서 효율적 운영이 중요성이 과거 그 어느 때보다도 강조되고 있다. 최근 고속도로망이 복잡해져 감에 따라 교통정보의 가치가 점점 높아지고 있고, 더욱이 정보통신기술의 급격한 발달과 함께, 휴대폰, PDA, PNS 등의 보급이 증가함에 따라 교통정보의 수요가 급증하고 있다. 특히, 통행시간 정보의 경우, 혼잡이 있는 도로망에서 최적의 경로선택을 하고, 경로간에 적절한 교통량분산을 통해 자원의 효율적 배분을 달성하는데 있어 필수적인 정보로서 그 중요성을 더해 가고 있는 실정이다.

본 연구에서는 고속도로 통행료수납시스템(TCS)에서 수집되는 톨게이트간 통행시간 데이터를 기반으로 TCS 통행시간 데이터의 속성과 시계열적 패턴을 규명하고, 이를 바탕으로 모듈라 신경망모형(Modular Neural Network Model)을 이용한 통행시간 예측모형을 개발하였다.

우선, 단거리(서울→수원)와 장거리(서울→대전) 그리고 평일과 주말로 구분하여 TCS 데이터에 대한 시계열 패턴 분석을 한 결과, 단거리와 장거리 공히 충분한 범위의 예측가능한 시간적 범위를 가지고 있으며, 복잡한 정도는 장거리가 높은 것으로 나타났다. 다만, 단거리구간이 장거리 구간에 비해 초기조건에 대한 민감성이 큰 이유로 상대적으로 장기예측이 어려운 것으로 분석되었다. 한편, 모형 적용 현장의 요구기능을 분석하여 모듈라 신경망 구조를 가진 예측모형을 개발하였으며, 최소한 약 80분 이상의 장기예측이 요구되는 서울→대전구간에 적용한 결과, 대부분 10분 이내의 낮은 오차를 보였다. 본 연구에서 개발된 모형은 예측범위가 고정적인 대부분의 시계열모형과는 달리 최소의 입력(3개)을 가지면서 하나의 신경망으로 학습한 최대/최소의 예측시간범위내에서 그 크기에 상관없이 거의 동일한 수준의 예측력을 보이는 장점을 가지고 있다.

1. 서론

1. 연구 배경 및 목적

최근 교통정보의 가치와 수요가 과거 어느때보다도 급증하고 있다. 특히, 점점 복잡해져 가는 도로망 확대로 동일한 출발/도착지를 갖는 통행에 있어 선택 가능한 경로의 수가 많아져 효과적인 "경로선택을 위한 교통정보"가 이용자 입장에서는 최적경로 선택, 운전자 입장에서는 교통량의 효과적인 분산 달성을 위하여 필수적인 정보로 자리 매김해 가고 있다.

통행시간 수집방법으로는 크게 지점검지기를 통해 얻어진 평균속도를 변환하여 통행시간을 간접적으로 얻어내는 방법과 주행중인 차량(Probe차량)의 위치를 실시간으로 파악하여 통행시간을 직접적으로 얻어내는 방법으로 나누어 볼 수 있다. 두 가지 방법중 어느 방법으로 이용하는가는 해당 인프라의 구축여부에 따라나, 일반적으로 후자의 경우가 통행시간 파악을 위해서는 가장 효과적인 것으로 알려져 있다. 그 이유는 통행시간 정보의 속성이 해당 구간의 시공간적 혼잡상태의 변화를 반영한 것이어야 한다는 것에 기인한다. 전자의 경우, 어느 시간단면에서 단위구간의 혼잡상태(또는 정체상태)를 단지 시간이라는 단위로 나타낸 것일 뿐, 혼잡상태가 고정적이지 않는 한, 이동에 소요될 시간을 의미한다고 보기 어렵기 때문이다.

여기서 Probe차량에 의한 통행시간 정보는 그 정보를 필요로 하는 차량입장에서는 자신보다 앞서 해당 구간을 이용한 차량이 경험한 통행시간이며, 교통망에서의 혼잡상태가 시공간적으로 변화한다는 점에서 예측의 과정없이 단지 수집된 데이터로부터 추정된 통행시간의 정확도는 해당 구간의 혼잡상태의 변화에 따라 달라 질 수 있다. 이는 혼잡의 정도가 어느 한 방향(증가 또는 감소)으로 진행하는 경우에는 더욱 큰 오차를 발생하게 된다. 따라서 통행시간 정보는 예측되어 제공되는 것이 바람직하다.

우리나라 고속도로의 경우, 2002년 현재 총 25개 선 약 2,600km로서 2001년말 서해안선을 포함한 5개 신설노선이 동시 개통됨으로써 과거와는 달리 고속도로만으로도 동일한 출발/도착지간에 여러개의 경로가 존재하는 소위 "다중경로"환경에 놓이게 되었다. 이로 인해 최적경로 선택과 효과적인 교통량 분산을 위해

통행시간 정보의 수요가 그 어느 때보다 급증하고 있는 실정이다.

고속도로 통행시간 정보는 루프 또는 영상검지기에 기반을 둔 차량검지시스템(VDS:vehicle detection system)과 출발 톨게이트(TG 또는 영업소) 및 도착 톨게이트 통과시간이 기록되는 통행료 수납시스템(TCS: toll collection system)에서 얻어지는 TG간 통행시간데이터를 이용하여 제공하고 있다. VDS를 통한 통행시간 정보는 서울 외곽순환고속도로와 같이 개방식 영업체제로 운영되고 있는 구간이나 단거리(TG기준 1, 2개구간정도) 위주로 제공되고 있고, 폐쇄식 운영 구간에서의 장거리 통행시간 정보는 TCS 데이터를 기반으로 제공하고 있다(한국도로공사, 2000).

본 연구는 TCS데이터를 기반으로 고속도로 출발/도착지 TG간 통행시간을 예측할 수 있는 모형을 개발하고자 하는 것을 목적으로 한다. 이를 위해 우선 고속도로 통행시간 예측을 위하여 모형이 갖추어야 할 요구기능을 정의하고, TCS 통행시간 데이터의 특성 및 시계열적 패턴분석을 통해 TCS 통행시간 데이터가 예측가능한 패턴을 가지고 있는지를 살피고, 마지막으로 모듈라 신경망구조를 이용한 예측모형과 그 평가결과를 제시하도록 한다.

2. 연구 범위 및 방법

고속도로 구간 중 TG간 통행시간 데이터를 얻을 수 있는 폐쇄식 운영구간을 대상으로 하며, 단기예측과 장기예측에서의 모형의 적응 성능을 평가하기 위해 서울→수원구간(13km)과 서울→대전구간(134km)을 분석대상으로 선정하였다. 예측의 재료로 쓰인 TCS 통행시간 데이터는 휴게소 체류시간, 갓길 정차 시간 등이 제거된 출발시각 기준 5분 단위의 TG간 순통행시간(주행시간:operating time)이다. 여기서 통행시간 예측단위로는 링크와 경로가 있으며, 본 연구의 제안모형은 경로단위의 통행시간 예측을 수행한다. 링크단위의 예측은 모든 OD간 예측을 수행하여야 하는 경우에 연산량에 있어 효율적이나, 링크단위 오차의 누적으로 예측의 정확도가 떨어질 수 있으며, 혼잡이 어느 한 방향으로(증가 또는 감소)로 진행하는 경우, 더욱 오차가 심해지는 경향을 갖는 단점이 있다.

한편, 예측모형을 크게 목적변수(통행시간)와 다른

변수와의 인과관계를 이용하는 '구조모형'과 과거의 추세로 미래를 설명하고자 하는 '시계열 모형'으로 구분할 때, 본 연구에서는 시계열 모형으로 접근하였다. 구조모형은 통행시간에 영향을 미치는 다양한 변수를 모두 추출하고 상호간 인과관계를 계량화하는 것이 한계가 있기 때문이다. 시계열 예측모형은 복잡한 패턴의 모델링에 유연하게 적응하며, 불규칙성을 어느 정도 포함하고 있는 입력에도 우수한 결과를 제공하는 특징이 있는 신경망 모형을 이용하였다.

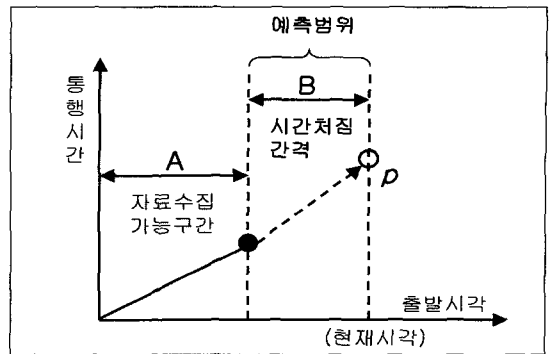
단, 과거 시계열적 패턴에 의한 예측을 하는 경우, 주어진 시계열 데이터가 유의한 패턴을 내재하고 있는지를 우선적으로 살피는 것이 중요하며, 이를 위해 본 연구에서는 모형개발에 앞서 시계열 패턴의 복잡성(complexity) 분석을 수행하였다. 이러한 과정은 예측모형의 적용이 유의한 범위의 예측결과를 보여준다는 여부에서 더 나아가 그 결과가 우연의 결과인지 아닌지 간접적으로 알려줄 뿐만 아니라, 적절한 예측모형을 선정하고 경우에 따라서는 모형구조를 설계할 때, 유용한 정보를 제공하여 준다.

II. TCS 데이터의 특성과 예측

TCS 데이터는 고속도로 이용차량의 도착 TG로부터 수집된다. 이는 TCS로부터 수집되는 통행시간 데이터는 모두 도착시점을 기준으로 수집되는 것임을 의미한다. 결국, 그 정보를 이용하고자하는 차량 입장에서는 비록 그 정보가 실시간으로 수집된다 하더라도 이미 '과거정보'라고 할 수 있다. 따라서 TCS 데이터는 출발시각에 따른 통행시간 측면에서 볼 때, 해당 구간을 이용한 가장 최근의 차량이 경험한 통행시간만큼 과거로 지연된 데이터이다. 이를 본 연구에서는 TCS 데이터의 '시간처짐' 특성이라 표현한다.

따라서 <그림 1>에서 보는 바와 같이 예측을 해야 할 시간적 범위는 현재시각을 출발시각으로 할 때, 시간처짐 간격(B)만큼이 되며, A영역내의 데이터를 이용하여 B의 시간범위만큼 떨어진 p 점을 예측하는 것이다. 문제는 시간처짐 간격인 B의 크기가 구간거리에 따라 커지며(즉, 예측해야할 시간범위가 커지며), 같은 구간이라 하더라도 혼잡에 따라 그 크기가 동적으로 변화한다는 것이다.

따라서 TCS 데이터를 이용한 통행시간 예측모형은 우선 대상 구간의 자유교통류 통행시간 크기 이상의



<그림 1> TCS 데이터의 시간처짐과 예측범위

장기예측이 가능해야 하며, 아울러 예측범위의 동적인 변화에 적응할 수 있어야 한다.

결론적으로 통행상황이 고정적이지 않고 혼잡 등으로 변화하는 여건하에서는 TCS 데이터로부터 수집된 통행시간 정보 자체는 운전자 입장에서 의미가 없을 수가 있으며, 시간처짐 간격만큼 예측된 정보이어야만 유용하다는 것이다.

III. 요구기능과 예측모형의 선정

다음은 본 연구에서 제안하는 예측모형이 갖추어야 할 요구기능과 그에 따른 모형선정의 근거이다.

첫째, 획득 가능한 자료만으로도 충분히 적용 가능한 모형이어야 한다. 예측의 대상이 되는 현상은 일반적으로 많은 변수가 관여한다. 그러나 많은 변수를 고려하여 아무리 좋은 결과를 도출하는 예측모형이라도 비현실적이거나 획득이 불가능한 입력변수를 요구하는 모형이라면 실제 현장 적용이 불가능하다.

둘째, 적절한 시간 범위 내에서 예측을 위한 연산이 이루어질 수 있어야 한다. 이를 위해 특히 유념해야 할 것은 가급적 예측모형에서 요구하는 입력 데이터를 최소화 해야한다는 것이다. 필요한 입력 데이터를 DB로부터 탐색하여 읽어오는 과정(Query)에서 많은 시간이 소요된다면 실제 현장적용은 불가능하다.

셋째, 모형구조가 단순하고, 다양한 현장여건의 차이에 영향을 크게 받지 않아야 한다. 복잡하고 정밀한 모형일수록 정확도는 높지만, 그 만큼 많은 정보량을 모형에 주어져야 함은 물론, 모형의 튜닝을 정밀하게 해주어야 하고, 현장여건의 차이에 민감하게 영향을 받을 수 있다.

넷째, 가장 중요한 것으로 예측범위가 변하는 것에 적용할 수 있도록 동적인 예측범위를 갖는 것이 바람

직하다. 앞서 살펴 본 바 있듯이 TCS 통행시간 데이터에는 "시간차집"이란 특성이 있으며 그 크기가 바로 예측범위가 된다. 문제는 그 시간차집이 혼잡수준에 따라 동적으로 변화하는데 있다.

다섯째, 장기예측이 가능하여야 한다. 지금까지의 대부분의 관련된 분야의 예측모형은 '단기예측'에 국한되어 있다. 그러나 TCS 데이터의 경우, 시간차집 특성상 구간거리가 멀면 멀수록 예측해야 할 미래의 시점은 더욱더 멀어지며, 따라서 장기예측이 필요하다(Al-Deck et al., 1998; Brian et al., 1994; Park, 1998).

마지막으로, 여건의 변화에 손쉽게 적용할 수 있는 모형이어야 한다. 통행시간 시계열 패턴은 구간별로 나뉘어 고유의 특성패턴을 지니고 있으며, 이는 고속도로의 확장이나 신설노선의 개통과 같은 물리적인 변화나 신도시의 건설, 공업단지의 입지 등 고속도로 주변 토지이용의 변화와도 밀접한 연관을 가지고 변화한다. 따라서 이러한 여건 변화에 따른 통행시간 패턴의 변화에 스스로 적응해 갈 수 있는 모형이 바람직하다.

본 연구에서는 이와 같은 요구기능을 만족시킬 수 있는 모형개발을 위해 연구과정에서 그간 시도되어온 다양한 예측모형(전통적인 통계모형, 신경망모형, 칼만필터, 패턴매칭 등)을 시험하였으며, 예측력 및 모형의 구조 등을 종합적으로 고려한 결과, 신경망 모형이 가장 적합하다는 결론을 얻었다.

신경망 모형은 해(solution)를 내놓는 과정이 명확하게 설명되지 않으며 최적 모형구조를 찾아내기 위해 경험적 실험방법에 의존해야 하는 단점을 가지고 있지만, 복잡한 패턴의 모델링에 유연하게 적용하며, 특히 일단 학습이 제대로 이루어지면, 불규칙성을 어느 정도 포함하고 있는 입력에도 우수한 결과를 제공하는 장점을 가지고 있다. 따라서, 다양한 인자가 관여하고 많은 외부 환경에 영향을 받아 불규칙한 패턴을 항상 내재하고 있는 통행시간 데이터에 적합하다. 더욱이 재학습을 통하여 언제든지 입력패턴의 변화에 모형의 적합(fitting)이 가능하여 TCS 통행시간 데이터와 같이 고속도로 확장 및 신설, 고속도로 주변 토지이용의 변화, 시계열 변동의 계절적 요인 등에 의해 그 패턴이 변화하는 입력에 우수한 적응력을 보여준다.

IV. TCS 데이터의 시계열 패턴분석

예측에 있어 주어진 데이터에 내재되어 있는 특성

이나 성질을 파악하는 것은 무엇보다 중요하다. 따라서 매우 불규칙하고 복잡하여 전혀 예측불가능할 것 같은 시계열데이터를 대상으로 어떤 일관성 있는 패턴이나 시계열이 형성되는 중요한 규칙이 있는지를 확인하는 과정은 시계열 예측에 있어 필수적이다.

본 장에서는 시계열 분석프로그램(Complexity ver. 1.0)을 이용하여 TCS 통행시간 데이터의 시계열적 특성을 규칙성과 복잡성 측면으로 나누어 분석한다.

1. 분석자료

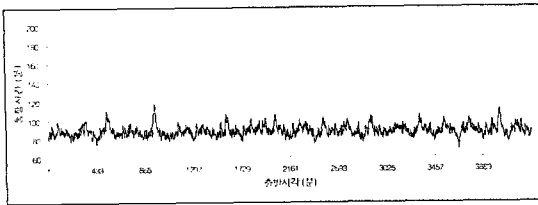
분석대상구간을 단거리 구간으로서 서울-수원구간을, 장거리 구간으로서 서울-대전구간을 선정하였다. 두 구간 모두 2000년 3월에서 5월중의 총 36일(평일 20일, 주말 16일-토, 일-)치 TCS 통행시간 시계열 데이터를 수집하여 분석하였다. 분석자료의 시계열 패턴은 <그림 2>의 (a)-(d)와 같다.

2. 시계열 분석 방법

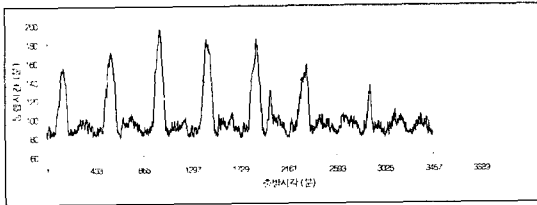
서울-수원과 서울-대전구간 TCS 데이터를 대상으로 다음과 같은 시계열 분석을 시행하였다. 먼저 시계열자료의 웨이블릿 스펙트럼(Wavelet Spectrum) 분석과 Self Information을 통하여 규칙성을 분석하였으며, 상관차원(Correlation Dimension) 및 리아프노프 지수(Lyapunov Exponent)를 이용하여 시계열자료의 복잡성을 분석하였다. 본 논문에서는 지면관계상 각 분석방법의 간단한 이론적 개요와 분석결과를 중심으로 살펴보고자 한다(Wei, 1990).

1) 웨이블릿 스펙트럼(Wavelet Spectrum) 분석

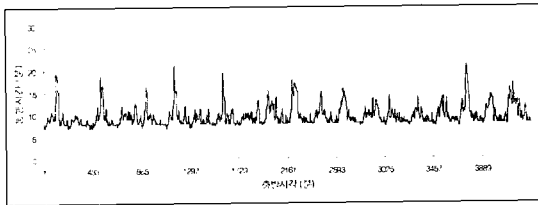
웨이블릿 변환(Wavelet Transform)은 기존의 데이터 공간(시계열데이터라면 time domain)을 직교 기저함수(orthogonal basis function)로 구성된 새로운 공간으로 변환시킨다. 웨이블릿 변환을 한 파수(wave number)가 높으면 높을수록 주어진 시계열의 진동성분의 주파수가 클을 의미하고, 이는 단위간격에서 변동이 빠르게 일어나는 시계열(변동이 잦음)이라고 할 수 있다. 한편, Wavelet amplitude의 경우, 그 값이 크면 클수록 그 변동의 크기가 크다는 것을 의미한다.



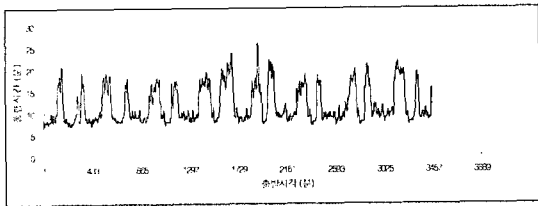
(a) 서울-대전, 평일(05:00~23:00, 20일치)



(b) 서울-대전, 주말(05:00~23:00, 16일치)



(c) 서울-수원, 평일(05:00~23:00, 20일치)



(d) 서울-수원, 주말(05:00~23:00, 16일치)

(그림 2) TCS 통행시간 시계열 패턴(단거리/장거리)

2) Self Information

Self Information은 원시계열($t_n, t_{n+1}, t_{n+2}, \dots$)과 일정 시간간격(d)만큼 지연된 시계열($t_{n+d}, t_{n+1+d}, t_{n+2+d}, \dots$)간의 Information 값이다. 즉 지연시간(d)을 0부터 5분단위로 늘려가면서 각 지연시간에 대해 계산한 Information값을 산출한다.

여기서 Information 값이 처음으로 최소값을 갖는데 걸리는 지연시간을 Deinformation Time이라 부르며, 이는 Information이 사라지는 데 걸리는 시간이다. Deinformation Time이 짧을수록 현재 데이터 값이 결정되기 위해 필요한 과거 데이터의 길이가 짧음을 의미하는 것으로, Deinformation Time

보다 적게 떨어진 과거 데이터들만이 현재 시점의 데이터와 연관을 갖고 있음을 나타낸다. 다시 말해 Deinformation Time 보다 더 먼 과거데이터는 현재 시점의 데이터를 결정하는 데 별 의미가 없다는 것이다. 이는 시계열분석에서 흔히 말하는 Memory effect와 같은 개념으로 주어진 시계열의 '예측가능범위'를 말해준다.

3) 상관차원(Correlation Dimension)

시스템의 동력학적 규칙에 대해서 아는 바가 없는 상태에서, 시스템에서 발생한 데이터를 이용하여 시스템에 대한 정보를 엿보려고 하기 위해 '상관차원(correlation dimension)'을 활용한다. 상관차원은 시계열로부터 위상공간(embedding space)을 구현한 후, 위상공간의 궤적이 차지하는 정도를 차원이라는 값으로 표시한다. 상관차원은 그 값이 크면 클수록 위상공간의 궤적이 더 복잡하다고 할 수 있으며, 더 나아가 시계열을 발생시킨 시스템의 구조가 더 복잡(Complex)함을 의미한다.

4) Largest Lyapunov Exponent

리아프노프 지수(Lyapunov Exponent)는 위상공간의 궤적이 시간에 따라 발산해 가는 특성을 정량화한 값으로 시계열의 비선형 동력학적 특징 중 계의 발산성(또는 초기조건에 대한 민감성, 불확정성)을 보여주는 대표적인 변수이다. 리아프노프 지수가 큰 값을 가질수록 시스템의 초기조건에 대한 민감성이 커지며, 이는 초기조건이 조금만 달라져도 시간이 지나면 시스템의 상태가 크게 변하기 때문에 장기예측이 매우 힘들다는 것을 의미한다.

3. 시계열 패턴분석

1) 웨이블릿 스펙트럼(Wavelet Spectrum) 분석

분석결과, 서울-대전구간 평일의 경우, 파수(Wave-number)가 5, 6인 구간에, 주말인 경우에는 파수가 2~3인 구간에 Wavelet amplitude가 높게 나오고 있음을 볼 수 있다. 이는 평일인 경우가 주말보다 통행시간의 변동이 심하다고 할 수 있으며, 변동의 크기는 Wavelet amplitude가 높은 주말이 큼을 말해준다. 쉽게 말해 평일의 TCS 통행시간 시계열에는 빠죽빠죽한 성분이 많고, 주말은 평일에 비해 그러한

것이 적지만, 크게 변한다는 것이다. 서울-수원구간의 경우에도 서울-대전구간과 마찬가지로 평일이 변동이 잦고, 주말이 변동이 크기가 큰 것으로 나타났다.

2) Self Information

Self Information 분석결과를 보면, 서울-대전구간의 Deinformation Time은 평일, 주말 모두 약 6시간을 보이고 있으며, 서울-수원구간의 경우, 평일은 약 7시간, 주말은 약 4.5시간을 나타내고 있다. 이 결과는 서울-대전구간의 경우, 6시간이상의 장기 예측은 아무런 의미가 없고, 그 이상의 예측은 맞는다 하더라도 우연에 해당된다는 것을 의미한다. 다시 말해 서울-대전구간은 약 6시간까지의 예측이 가능하다는 것이다. 서울-수원구간의 경우, 주말(4.5h)이 평일(7h)보다 작은 예측범위를 보이고 있는데, 이는 주말의 통행시간 시계열이 평일보다 예측하기 어려운 패턴을 가지고 있음을 의미한다.

3) 상관차원(Correlation Dimension)

상관차원은 그 값이 크면 클수록 위상공간의 궤적이 더 복잡하다고 할 수 있으며, 더 나아가 시계열을 발생시킨 시스템의 구조가 더 복잡(Complex)함을 의미한다. 분석자료의 상관차원을 구해보면, 서울-대전구간의 경우, 5, 6차원(소수점 올림)으로 계산되었고, 서울-수원구간이 이보다 작은 4차원으로 나타났다. 이는 장거리구간일수록 통행시간 패턴이 복잡함을 의미한다. 그러나 다양한 인자가 통행시간 패턴에 복합적으로 영향을 미치기 때문에 반드시 그렇다고 확신할 수는 없다. 다만, 구간거리가 패턴의 복잡함에 영향을 미칠 수 있는 가능성을 나타낸다고 보는 것이 정확한 표현이다. 참고로, 어떤 시계열의 차원값은 그 시계열을 생성하도록 영향을 준 변수의 개수를 의미하기도 한다. 즉, 시계열데이터가 그러한 패턴을 갖도록 한 영향인자가 무엇인지는 모르지만, 적어도 차원 수 만큼의 변수가 관여를 한다는 것이다. 이는 인과모형을 기술할 때, 매우 중요한 정보가 된다.

4) Largest Lyapunov Exponent

리아프노프 지수를 살펴보면, 장거리인 서울-대전구간은 1에 가까운 값을 갖고, 단거리인 서울-수원은 이보다 큰 2에 가까운 값을 보이고 있다. 따라서 두 구간모두 어느 정도의 범위 내에서 예측이 가능한 것

으로 판단되며(여기서 그 범위는 앞서 Self Information 분석을 통해 살펴본 바 있다). 다만, 서울-대전구간보다 서울-수원구간의 통행시간이 초기조건에 민감하여 상대적으로 장기 예측이 어려운 것으로 분석된다. 그러나 TCS 데이터의 시간차집 간격이 작은 단거리 구간일수록 장거리 구간에 비해 짧은 범위를 예측하면 되므로 단거리 구간이 장거리 구간에 비해 장기예측이 힘들다는 것은 크게 문제가 되지 않는다.

4. 분석결과 종합

이상으로 단거리(서울-수원)구간과 장거리(서울-대전)구간으로 나누어 살펴본 TCS 통행시간 시계열 패턴 분석결과를 정리하면 다음과 같다.

- ① 평일은 주말에 비해 동일한 범위에서 변동이 잦고(삐죽삐죽한 성분이 많고), 주말은 평일에 비해 잦은 변동이 덜하지만 변동의 크기가 크다.
- ② 예측가능범위로서 서울-대전구간은 평일/주말 모두 6시간으로 나타났으며, 서울-수원구간은 평일 7시간, 주말 4.5시간으로 나타났다. 이는 구간거리에 따른 TCS 데이터의 시간차집간격을 감안할 때, 충분한 범위로 판단된다.
- ③ 장거리구간 일수록 상대적으로 통행시간 패턴이 복잡한 것으로 나타났다. 즉, 시계열 패턴을 발생시킨 시스템의 구조가 장거리 구간이 단거리에 비해 복잡함을 의미한다.
- ④ 각 구간은 어느 정도 예측가능한 규칙성을 가지고 있으며(즉, 무작위적이지 않으면) 단거리 구간의 경우, 장거리구간에 비해 '초기조건에 대한 민감성'이 커 장기예측의 경우 오히려 장거리 구간에 비해 예측이 더 어려운 것으로 분석되었다.

V. 고속도로 통행시간 예측모형 개발

1. Modular 신경망 예측모형의 개요

Modular 신경망모형은 다층 퍼셉트론 신경망으로 대규모 입력자료를 학습하려고 할 때에 학습시간이 너무 길어지는 것을 방지하기 위해 비슷한 입력자료별로 나누어서 학습을 진행시키기 위하여 만들어진 신경망 네트워크이다. 그러나 본 연구에서는 대규모 입력자료의

효율적인 학습을 위한 것이 아니라, 2가지 성질이 다른 (하나는 과거의 시계열 자료, 또 하나는 예측점의 출발 시각) 자료를 하나의 신경망내에서 학습이 이루어지도록 하기 위하여 Modular 신경망을 설계하였다.

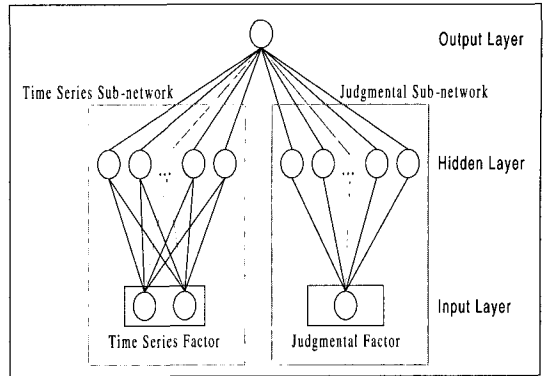
일반적으로 Modular 신경망모형은 입력층 퍼셉트론(input multilayer perceptron)과 결정층 퍼셉트론(decision multilayer perceptron)으로 구분된다. 이러한 경우에 역전파 학습법을 이용한 학습알고리즘은 2단계로 나누어서 실시되는데 첫째는 입력층에 해당하는 각 부분 네트워크(sub-network)내에서 각각 학습을 위한 전방향(feedforward) 진행을 수행한다. 이때 각각의 부분 네트워크가 서로 독립적이기 때문에 각 부분 네트워크는 동시에 각각 학습을 수행할 수 있으며 이로 인해서 수행시간이 현저히 감소될 수 있다. 두 번째 단계로서는 결정층에 해당하는 부분 네트워크가 학습을 진행하는데 입력층의 각 부분 네트워크에서 출력값을 다시 결정층의 입력값으로 받아서 감독 하의 학습이므로 출력층의 목표값과 비교하여 연결강도를 조절하고 이를 역방향으로 입력층까지 전달하여 에러를 최소화시키는 방식으로 연결강도로 조정하게 된다.

본 연구에서 개발된 Modular 인공 신경망은 출력층 값이 발생한 시점의 절대위치를 예측에서 고려할 수 있게 하는 판단 인자(judgemental factor)를 신경망 결정층 입력노드로 추가하였다.

2. 개발모형의 기본구조

본 연구에서 제안한 Modular 신경망모형의 구조는 다음의 3가지 특징으로 요약될 수 있다. 먼저 아래의 그림과 같이 신경망의 구조가 시계열적인 자료(time-series factor)와 판단 인자(judgemental factor)를 각각 독립적으로 고려하는 형태를 가진다. 즉, 입력층과 은닉층의 모든 유니트들간을 모두 연결하는 기존의 신경망 네트워크와는 달리 Modular 네트워크는 기존의 네트워크를 부분적인 네트워크(sub-network)로 분류하여 각각의 입력변수를 입력을 받아 은닉층까지 따로따로 학습을 진행시키는 구조를 가진다.

또 다른 특징은 위에서도 설명하였듯이 입력층은 시계열 부분과 판단 부분(judgemental part)으로 나뉘어지는데 시계열 부분(time-series part)에는 TCS 통행시간 시계열데이터로, 판단 부분에는 출력



〈그림 3〉 Modular 신경망 예측모형의 기본구조

치의 출발시각으로 구성된다. TCS 통행시간 데이터는 현재의 출발시각을 S라고 하고 예측시간간격을 p라 한다면 출발시각 S-p의 TCS 자료와 S-p-1 시간대의 TCS 자료가 입력되는 2개의 유니트로 시계열 부분이 구성되며, 또한 현재 출발시각 S에서 입력되는 1개의 유니트로 판단 부분이 구성된다. 출력은 현재 출발시각 S에서의 통행시간인 t(S)가 된다.

3. 모형의 최적구조 설계

본 연구에서는 최적의 모듈라(Modular) 신경망의 구조를 찾아내기 위하여 우선 시계열 부분(time series sub-network)과 판단 부분(judgemental sub-network)의 은닉층(hidden layer)의 최적조합을 검토하였다.

은닉층의 변화에 따른 신경망 모형의 예측성능 시험은 시계열부분과 판단부분의 은닉층의 갯수를 다음과 같이 크게 3개의 경우로 나누어 살펴보았다.

- ① 시계열부분(HL) = 판단부분(HR) : 6-6, 3-3
- ② 시계열부분(HL) > 판단부분(HR) : 6-3
- ③ 시계열부분(HL) < 판단부분(HR) : 3-6, 6-9

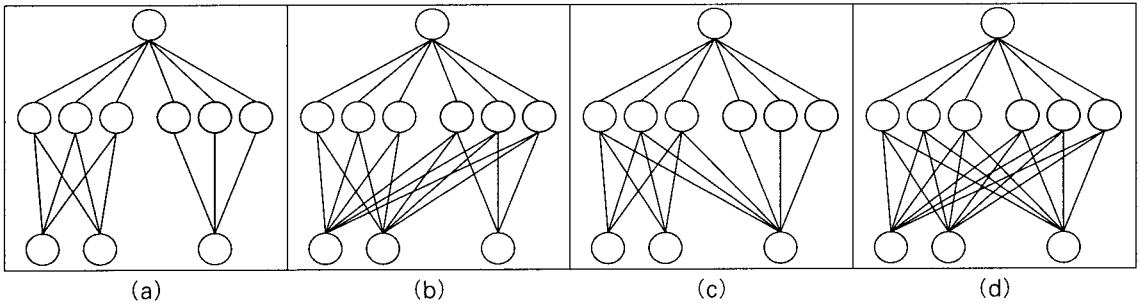
서울-대전구간의 평일 TCS 통행시간 데이터 10일치(2주)를 대상으로 은닉층의 노드 조합의 변화에 따른 모듈라 신경망모형의 예측결과 비교는 다음 〈표 1〉과 같다.

〈표 1〉에서 보는 바와 같이 양쪽 모두 3개씩인 경우(HL:HR=3:3)가 대부분의 예측범위에서 가장 안정적으로 좋은 결과를 보이고 있다. 은닉층의 수가 작으면 작을수록 학습에 소요되는 시간이 적게 걸림을 감안할 때, 이는 매우 바람직한 결과라고 할 수 있다.

〈표 1〉 모듈라 신경망의 은닉층 노드조합의 변화에 따른

은닉층 노드의 조합		60분 예측오차		90분 예측오차		120분 예측오차	
H _L	H _R	MARE	RRSE	MARE	RRSE	MARE	RRSE
6	6	0.0819	0.1069	0.0883	0.1282	0.0854	0.1166
3	3	0.0840	0.1087	0.0832	0.1066	0.0812	0.1028
6	3	0.0838	0.1060	0.0885	0.1195	0.1085	0.1864
3	6	0.0826	0.1099	0.0861	0.1211	0.1015	0.1730
6	9	0.0815	0.1080	0.0954	0.1855	0.0947	0.1370

註) 가장 작게 나온 부분을 음영으로 별도 표시하였음.



〈그림 4〉 신경망의 입력층과 은닉층 연결방법의 조합 예측성능비교

다음으로 은닉층의 개수를 3:3으로 한 구조에서 입력층과 은닉층의 연결방법의 조합을 〈그림 4〉와 같이 4개의 경우로 나누어서 살펴본 결과, 시계열부분과 판단부분을 은닉층 수준(level)까지 완전히 독립시킨 것이(유형 a) 전반적으로 가장 좋은 결과를 보였다.

따라서 본 연구에서는 이러한 시험결과를 바탕으로 모듈라 신경망의 구조를 다음 그림과 같이 설계하였다. 입력층에 있어 시계열부분의 노드에는 해당 구간에서 가장 최근까지 들어온 통행시간 데이터의 출발 시각 S_n 을 기준으로한 통행시간 $t(S_n)$ 과 직전 통행시간 $t(S_{n-1})$ 을 입력하고, 판단부분에는 예측범위 d

만큼 지연시킨 출발시각 S_{n+d} 를 입력한다. 또한 출력층에는 출발시각 S_{n+d} 에서의 통행시간인 $t(S_{n+d})$ 를 대응시킨다.

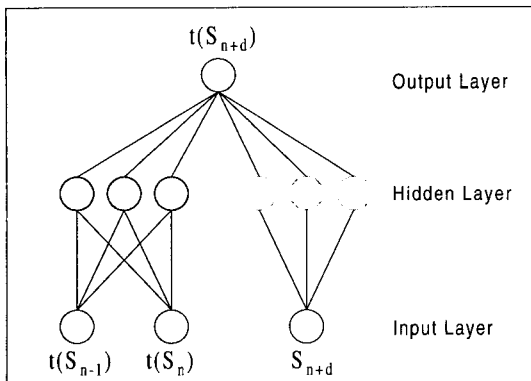
여기서 판단부분에 입력되는 S_{n+d} 는 통행시간시계열패턴에서 출력노드 출발시각의 절대위치를 의미하는 것으로, 출력노드에 대한 입력노드의 패턴을 고유하게 특성화 시켜주는 역할을 한다. 이는 매우 중요한 부분으로서 앞서 언급한 바 있듯이 학습자료에서 입력패턴과 출력패턴간의 1:N의 사상으로 인한 신경망 학습실패를 방지한다.

4. 학습자료의 구성

1) 학습자료

모듈라 신경망 예측모형의 수행을 위한 학습자료는 다음과 같다.

- ① 대상구간 : 경부선 하행 서울-대전(134km)구간, 서울-수원(13km)구간
- ② 학습자료 : 평일 20일(4주), 주말 16일(8주)
- ③ 검증자료 : 평일 10일(2주), 주말 8일(4주)



〈그림 5〉 모듈라 신경망 예측모형의 최적구조

여기서, 대상구간은 시계열패턴분석을 통해 구간계

리에 따라 시계열 패턴의 특성이 다름을 감안하여 단거리와 장거리 2개구간으로 하였으며, 각 구간에 대해서 시계열 패턴의 분명히 구분되는 평일과 주말로 다시 나누었다. 또한 학습자료는 2000년 3월~6월 기간중의 평일 20일, 주말 16일치를 대상으로 하였다. 평일과 주말의 학습자료의 양을 이렇게 설정한 것은 신경망 학습시 가급적 계절적 요인에 의한 영향을 최소화하고, 아울러 패턴의 변동으로 인한 잦은 재학습을 방지하기 위한 것이다. 실제로 학습자료의 규모에 따른 예측력 시험결과, 평일의 경우 2주 정도의 학습량만으로도 우수한 예측력을 보였다. 그러나 계절적 요인에 의해 시계열 패턴의 구조가 바뀌는 변환기간, 즉 방학시작과 종료시점인 12월과 3월, 장마철인 7, 8월에 학습기간이 걸치는 경우, 충분한 학습이 이루어질 필요가 있기 때문에 평일 4주, 주말 8주의 기간을 학습자료의 범위로 하였다. 한편, 학습결과(또는 예측결과) 검증하기 위한 자료로는 평일 2주, 주말 4주의 데이터를 이용하였다. 일반적으로 학습된 신경망 예측모형을 사전평가하기 위한 자료의 규모로는 학습자료와 2:1의 비율로 구성하는 것이 적절한 것으로 알려져 있다.

2) 학습자료의 구성

학습자료를 구성하기 위해서는 우선 모듈라 신경망 입력층의 판단부분에 들어갈 출발시각 $S_n + d$ 에서 d 를 결정할 필요가 있다. 여기서 d 는 해당 구간에서 가장 최근에 수집된 통행시간 데이터의 출발시각(S_n)과 현재시각(예측시점)과의 차이로서 TCS 데이터가 갖는 '시간차집'의 간격인 동시에, 예측범위가 된다. 따라서 기본적으로 d 는 예측범위를 몇 개로 두느냐에 따라 결정된다. 예를 들어 어느 구간의 시간차집 간

격(TCS 데이터 수집시각과 수집된 TCS 통행시간 데이터의 출발시각)이 20분에서 60분까지 나타난다면, d 는 시계열부분 입력노드와의 출발시각 지연간격을 20분~60분내의 범위로 설정해야 한다는 것이다. 여기서 d 의 개수를 결정하는 것은 시간차집 간격의 범위(20~60분)를 나누는 단위간격이다. 만약 10분 간격으로 나눈다면, 20, 30, 40, 60분으로서 d 는 총 4가지가 되는 것이다.

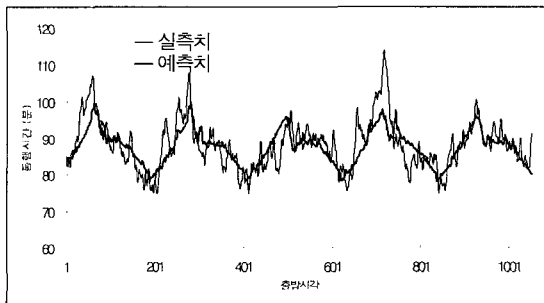
본 연구에서는 학습자료 구성시 d 의 크기와 개수는 각 구간의 통행시간 히스토그램(histogram)을 토대로 서울-수원구간의 경우, "d=10분, 15분, 20분"으로 하였으며, 서울-대전구간의 경우, "d=80분, 100분, 120분, 140분"으로 설정하였다. 다만, 시험결과, 모든 범위(모든 d)를 학습한 것과 최소 예측범위와 최대예측범위(즉, 예를 들어 서울-대전의 경우, $d=80, 140$ 분 예측만을 학습)만을 학습한 것의 예측성능의 차이가 거의 없어 학습자료는 서울-수원은 $d=10, 20$ 분, 서울-대전은 $d=80, 140$ 만으로 구성하였다.

5. 모형의 수행결과 및 평가

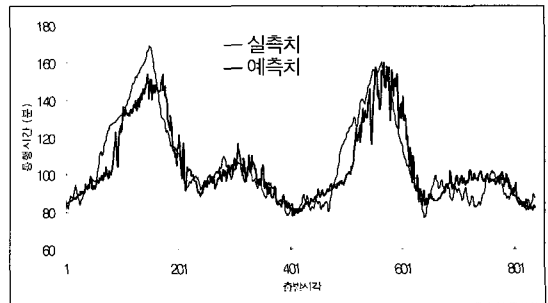
1) 모형의 수행결과

다음 <그림 7>의 (a)-(d)는 단거리구간(서울-수원)과 장거리구간(서울-대전)을 대상으로 단기예측(10분, 15분, 20분)과 장기예측(80분, 100분, 120분, 140분)에 대한 모형의 수행결과이다. 그림과 <표 2>에서 나타난 바와 같이 수행모형은 장단기 예측에서 전반적으로 모두 합리적인 예측성능을 보이는 것으로 나타났다.

여기서 주목할 만한 결과는 본 연구에서 설계된 모듈라 신경망모형이 모든 예측범위를 학습하지 않고,

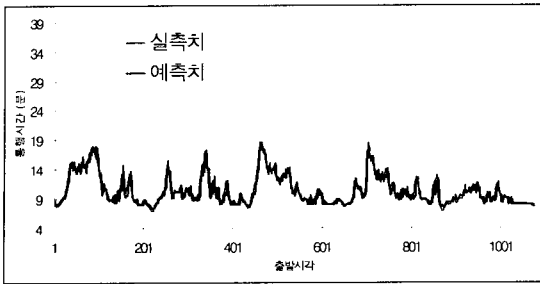


(a) 평일

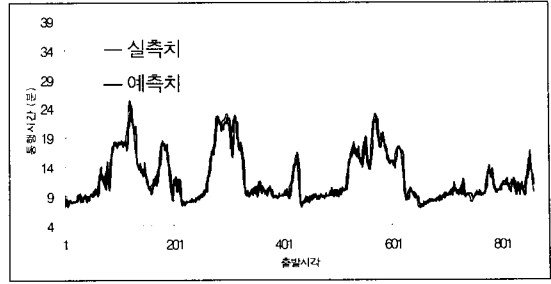


(b) 주말

<그림 6> 서울-대전 140분예측 결과



(a) 평일



(b) 주말

〈그림 7〉 서울-수원 20분 예측결과

〈표 2〉 예측모형의 오차 평가(서울-대전)

학습자료	80분		100분		110분		120분		140분	
	MARE	RRSE	MARE	RRSE	MARE	RRSE	MARE	RRSE	MARE	RRSE
평일	0.0361	0.0472	0.0361	0.0473	0.0366	0.0476	0.0368	0.0478	0.0373	0.0487
주말	0.0526	0.0680	0.0560	0.0715	0.0579	0.0738	0.0605	0.0767	0.0675	0.0862
평일/주말	0.0509	0.0703	0.0535	0.0753	0.0545	0.0780	0.0556	0.0807	0.0577	0.0839

최소와 최대예측범위만 학습을 해도 거의 동일한 예측력을 얻을 수 있었다는 것이다. 이는 중요한 의미를 갖는 결과로서 본 연구에서 개발된 예측모형이 동적으로 변하는 예측범위에 잘 적용한다는 것이다. 지금까지의 대부분의 시계열 예측모형은 예측범위가 고정적이며, 특히 신경망 시계열 예측모형의 경우, 하나의 예측범위에 대해서만 학습하는 구조를 유지해 왔다. 그러한 기존모형의 한계를 극복하고, 예측범위의 변화에 하나의 신경망을 가지고 우수한 예측력을 유지한다는 것은 본 모형의 중요한 장점이라고 할 수 있다.

〈표 3〉 예측모형의 오차 평가(서울-수원)

학습자료	10분		15분		20분	
	MARE	RRSE	MARE	RRSE	MARE	RRSE
평일	0.0370	0.0478	0.0514	0.0678	0.0612	0.0833
주말	0.0379	0.0496	0.0551	0.0735	0.0698	0.0967
평일/주말	0.0367	0.0474	0.0522	0.0696	0.0643	0.0884

VI. 모형의 평가 및 분석

1. 누적확률분포를 이용한 오차분석

예측오차의 누적확률분포를 살펴보면(〈표 4〉), 서울-대전구간의 경우, 평일에는 전체 예측치의 97%가, 주말에는 약 88%가 10분 이내의 오차를 보였다.

〈표 4〉 예측오차의 누적확률분포를 이용한 오차분석

구간	요일	예측 범위	최대 오차 (분)	오차 누적 분포 (%)			
				5분 이내	10분 이내	15분 이내	20분 이내
서울-대전	평일	80분	17	84	97	99	100
		140분	18	82	97	99	100
	주말	80분	23	69	88	96	99
		140분	37	61	78	90	95

구간	요일	예측 범위	최대오차(분)	오차 누적 분포(%)
				5분이내
서울-수원	평일	10분	10	99.8
		20분	6	99.9
	주말	10분	7	99.7
		20분	11	99.4

서울-수원구간의 경우, 평일/주말 거의 모두 5분 이내의 오차만을 보였으며, 그 중 최대 90~97%이내 가 2분 이내의 오차를 나타냈다. 여기서 최대오차는 다소 크게 나타나고 있음을 볼 수 있는데, 그 빈도가 희박함을 감안할 때, 크게 문제가 되지 않는다고 볼 수 있다. 다만, 실험데이터 범위내에서 예측실패가 어느 정도 크게 나타날 수 있는지를 보여주고자 한 것이다.

결론적으로 서울-대전구간의 경우, 평일과 주말에 다소 차이를 보이고 있지만, 평균적으로 대부분 90%이상의 예측 빈도에서 5~10분 이내의 오차를 가지며, 서울-수원구간의 경우, 99%이상이 5분 이내의 오차를 갖는다고 할 수 있다.

2. 실제치-예측치 변동 분석

실제치-예측치 변동그래프는 실제치의 변동과 실제치에 대한 예측치의 변동에 의해서 결정된 오차의 형태를 보여준다. <그림 8>에서 가로축은 실제치의 변동을 나타내고 세로축은 실제치에 대한 예측치의 변동을 나타낸다. 양의 45도 기울기를 갖는 점선은 '완전예측선(perfect forecast line)'으로서 이 선상의 점들은 실제치와 예측치간의 완벽한 일치성을 의미한다. 변동그래프에서는 6가지 형태의 오차를 찾아볼 수 있다. 즉, 완전예측선을 기준으로 A영역에 나타나는 점은 실제치와 예측치의 변동방향이 모두 양(+) 이나 실제치의 변동보다 예측치의 변동이 작아 과소 예측이 되었음을 보여주고, 반대로 B 영역에 있는 점들은 양(+)변동의 과대예측을 나타낸다. 또한 D와 E영역에 있는 점들은 실제치와 예측치가 같은 음(-)의 방향으로 변동하였으나 과대 또는 과소 추정되었음을 의미한다. C와 F에 존재하는 점들은 실제치와 예측치가 서로 반대방향으로 변동한 것을 나타낸다.

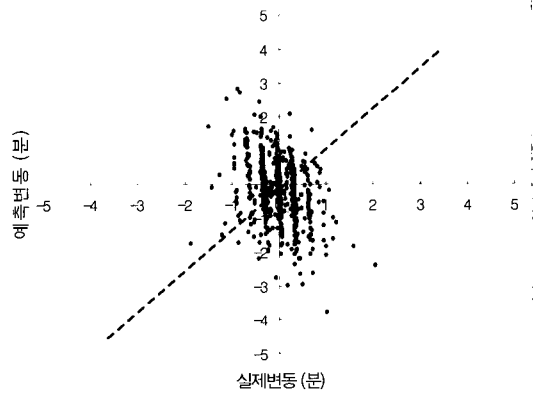
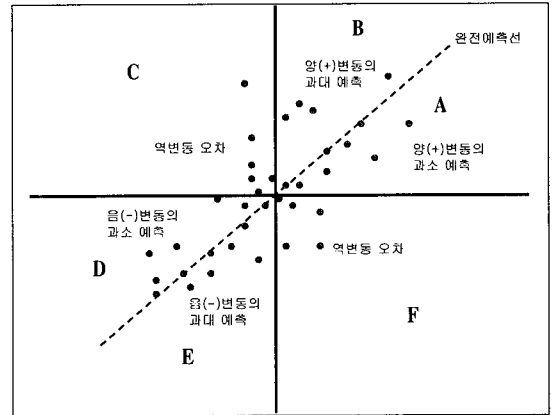
이와 같은 실제치-예측치 변동그래프를 통해 서울-대전과 서울-수원구간의 예측결과를 살펴보기로 한다. 여기서 분석은 상대적으로 오차가 큰 서울-대전 구간 140분 예측, 서울-수원구간 20분 예측결과를 평일과 주말로 나누어 수행하였다.

분석결과, <그림 8> (a)-(d)에서 나타난 바와 같이 서울-수원구간의 경우, 주로 중앙집중 되어 있음을 볼 수 있으며, 전체적인 형태는 2, 4상한에 보다 많이 분포하는 특성을 보이고 있다. 이는 예측치의 변동방향이 실제치의 변동방향을 잘 반영하지 못하고 있음을 나타낸다. 그러나 앞서 언급한 바 있듯이 그로 인하여 예측오차가 커졌다 작아졌다를 반복하는 형태를 갖지만, 대부분 그 범위가 3분 이내의 범위로 분포하고 있어 예측결과는 우수하다고 할 수 있다.

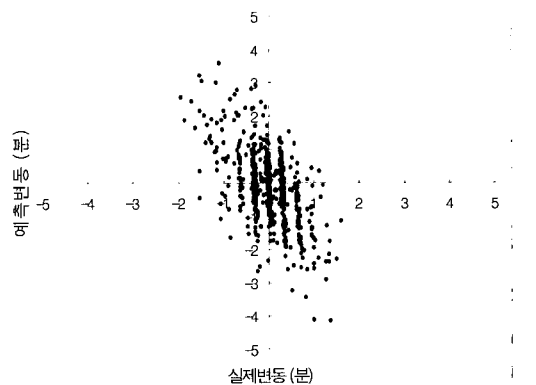
실제로 통행시간 예측에 있어 변동량의 차이가 크지 않다면, 변동방향의 일치여부는 그리 중요하지 않다.

다음으로 서울-대전구간의 경우, 평일과 주말 모두 세로축 주변으로 길쭉하게 분포하고 있으면서 분포범위가 평일은 10분, 주말은 약 15분의 범위내에서 대부분 나타나고 있음을 볼 수 있다. 다만, 주말의 경우에는 평일보다 더욱 세로축으로 길게 분포하고 있어 평일에 비해 예측력이 다소 떨어지고 있음을 보여

주고 있다. 그러나 그 차이는 5분~10분 이내로서 주말이 평일에 비해 혼잡의 영향으로 통행시간이 크다는 점을 감안할 때, 충분히 허용할 만한 차이를 보이고 있다고 판단된다.

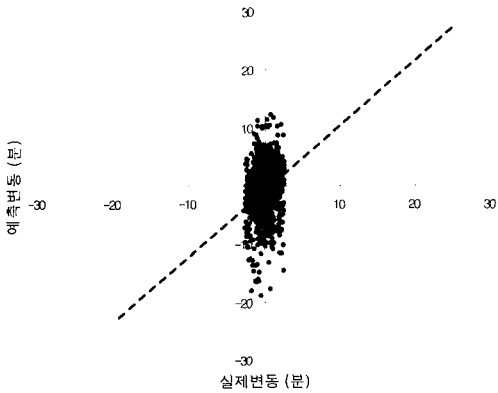


(a) 평일(서울-수원, 20분 예측)

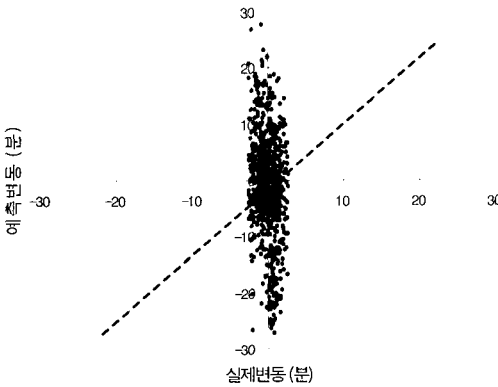


(b) 주말(서울-수원, 20분 예측)

<그림 8> 실제치-예측치 변동그래프를 통한 오차분석(계속)



(c) 평일(서울-대전, 140분 예측)



(d) 주말(서울-대전, 140분 예측)

<그림 8> 실제치-예측치 변동그래프를 통한 오차분석

Ⅶ. 결론

최근 교통정보의 필요성이 더욱 더 강조되는 한편, 정보의 정확도와 신뢰성에 대한 요구 또한 높아져 가고 있는 현실에서 이용자가 효과적인 통행의사결정을 하기 위한 진정한 가치를 지닌 '정보다운 정보'를 만들어내는 일이야말로 매우 중요한 일이다. 그러한 관점에서 가능한 한, 정확하고 충분히 대표성 있는 교통정보를 실시간으로 산출하고, 이를 토대로 신뢰성 있는 예측정보를 제공하는 기능은 교통관리시스템이 갖추어야 할 중요한 기능이다.

본 연구에서 개발한 시스템은 현장에서 활용될 것을 목적으로 한다. 따라서 개발과정에서 가장 중요시 다루어진 것은 "단순성"의 추구이다. 즉, 현장에 적용될 모형이나 알고리즘은 그 구조가 단순하고, 특히 실시간 데이터 수집환경에 적용이 용이해야 한다는 것이다. 본 연구에서는 다양한 실험을 통하여 개발된

모듈라(modular) 신경망 모형구조를 바탕으로, 최소한 약 80분 이상의 장기예측이 요구되는 서울-대전 구간에서 거의 대부분 10분 이내의 오차를 보이는 최적모형을 개발하였다. 특히, 대부분의 시계열 예측모형이 예측범위가 고정적인 반면에 개발모형은 하나의 신경망모형으로 학습한 범위내의 예측범위에 대해서 그 크기에 상관없이 거의 동일한 수준의 예측력을 보여준다는 점은 중요한 장점이라 할 수 있다.

참고문헌

1. 한국도로공사 도로연구소(2000), "고속도로 통행 시간 예측시스템 개발 연구".
2. Al-Deek, H. M., M. P. D'Angelo, and M. C. Wang(1998), "Travel Time Prediction with Non-Linear Time Series", Proceedings of the 5th International Conference on Application of Advanced Technologies in Transportation, April 26-29, Newport Beach, California, pp.317~324.
3. Alex Aussem(1998), "Training Dynamical Recurrent Neural Networks with the Temporal Recurrent Back-Propagation Algorithm: Application to Time Series Prediction and Characterization".
4. Brian L. Smith, and Michael J. Demetsky (1994), "Short-Term Traffic Flow Prediction: Neural Network Approach", Transport Research Record 1453, pp.98~104.
5. Dongjoo Park, and Laurence R. Rilett(1998), "Forecasting Multiple-Period Freeway Link Travel Times Using Modular Neural Network", Transport Research Record 1617, pp 163~1617, No. 98-0743.
6. Wei W. S.(1990), "Time Series Analysis", Addison Wesley.

✉ 주 작 성 자 : 남궁성
 ✉ 논문투고일 : 2002. 6. 29
 논문심사일 : 2002. 8. 3 (1차)
 2002. 8. 8 (2차)
 심사판정일 : 2002. 8. 8
 ✉ 반론접수기간 : 2002. 12. 31