응용통계연구 (2009)
22(6), 1177–1190

교통 통계 정보를 이용한 속도 패턴 예측에 관한 연구

최보승[1] · 강현철[2] · 이성건[3] · 한상태[4]

[1]고려대학교 경제연구소, [2]호서대학교 정보통계학과, [3]성신여자대학교 통계학과
[4]호서대학교 정보통계학과

(2009년 8월 접수, 2009년 10월 채택)

요 약

도로의 성능을 측정는데 있어서, 주행속도는 가장 중요한 정보가 된다. 또한 도로 교통의 정보를 제공하는데 있어서 현 시점의 교통정보와 더불어 향후 예측되는 교통정보를 함께 제공하는 것은 보다 정확한 예측 시간과 구간을 제공하기 위한 차별화된 기능이라 할 수 있다. 본 연구에서는 그 동안 축적된 도로 구간별 속도 자료를 이용하여 속도 패턴을 다양하게 분석하고 퓨리에 변환 및 삼각함수를 설명변수로 하는 시계열 회귀모형을 이용한 예측모형을 개발하여 구간별 및 시간대별 평균 속도를 예측하였다. 이와 더불어 보다 정확한 예측을 위하여 결측치에 대한 대체 방법 및 특이치 처리 방법을 함께 고려하였고 방대한 데이터에 대한 효율적인 분석을 위하여 유사 속도 구간에 대한 그룹핑(grouping) 방법도 제안하였다.

주요용어: 주행속도, 예측, 퓨리에 변환, 시계열 회귀모형, 속도 그룹핑(Grouping).

1. 서론

90년대 이후 산업발달과 생활수준의 향상은 자동차 수의 급격한 증가를 가져왔다. 자동차 수의 증가에 따라 도로의 확충도 함께 이루어 지고 있으나 도로의 확충 속도는 자동차 수의 증가를 따라가고 있지 못하는 실정이다. 그로 인하여 전국의 대부분의 도로들은 극심한 교통 정체를 겪고 있으며 이에 따른 도로의 평균 속도는 해마다 줄어들고 있는 실정이다. 도로의 속도는 도로의 성능을 측정하는 가장 중요한 요인 가운데 하나이며 차량의 주행 속도를 정확히 예측하는 것은 도로계획, 설계, 교통운영 및 안정성을 평가하는데 매우 유용한 수단이 될 수 있다 (이종필과 김성호, 2002).

최근 정보통신 산업의 발달과 더불어 도로의 주행속도와 밀접한 연관을 가지는 장치는 일명 네비게이션이라 불리우는 자동항법장치이다. 원래는 군사적인 목적으로 개발된 장치였으나 현재는 대부분 차량들이 사용하는 보편적인 장치라 할 수 있다. 네비게이션의 성능을 측정하는 기본적인 요인은 정확한 도로의 안내이지만 이와 더불어 네비게이션의 기능을 향상시키는 중요한 요인은 정확한 속도 예측을 반영한 길 안내라 할 수 있다.

주행속도의 정확한 정의는 “도로 구간별 설계속도에 의해 결정되는 안전속도를 초과하지 않고 양호한 기상조건하에서 운전자가 주어진 도로구간을 주행할 수 있는 최대속도”라 할 수 있다 (AASHTO, 1994). 본 연구의 주된 목적은 과거 일정 시간 동안에 측정된 도로의 주행속도 자료를 이용하여 향후 시점에서의 주행속도에 대한 보다 정확한 예측을 위한 통계적인 모형의 개발이다. 통계 모형을 이용하여

[4]교신저자: (336-795) 충청남도 아산시 배방면 세출리, 호서대학교 정보통계학과, 교수.
E-mail: sthan@hoseo.edu

표 **2.1.** 교통 통계 자료 정리

	서울시내자료	고속도로자료
총 구간 수	8,953개	1,016개
관찰시점	2005년 4월1일∼2007년 3월 31일	2006년 8월 1일∼2007년 10월 11일
속도측정단위	5분간격	15분간격 또는 10분 간격

주행속도를 예측하고자 하는 대부분의 연구들은 주로 선형 회귀모형을 이용한 연구방법 들이다. 이점호 등 (2006)은 외국의 사례를 중심으로 주행속도를 예측하는 모형을 제안하였다. 이점호 등 (2006)은 도로의 여러 기하적인 특성, 예를 들면 직선의 길이, 곡선의 길이, 곡선의 반경, 편각, 차로폭, 차선 등을 이용하여 적절한 변환을 수행한 후에 이를 설명변수로 하고 사전에 측정된 속도를 반응변수로 하는 회귀모형을 구축하였다. 이종필과 김성호 (2002)는 회귀모형이 가지는 제약과 한계를 극복하기 위하여 신경망 모형을 이용하여 주행속도를 예측하고자 하였다. 허태영 등 (2007)은 공간통계 기법을 이용하여 여러 지점간의 교통량을 예측하는 방법을 제안하였다.

본 연구에서 제안하고자 하는 주행속도 예측모형이 기존의 방법들과 가지는 가장 큰 차별점은 주행속도를 측정하는 자료에 있다. 기존에 제안된 방법들은 속도 이외의 다양한 기하적 변수들을 설명변수로 하여 주행속도 예측모형을 구축하였고 특정 구간의 자료를 적용하여 모형 구축을 수행하였다. 그러나 이러한 방법들은 다양한 도로의 특성을 일일이 반영하여 주행속도 예측 모형을 구축하는 방법으로 구축된 모형의 일반화 및 적용에 많은 어려움이 따른다. 특히 네비게이션과 같은 장비에 적용하기 위해서는 모든 도로에 따른 모형이 적용되어야 하고 이에 따라 방대한 추가정보가 필요하게 되고 분석을 위한 시간도 추가적으로 요구된다. 이에 본 연구에서는 상대적으로 단순하면서 추가적인 설명변수를 고려하지 않은 상황하에서의 주행속도 예측 모형을 구축하고자 하였다. 이를 위하여 속도 이외의 추가 변수를 고려하지 않은 상황하에서 상대적으로 단순한 모형을 적용하여 예측 모형을 구축하고자 하였다. 본 연구에서는 단위시간 동안 측정된 속도자료만을 이용하여 퓨리에 변환을 적용하고 삼각함수를 설명변수로 가지는 시계열 회귀모형을 이용하여 주행속도를 예측하는 모형을 구축하였다.

주행 속도 자료를 이용하여 주행속도 예측 모형을 구축하는데 있어서 추가적인 고려사항은 자료가 가지는 방대함이다. 속도 자료는 주어진 10km 이내의 일정 구간내에서 5분 내지 10분간격으로 중단없이 측정되는 시계열 자료이다. 이에 따라 전체 자료는 방대해 질 수 밖에 없으며 측정장비의 오작동이나 수리, 교체 등에 따라 빈번한 자료의 결측이 발생하게 된다. 그리고 갑작스러운 기상 상황의 변화와 교통사고 또는 도로 공사와 같은 상황에 의한 급격한 속도의 저하를 볼 수 있다. 본 연구에서는 이와 같은 속도자료가 가지는 결측치와 특이치 처리 방법을 함께 고려하고자 하였다. 이와 함께 보다 효율적인 모형 구축을 위하여 비슷한 속도 유형을 보이는 자료들을 군집화 하는 속도 그룹핑(grouping) 방법을 함께 제안하였다.

본 논문의 구성은 다음과 같다. 2절에서는 자료의 설명과 함께 자료의 정제 방법 및 평균속도 산출 방법에 대하여 설명하였다. 3절에서는 주행속도의 예측모형 구축 방법과 속도 그룹핑 방법에 대하여 설명하였고, 마지막 4절에서 결론을 통해 본 연구의 의의를 정리하였다.

2. 교통 통계 정보

2.1. 교통 통계 정보 자료의 특징

주행속도 예측을 위하여 이용된 자료는 서울 시내 구간별 속도 자료와 고속도로 구간별 속도 자료로 구분할 수 있다. 각 자료의 기본 사항은 표 2.1과 같다.

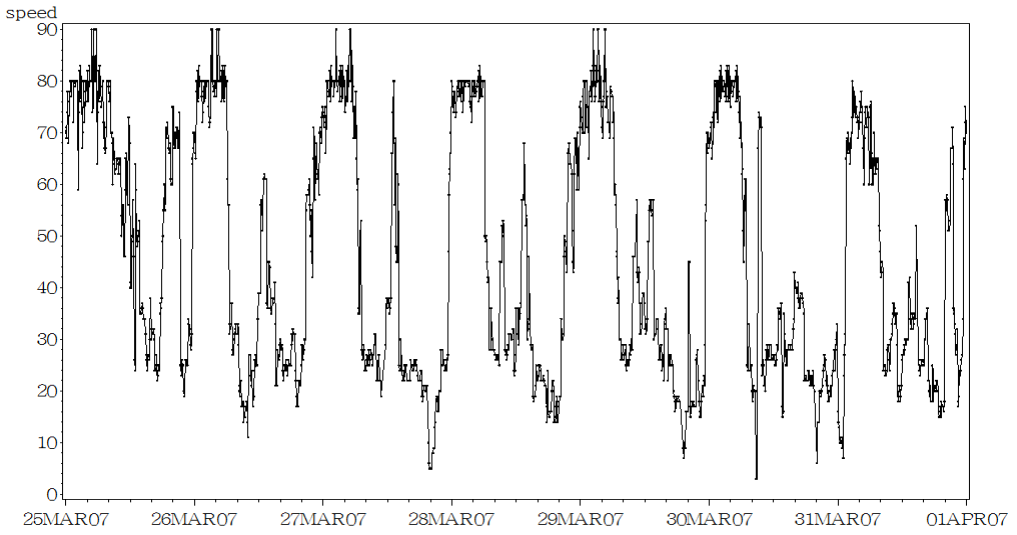

그림 2.1. 서부간선도로 상행 오금교~목동교 구간 시도표

교통정보 자료의 특징을 살펴보자. 그림 2.1은 서부간선도로 상행선 '오금교~목동교' 구간의 2007년 3월 25일 0시부터 2007년 3월 31일 오후 23시 55분까지 일주일간 5분 간격으로 측정된 속도에 대한 시도표 이다. 시도표를 살펴 보았을 때 하루 24시간을 단위로 하여 일정한 패턴을 보이고 있다. 새벽시간인 자정부터 오전 6시까지 전반적으로 가장 높은 속도를 보이며 한 낮 시간인 13시부터 14시 사이에 두 번째 속도의 피크를 보이고 있다. 간선도로 임에도 불구하고 그 외의 시간에서는 평균속도가 30km 이하로 극심한 정체를 보이고 있다고 할 수 있다. 그러나 조금 더 세밀하게 살펴본다면 속도의 변화는 매우 빠르게 변하고 있음을 확인할 수 있다. 새벽 시간대에서 출근시간인 오전 시간으로 옮겨가는 경우에 속도의 변화는 매우 극심하며 같은 시간 안에서도 5분간격으로 측정된 속도는 매우 급격하게 속도의 급락을 보이고 있다. 이와는 반대로 밤 11시 경부터는 다시 급격한 속도의 증가를 보이고 있다. 보다 자세하게 살펴보기 위하여 요일별로 속도의 흐름을 살펴보기로 하자. 그림 2.2는 같은 오금교~목동교 구간에서 토요일의 5분단위로 측정된 속도에 대한 시도표를 나타낸다. 새벽시간인 자정부터 약 70km 정도의 속도를 유지하고 있고 7시 까지 이러한 높은 속도가 유지 된다. 그리고 그 외의 시간에서는 평균 속도가 30km 이하를 보이고 있다. 그리고 오후에 약간의 속도 상승이 보이나 그 정도는 새벽시간에 미치지 못하고 있다 할 수 있다.

속도 자료가 가지는 또 하나의 특징이자 어려움은 빈번한 결측치의 발생과 다양한 특이치의 발생이다. 그림 2.2에서 15시 경 속도가 0km를 나타내고 있으나 이는 결측치가 발생한 경우이다. 이러한 결측치의 발생은 거의 모든 구간의 자료에서 빈번하게 발생하고 있다. 본 연구에서는 스플라인(spline) 기법을 이용한 결측치 대체를 수행하였다. 속도자료가 시간의 흐름에 따른 연속적인 자료임을 고려한다면 이러한 스플라인 기법의 도입은 타당하다고 할 수 있다.

다음으로 특이치 처리에 대한 문제를 살표보자. 그림 2.2의 토요일 시도표에서 자정부터 새벽 1시 30분까지 속도가 20km 이하이다. 일반적으로 새벽시간에는 70km 이상의 속도를 보이고 있으나 이와는 큰 차이를 보이고 있다. 도로공사나 사고와 같은 요인에 의하여 속도가 극격히 감소했다고 볼 수 있다. 이와 같이 속도의 특이치가 발생하는 경우 이에 의하여 예측된 속도가 영향을 받을 수 있다. 평균 속도를

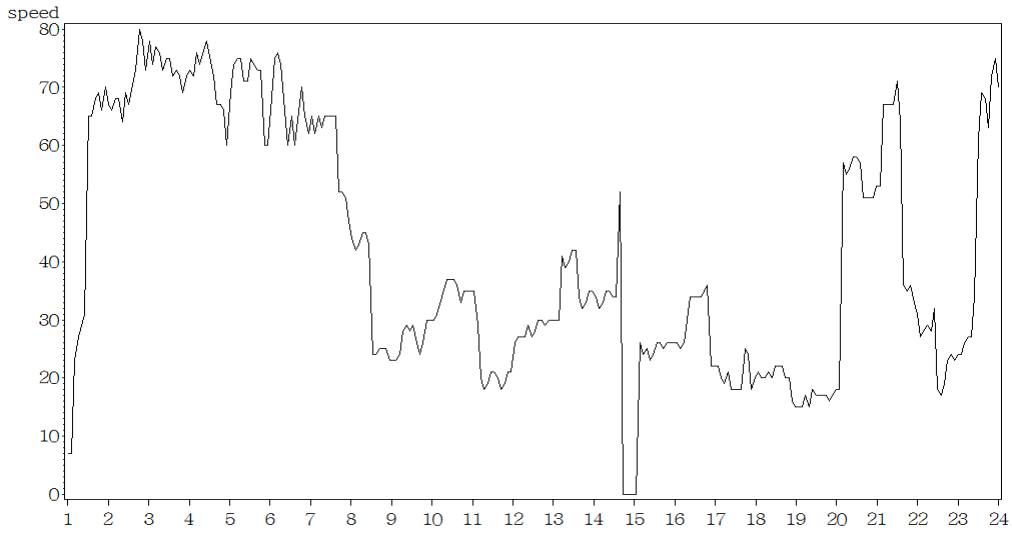

그림 2.2. 서부간선도로 상행 오금교~목동교 구간 토요일 시도표

이용하여 예측을 수행하는 경우 특이치에 의하여 평균 값의 쏠림현상이 발생할 수 있다. 그림 2.2의 경우와 같이 특이치가 발생하는 경우 평균속도가 실제평균보다 작은 값을 가지는 현상이 발생하게 된다. 이와 같은 현상을 줄이기 위하여 본 연구에서는 사분위 범위를 반복적으로 적용하여 특이치를 제거하는 방법을 이용하였다. 제거된 특이치는 다시 데이터 정제에서 사용된 스플라인 기법을 사용하여 데이터 대체를 수행하였다.

2.2. 데이터 정제 및 평균속도 산출

주행속도 예측 모형 구축을 위한 데이터 마트의 생성은 크게 3가지 단계로 나뉜다. 첫 번째는 스플라인 기법을 이용하여 결측치를 대체하는 단계이고, 두 번째는 각 속도자료를 비슷한 패턴을 보이는 요일별로 범주화하는 단계이다. 마지막 세번째 단계는 요일별로 범주화된 자료들의 특이치를 제거하여 요일 범주별 평균 속도를 계산하는 단계이다.

결측치 대체

5분 또는 10분 단위로 측정된 속도자료는 기본적으로 시간의 흐름에 따라 관찰된 시계열 자료이다. 이와 같은 시계열자료의 특성을 살려 예측을 위한 모형을 구축하기 위해서는 시간 흐름에 따라 중간 중간에 발생하는 결측치의 대체가 반드시 선행 되어야 한다. 본 연구에서는 이와 같은 결측치를 대체하기 위하여 스플라인 기법을 이용하였다. 스플라인은 구분적(piecewise) n차 다항식 모형이라 할 수 있다. 각 구간의 연결점을 넛(knot)이라 부르며 일반적인 다항식 모형은 넛이 없는 스플라인 모형의 특수한 경우라 할 수 있다. 구간에서 다항식의 차수와 넛의 수는 상황에 따라 다르다. 본 연구에서 사용된 스플라인은 3차 스플라인 곡선(cubic spline)이며 이는 점들을 보간하는 경우, 여러 개의 곡선 세그멘트(segment)를 이어서 복합곡선(composite curve)을 형성한다. 공간상의 곡선을 정의하는 최소 차수의 다항식은 3차곡선이므로, 여러 개의 허밋 커브(Hermite curve)를 이어서 적절한 복합 곡선을 만들 수 있다. 그러나 각 점의 접선 벡터값을 입력하는 것은 실제로는 매우 불편하므로 곡선의 중간 점에서

1차 및 2차 미분 연속조건을 사용하는 것이 보편적이다. 스플라인 함수는 하나의 점 x에서 다음과 같은 식을 통해 계산된다.

$$f(x) = a_i(x - x_i) + b_i(x - x_i) + c_i(x - x_i) + d_i,$$

여기서 세그먼트 i는 다음과 같이 선택된다.

$$i = \begin{cases} i, & x < x_1, \\ 1, & x_i \le x < x_{i+1}, \quad 1 \le i < n, \\ n, & x \ge x_n. \end{cases}$$

시계열자료의 중간에 결측 자료가 생기는 경우 스플라인 기법을 이용하여 대체를 수행하여 결측이 발생하는 시점에서의 주변 자료의 흐름에 따라 매끄러운 곡선의 형태로 결측자료를 대체하게 된다. 만약 결측이 발생되는 시점에서 사고와 같은 특정한 상황이 발생하지 않는다면 속도의 변화는 크지 않을 것이며 스플라인 기법을 사용하여 결측을 대체하는 것이 타당하다 할 수 있다.

속도패턴의 요일별 범주화

속도 자료는 기본적으로 요일에 따라 일정정도 유사한 패턴을 보인다. 각 구간별 속도를 예측하는데 있어서 요일별로 범주를 구성하고 요일별로 구분된 속도 예측 모형을 생성하는 것이 효율적이라 할 수 있다. 요일별 범주화에 함께 고려하여야 할 사항은 추석과 설과 같은 휴일효과의 반영이다. 우리나라의 문화적 특성상 추석및 설과 같은 공휴일에는 다른 날과 다른 교통흐름의 패턴을 보인다. 설과 추석을 전후하여 서울 시내의 교통흐름과 고속도로의 교통흐름이 매우 다르게 나타난다. 이를 반영하여 요일구분에 설과 추석을 별도의 명절로 범주화 하여 요일 범주에 추가 하였다. 우리나라에서는 추석과 설이 각각 3일 연휴로 구성되어 있는데 교통흐름을 함께 고려한다면 대부분의 경우에 명절연휴 전날과 명절연휴 이후에서도 명절의 교통흐름과 비슷한 패턴을 보이는 것을 알 수 있다. 따라서 이러한 효과는 추석과 설 연휴 기간의 바로 전날과 다음날도 그 효과가 미친다는 판단하에 추석과 설연휴에 함께 포함하였다. 다음으로 고려할 수 있는 상황은 기타 공휴일이다. 공휴일의 경우 평일이라 하더라도 일요일과 일정 정도 유사한 패턴을 보일 수 있다. 마지막으로 금요일 효과를 고려하였다. 주 5일제의 시행과 정착에 따라 금요일 저녁에는 모든 도로에서 극심한 정체상황을 보이고 있다. 이와 같은 금요일 효과는 다른 공휴일 전날에도 유사한 패턴을 보이고 있다. 이를 위하여 금요일 오후 6시부터 자정까지를 금요일 효과라 하고 일요일을 제외한 기타 공휴일 전날의 오후 6시부터 자정까지를 금요일과 같은 범주로 하여 금요일 효과를 반영하도록 하였다. 이상을 정리하여 전체 자료에 대한 구분은 일요일부터 토요일까지 7개의 범주에 추석과 설을 반영하는 명절, 마지막으로 기타 공휴일을 반영하는 공휴일까지 총 9개 범주로 구성하였으며 이 구성된 9개의 요일 범주에 따라 구간별 속도를 가지고 예측모형을 구축하였다.

특이치 제거와 평균속도 구축

AASHTO (1994)의 정의에 따라 주행속도는 정상적인 도로상황에서의 최고속도라 할 수 있다. 그러나 속도 자료는 비정상적인 기상상황이나 도로 공사, 사고 등에 의하여 일시적인 속도 저하를 보일 수 있으며 이러한 급격한 속도의 저하는 모두 특이치로 고려될 수 있다. 반대로 갑작스럽게 속도가 증가하는 경우도 발생할 수 있다. 이러한 예로써 금요일 오후 5시 정각 서울시내 서부간선도로 오목교~목동교 구간의 평균 속도를 계산한 결과 시속 28km로 계산되었으나 일부 자료가 60km 이상의 속도를 보였다. 이는 일반적인 특성과 다른 특이치로 고려할 수 있다. 이러한 특이치를 고려하지 않고 속도의 평균을 구하게 되면 이와 같은 특이치에 의하여 상대적으로 큰 평균을 보이게 되고 실제 운전자들이 느끼는 속도보다 큰 값을 가지게 된다. 이러한 특이치를 제거하기 위하여 다음과 같은 반복작업을 수행하였다.

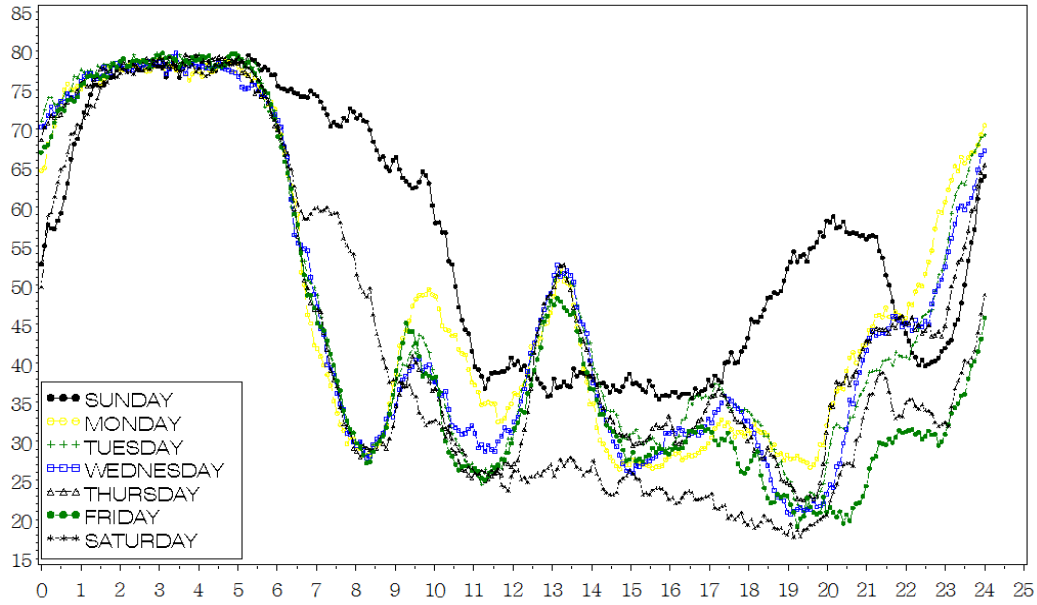

그림 2.3. 서부간선도로 상행 오금교~목동교 구간 요일별 평균 속도 시도표

단계 1: 원자료를 이용하여 평균을 계산하고 표본수를 계산하고 이를 n_b로 한다.

단계 2: 사분위범위(inter quartile range; IQR)를 계산하고 1사분위수 $-1.5 \times$ IQR 보다 작거나 3사분위수 $+1.5 \times$ IQR 보다 크면 특이치로 고려하여 제거한다. 특이치를 제거하고 표본수를 다시 계산하여 이를 n_a로 한다.

단계 3: 단계 1에서 단계 2를 반복적으로 수행하여 특이치 제거 이전과 특이치 제거 이후의 표본 수의 차이$(= n_b - n_a)$가 더이상 변동되지 않을 때까지 반복한다.

이와 같은 과정을 거쳐 특이치가 제거된 자료가 생성되면 이 자료를 가지고 구간별, 요일 범주별 그리고 측정된 시간대 별로 평균값을 계산한 후 주행속도 예측 모형을 위한 분석용 데이터마트를 구축하게 된다. 그림 2.3은 서울시내 서부간선도로 상행 오금교~목동교 구간에서 5분단위로 평균속도를 계산하여 요일별로 작성된 24시간 시도표이다. 시도표를 보면 출퇴근 시간과 같은 특정 시구간에 속도저하가 일어나며, 또한 속도가 회복되고 다시 정체되는 패턴이 반복적으로 일어남을 알 수 있다. 또한 요일별로 정상적인 근무일(월~금)과 토요일 그리고 일요일의 속도패턴이 뚜렷히 구분되는 것을 확인할 수 있다. 새벽시간의 속도는 출근시간이 시작되면서 급격하게 감소하고 있음을 볼 수 있다. 그러나 일요일과 토요일에는 이러한 속도의 감소가 상대적으로 완만함을 확인할 수 있다. 그리고 19시 이후 가장 작은 속도를 보이는 구간은 금요일으로써 저녁시간대의 금요일 효과를 확인할 수 있다. 표 3.2는 같은 구간에서 20시 정각의 자료로부터 특이치를 제거하고 계산된 통계량을 요일별로 정리한 표이다. 금요일의 경우 원자료의 수는 51개 였으나 특이치의 제거에 의하여 7개의 자료가 삭제되었으며 최종적으로 44개의 자료를 이용하여 통계량이 계산되었다. 요일별 평균값을 비교하였을 때 금요일의 평균속도가 다른 요일에 비하여 현저하게 작은 것을 확인할 수 있다. 토요일 역시 금요일 다음으로 작은 값을 보이고 있으며 일요일의 경우 매우 높은 평균 속도를 보이고 있음을 확인할 수 있다.

표 **3.1.** 오금교~목동교 구간 20시 정각 요일별 통계량

요일구분	평균	표준편차	최소값	최대값	변이계수	제거된 자료수	전체자료수
일요일	61.1	8.4	41.0	79.0	13.8	6	48
월요일	33.2	19.2	6.0	69.0	57.8	0	42
화요일	28.9	20.2	5.0	75.0	70.0	0	46
수요일	29.6	18.4	5.0	72.0	62.3	1	48
목요일	31.9	21.8	5.3	93.5	68.4	0	47
금요일	12.6	5.6	0.2	26.0	44.5	7	51
토요일	15.7	6.0	3.7	31.0	38.4	3	48

3. 주행속도 예측 모형 구축

3.1. 삼각함수를 이용한 시계열 회귀모형

통계 교통정보 데이터를 이용하여 도로구간별 평균 주행속도 예측 방법을 알아보자. 그림 2.3에서 살펴본 바와 같이 속도는 요일과 교통 상황에 따라 속도의 하강과 상승의 패턴이 반복적으로 나타나는 것을 확인할 수 있다. 그리고 도로 주행 중 자동차의 참(true) 속도는 연속적인 값이므로 그림 2.3과 같이 불규칙하게 나타나지 않고 부드러운(smooth) 곡선으로 표현되는 것이 타당할 것이다. 속도의 참 함수(true function)를 예측하기 위한 방법으로 퓨리에 변환을 고려할 수 있다. 퓨리에 변환은 삼각함수를 이용하여 함수를 표현하고자 제안된 방법이다. 따라서 평균 속도의 24시간 패턴이 삼각함수로 잘 적합될 수 있다 판단되므로 평균 속도 데이터에 퓨리에 변환 방법을 적용하고자 한다.

일반적인 퓨리에 급수를 이용한 밀도함수 추정에 있어서 함수 $f(x)$는 다음과 같이 표현된다.

$$f(x) = \sum_{k=-\infty}^{\infty} B_k \Phi_k(x),$$

여기서 $\Phi_k(x)$는 정규 직교의 조건을 만족하는 급수이고, $B_k = \int \Phi_k(x) f(x) dx$로 일반화된 퓨리에 계수들이다. 함수의 추정에 있어서 크게 두 가지 전략적인 방법이 이용될 수 있다. 첫 번째 방법은 정지규칙(stopping rule), 밀도 함수를 최적화 시키는 평활모수(smoothing parameter)를 찾는 방법으로 다음과 같은 추정함수로서 표현된다.

$$\hat{f}(x) = \sum_{k=-\infty}^{\infty} \hat{B}_k \Phi_k(x),$$

여기서 m은 평활 모수의 수이고, $\hat{B}_k$는 표본으로부터 추정된 계수로서 $\hat{B}_k = n^{-1} \sum \Phi k(x_i)$이고 B_k의 불편추정량이 된다. Tarter와 Kronmal (1976), Hart (1985), Diggle과 Hall (1986) 등에 정지규칙에 관한 다양한 연구결과들이 제시되어 있다.

두 번째 방법은 함수에 승수(multiplier), b_k라고 하는 가중모수를 곱하여 함수를 최적화 시키는 b_k를 찾는 전략이다. 즉 승수 b_k는 k가 증가함에 따라 '0'에 수렴함으로써 함수를 추정할 수 있다. 추정된 함수의 표현식은 다음과 같다.

$$f^*(x) = \sum_{k=-\infty}^{\infty} b_k \hat{B}_k \Phi_k(x), \quad 0 < b \leq 1.$$

이에 대한 연구는 Watson (1969), Fellner (1974), Wahba (1981) 등에 의해 연구되어져 왔다. 이러한 추정치들의 정확도를 평가하기 위한 판단 기준으로 평균 누적제곱오차, $R = E[\hat{f}(x) - f(x)]^2$가 보편적으로 많이 사용된다.

표 **3.2.** 주기도 검정결과

주기	$df1$	$df2$	F-통계량	p-value
288	2	285	287.36	$p < 0.0001$
144	2	285	1.03	$p < 0.0001$
48	2	285	4.89	0.0082

다음으로 정지 규칙에 의한 밀도함수의 추정 방법에 대하여 알아보자. 밀도함수를 코사인(cosine) 퓨리에 급수를 통하여 다음과 같이 정의하자.

$$f(x) = \sum_{k=-\infty}^{\infty} a_k \Phi_k(x), \quad 0 < x < 1, \tag{3.1}$$

여기서 $a_k = \int_0^1 \Phi_k(x) f(x) ds$이고 $\Phi_k(x) = \sqrt{2}\cos(k\pi x)$, $k = 1, 2, 3, \ldots$라고 할 때 $[\Phi_k(x)]_{k=0}^{\infty}$는 $L^2(0,1)$에서 정규직교 기저(orthonormal basis)이다. 이러한 코사인 수열의 사용은 계산을 간편하게 할 뿐만 아니라 곡선의 변동을 잘 관찰할 수 있는 이점이 있다. 식 (3.1)에서 a_k의 불편추정량 $\hat{a}_{kn} = (1/n)\sum_{i=}^{n} \sqrt{2}\cos(k\pi x_i)$를 대입하여 함수 추정량 $\hat{f}_m$을 다음과 같이 얻을 수 있다.

$$\hat{f}_m = \sum_{k=-\infty}^{m} \hat{a}_{kn} \sqrt{2}\cos(k\pi n),$$

여기서 m은 절단점이다. 정지규칙에 있어서 최적화된 함수를 찾는 방법은 함수 f와 추정량 $\hat{f}_m$에 대한 차이를 최소화하는 m을 선택하는 것이다. m에 대한 결정방법 가운데 하나로써 주기와 빈도를 이용한 스펙트럼 분석을 수행한 후에 주기도 검정을 수행한 후 유의미한 주기를 찾는 방법을 이용할 수 있다. 구체적인 방법은 전새봄 등 (2009)과 Wei (2006)을 참조할 수 있다. 본 연구에서는 보다 간단한 방법을 이용하여 결정하고자 하였다. 먼저 예측 모형 구축을 위한 통계적인 모형으로 퓨리에 변환을 설명변수로 적용할 수 있도록 삼각함수를 설명변수로 하는 시계열 회귀모형을 이용하였다. 일일 24시간을 5분단위로 계산된 전체 288개 시구간에 대하여 총 15개의 주기를 삼각함수를 통하여 반영하도록 하였다. 5분단위 속도를 z_t라 하고 전체 구간 길이 288을 L이라 할 때 함각함수를 설명변수로 하는 시계열 회귀모형은 다음과 같다.

$$\begin{aligned} z_t = \beta_0 &+ \beta_1 \sin\left(\frac{2\pi t}{L}\right) + \beta_2 \cos\left(\frac{2\pi t}{L}\right) + \beta_3 \sin\left(\frac{4\pi t}{L}\right) + \beta_4 \cos\left(\frac{4\pi t}{L}\right) \\ &+ \cdots + \beta_{29} \sin\left(\frac{30\pi t}{L}\right) + \beta_{30} \cos\left(\frac{30\pi t}{L}\right) + \epsilon_t, \quad t = 1, 2, \ldots, 288. \end{aligned} \tag{3.2}$$

위의 식 (3.2)를 기본 모형으로 하여 각 도로 구간별 모형을 적용하여 m을 결정하였다. 본 연구에서는 m을 결정하기 위하여 두 가지 방법을 고려하였다. 첫 번째 방법은 전새봄 등 (2009)에서 제시된 방법으로 식 (3.2)에 스펙트럼 분석을 수행하여 주기도를 계산한다. 모형에서 고려하고 있는 전체 15개의 주기별로 주기도를 계산한 후 각 주기별로 주기도에 대한 유의성 검정을 위하여 F-검정을 수행한 후 유의한 주기만을 고려하여 m을 찾는 것이다. 다음 표 3.2은 서부간선도로 상행 오목교~목동교 구간의 분석결과로써 전체 주기 가운데 유의수준 5%에서 유의한 결과만을 정리한 것이다. 제1주기인 288, 제2주기인 144, 제6주기인 48에서 유의한 주기를 볼 수 있다.

다음으로 식 (3.2)에 대한 회귀분석을 수행한 후 회귀모형의 변수 선택 방법 가운데 하나인 후향소거법(backward elimination)을 이용하여 m을 결정하는 것이다. 다음 표 3.3는 서부간선도로 상행 목동

표 **3.3.** 추정된 회귀계수

Variable	Estimate	Standard Err.	t-value
β_0	46.3141	0.0560	827.18****
β_1	16.8543	0.0792	212.85****
β_2	16.0125	0.0792	202.22****
β_3	14.3158	0.0792	180.79****
β_4	3.2429	0.0792	40.95****
β_5	−1.7769	0.0792	−22.44****
β_6	−2.0543	0.0792	−25.94****
β_7	2.9052	0.0792	36.69****
β_8	−0.8947	0.0792	−11.30****
β_9	1.8165	0.0792	22.94****
β_{10}	−0.3586	0.0792	−4.53****
β_{11}	4.2644	0.0792	53.85****
β_{12}	−2.9835	0.0792	−37.68****
β_{13}	2.1136	0.0792	26.69****
β_{14}	−1.6832	0.0792	−21.26****
β_{15}	1.4457	0.0792	18.26****
β_{16}	0.6192	0.0792	7.82****
β_{17}	0.9156	0.0792	11.56****
β_{18}	−0.8919	0.0792	−11.26****
β_{25}	0.6149	0.0792	7.76****
β_{26}	0.6182	0.0792	7.81****
β_{27}	−0.4407	0.0792	−5.57****
β_{28}	−0.2520	0.0792	−3.18***

$* < 0.05$, $** < 0.01$, $*** < 0.001$, $**** < 0.0001$

교∼오목교 구간 자료를 이용하여 단계적 방법에 의하여 m을 결정한 후 최종적으로 적합된 모형에 대한 적합결과이다. 이 경우에도 유의수준은 5%로 하였다. 전체 15개의 주기 가운데 제10, 제11, 제12, 제15주기를 제외한 모든 주기가 모형에 선택되었다.

표 3.2과 3.3의 결과를 비교하여 보면 회귀분석의 후향소거법에 의하여 m을 결정하는 것이 주기도의 검정에 의하여 선택하는 것보다 항상 많은 변수를 선택한다. 그리고 주기도 검정에서 유의한 주기는 후향소거법에서도 가장 유의한 변수로 선택되어 진다. 모형의 간결성의 측면에서 보자면 주기도 검정에 의하여 m을 결정하는 것이 효율적일 수 있다. 그러나 실제 컴퓨터를 이용하여 계산하는 입장에서 본다면 단계적 방법에 의한 변수 선택이 보다 효율적이다. 주기도 검정을 위해서는 스펙트럼 분석, 주기도 검정, 회귀분석을 거쳐 결정이 이루어 지지만 후향소거법은 회귀분석 한번에 의하여 결정되어 진다. 이에 최종 모형은 삼각함수들을 후향소거법을 통한 통계적으로 유의한 변수들만을 포함하는 모형으로 결정하였다. 이와 같은 과정을 전체 9,969개 도로간에 적용하여 최종 모형을 구축하게 된다.

그림 3.1은 표 3.3에 의한 결과로써 서부간선도로 상행선 오목교∼목동교 구간의 화요일 평균속도 자료를 이용하여 예측 모형을 구축하고 실제 평균 속도와 예측된 속도를 표시한 시도표이다. 시도표에서 점으로 표시된 부분이 5분단위로 계산된 평균 속도이고 곡선으로 표시된 부분이 예측 모형에 의하여 적합된 결과이다. 예측 모형이 원자료를 잘 설명하고 있음을 볼 수 있으며, 적합된 회귀 모형의 결정계수는 약 99.7%이다.

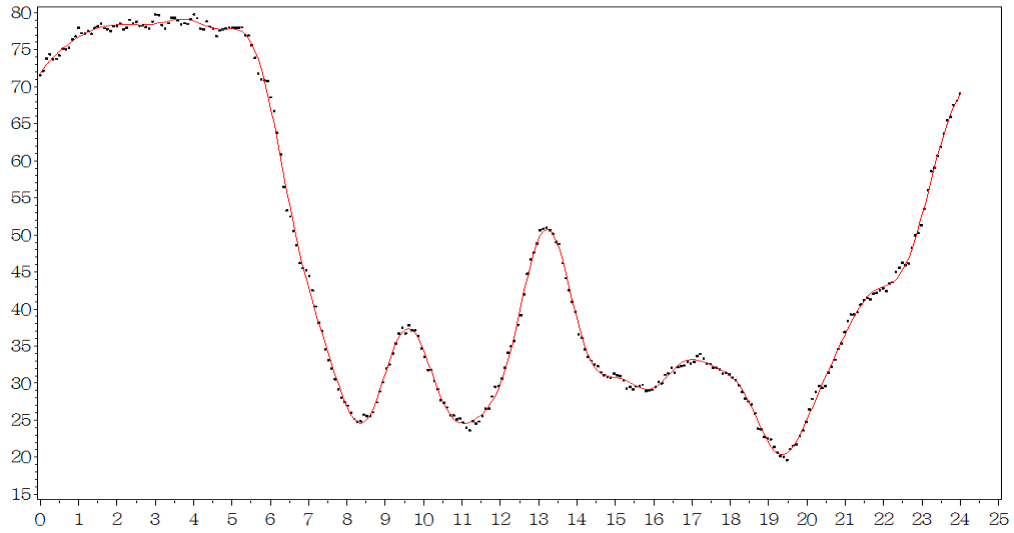

그림 3.1. 예측값의 시도표: 서부간선 상행 오목교~목동교 (화요일)

3.2. 예측 모형 구축을 위한 과거자료 구간 설정

평균 속도 예측을 위해 분석 시점에 사용 가능한 자료로서 본 연구에서 이용 가능한 자료는 표 2.1과 같이 서울시내 교통자료 2년, 고속도로 교통자료 1년치 이다. 여기서 새로운 도로의 개통, 신도시의 건설 등과 같이 도로의 환경이 급격히 변화하는 우리나라의 상황을 고려할 때 과거 축적된 데이터에서 가능한 최신의 데이터를 활용하는 것이 타당하다. 그러나 도로의 상황이 자주 변한다 하여 현재 시점에서 너무 가까운 최근 데이터만을 사용한다면 가용할 수 있는 데이터 수가 적어짐으로 인한 예측 결과의 통계적 유의성을 담보하기 어려운 문제가 발생한다. 이에 적절한 데이터 기간을 결정하는 것은 매우 중요한 문제 가운데 하나라 할 수 있다. 본 연구에서는 평균 속도 예측에 사용될 적절한 데이터 기간을 결정하기 위한 모의실험을 다음과 같이 설계하였다.

검증데이터(1개월):

2007년 3월 1일~2007년 3월 31일 사이의 평균 속도

모의실험 분석데이터:

1. 최근 3개월: 2006년 12월 1일~2007년 2월 28일
2. 최근 6개월: 2006년 9월 1일~2007년 2월 28일
3. 최근 1년: 2006년 3월 1일~2007년 2월 28일
4. 최근 2년: 2005년 4월 1일~2007년 2월 28일(23개월)

모의실험 도로구간:

100개의 도로구간을 임의 추출함(random sampling)

적중의 비교기준:

1. 95% 신뢰구간 Hit 비율: 예측 평균의 95%신뢰구간 포함 여부

표 **3.4.** 모의실험 결과 - 각 비교 기준별 적중률

데이터기간	변수	평균	중위수	최소값	최대값
최근2년	MAE	0.138	0.121	0.031	0.416
	95%신뢰구간	**0.957**	0.994	0.583	1.000
	±5km	0.718	0.768	0.270	0.956
	±10km	0.914	0.974	0.460	1.000
최근1년	MAE	**0.134**	0.124	0.031	0.417
	95%신뢰구간	0.951	0.990	0.581	1.000
	±5km	**0.732**	0.776	0.273	0.957
	±10km	**0.919**	0.976	0.457	1.000
최근6개월	MAE	0.135	0.124	0.031	0.437
	95%신뢰구간	0.938	0.979	0.455	0.997
	±5km	0.725	0.772	0.269	0.948
	±10km	0.916	0.974	0.433	1.000
최근3개월	MAE	0.147	0.134	0.033	0.498
	95%신뢰구간	0.911	0.953	0.442	0.993
	±5km	0.698	0.736	0.131	0.936
	±10km	0.904	0.963	0.345	1.000

2. ±5km Hit 비율: 예측평균의 ±5km 이내 포함 여부
3. ±10km Hit 비율: 예측평균의 ±10km 이내 포함 여부
4. MAE: Mean Absolute Error

모의실험은 2007년 3월 1일~2007년 3월 31일 사이의 평균속도를 검증데이터로 하여 각각 최근 3개월, 6개월, 12개월, 1년, 2년을 데이터 기간을 기반으로 개발된 모형의 적중률을 비교하여 보았다. 적중률은 4개 기준을 이용하여 비교 하였으며 도로구간의 표본으로 100개를 단순임의추출(SRS; simple random sampling)방법을 이용하여 추출하였다. 모의실험 결과는 표 3.4에 나타나 있다. 표 3.4에서 확인할 수 있듯이 각 4개의 구간을 가지고 예측모형을 구축하였을 때 검증자료에 대한 적중률은 최근 2년의 자료를 이용하였을 때가 95.7%로 가장 높음을 볼 수 있으며 다른 기준을 적용하였을 때는 모두 과거 1년의 자료를 이용하였을 때 가장 높은 적중률을 보였다. 이 결과를 바탕으로 평균속도 예측을 위하여 최근 1년 데이터를 이용하는 것이 가장 적절하다고 결론을 지울 수 있었으며, 실제 각 구간에 대한 모형구축을 위하여 모두 과거 1년 데이터를 이용하여 예측 모형을 구축하게 되었다.

3.3. 속도 그룹핑

통계 교통정보 데이터는 최소 5분단위로 속도값이 측정되어 있으며 이를 바탕으로 평균속도를 예측하였다. 또한 회귀모형을 이용한 예측값은 1분 이하에 대한 속도값도 산출이 가능하다. 그러나 속도의 편차가 거의 없는 구간에 대해 5분 단위 이하의 속도예측은 효율적이지 않으므로 이에 대한 보정이 필요하다 할 수 있다. 예를 들어 서울시내의 경우 낮에 비해 심야에는 속도 편차가 크지 않으므로 이를 30분 혹은 1시간 이상의 단위로 예측하는 것이 타당하며 더욱이 소통이 원활한 고속도로의 경우에는 24시간 거의 같은 속도값을 갖는 경우가 대부분이다. 이에 편차가 작은 시구간을 그룹핑함으로써 속도 예측값 활용의 효율화를 높일수 있다. 그룹핑의 과정은 속도 편차가 작은 구간만을 찾아내어 그룹핑을 수행하고 이를 최종 예측치에 반영한다. 속도 그룹핑 과정은 다음과 같다. 속도 그룹핑은 회귀모형을 통해 예측된 속도 예측값에 대해, 먼저 하루 24시간의 최대 속도 변화를 관찰하여 그 변화값이 5km 미만이면 전

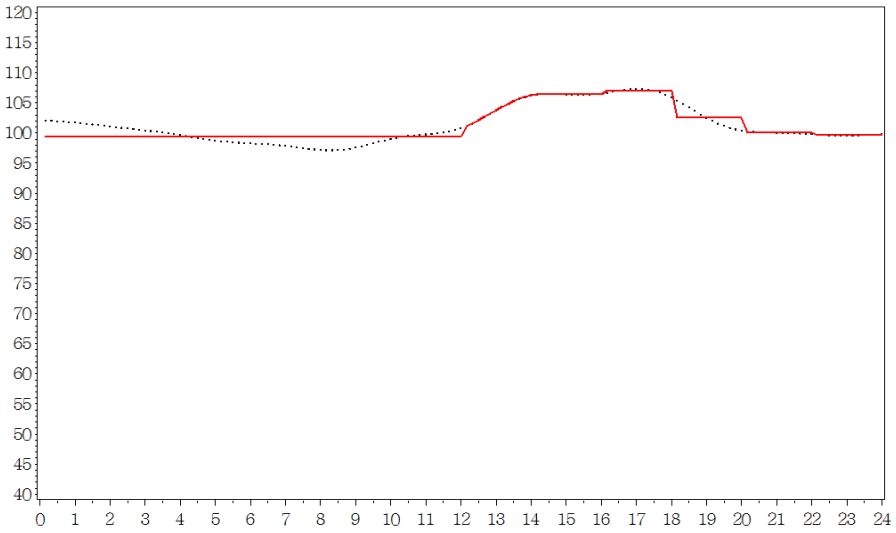

그림 3.2. 속도 그룹핑 결과: 고속도로 구간

체 데이터에 대해 하루 평균값으로 대체하고, 5km 이상이면 12시간 중 최대 속도변화를 관찰하여 변화값이 5km 미만이면 12시간 평균으로 대체하고 그렇지 않으면 6시간의 속도변화를 관찰한다. 이러한 단계를 1시간 중 최대 속도 변화량까지 관찰한 후 변화값이 5km 이상이면 원래의 예측값을 사용하고 그렇지 않으면 1시간의 평균값으로 대체하는 것이다. 이 과정을 적용한 예제 구간이 그림 3.2에 나타나 있다. 이 구간은 경부고속도로 하행선 구간으로 하루 중 속도 변화가 거의 없는 고속도로 구간에 대한 예이다. 그림에서 점선으로 표시된 부분이 10분 간격으로 측정된 실제 주행속도이고 실선으로 표시된 부분이 그룹핑의 결과를 나타낸다. 실선이 올라가거나 내려가는 변화 없이 시간 축(가로축)에 따라 평행하게 연결되어 있는 부분이 같은 속도로 그룹핑 된 부분으로 볼 수 있다. 이러한 구간에서는 5분 또는 10분 단위의 예측값을 제공할 필요는 없을 것이다.

4. 결론

본 연구는 우리나라 도로의 주행속도 패턴을 분석하여 주행속도에 대한 최적의 예측 모형을 구축하였다. 8,953개의 5분 단위 서울 시내 도로 속도 정보와 1,016개의 고속도로 구간에 대한 예측을 수행하였다. 평균 속도의 예측을 위한 분석에 앞서 오류 데이터 유무를 기초 통계 분석을 통하여 탐색하고 외부 정보를 전혀 고려하지 않고 순수한 속도 데이터로만 분석이 진행되었으므로 외부 환경 요인에 영향을 받은 속도값은 특이값 탐색을 통하여 가능한 배제하였다. 이는 본 연구에서 고려하고 있는 요일에 대한 순수한 효과만이 예측 결과에 반영되는 것이 타당하기 때문으로 판단된다. 또한 공휴일 및 명절을 하나의 다른 요인으로 고려함으로써 예측력의 향상을 꾀하였다. 요일 이외에 속도에 영향을 줄 수 있는 다른 요인으로 월/해뜨는 시간/노는 토요일 등도 함께 고려해 보았으나 유의한 결과는 얻을 수 없었다. 평균 속도 예측 분석에 적합하게 데이터 마트를 구축한 후에는 모의실험 결과 예측력이 가장 좋다고 판단되는 최근 1년치의 데이터를 이용하여 요일별 5분 또는 10분 단위 속도에 대한 평균값을 산출하였다. 평균값에 대한 분석을 통해 요일별로 속도의 패턴이 다름을 알 수 있었고, 따라서 요일별로 평균 속도값 예측이 이루어졌다. 그러나 산출된 평균값은 바로 예측값으로 사용되지 않았다. 속도 패턴의 참 함수(true

function)를 산출하기 위한 기초 자료로써 이러한 평균값이 사용되었고 이를 이용하여 통계적 모형을 구축하였다. 통계적 방법론은 퓨리에 변환과 삼감함수를 설명변수로 하는 시계열 회귀 모형이 사용되었다. 자료의 방대함과 분석의 처리속도를 고려하여 상대적으로 모형 구축이 쉽고 분석시간이 빠른 방법으로 이용하고자 하였다. 모형 적합 결과 우수한 설명력을 보였으며 평균값을 그대로 사용하는 것보다 활용성과 효율성 관점에서 많은 장점이 있음을 보았다. 통계적 예측 모형을 통해 계산된 속도 예측값은 1분 이하 단위의 속도값도 산출할 수 있으나 속도의 변화가 작은 시구간에 대해서는 효율성이 떨어지므로 비슷한 속도값을 갖는 시간 단위에 대해 적절한 그룹핑을 수행하여 예측의 효율성 증대를 꾀하였다.

현재 수행된 연구의 결과 보다 향상된 예측을 위해서는, 도로의 외부 환경요인 정보, 즉 교통사고 정보, 공사 정보, 기상 정보 등이 중요 요인으로 함께 고려되어야 할 것이다. 또한 도로 정체 상황의 전파 패턴을 분석함으로써 정체구간과 주변 구간에 대한 속도예측 모형도 별도로 고려해 볼 수 있을 것이다.

참고문헌

이점호, 홍다희, 이수범 (2006). 고속도록 교통운영 특성 및 도로선형요소를 반영한 주행속도 예측모형 개발, <대한교통학회지>, **20**, 131–139.

이종필, 김성호 (2002). 주행속도 예측을 위한 모형 개발(2차로 지방부 도로를 중심으로), <대한교통학회지>, **24**, 109–121.

전새봄, 최보승, 김성용 (2009). 스펙트럼 분석을 이용한 시계열 자료의 패턴 분류, *Journal of the Korean Data Analysis Society*, **11**, 349–359.

허태영, 박만식, 엄진기, 오주삼 (2007). 최단경로 기반 교통량 공간 예측에 관한 연구, <응용통계연구>, **20**, 459–473.

AASHTO (1994). *A Polcy on Geometric Design of Highway and Streets*, Washington, D.C.

Diggle, P. J. and Hall, P. (1986). The selection of terms in an orthogonal series density estimator, *Journal of the American Statistical Associaton*, **81**, 203–233.

Fellner, W. H. (1974). Heuristic estimation of probability densities, *Biometrika*, **61**, 485–492.

Hart, J. D. (1985). On the choice of a truncation point in fourier series density estimation, *Journal of Statistical Computation and Simulation*, **21**, 95–116.

Tarter, M. E. and Kronmal, R. A. (1976). An introduction to the implementation and theory of nonparametric density estimation, *The American Statistician*, **30**, 105–112.

Wahba, G. (1981). Spline interpolation and smoothing on the sphere, *SIAM Journal of Scientic and Statistical Computing*, **2**, 5–16.

Watson, G. S. (1969). Density estimation by orthogonal series, *The Annals of Mathematical Statistics*, **40**, 1496–1499.

Wei, W. S. (2006). *Time Series Analysis*, 2nd Ed., Pearson Education, the United States of America.

A Study for Traffic Forecasting Using Traffic Statistic Information

Boseung Choi[1] · Hyuncheol Kang[2] · Seong-Keon Lee[3] · Sang Tae Han[4]

[1]Institute of Economics, Korea University
[2]Department Informational Statistics, Hoseo University
[3]Department of Statistics, Sungshin Women's University
[4]Department Informational Statistics, Hoseo University

(Received August 2009; accepted October 2009)

Abstract

The traffic operating speed is one of important information to measure a road capacity. When we supply the information of the road of high traffic by using navigation, offering the present traffic information and the forecasted future information are the outstanding functions to serve the more accurate expected times and intervals. In this study, we proposed the traffic speed forecasting model using the accumulated traffic speed data of the road and highway and forecasted the average speed for each the road and high interval and each time interval using Fourier transformation and time series regression model with trigonometrical function. We also propose the proper method of missing data imputation and treatment for the outliers to raise an accuracy of the traffic speed forecasting and the speed grouping method for which data have similar traffic speed pattern to increase an efficiency of analysis.

Keywords: Traffic operating speed, forecasting, fourier transformation, time series regression, grouping.

[4]Corresponding author: Professor, Department Informational Statistics, Hoseo University, Asan 336-795, Korea. E-mail: sthan@hoseo.edu