이 성 건 교수지도

석사학위청구논문

통계적 방법을 활용한 교통속도 패턴에 관한 연구

2009

성신여자대학교 대학원

통 계 학 과

박 애 란

통계적 방법을 활용한
교통속도 패턴에 관한 연구

이 성 건 교수지도

이 논문을 석사학위논문으로 제출함

2009년 5월

성신여자대학교 대학원

통 계 학 과

박 애 란

인 준 서

박 애란의 석사학위 논문으로 인준함.

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

성신여자대학교 대학원

논 문 개 요

오늘날 우리나라의 도로이용자들에게 급증하는 교통 수요로 인한 실생활의 불편이 초래되고 있다. 현재 이러한 도로 교통 상황에 대해서 각 분야에서 축척해 온 도로 속도 정보를 이용하여 도로이용자들에게 보다 유익한 정보를 제공하고자 많은 연구 및 개발이 시도되고 있다.

본 논문은 서울시내 올림픽대로와 강변북로 2005년-2007년의 속도자료를 통계적 분석에 이용하여 교통 정보를 제공하고자 한다. 먼저 각 시구간에 대한 평균속도 예측모형을 개발하고, 그 다음으로 각 구간의 평균속도 패턴이 유사한 것들끼리 군집화하는 군집분석을 수행하였다.

속도예측을 위한 통계모형구축을 위해서 본 논문에서는 스플라인 회귀모형을 제안한 후, 기존의 연구된 퓨리에 급수를 이용한 회귀모형과 결정계수를 이용하여 비교하였다.

군집분석은 k-평균 군집화를 통해 24시간 동안의 교통속도 패턴이 비슷한 군집을 구축하였다. 와드(Ward)의 방법으로 나온 결과를 이용하여 군집수와 초기값을 정하고, k-평균 군집분석을 실시한 뒤 도로구간의 각 군집별의 패턴을 그래프로 나타내어 보았다. 생성된 군집의 특성분석을 통해 전반적인 교통속도 패턴을 이해하는데 도움을 주고자 한다.

목 차

제1장 서 론

오늘날 우리나라의 도로이용자들은 급증하는 교통 수요에 따른 도로지체시간의 증가로 인하여 불만이 고조되고 있다. 특히 서울시내 도로는 서울시 및 수도권 내에 인구밀도가 높아짐에 따라 정체가 심화되고 있고, 주5일제 실시 이후 여가를 즐기기 위한 도로 이용자가 늘어남에 따라 정체시간이 길어지고 있는 실정이다.

1.1 연구배경 및 목적

현재 이러한 도로 교통 상황에 대해서 각 분야에서는 축척해 온 도로 속도 정보를 이용하여 도로이용자들에게 보다 나은 정보를 제공하고자 많은 연구 및 개발을 시도하고 있다. 길찾기 서비스를 제공하고 있는 네비게이션(navigation)이 그 한 예라고 할 수 있다. 현재 상용화되고 있는 네비게이션에서의 도로 속도 정보는 단순히 실시간 교통상황을 보여주는 것에 그치지만, 알고자 하는 교통정보에 통계적 분석 방법을 이용한다면 좀 더 유익한 정보를 제공할 수 있을 것이다.

본 연구의 목적은 도로교통상황에서의 돌발상황 즉 공사, 사고, 기후등에 영향을 받지 않는 순수한 속도자료만으로 도로구간별 시간대별 평균속도를 예측할 수 있는 통계적 모형을 구축하고, 교통 속도의 패턴이 비슷한 군집으로 묶은 후 유사한 패턴을 갖는 구간의 특징을 살펴보고자 함이다.

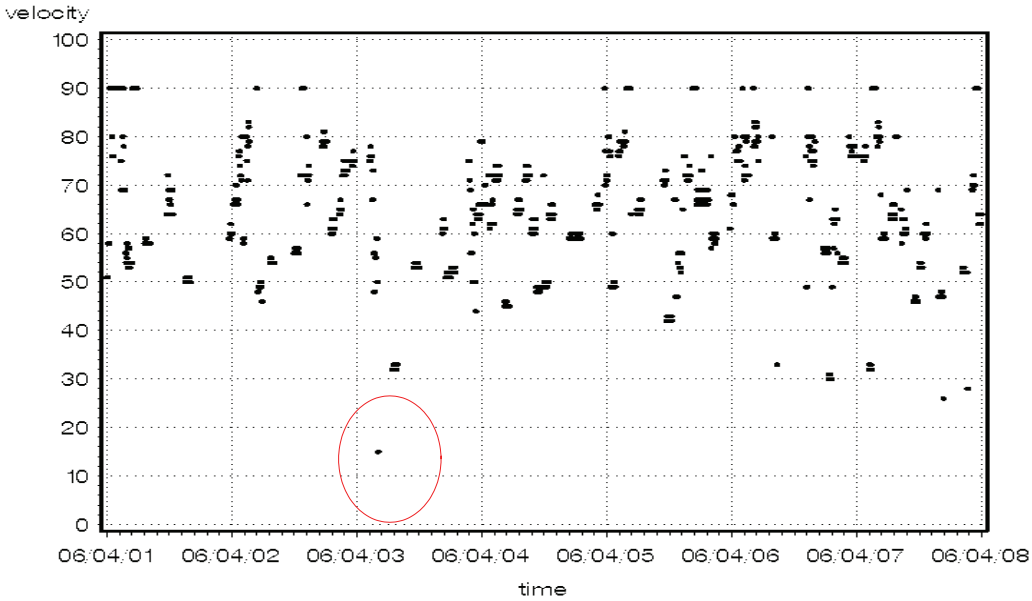
1.2 교통통계정보 자료의 소개

본 논문에 이용한 자료는 서울시 속도 자료중 올림픽대로와 강변북로의 5분 단위 속도측정 자료이다. 총 도로 구간 수는 304개로 2005년 4월1일부터 2007년 3월 31까지의 자료이다(한상태 등, 2007).

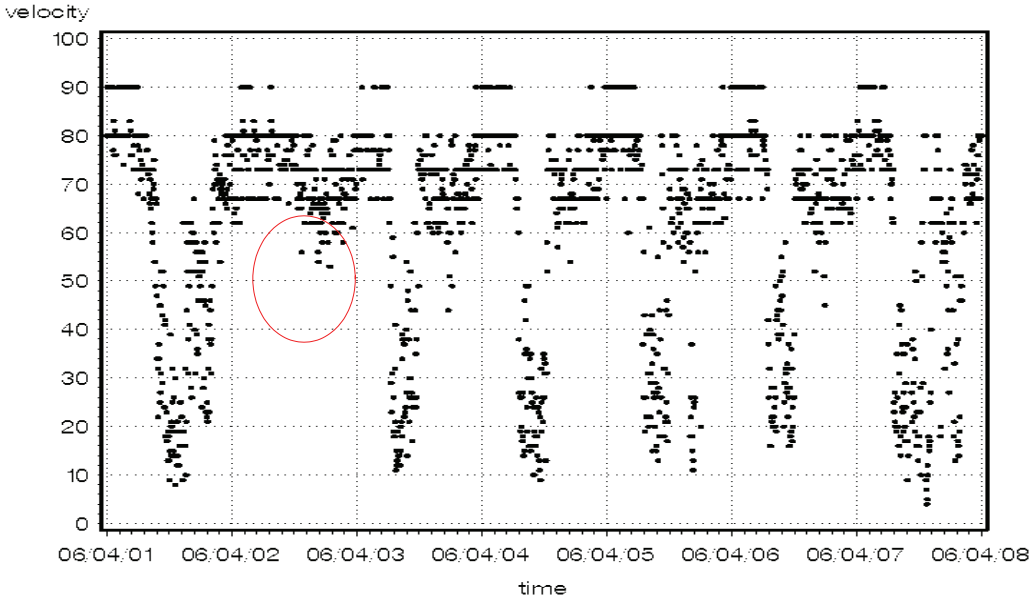
자료의 이해를 위하여 원자료를 살펴보자. 서울시 자료중 올림픽대로 강동대교-천호대교(김포방향)과 강변북로 성산대교-양화대교(구리방향)의 일주일 동안의 속도변화의 시도표를 살펴보았다. 이 때 자료는 2006년 4월 1일(금요일) 0시부터 2006년 4월 7일(목요일) 11시 55분까지의 5분 간격으로 측정된 것을 이용하였다.

먼저, <그림 1.1>는 올림픽대로 강동대교-천호대교(김포방향) 구간의 시도표이다. 30km에서 90km의 속도 변화가 있음을 알 수 있고 눈에 띄게 그래프가 촘촘하지 않은 것을 보아 결측치가 많은 것을 알 수 있다. 구체적으로 2006년 4월 1일 (금요일)의 속도값에 대한 결측치를 살펴보면 0시10분경, 1시 35분경, 5시 50경, 8시 20경, 12시 45분경, 16시 5분경에 값이 없는 것으로 확인되었다. 또한 2006년 4월 3일(일요일)에 4시 5분경의 속도는 약 15km로 특이치이거나 일요일에만 보이는 요일별 특징패턴으로 볼 수 있다.

그 다음, <그림1.2>의 강변북로 성산대교-양화대교(구리방향) 구간의 시도표를 살펴보면 속도변화의 패턴을 살펴보면 새벽시간인 자정부터 감소하다가 낮 12시쯤에 최소 속도를 보이고 다시 서서히 증가하는 패턴이 일주일 동안 반복되는 패턴을 보이고 있다. 2006년 4월 2일(토요일) 오전에는 다른 요일의 낮 시간대의 속도보다 높은 것을 알 수 있다.



<그림 1.1> 올림픽대로 강동대교-천호대교(김포방향) 구간 시도표



<그림 1.2> 강변북로 성산대교-양화대교(구리방향) 구간 시도표

이와 같이 살펴본 자료에 결측치와 특이치가 다수 포함되어 있으므로, 이를 분석에 사용하기 전에 자료를 정제한 후 구간별 5분 단위속도에 대한 평균값을 산출하여 통계적 분석을 수행하였다.

본 논문의 구성은 다음과 같다. 2장에서는 자료를 분석하기 전에 자료의 오류에 대한 검토를 마친 뒤 특이치(outlier)와 결측치(missing value)를 처리한 후 요일 변수 등 기초 분석에 필요한 변수를 만든다. 마지막으로 시구간(time interval)별 평균값을 산출하여 분석자료를 생성한다. 3장에서는 도로 각 구간별 시간대별 평균속도를 예측하기 위해서 기존에 연구된 퓨리에 급수(Fourier series)를 이용한 회귀모형과 본 논문에서 제시한 스플라인 회귀(spline regression)모형에 적합시킨 후 두 모형의 성능을 비교해 본다. 4장에서는 24시간 동안의 속도 패턴(traffic velocity pattern)이 유사한 도로를 군집화 하기 위해 k-평균(k-means) 군집 방법을 이용하여 군집을 구축한다. 5장에서는 결론 및 향후 연구 과제를 살펴본다.

제2장 분석 자료의 생성

본 장에서는 평균속도를 예측하고 패턴이 유사한 도로 구간을 군집화 하기 위해서 분석에 사용할 구간별 5분 단위 평균 속도값을 생성하고자 한다. 먼저, 자료의 오류 여부를 검토하였다. 특이치는 1.5 *IQR(Inter Quantile Range) 방법을 반복 적용하여 제거하였다. 결측치는 스플라인(spline)기법을 이용하여 보간(interpolation) 한 후 요일별, 명절, 공휴일효과의 특징을 살펴보기 위해서 요일 변수의 범주를 9개로 나누었다.

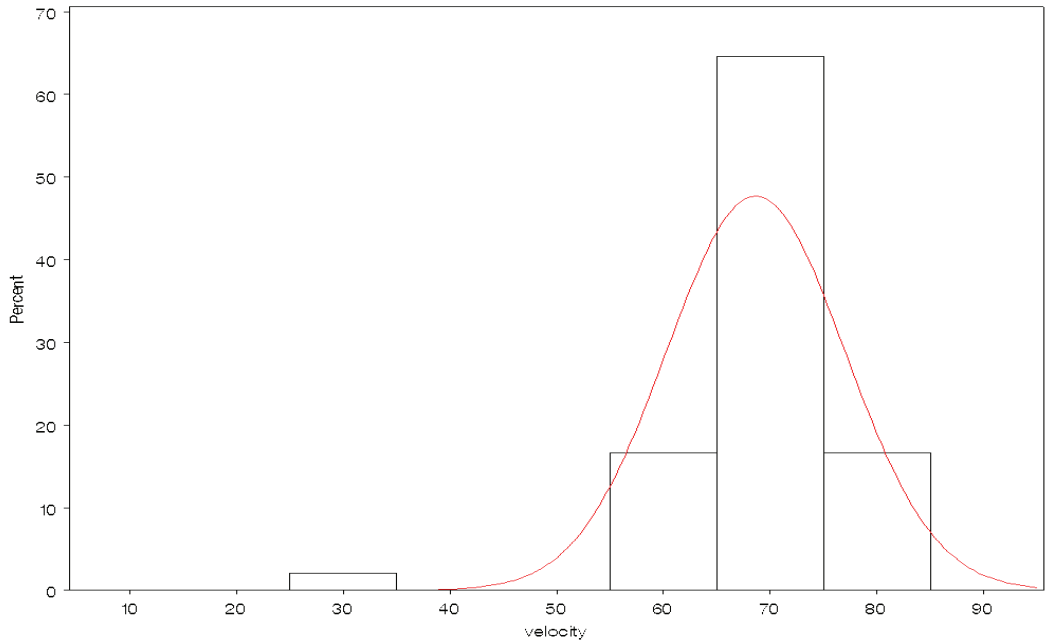
2.1 자료의 오류 검토

자료의 원천은 서로 다른 관계형 데이터 베이스(relational data base)에서부터 파일 또는 이메일(e-mail) 자료 등 다양할 수 있으며 자료 형식도 다양할 수 있다. 이 과정에서 오류를 보정하고 포맷(format)을 통일시키고 자료의 일관성을 유지하며 스키마(schema)를 통합하는 등의 작업을 수행한다. 본 논문에서 쓰인 자료는 서울시의 올림픽대로와 강변북로의 자료로 오류가 있는지를 검토하였다. 원자료에서 -1 값으로 표기 된 것은 자료입력이 되어 있지 않은 값으로 결측치로 처리하였다. 또한 일정시간 동안 같은 값을 갖는 구간이나 특정구간에 누락된 자료가 있는지를 확인하였다.

2.2 특이치(outlier) 제거 및 시구간(time interval)별 평균속도 산출

앞서 언급하였듯이 도로교통상황에서의 돌발상황 즉 공사, 사고, 기후 등에 영향을 받지 않는 순수한 속도 자료만으로 분석을 실시하고자 하므로 일상적이지 않은 특이치를 제거하기로 한다. 이러한 특이치가 자료 전체에 큰 영향을 미치기 때문에 본 논문에서는1.5*IQR(Inter Quantile Range) 방법을 반복적으로 수행하여 이를 제거하고자 한다.

특이치가 포함된 속도자료가 가지는 문제점을 살펴보면 다음과 같다. 다음 <그림 2.1>는 올림픽대로 행주대교-김포대교(김포방향) 구간 금요일 오후 4시 30분일 때의 2005년 4월 1일부터 2007년 3월 31일까지 2년간의 원자료를 이용하여 작성한 히스토그램이다. 그림에서 보는 바와 같이 전체적인 자료의 평균은 대략 68km에서 형성되어 있다. 그러나 일부 자료에서 시속 30km의 값을 보인다. 금요일 오후 4시 30분이라는 시간대를 고려하면 이는 보통의 도심 도로에서 퇴근 정체가 서서히 시작되는 시간이라 할 수 있지만 이 구간은 도심외곽구간이므로 돌발상황으로 인한 특이치로 추정된다. 이러한 특이치를 고려하지 않고 속도의 평균을 구하게 되면 이에 의하여 상대적으로 작은 평균을 보이게 되고 실제 도로이용자들이 느끼는 속도보다 작은 값을 가지게 된다.



<그림 2.1> 올림픽대로 행주대교-김포대교(김포방향) 구간
금요일 오후 4시 30분 히스토그램

이러한 특이치를 제거하기 위하여 본 연구에서는 1.5*IQR 방법을 반복적으로 수행하는 프로세스를 구축하였다. 먼저 사분위 범위에 대하여 설명을 하면 다음과 같다. 상자도표에서 가운데 상자는 전체자료를 크기순으로 나열한 후 4등분한 4분위수를 이용하여 전체자료의 하위 25%에 해당하는 1분위수와 75%에 해당하는 3사분위수로 그린다. 이때 사분위 범위(Inter Quantile Range)는 다음과 같다.

- 1사분위수 : $(n+1)\times\frac{1}{4}$ 번째 자료
- 3사분위수 : $(n+1)\times\frac{3}{4}$ 번째 자료

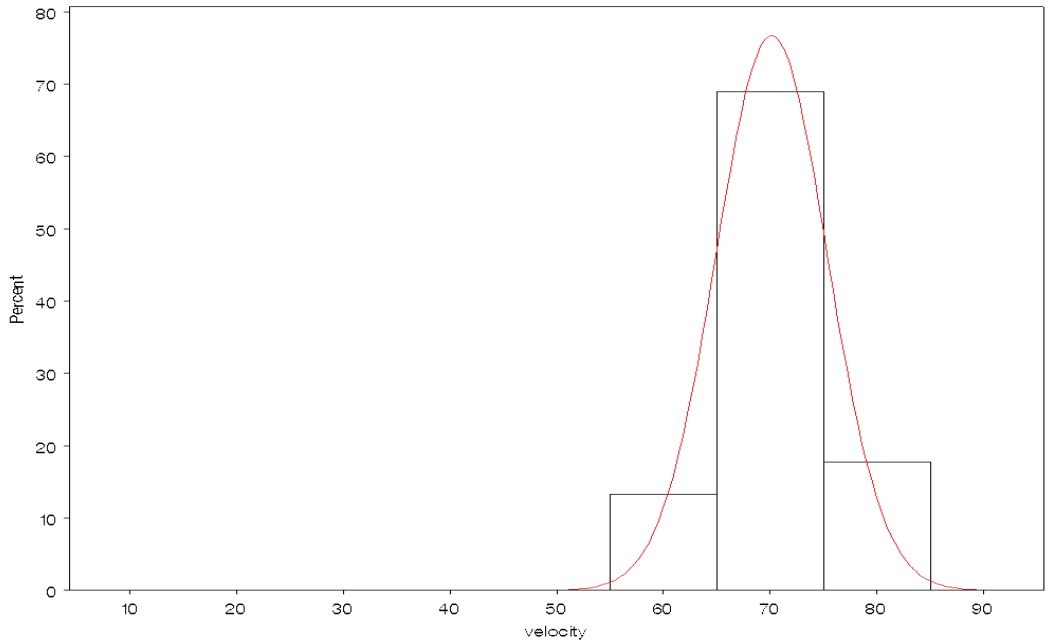
▸ IQR(Inter Quantile Range) = 3사분위수 - 1사분위수

이러한 사분위범위의 1.5배를 구하여 속도 값이 '3사분위수+1.5×IQR'보다 크거나 '1사분위수-1.5×IQR'보다 작으면 특이치로 고려한다. 위의 자료를 다음과 같은 1.5*IQR 방법을 통한 특이치 제거를 반복적으로 수행하여 평균을 계산하는 프로세스는 다음과 같다.

<표 2.1> 특이치 제거를 위한 프로세스

단계 1	원자료를 이용하여 평균을 계산한다. 이 때 평균과 함께 1사분위수, 3사분위수 를 계산한다.
단계 2	3사분위수+1.5×IQR보다 크거나 1사분위-1.5×IQR 보다 작으면 특이치로 고려한다.
단계 3	특이치를 제거한 후 자료의 수를 계산하여 특이치 이전의 자료의 수와 비교한다.
단계 4	단계 1에서 단계 3을 반복적으로 수행하여 특이치 제거이전과 특이치 제거이후의 표본의 수가 변동되지 않을 때까지, 즉 더 이상의 특이치가 검출되지 않을 때까지 프로세스를 반복한다.

위의 구간을 특이치 제거한 후의 히스토그램을 다시 그려보면 아래 <그림 2.2>와 같이 평균 속도의 분포가 정규분포에 가깝게 변화한 것을 알 수 있다. 이와 같은 과정을 거친 후 각각 올림픽대로와 강변북로의 2005년 4월 1일부터 2007년 3월 31까지의 자료를 이용하여 24시간 동안의 5분 단위 평균속도를 산출한다.



<그림 2.2> 특이치가 제거된 올림픽대로 행주대교-김포대교(김포방향) 구간 금요일 오후 4시 30분 히스토그램

2.3 결측치(missing value) 대체

속도자료는 기본적으로 시간의 흐름에 따라 관찰된 시계열자료이다. 이와 같은 시계열 자료의 특성을 살려 예측을 위한 모형을 구축하기 위해서는 시간흐름에 따라 중간에 발생하는 결측치의 대체가 선행되어야 한다. 이와 같이 결측이 발생한 부분을 보정하기 위해서 자료에 대한 보간(interpolation)이 수행되어야 한다. 이와 같은 결측치를 보간하는 방법으로 매끄러운 곡선 형태로 대체하기 위하여 3차 스플라인(spline) 곡선을 이용하였다.

스플라인 함수보간법은 다음과 같다(박전수, 2007). 자료의 점들의 부분집합에 저차 다항식을 적용시켜 나가는 연결 다항식을 스플라인 함수라고 한다. 스플라인 함수는 일반적으로 완만하게 변하지만 특정 구역에서는 급격하게 변하는 경우도 있다. 스플라인 함수는 국부적으로 급격하게 변하는 움직임에 우수한 근사값을 제공한다.

주어진 구간 $[a,\ b]$가 $a=x_1<x_2<x_3\cdots<x_n=b$와 같이 $n-1$ 개의 소구간으로 이루어졌을 때 차수가 k인 스플라인함수 $s(x)$ 는 다음 조건을 만족해야 한다.

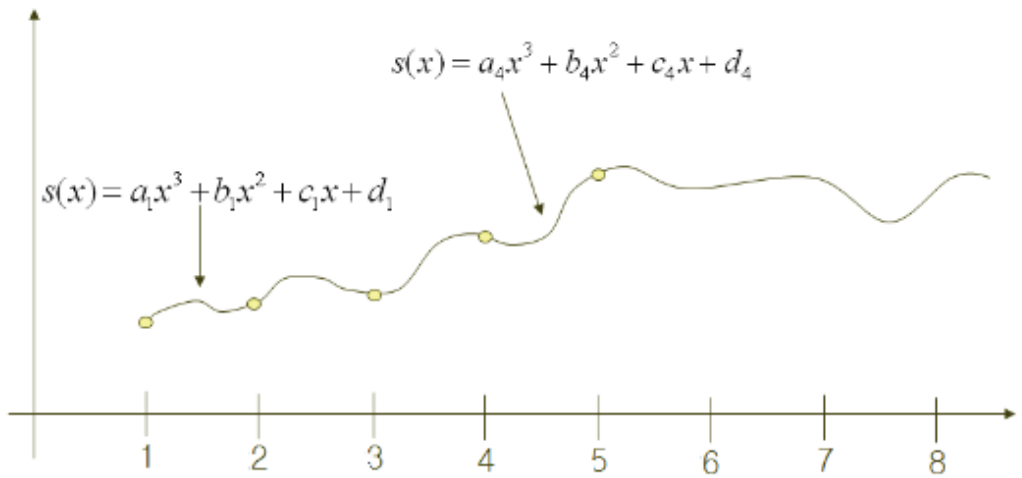
▸조건 1: $s(x)$ 는 $i=1,2,...,n$에 대한 소 구간 $[x_i,\ x_{i+1}]$에서 k차 이하의 다항식으로 표현된다.

▸조건2: $s(x),\ s'(x),\ s''(x),\ \cdots,\ s^{k-1}(x)$등의 도함수들은 구간 $[a,\ b]$에서 연속이어야 한다.

두 조건을 만족하는 최소의 차수는 3차로서 모든 소구간$[x_i,\ x_{i+1}]$에서 3차 다항식으로 표시되는 함수를 3차 스플라인 함수라고 한다.

$$s(x)=y=a_i(x-x_i)^3+b_i(x-x_i)^2+c_i(x-x_i)+d_i\ (\text{단},\ x_i<x<x_{i+1}) \tag{2.1}$$

위와 같은 3차 스플라인 함수를 이용하여 결측치를 보간하였다.

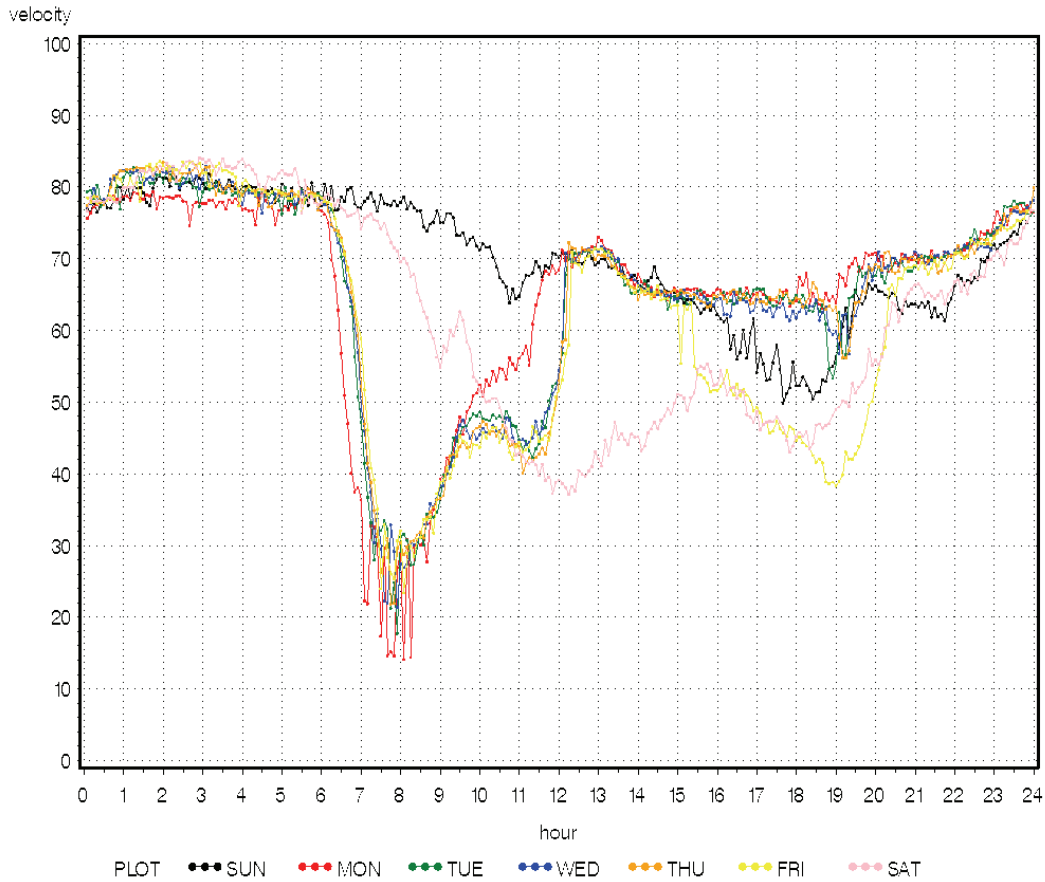


<그림 2.3> 스플라인을 이용한 결측치 보간

2.4 요일 변수 생성

이 절에서는 자료의 기본적이 패턴을 살펴 보기 위하여 7개의 요일효과, 명절효과, 그 밖에 공휴일변수를 생성하였다. 먼저, 도로 속도는 권성진(1999)에 의하여 일주일간에 평균 통행발생비율을 군집분석을 한 결과 수요일, 주중, 주말로 서로 다른 패턴을 가지고 있다고 하였다. 이 때 평균 통행발생비율이란 일주일간 총통행수에 대한 요일별 총통행수 비율을 나타낸다. 이에 본 논문에서는 요일별 도로 속도의 패턴이 다를 것을 예상하고 각 요일 효과를 반영하기 위해서 일, 월, 화, 수, 목, 금, 토요일등 요일 변수를 생성하였다. 강변북로 성산

대교-양화대교(구리방향) 구간으로 일주일 동안의 요일별 속도 변화를 그려보면 <그림 2.4>와 같이 요일마다 다른 패턴을 보인다. 특히 토요일과 일요일의 속도 패턴의 변화는 다른 요일과는 달리 매우 다름을 알 수 있다.



<그림 2.4> 강변북로 성산대교-양화대교(구리방향) 구간의 요일별 통행 패턴

다음, 명절효과를 살펴보기 위하여 우리나라의 고유의 명절인 설과 추석으로 명절변수를 만들었다. 명절에는 전국적인 민족의 대이동으로 교통흐름이 주말과는

매우 다를 것으로 예상된다. 여기에 추석과 설 전날과 다음날에도 이 효과가 미친다고 판단하여 같은 명절효과의 범주로 분류하였다.

마지막으로 생성한 변수는 기타 공휴일변수로 국정 공휴일 등을 포함하였다. 여기서 국정 공휴일은 2005년부터 2007년까지의 국정 공휴일 날을 하나하나 지정하여 분류하였고, 여름휴가기간을 7월 21일에서 8월 10일 정도로 예상하여 전체 연휴기간을 정하여 기타 공휴일 효과의 범주에 분류하였다. 전날에는 금요일처럼 연휴 전날 효과가 있을 것으로 보여 기타 공휴일 효과의 저녁 6시 이후는 금요일 효과에 포함시켰다.

제3장. 속도예측을 위한 통계모형 구축

평균 속도예측을 위해서 이 장에서는 현재 연구된 퓨리에 급수(Fourier series)를 이용한 회귀모형과 본 논문에서 제시한 스플라인 회귀(spline regression) 모형의 성능을 비교해 보고자 한다. 분석에 이용한 자료는 올림픽대로 강동대교-천호대교(김포방향) 구간, 강변북로 하행 성산대교-양화대교(구리방향) 구간의 5분 간격 평균속도 자료로 구간별 24시간의 평균 속도를 예측하고자 한다.

3.1 퓨리에 급수(Fourier series)를 이용한 회귀모형

<그림 1.4>에서 24시간동안의 평균속도 패턴을 살펴보면 출퇴근 시간과 같은 특정 시간에 속도저하가 일어나며, 또한 속도가 회복되고 다시 정체되는 패턴이 반복적으로 일어남을 알 수 있다. 이러한 반복적인 패턴의 적합을 위해서는 삼각함수로 만든 퓨리에 급수를 이용한 회귀모형을 이용하고자 한다.

먼저 퓨리에 급수란 다음과 같다(이해기, 2009). 퓨리에 급수란 주기함수를 코사인(cosine)과 사인(sine)의 합으로 표현하는 무한급수이다. 이때 주기 함수란 다음과 같은 식을 만족한다.

$$f(x+p)=f(x) \tag{3.1}$$

위의 함수의 일종인 삼각함수들로부터 다음과 같은 급수를 얻을 수 있다.

$$a_0 + a_1\cos x + b_1\sin x + a_2\cos 2x + b_2\sin 2x + \ldots$$

$$= a_0 + \sum_{n=1}^{\infty}(a_n\cos nx + b_n\sin nx) \tag{3.2}$$

여기서 a_0, a_1, b_1, a_2, b_2,…는 상수이며 이 급수의 계수를 퓨리에 계수(Fourier coefficients)라고 부른다. 각 항의 주기는 2π이다. 급수가 수렴하면 그 합은 주기가 2π인 주기함수가 된다. 이 급수의 중요한 특징은 구간 $-\pi \le x \le \pi$에서 직교성을 보인다는 것이다.

함수의 추정에 있어서 크게 두 가지 정지규칙(stopping rule)으로 밀도 함수를 최적화 시키는 평활모수(smoothing parameter)를 찾는 방법과 함수에 승수(multiplier)를 곱하여 함수를 최적화 시키는 가중모수를 찾는 방법이 있다. 여기서는 함수를 추정하기 위해서 전자를 사용하였다(김종태, 1997). 여기서 밀도 함수는 주기가 2L인 우함수의 퓨리에 급수인 퓨리에 코사인 급수라고 하자.

$$f(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos\frac{n\pi}{L}x \tag{3.3}$$

여기서 계수는 다음과 같다.

$$a_0 = \frac{1}{L}\int_0^L f(x)dx, \quad a_n = \frac{2}{L}\int_0^L f(x)\cos\frac{n\pi x}{L}dx, \quad n = 1,2,\ldots \tag{3.4}$$

이러한 코사인 수열의 사용은 계산을 간편하게 할 뿐만 아니라 곡선의 변동을 잘 관찰할 수 있는 이점이 있다. 위 식에서 a_k의 불편추정량 $\hat{a_n} = (\sum_{i=1}^{n} 1/n)\sqrt{2}\cos(k\pi x_i)$를

대입하여 함수 추정량 을 $\widehat{f_m}$ 다음과 같이 얻을 수 있다.

$$\hat{f}(x) = \sum_{k=-\infty}^{m} \widehat{a_m} \sqrt{2}\, cos(k\pi n) \tag{3.5}$$

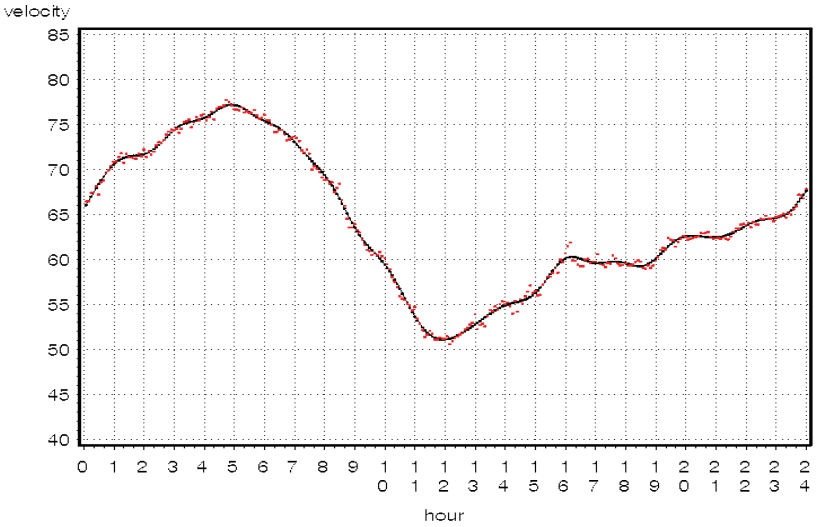
여기서 m은 절단점이다. 정지규칙에 있어서 최적화된 함수를 찾는 방법은 함수 f와 추정량 $\widehat{f_m}$에 대한 차이를 최소화하는 m을 선택하는 것이다. 본 연구에서 m은 회귀모형을 통한 설명력 및 시스템 메모리 상황을 감안하여 30으로 고려하였다.

올림픽대로 강동대교-천호대교(김포방향) 구간의 퓨리에 급수로 적합된 회귀모형은 <표 3.1>의 회귀분석 결과를 보면 전체모형이 유의(p-값<0.0001)하며 결정계수(R-squared)가 0.9974로써 전체 변동의 약 99.7%를 설명하는 모형으로 나타났다. 추정된 회귀모형의 식은 다음과 같다.

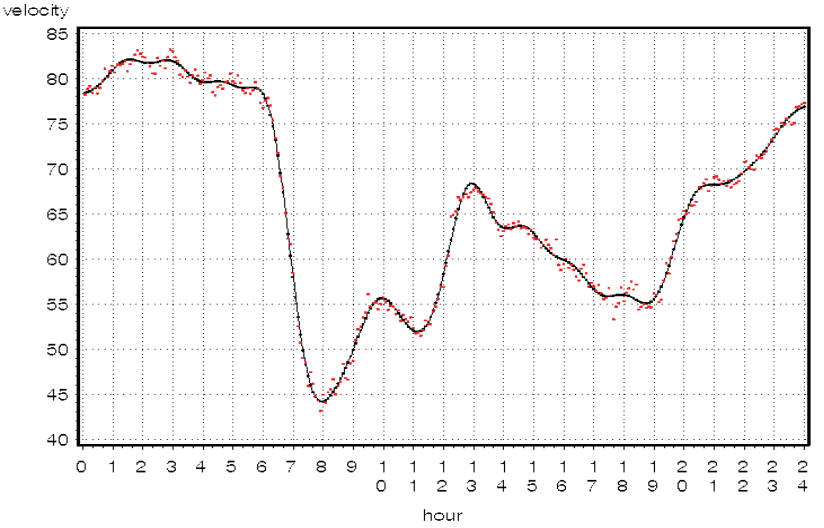
$$speed = 63.07 + 0.3327t + 6.5950\cos(t) + 7.8519\sin(t)$$

$$+ \ldots + 0.0887\cos(15t) + 0.0377\sin(15t) \tag{3.6}$$

또한 <표 3.2>에서도 볼 수 있듯이 강변북로 성산대교-양화대교(구리방향) 구간의 결정계수는 98.8%이다.



<그림 3.1> 올림픽대로 강동대교-천호대교(김포방향) 구간의 퓨리에 급수를 이용한 회귀모형



<그림 3.2> 강변북로 성산대교-양화대교(구리방향) 구간의 퓨리에 급수를 이용한 회귀모형

<표 3.1> 회귀 분석 결과 : 올림픽대로 강동대교-천호대교(김포방향) 구간

The REG Procedure
Model: MODEL1
Dependent Variable: at202529

Number of Observations Read 288
Number of Observations Used 288

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	31	16865	544.03384	2421.09	<.0001
Error	256	57.52468	0.22471		
Corrected Total	287	16923			

Root MSE	0.47403	R-Square	0.9966
Dependent Mean	64.12434	Adj R-Sq	0.9962
Coeff Var	0.73924		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	63.07538	0.24794	254.39	<.0001
t	1	0.33274	0.07815	4.26	<.0001
c1_t	1	6.59503	0.03954	166.80	<.0001
s1_t	1	7.86191	0.16121	48.77	<.0001
c2_t	1	-4.23868	0.03954	-107.20	<.0001
s2_t	1	1.59433	0.08755	18.21	<.0001
c3_t	1	0.98281	0.03954	24.86	<.0001
s3_t	1	-0.33032	0.06537	-5.05	<.0001
c4_t	1	-0.58464	0.03954	-14.79	<.0001
s4_t	1	0.22272	0.05555	4.01	<.0001
c5_t	1	0.12683	0.03954	3.21	0.0015
s5_t	1	0.41245	0.05036	8.19	<.0001
c6_t	1	0.06182	0.03954	1.56	0.1192
s6_t	1	0.42643	0.04730	9.02	<.0001
c7_t	1	0.14896	0.03954	3.77	0.0002
s7_t	1	-0.07581	0.04535	-1.67	0.0959
c8_t	1	-0.42755	0.03954	-10.81	<.0001
s8_t	1	0.18545	0.04405	4.21	<.0001
c9_t	1	0.05209	0.03954	1.32	0.1889
s9_t	1	0.08134	0.04313	1.89	0.0604
c10_t	1	-0.17342	0.03954	-4.39	<.0001
s10_t	1	0.22441	0.04246	5.29	<.0001
c11_t	1	-0.07247	0.03954	-1.83	0.0680
s11_t	1	0.15850	0.04196	3.78	0.0002
c12_t	1	0.08781	0.03954	2.22	0.0272
s12_t	1	0.04064	0.04157	0.98	0.3292
c13_t	1	-0.16033	0.03954	-4.05	<.0001
s13_t	1	-0.13293	0.04127	-3.22	0.0014
c14_t	1	0.03127	0.03954	0.79	0.4297
s14_t	1	-0.01663	0.04103	-0.41	0.6856
c15_t	1	0.08878	0.03954	2.25	0.0256
s15_t	1	0.03773	0.04083	0.92	0.3564

<표 3.2> 회귀 분석 결과 : 강변북로 성산대교-양화대교(구리방향) 구간

The REG Procedure
Model: MODEL1
Dependent Variable: bt203355

Number of Observations Read 288
Number of Observations Used 288

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	31	35636	1149.55413	1801.95	<.0001
Error	256	163.31527	0.63795		
Corrected Total	287	35799			

Root MSE	0.79872	R-Square	0.9954
Dependent Mean	65.69628	Adj R-Sq	0.9949
Coeff Var	1.21577		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	66.42986	0.41777	159.01	<.0001
t	1	-0.23270	0.13168	-1.77	0.0784
c1_t	1	10.99684	0.06662	165.06	<.0001
s1_t	1	3.40172	0.27163	12.52	<.0001
c2_t	1	2.89226	0.06662	43.41	<.0001
s2_t	1	7.57915	0.14752	51.38	<.0001
c3_t	1	-3.13333	0.06662	-47.03	<.0001
s3_t	1	-3.59674	0.11014	-32.66	<.0001
c4_t	1	0.83162	0.06662	12.48	<.0001
s4_t	1	-0.78271	0.09359	-8.36	<.0001
c5_t	1	1.08889	0.06662	16.34	<.0001
s5_t	1	1.79383	0.08485	21.14	<.0001
c6_t	1	-1.72078	0.06662	-25.83	<.0001
s6_t	1	1.56991	0.07969	19.70	<.0001
c7_t	1	0.79025	0.06662	11.86	<.0001
s7_t	1	-2.34830	0.07642	-30.73	<.0001
c8_t	1	0.96761	0.06662	14.52	<.0001
s8_t	1	0.84018	0.07422	11.32	<.0001
c9_t	1	-0.55664	0.06662	-8.36	<.0001
s9_t	1	-0.45308	0.07267	-6.23	<.0001
c10_t	1	-0.58779	0.06662	-8.82	<.0001
s10_t	1	0.10673	0.07154	1.49	0.1370
c11_t	1	0.23421	0.06662	3.52	0.0005
s11_t	1	-0.34424	0.07070	-4.87	<.0001
c12_t	1	0.07980	0.06662	1.20	0.2321
s12_t	1	0.50514	0.07005	7.21	<.0001
c13_t	1	0.00202	0.06662	0.03	0.9759
s13_t	1	-0.22584	0.06953	-3.25	0.0013
c14_t	1	-0.18391	0.06662	-2.76	0.0062
s14_t	1	-0.36263	0.06913	-5.25	<.0001
c15_t	1	0.20302	0.06662	3.05	0.0026
s15_t	1	0.03012	0.06880	0.44	0.6619

3.2 스플라인 회귀(spline regression) 모형

이 절에서는 먼저 스플라인(spline)의 개념을 살펴보고 앞장에서 설명한 스플라인 보간(interpolation)과 비교하여 스플라인 회귀(spline regression)를 설명하고자 한다(Guo & White, 2004).

곡률표현 때문에 수학적 함수를 이용해 그려지는 선을 단순히 라인(line)이라고 하지 않고 스플라인이라고 부른다. 얇고 유연한 가늘고 긴 조각을 이용하여 운형의 자를 이용했는데 이것을 스플라인이라 불렀다. 통계학에서 스플라인은 구분적 n차 다항식이라고 할 수 있다. 각 구간의 연결점을 넛(knot)이라고 부르며 일반적인 다항식 모형은 넛이 없는 스플라인모형의 특수한 모형이라고 할 수 있다.

먼저, 스플라인 보간을 살펴보자. $(x_{1,}y_1)$, $(x_{2,}y_2)$, ... ,$(x_{m,}y_m)$인 m개의 점이 있는데 부드러운 곡선으로 점을 연결하고자 한다. $m-1$개의 간격으로 나누어진다. 각각의 구간은 3 차 다항식으로 다음과 같이 표현된다.

$$y = a_i + b_i x + c_i x^2 + d_i x^3 \tag{3.7}$$

위의 x는 독립변수, y는 종속변수로 $y = a_i + b_i x + c_i x^2 + d_i x^3$는 i번째 모수이다. $m-1$의 구간을 완벽하게 정의하기 위해서 $4*(m-1)$모수가 결정 되어야 한다. 즉 $4*(m-1)$개의 방정식이 필요하다. 각각 두 번 사용되므로 식(3.7)에서 모든 점이 유효하므로 $2*m$의 방정식을 세울 수 있다. 독립변수가 각 값의 연속이므로 종속변수의 한 번, 두 번 미분 값 모두 가정함으로서 $2*(m-2)$의 방정식을 구체화할 수 있다. $(x_{1,}y_1)$, $(x_{m,}y_m)$의 두 점에서 종속변수의 두 번 미분 값을 구하면서 두 개의 방정식

을 도출 할 수 있다. $4*(m-1)$개의 모수가 일단 결정되면, 간격 안의 어떤 점을 보간하기 위해서 식(3.7)을 사용한다. 이 과정을 흔히 3차 스플라인 보간이라고 한다. 3차 스플라인 보간 방법은 데이터 값이 적을 때 유리하므로 수가 많아지면 보간 방법은 비효율적이다.

3차 스플라인 회귀는 3차 스플라인 보간과 비슷하다. 3차 스플라인 회귀는 각 구간에 또한 3차 다항식이 사용되고 종속변수와 한 번, 두 번 미분값이 모든 넛(knot)에서 연속이어야 한다. 다음은 3차 스플라인 회귀식으로 다음과 같이 쓸 수 있다.

$$y = a + bx + cx^2 + dx^3 + \sum_{i=1}^{k} D_i e_i (x - x_i)^3 \tag{3.8}$$

여기서 a, b, c, d, e는 모수이고 x_i는 넛의 위치이고 k는 넛의 수이다. D_i는 더미변수이다. 식(3.8)는 한 회귀모형구간과 다음 구간의 모형에 대한 일반적인 식이다. 첫 번째 구간에 대한 식은 다음과 같다.

$$y = a + bx + cx^2 + dx^3 \tag{3.9}$$

D_i가 첫 번째 구간에서 0과 같다. 두 번째 구간에서 모형식은 다음과 같다.

$$y = a + bx + cx^2 + dx^3 + e_1 (x - x_1)^3 \tag{3.10}$$

여기서 두 번째 구간에서 $D_1 = 1$이다. 세 번째 구간에서 모형식은 다음과 같다.

$$y = a + bx + cx^2 + dx^3 + e_1(x - x_1)^3 + e_2(x - x_2)^3 \quad (3.11)$$

여기서 D_1과 D_2는 1이다. 식 (3.9)- (3.11)에서 보았듯이 두 인접한 구간의 대한 사용된 방정식은 한 구간당 다르다. 양쪽 방정식에 대한 인접한 넛에서 미분가능하고 연속이라는 조건에 의해 x_1은 첫 번째과 두 번째 구간에서 생겨나고 x_2는 두 번째와 세 번째 구간에서 각각 구해진다. 이것 때문에 곡선은 3차 스플라인 회귀곡선에의 의해 예측된 곡선은 매우 부드러운 곡선으로 예측된다.

3차 스플라인 회귀와 3차 스플라인 보간의 차이는 데이터 궤적의 곡률의 변화에 기반한 3차 스플라인 회귀에 이용되는 모수의 수가 적다는 것이다. 예를 들면 넛의 수에 따라 추정할 모수를 각각 넛의 수가 4일 때를 살펴보면, 3차 스플라인 보간시에는 4*(넛의수 -1)=4*(4-1)=12이고 3차 스플라인 회귀 분석시에는 4+더미변수의 갯수+(넛의 수-2)*2=4+2+(4-2)*2=10이 된다. 아래에의 <표 3.3>는 3차 스플라인 회귀 분석시 넛의 수에 따른 추정할 모수의 수를 나타낸다.

<표 3.3> 넛의 수에 따른 3차 스플라인 회귀 분석시 추정할 모수의 수

넛의수	2	3	4	5	6	7	8	9	10
3차 스플라인 회귀분석시 추정할 모수의 수: 4+더미변수갯수+(넛의수-2)*2	4	7	10	13	15	17	19	21	23

스플라인 회귀는 제약이 좀 더 가해진 가변수 모형이라고 할 수 있다. 보통 다항모델보다 다중공선성 문제에 있어서 더 자유로운 장점이 있다. 스플라인회귀에 종류에는 선형(linear) 스플라인, 이차(quadratic) 스플라인, 삼차(cubic) 스플라인, 혼

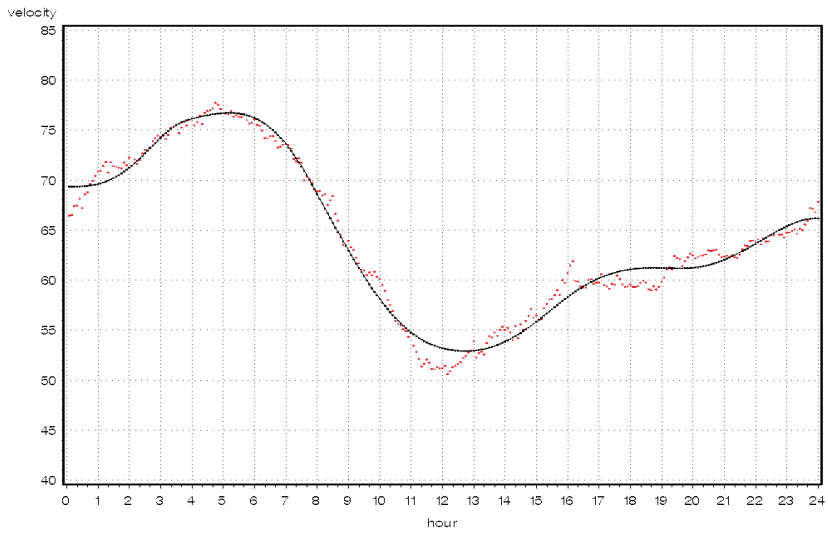
합(hybrid) 스플라인 등이 있다. 스플라인 기법을 이용할 때 고려해야 할 점은 넛의 수, 넛의 위치, 적합함수의 선택이다. 여기서 위의 세 가지를 아는지 여부에 따라 보통 넛의 수와 넛의 위치를 아는 경우, 넛의 수는 알고 있으나 넛의 위치를 모르는 경우, 넛의 수조차 모르는 경우가 있다(Marsh, 2002).

먼저 올림픽대로 강동대교-천호대교(김포방향) 구간을 스플라인 회귀모형에 적합을 시킨 결과는 다음과 같다. 본 논문에서 이용한 자료에서는 넛의 수와 넛의 위치를 모르는 경우에 해당하므로 넛의 수를 5개, 10개, 25개로 늘려가면서 적합을 시켜보았다. 넛을 5개 놓았을 때 결정계수(R-squared)는 0.9795으로 <표 3.4>의 결과와 같고 적합시킨 결과를 그래프로 나타내면 <그림 3.3>으로 나타난다.

<표 3.4> 회귀 분석 결과 - 일부 :

올림픽대로 강동대교-천호대교(김포방향) 구간(KNOT 5)

Iteration Number	Average Change	Maximum Change	R-Square	Criterion Change	Note
0	0.96056	2.89163	0.08024		
1	0.00000	0.00000	0.97953	0.89929	Converged

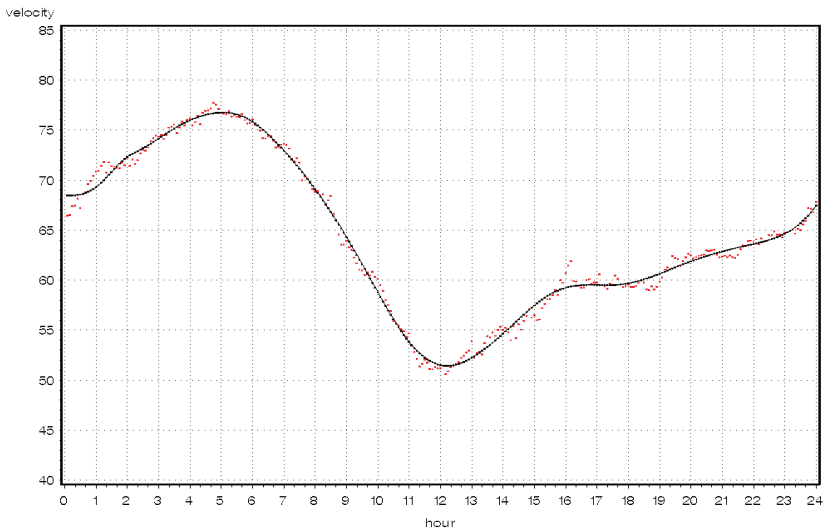


<그림 3.3> 올림픽대로 강동대교-천호대교(김포방향) 구간자료의
스플라인을 이용한 회귀모형(넛의 수 5)

넛을 10개 놓았을 때 결정계수가 0.9918으로 <표 3.5>의 결과와 같고 적합시킨 결과를 그래프로 나타내면 <그림 3.4>으로 나타난다.

<표 3.5> 회귀 분석 결과 - 일부 :
올림픽대로 강동대교-천호대교(김포방향) 구간(넛의 수 10)

Iteration Number	Average Change	Maximum Change	R-Square	Criterion Change	Note
0	0.94566	3.05838	0.08024		
1	0.00000	0.00000	0.99181	0.91157	Converged

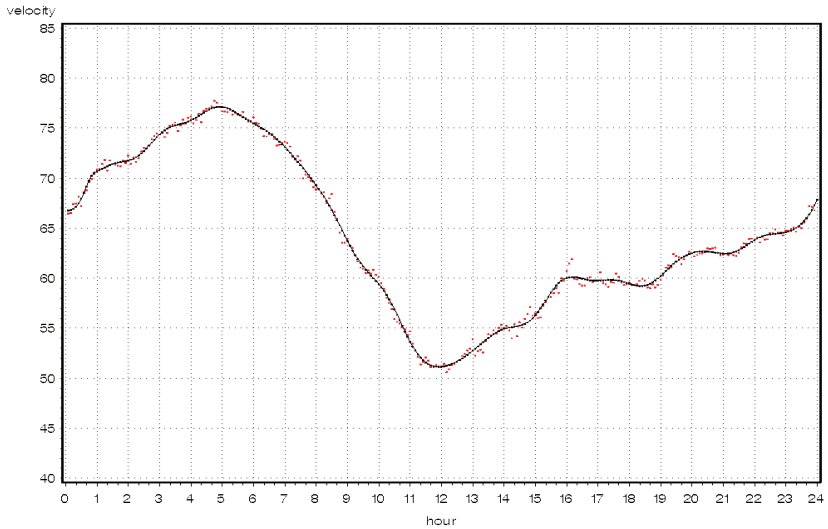


<그림 3.4> 올림픽대로 강동대교-천호대교(김포방향) 구간자료의
스플라인을 이용한 회귀모형(넛의 수 10)

넛을 25개 놓았을 때 결정계수 0.9965으로 <표 3.6>의 결과와 같고 적합시킨 결과를 그래프로 나타내면 <그림 3.5>으로 나타난다. 이와 같은 결과로 보아 넛의 수를 5, 10, 25개로 늘였을 때 점점 모형의 결정 계수값이 늘어남을 알 수 있다.

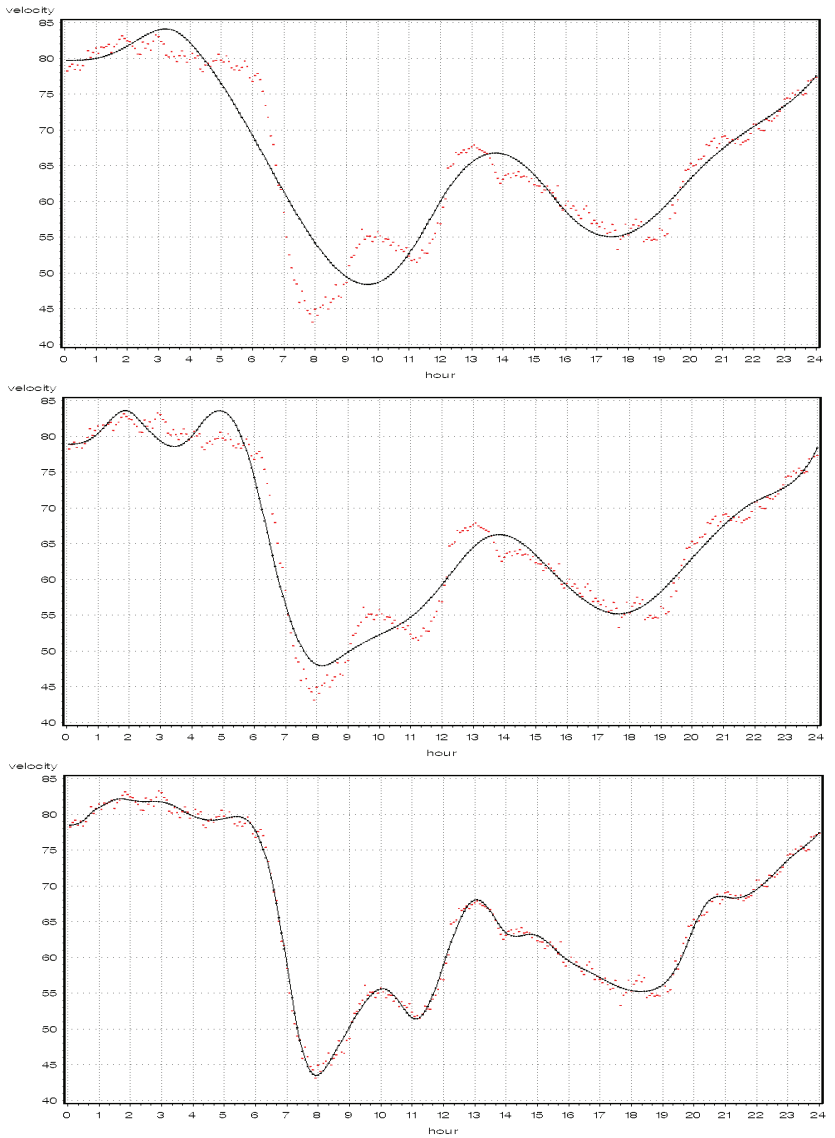
<표 3.6> 회귀 분석 결과 - 일부 :
올림픽대로 강동대교-천호대교(김포방향) 구간(KNOT 25)

Iteration Number	Average Change	Maximum Change	R-Square	Criterion Change	Note
0	0.93956	3.10725	0.08024		
1	0.00000	0.00000	0.99654	0.91630	Converged



<그림 3.5> 올림픽대로 강동대교-천호대교(김포방향) 구간자료의 스플라인을 이용한 회귀모형(넛의수 25)

마찬가지로 강변북로 성산대교-양화대교(구리방향)구간의 자료도 위의 구간과 같이 스플라인 회귀모형에 적합시킨 결과 넛을 5, 10, 25개 놓았을 때 결정계수는 각각 89.2, 95.3, 99.4로 적합 시킨 결과를 그래프로 나타내면 <그림 3.6>과 같다.



<그림 3.6> 강변북로 하행 성산대교-양화대교자료의 스플라인을 이용한 회귀모형(넛의 수 5, 10, 25개로 변화)

결정계수를 통하여 올림픽대로 강동대교-천호대교(김포방향), 강변북로 성산대교-양화대교(구리방향)의 두 구간의 평균 속도 자료를 가지고 적합시킨 퓨리에 급수를 이용한 회귀모형과 스플라인 회귀모형과 비교하고자 한다. 삼각함수의 개수가 30인 퓨리에 급수를 이용한 회귀모형의 결정계수는 각각 99.6, 99.5로 매우 높은 값을 갖는다. 스플라인 회귀모형은 넛의 수가 5, 10, 25로 변함에 따라 각각 97.9, 99.1, 99.6으로 증가하고 강변북로 구간 또한 89.2, 99.1, 99.6으로 결정계수의 값이 증가함을 알 수 있다. 퓨리에 급수를 이용한 회귀모형에서의 결정계수 값과 비슷한 값을 얻기 위해서는 스플라인 회귀모형에서 넛의 수를 25개 이상 충분히 두는 것이 좋다. 그러나 네비게이션 등 메모리의 효율성을 감안할 때 넛의 수를 5-10정도로 고정시켜서 적합시키는 것이 적당하다. 왜냐하면 모수의 절약 면에서 퓨리에 급수에서 추정할 모수가 30개인 반면, 스플라인 회귀모형에서 5-10정도의 넛의 수를 갖는다면 추정할 모수가 13-23개이기 때문이다. 스플라인 회귀모형은 현실적으로 수많은 구간의 회귀모형을 만들어 적합시키기 위해서는 모형의 설명력은 80-90으로 떨어지나 좀 더 효율적인 모형이라고 할 수 있다.

제4장 교통속도패턴(traffic velocity pattern)에 관한 군집화(clustering)

24시간 동안의 각 구간별 속도 패턴을 알아보기 위해 k-평균 (k-means) 군집방법를 통해 비슷한 패턴을 갖는 구간끼리 군집화 하였다. 이 때 이 분석에 사용한 데이터는 올림픽대로 전구간인 152구간으로 5분 간격 평균속도 자료이다.

4.1 k-평균(k-means)군집 방법

군집분석(cluster analysis)이란 각 객체의 유사성을 측정하여 유사성이 높은 대상 집단을 분류하고, 군집에 속한 객체들의 유사성과 서로 다른 군집에 속한 객체간의 상이성을 규명하는 통계분석 방법이다. 대상들을 분류하기 위한 명확한 기준이 존재하지 않거나 기준이 밝혀지지 않은 상태에서 다양한 특성을 지닌 전체를 군집으로 분류하는데 사용되는 기법이다(허명회, 2005).

군집분석을 하는 데는 3가지 중요한 과제가 있는데, 첫 번째는 설명변수를 선정하는 것으로 군집의 특징을 나타내는 변수를 잘 선택하는 것이 중요하다. 두 번째는 유사성의 측정방법으로 제곱유클리드 거리(Squared Euclidean Dis tance), 유클리드 거리(Euclidean Distance), 민코브스키 거리(Minkowski Distan ce), 마할라노비스 거리(Mahanlanobis Distance) 등이 있다. 마지막은 군집화의 방법을 선택하는 것이다. 군집분석은 기법에 따라 계층적 방법(hierachical)과 비계층적(nonhierachical) 방법인 두 가지로 분류된다. 먼저, 하나는 계층적방법이다. 이 방법에서는 군집

의 위계가 있어서 일단 한 군집에 속하게 된 두 개체는 다시 흩어지지 않는다. 이 방법에는 단일결합법(single-linkage), 완전결합법(complete-linkage), 평균결합법(average-linkage), 와드법(Ward's method), 중심연결법(centroid method) 등이 있다. 다른 하나는 비계층적 방법이다. 이 방법에서는 군집이 형성된 이후에도 일정기준에 따라 개체들이 이합집산을 되풀이하는 과정을 거친다. 이 방법에는 k-평균 군집방법이 대표적이다.

본 논문에서는 설명변수는 288개의 5분 간격의 시간 변수를 가지고, 유사성의 척도는 제곱 유클리드 거리로 한 후, 초기에 부적절한 병합 또는 분리가 회복될 수 없는 계층적 군집과 달리 개체의 재할당이 가능한 k-평균 군집방법을 실시하였다. 이때 288개의 변수는 5분 간격의 자료이므로 24시간 동안의 시간변수는 12*24=288이 되기 때문에 생성된 것이다.

4.1.1 k-평균(k-means)군집 알고리즘(algorithm)

k-평균 군집은 위에서 설명한 바와 같이 비계층적 군집분석의 한 방법으로서 다음과 같은 과정을 통해서 군집을 분류한다.

<표 4.1> k-평균 군집화 알고리즘

단계1	k 개의 각 군집에 초기값을 설정한 후
단계2	모든 개체를 각각 가장 가까운 군집중심을 찾아 할당한 후
단계3	군집의 모든 개체의 평균을 다시 계산 한 후 다시 2단계로 간다.
단계4	변화가 없을 때까지 단계2와 단계3을 반복한다.

각 개체는 k개의 초기 값에 가까운 군집에 일단 할당되고, 그 군집의 중심은 추가적으로 할당된 구성원에 의해 다시 계산된다. 이러한 할당과정에서 군집 내 제곱합이 큰 군집은 다시 분리되고 작은 군집은 서로 병합되는 형식으로 군집이 형성된다. 이때 군집의 개수인 k의 값이 변화될 수도 있다.

4.1.2 군집수 결정 방법

위의 k-평균 군집 알고리즘에서도 볼 수 있듯이 미리 군집수를 지정해야 한다. k-평균 군집화는 군집수 k를 얼마로 지정하느냐에 따라 최종결과가 크게 달라질 수 있다. 또한, 군집화의 결과가 단계1에서의 초기값에 크게 좌우될 수 있다. 다시 말하여, 처음에 어느 개체가 가장 먼저 군집에 진입하느냐에 따라 최종결과가 영향을 받기 쉽다. 때문에 초기값을 임의로 하는 것은 좋지 않은 것으로 알려져 있다. 좀 더 객관적이고 합리적인 군집 수 k와 초기값 선정방법은 계층적 군집화의 결과를 이용하는 것이 좋다(김기영 등, 1990).

여기서 군집수를 결정하기 위한 판정기준으로 다음과 같은 6가지 방법이 있다. (이성규, 2006).

(1) RMSSTD(Root Mean Square Standard Deviation)의 최소화

새로운 형성된 군집의 RMS 표준편차로 값이 작을수록 군집의 균질성이 높다고 할 수 있다.

$$\text{RMSSTD}=\sqrt{\frac{W_K}{v(N_K-1)}} \tag{4.1}$$

$W_K=\sum_{i=C_K}\| X_i-\overline{X_K}\|^2$: K군집 평균벡터 v: 변수 개수 N_K: K군집의 개수

(2) PST2(Pseudo T Squared Statistic)의 최대화

군집수의 변화에 따라 PST2의 변화가 전구간보다 급격히 상승하는 구간의 바로 전 구간을 최적의 군집수로 결정한다. PST2는 단일 변수에 대한 집단 간 평균차이를 검정하는 T검정의 확장된 형태로 여러 변수들 간의 집단 간 차이가 있는지 검정하는 다변량 분산분석과 비슷한 개념이다.

$$\text{PST2}=\frac{B_{KL}}{(W_K+W_L)/(N_K+N_K-2)} \tag{4.2}$$

$$B_{KL}=W_M-W_K-W_L \quad \text{if} \quad C_M=C_K\cup C_L \quad C_K\text{: } K\text{군집}$$

(3) 결정계수(R- squared)의 최소화

결정계수(RSQ)는 다중상관계수의 제곱으로 군집 간 제곱합을 전체 제곱합으로 나눈 것이다. 결정계수는 군집의 분리정도를 나타내는 하나의 척도로써 값이 크면 군집간의 이질성이 높음을 나타낸다.

$$\text{RSQ}=1-\frac{P_G}{T} \tag{4.3}$$

$$P_G=\sum W_J \quad G\text{:군집 수} \quad T=\sum_{i=1}^{n} \| X_i-\overline{X} \|^2$$

X_i: i번째 관측치 $\overline{X}$:표본평균벡터 $\overline{X_K}$: K군집 평균벡터 $|X|$: X벡터의 유클리드거리

(4) 준 부분결정계수(Semipartial R -squared)의 최대화

준 부분결정계수는 동질성 손실(통합전후의 제곱합의 차이)정도를 전체 제곱합으로 나눈 비율로서 값이 작을 수록 군집의 동질성이 크다고 할 수 있다.

$$\text{SPR}=\frac{B_{KL}}{T} \tag{4.4}$$

(5) PSF(Pseudo F Squared Statistic)이 국부적인 최대값

PSF가 군집수의 변화에 따라 국부적 최고점을 보이는 곳에서 최적의 군집수를 결정하는 방법으로 모든 군집들 간의 분리정도를 측정하여 값이 크면 더 이상 군집을 합치는 것이 의미가 없다고 본다.

$$\text{PSF}=\frac{(T-P_G)/(G-1)}{(P_G)/(n-G)} \tag{4.5}$$

n: 관측치수

(6) CCC(Cubic Clustering Criterion)이 국부적인 최대값

군집의 개수에 대한 판정기준으로서 CCC를 시험적용 시키고 있다. 이는 우선 초사각형(hyper-rectangle)상에서 균일분포를 따르고 있다고 여겨지는 점들이 만약 어떤 군집들을 이루고 있다면 이들은 대체적으로 초입방체 형태로 구분되어 있을 것으로 가정하고 있다. 이와 같은 가정 하에서 유도된 CCC 판정기준은 2 내지 N/10정도의 군집수를 CCC의 값에 대해 플롯 했을 때 적절한 군집의 개수와 자료의 구조에 관해 유익한 정보를 제공하고 있음이 알려져 있다. 이를 요약하면 다음과 같다.

ⅰ) 군집당 평균 표본단위의 개수가 10이하이면 CCC의 변동은 매우 자유분방하다.

ⅱ) CCC>2 혹은 3일 때 국부적 최고점이 있으면 이 점에 대응되는 군집의 수가 적절하다.

ⅲ) 계층형 자료에서는 비교적 여러 개의 국부적 최고점이 있게 된다.

ⅳ) 비계층형 자료의 경우에는 국부적 최고점 전에 급격한 증가를 보이고 국부적 최고점 이후에는 완만한 감소를 보인다.

ⅴ) 0< CCC <2구간에는 국부적 최고점이 있으면 가능한 군집이 있으므로 주의 깊게 고찰해야 한다.

ⅵ) 둘이상의 군집의 개수에 대응되는 CCC가 음(-)의 값을 취하고 점차 감소하고 있으면 이는 단봉분포를 의미한다.

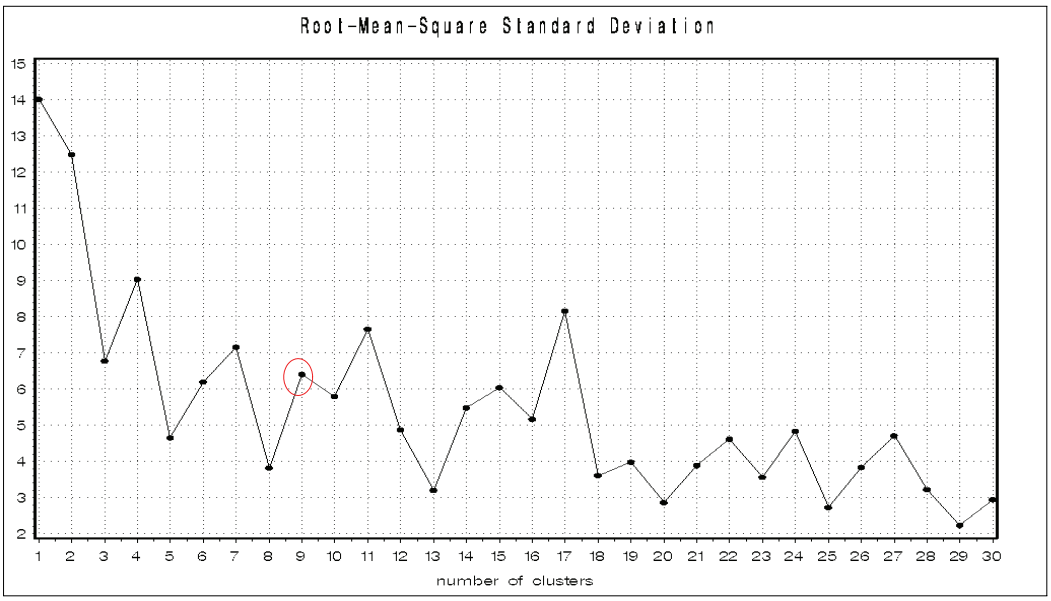
ⅶ) 군집의 개수가 증가함에 따라 CCC의 값이 단조증가하면 이는 지나치게 자료를 반올림했던 가 소수점이하의 값을 너무 많이 잘라 버렸을 때 일어나는 현상이다. 이와 같은 성질을 가지는 CCC 가 군집의 개수를 결정하기 위해 적용되었을 때 가지는 판별력은 적어도 2차원에서 100개의 객체들을 눈으로 가늠한 것만큼 좋은 것

으로 알려져 있다.

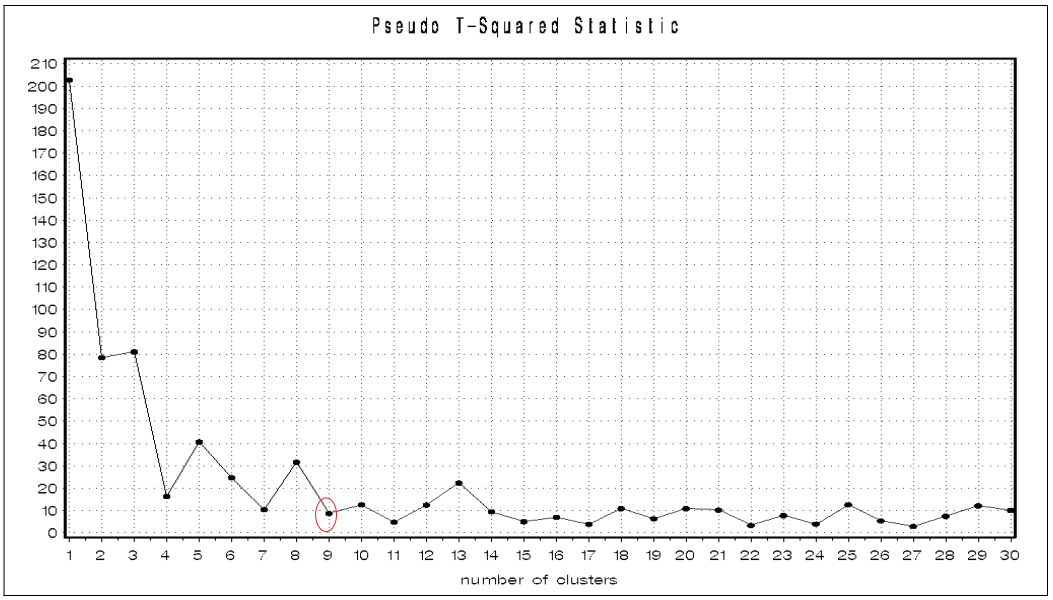
본 논문에서는 군집수의 초기값과 군집수를 정하는 방법에 계층적 방법인 와드의 방법의 결과를 이용하였다. 와드의 방법은 총 군집 내 제곱거리의 오차 제곱합을 최소화하도록 군집끼리 합병한다. 이 방법으로 결정된 초기값과 군집수를 이용하여 k-평균 군집방법을 실시하고자 한다.

4.2. k-평균군집 분석 결과

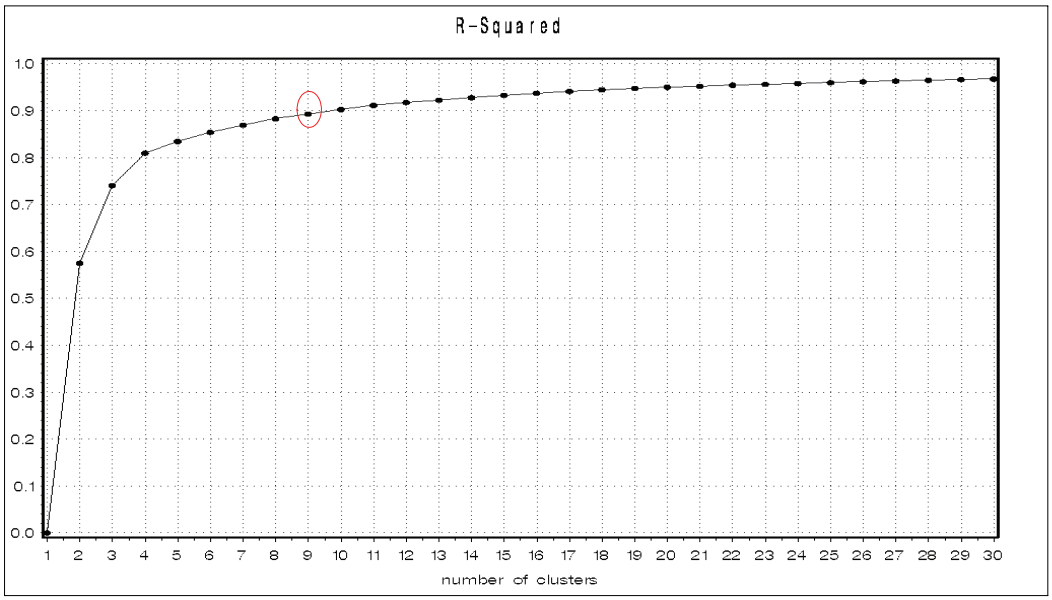
서울시내 자료 중 올림픽 대로의 152구간 자료를 이용하여 군집분석을 실시하였다. 초기 값과 군집 수는 와드의 방법의 결과를 이용하여 구한 뒤, 그 결과를 토대로 k-평균 군집분석을 수행하였다. 와드의 방법으로 초기값과 군집 수를 구하기 위하여 앞 절에서 설명한 6가지 기준 통계량을 살펴보아야 한다. 즉, RMSSTD, PST2, 결정계수, 준 부분결정계수, PSF, CCC 플롯(plot)을 각각의 기준에 따른 통계량을 군집 수에 따라 아래 <그림 4.1>-<그림 4.6>과 같이 그려보았다. 최대의 군집수를 30개까지 살펴보았지만 실제로 10개 이상 나누게 되면 의미를 해석하기도 쉽지 않으므로 5개에서 10개 사이에서 군집수를 정하는 것이 적절하다. 그러므로 5-10사이의 군집수사이에서 각각 최적의 군집수를 살펴보면 RMSSTD는 5, 8, PST2는 5, 8, RSQ는 5, 6, SPR는 5, 6, PSF는 5, 6과 CCC는 9, 10 으로 주로 5, 8이 최적의 군집수로 나타났다. 그러나 본 논문에서는 시행착오 결과 군집의 의미를 부여하는 것이 용이한 군집수가 9개였기 때문에 군집수를 9개로 정하게 되었다.



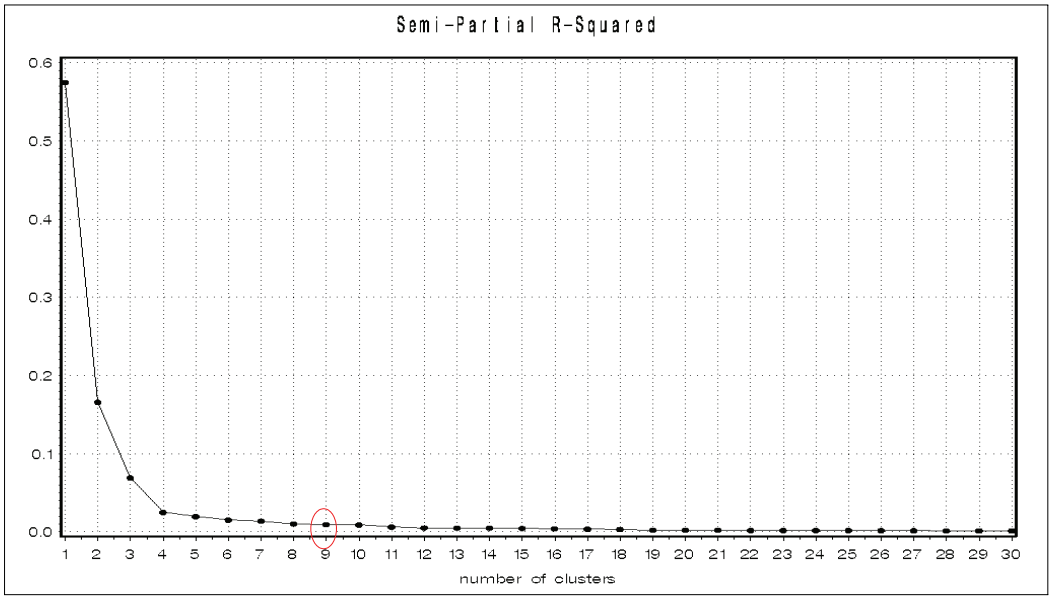
<그림 4.1> RMSSTD(Root-Mean-Squar Standard Deviation)



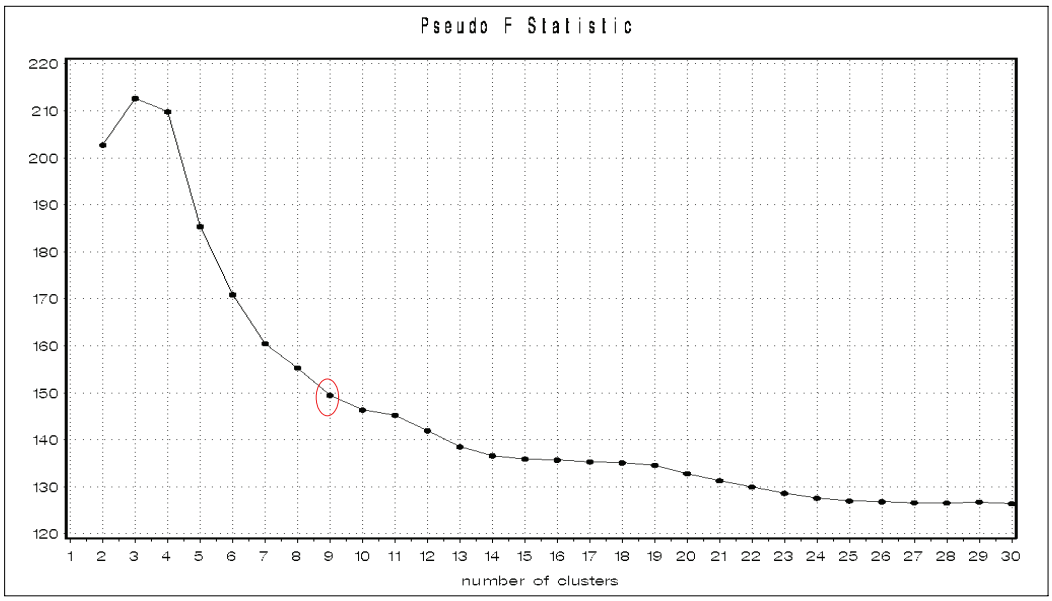
<그림 4.2> PST2(Pseudo T Squared Ststistic) 플롯



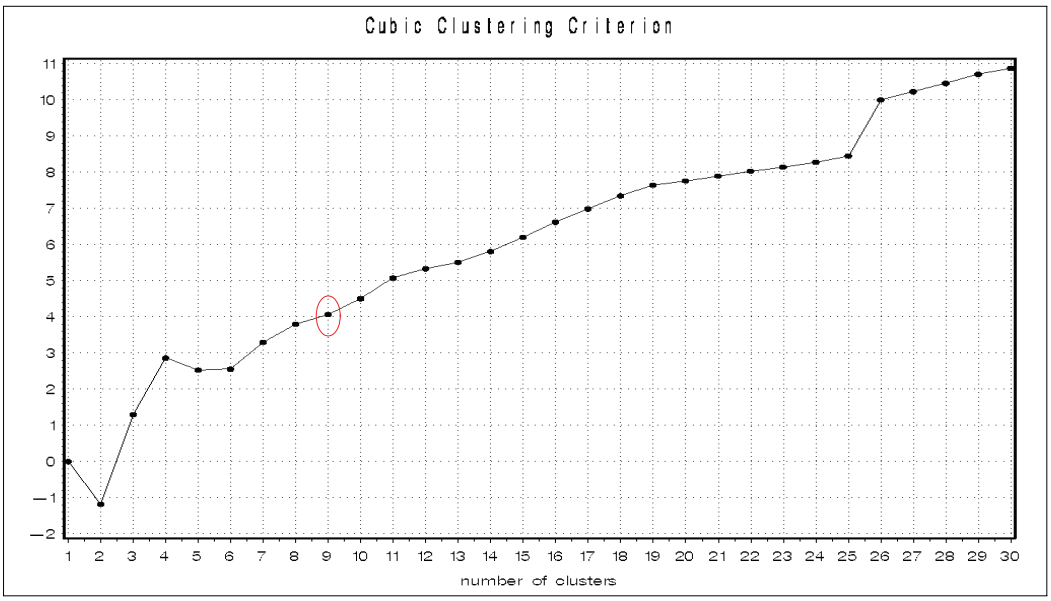
<그림 4.3> 결정계수(RSQ) 플롯



<그림 4.4> 준 부분결정계수(SPR) 플롯



<그림 4.5> PSF(Pseudo F squared Ststistic) 플롯



<그림 4.6> CCC(Cubic Clustering Criterion) 플롯

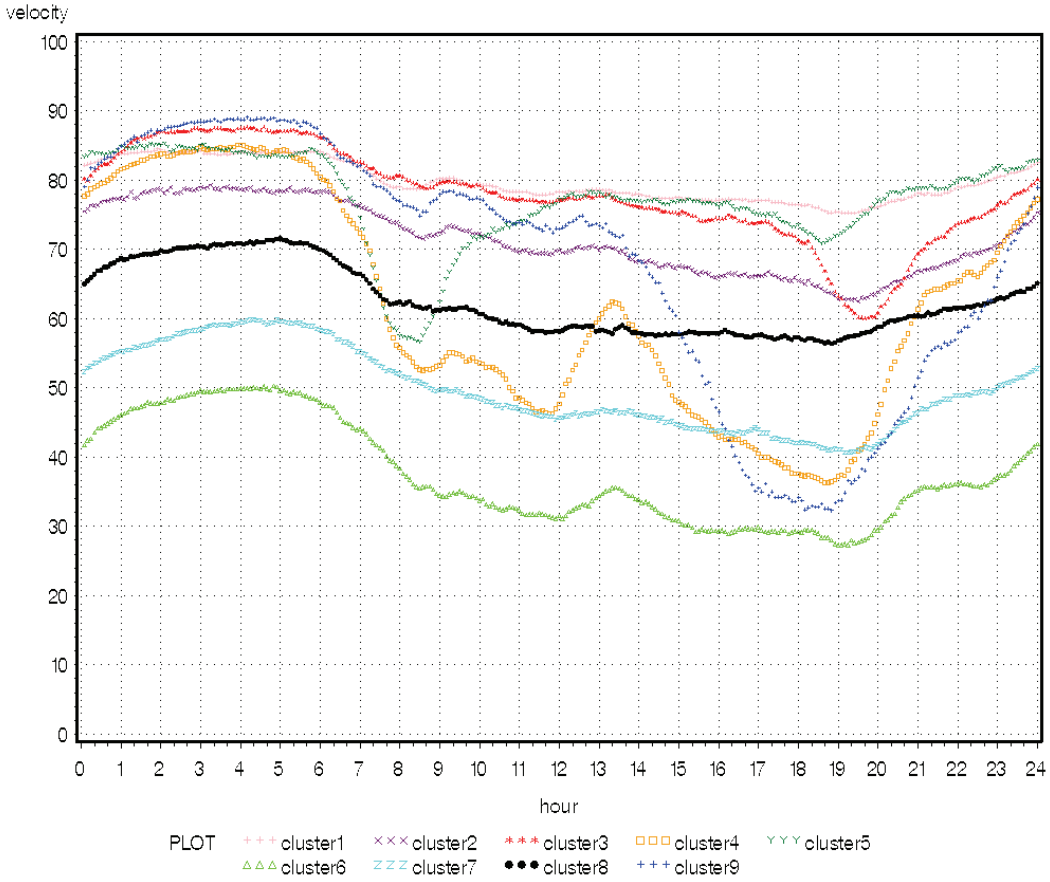
와드의 방법으로 나온 결과를 이용하여 군집수를 9로 정하고, k-평균 군집분석을 실시한 뒤 올림픽대로의 각 군의 패턴을 그래프로 나타내어 보면 <그림 4.7>와 같다. 평균 시속 80km정도에서 속도 증감 폭이 크지 않은 군집 1이 있고 평균 시속 70-80km 정도의 패턴을 나타내는 군집 2, 시속 60-70km정도에서 군집 8, 시속 50km정도에서 군집 7, 시속 30-40km정도에서 군집 6으로 오전부터 낮 시간대인 10시-20시정도 사이에 약간 속도가 떨어지는 특징이 있다. 군집 3은 보통 시속80km정도로 시작하여 오전부터 서서히 속도가 떨어지다가 저녁 7시 무렵 급격히 20km정도 떨어지는 특징이 있다. 군집 4는 오전, 오후가 최대 40-50km까지 급격한 속도 변화를 보이는 곳으로 특히 출퇴근시간에 급격한 속도 저하가 나타난다. 군집 9도 오전, 오후가 최대 50km까지 급격한 속도 변화를 보이는 곳으로 오전 출근 시간보다 퇴근시간에 더 심각한 정체를 보이게 된다. 군집 5도 군집 4, 군집 9와 같이 오전, 오후에 속도 변화가 있는 패턴으로 오후보다는 오전에 최대 25km 까지 변화가 있는 군집이다. 이와 같은 패턴을 보이는 각각의 군집에 따라 군집이름을 붙일 수 있다. 예를 들면 군집4는 출근과 퇴근시간 모두정체를 보이는 구간으로 군집이름을 '출퇴근 정체구간'으로 명명할 수 있다. 마찬가지로 도로구간의 특징에 따라 <표 4.2>와 같이 군집이름을 정할 수 있다.

<그림 4.8>은 올림픽대로의 군집결과를 지도에 나타내 보면 다음과 같다. 각각 김포방향과 하남방향의 구간의 예를 들어 <표 4.2>에 나타내었다. 군집 1이 나누어진 부분을 살펴보면 42개의 구간으로 방화-행주, 가양-방화대교방면, 올림픽-천호대교방면, 군집 2가 나누어진 부분을 살펴보면 13개의 구간으로 한남-반포대교방면, 동작-반포대교방면, 군집 3이 나누어진 부분을 살펴보면 16개의 구간으로 성산-가양대교방면, 여의상류IC-한강대교방면이다. 군집 4가 나누어진 부분을 살펴보면 30개의 구간으로 동작-한강대교방면, 한강-동작대교방면, 군집 5로 나누어진 부

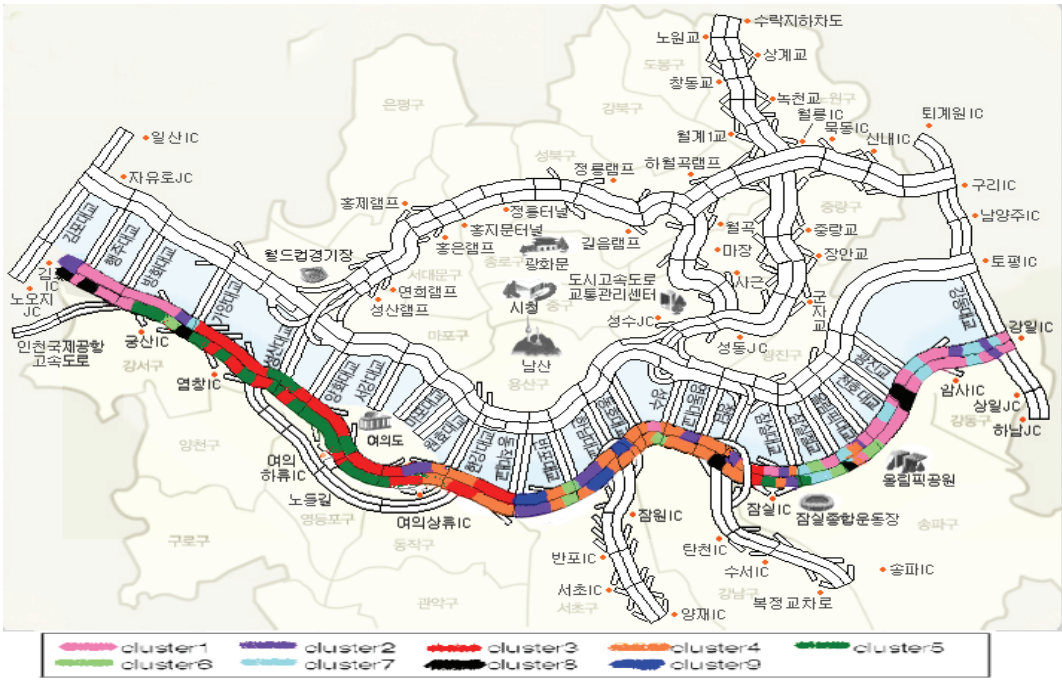
분을 살펴보면 19개의 구간으로 양화-성산방면, 양화-여의하류IC, 군집 6이 나누어진 부분을 살펴보면 8개의 구간으로 반포-동작대교방면, 동작-반포대교방면, 군집 7이 나누어진 부분을 살펴보면 12개의 구간으로 잠실철교-잠실대교방면, 잠실대교-잠실철교방면, 군집 8이 나누어진 부분을 살펴보면 9개의 구간으로 광진교-천호대교방면, 천호-광진교방면이다. 군집 9가 3개의구간으로 동호-한남대교방면, 반포-한남대교방면이다. <그림 4.9>-<그림 4.17>까지는 각 군집별 24 시간 속도패턴과 지도에서 어느 구간이 각 군집에 속하였는지를 나타낸 결과이다.

<표 4.2> 올림픽 대로 자료의 군집분석결과

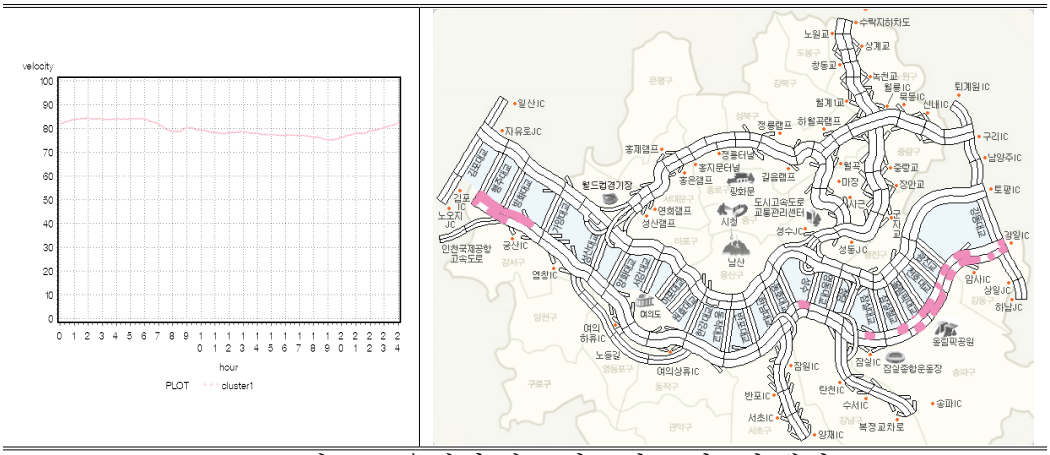
군집	군집 이름	구간 수	군집의 패턴	김포방향 예)	하남방향 예)
1	80km 속도 구간	42	시속 80km정도에서 속도 증감 폭이 크지 않은 특징	방화-행주 가양-방화	올림픽- 천호
2	70-80km 속도구간	13	평균 시속 70-80km	한남-반포	동작-반포
3	미약한퇴근 정체구간	16	보통 시속 80km정도로 시작하여 오전부터 서서히 속도가 떨어지다가 저녁 7시 무렵 20km 떨어지는 특징	성산-가양	여의상류 IC -한강
4	출퇴근정체 구간	30	오전, 오후가 최대 40-50km까지 급격한 속도 변화를 보이는 특징	동작-한강	한강-동작
5	미약한출근 정체구간	19	오후보다는 오전에 최대 25km 까지 변화가 있는 특징	양화-성산	양화- 여의하류 IC
6	30-40km 속도구간	8	평균 시속 30-40km	반포-동작	동작-반포
7	50km 속도 구간	12	평균 시속 50km	잠실철교- 잠실대교	잠실대교- 잠실대교
8	60-70km 속도구간	9	평균 시속 60-70km	광진교-천호	천호-광진교
9	극심한퇴근 정체구간	3	오후퇴근시간에 최대 50km까지 급격한 속도 변화를 보이는 구간으로 오전 출근 시간보다 퇴근시간에 더 심각한 정체 특징	동호-한남	반포-한남



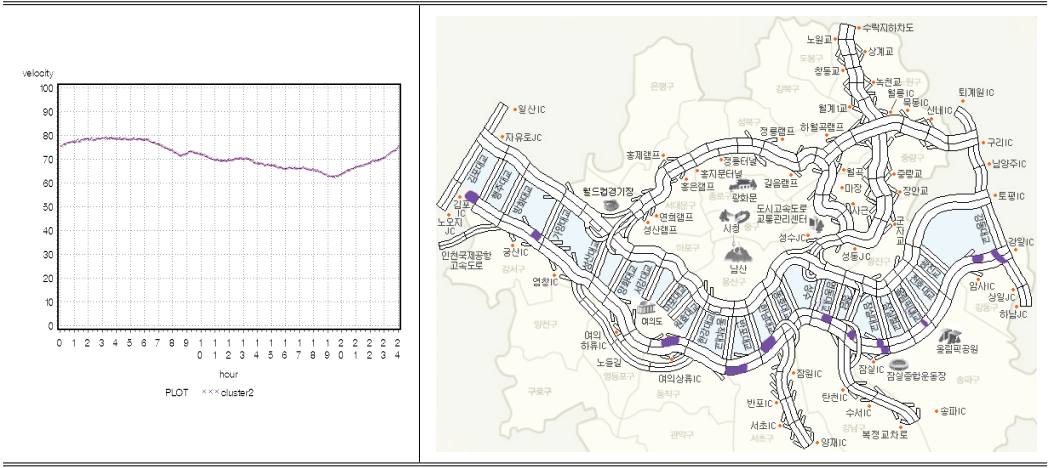
<그림 4.7> 올림픽 대로 자료의 군집분석결과 그래프패턴



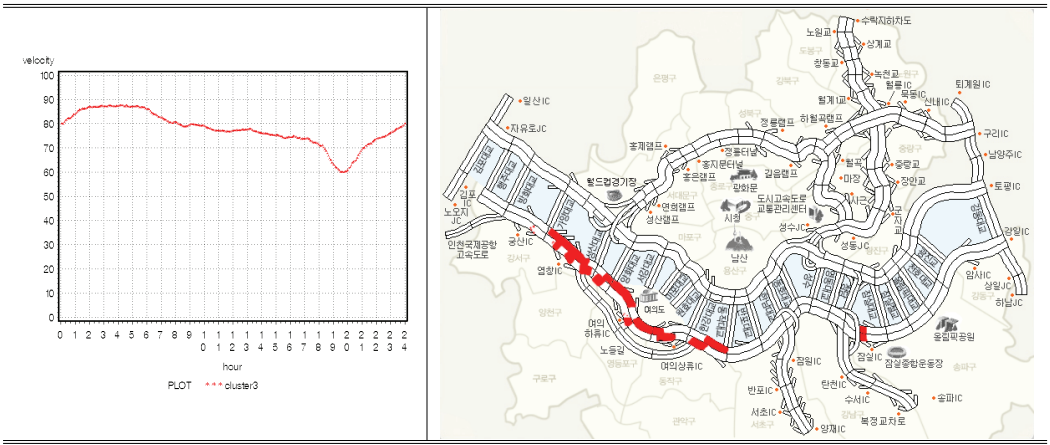
<그림 4.8> 올림픽 대로 자료의 군집분석결과 지도



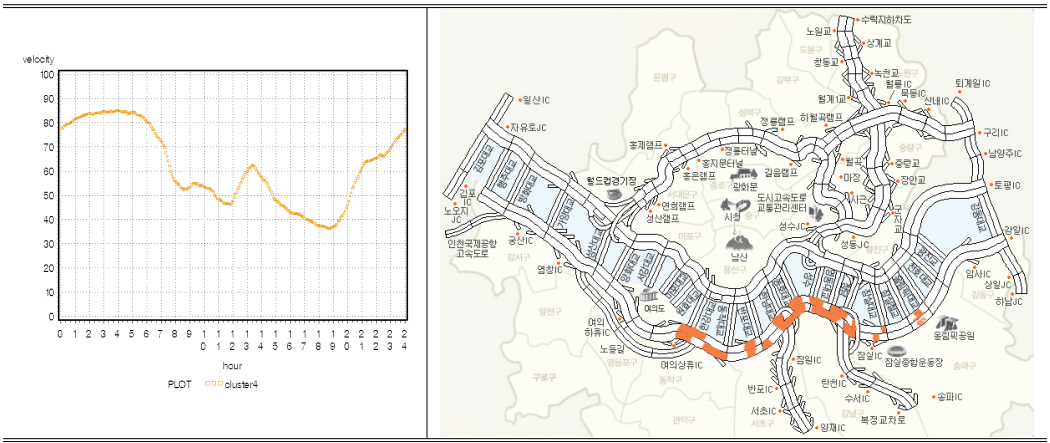
<그림 4.9> 올림픽 대로 자료의 군집 1의 패턴



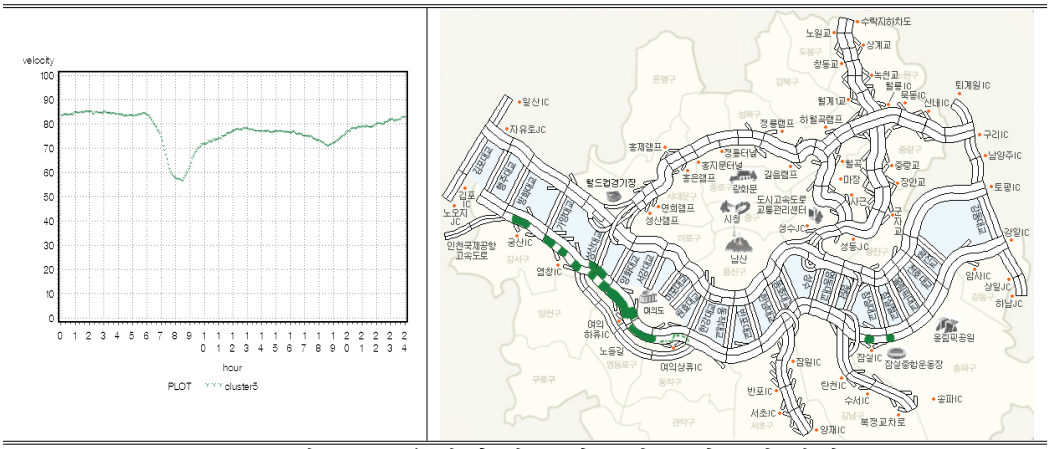
<그림 4.10> 올림픽 대로 자료의 군집 2의 패턴



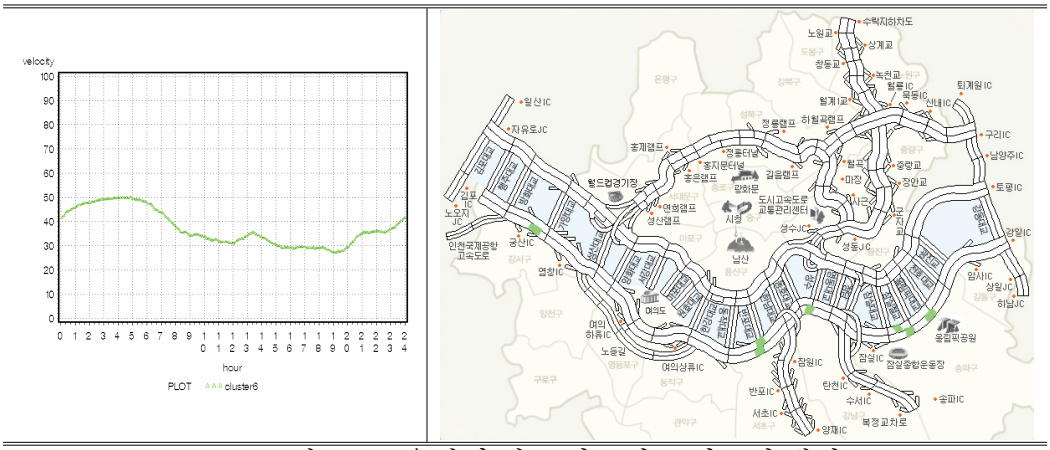
<그림 4.11> 올림픽 대로 자료의 군집 3의패턴



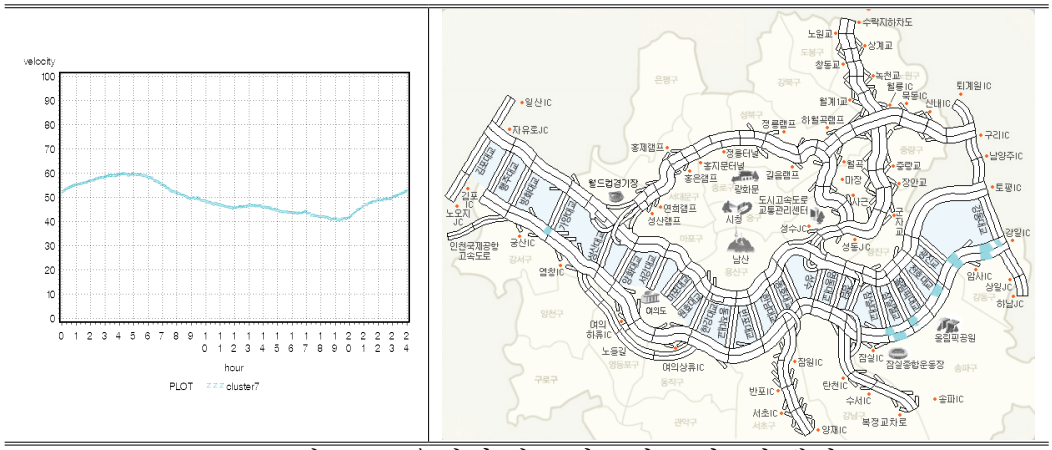
<그림 4.12> 올림픽 대로 자료의 군집 4 의 패턴



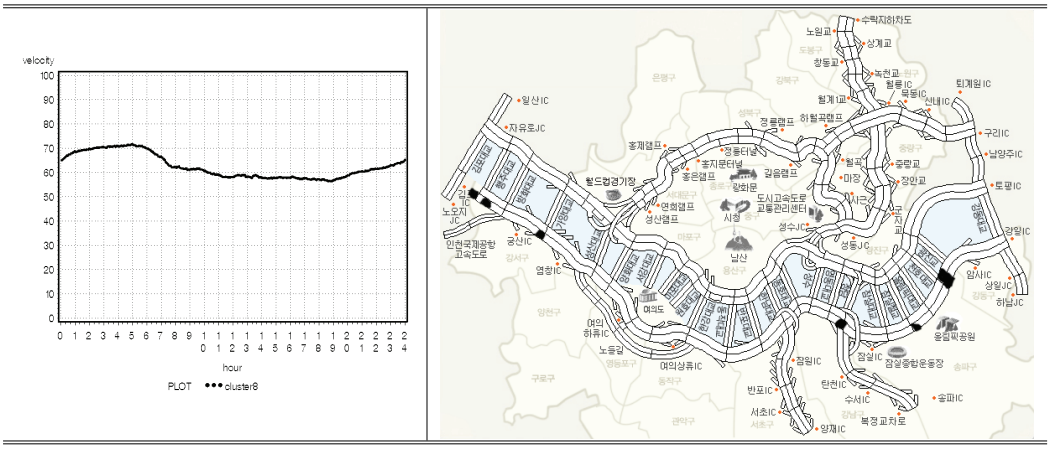
<그림 4.13> 올림픽 대로 자료의 군집 5 의 패턴



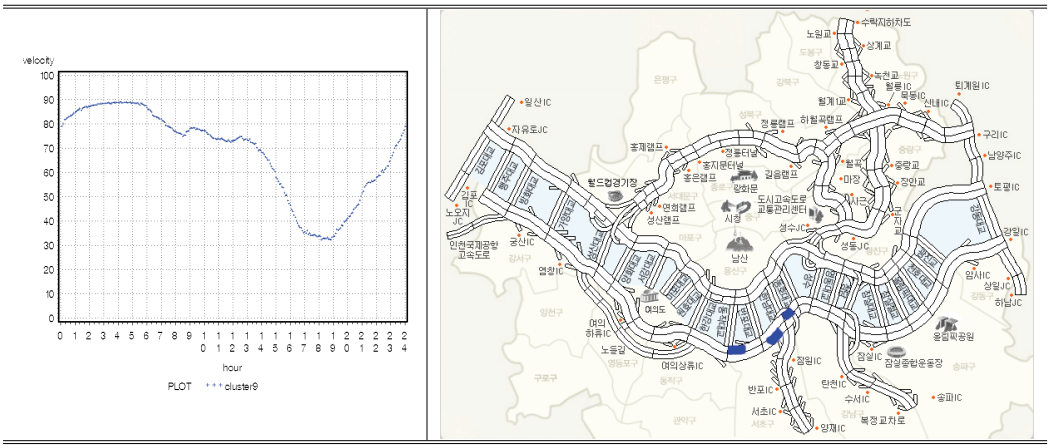
<그림 4.14> 올림픽 대로 자료의 군집 6 의 패턴



<그림 4.15> 올림픽 대로 자료의 군집 7의 패턴



<그림 4.16> 올림픽 대로 자료의 군집 8의 패턴



<그림 4.17> 올림픽 대로 자료의 군집 9의 패턴

제5장. 결론 및 향후 연구과제

본 논문에서는 도로 속도 정보를 이용하여 도로이용자들에게 보다 유익한 정보를 제공하고자 회귀모형을 통한 평균 속도 예측과 비슷한 24시간 속도 패턴을 군집화하기 위한 군집분석을 실시하였다.

먼저, 속도예측모형으로 퓨리에 급수를 이용한 회귀모형과 스플라인 회귀모형을 결정계수를 통하여 비교하였다. 자료는 올림픽대로 강동대교-천호대교(김포방향), 강변북로 성산대교-양화대교(구리방향)의 두 구간에서 24시간 동안의 5분 간격 평균속도 자료를 이용하였다. 삼각함수가 30개인 퓨리에 급수를 이용한 회귀모형의 결정계수는 99.6, 99.5로 매우 높은 값을 갖는다. 반면, 스플라인 회귀모형은 넛의 수가 25이상이면 퓨리에 급수를 이용한 회귀모형의 설명력과 비슷한 결과를 보였다. 본 논문에서는 모형의 설명력을 대략 80-90으로 고정시킬 경우, 각 구간의 각각의 예측모형을 추정하는 데에 소요되는 메모리의 효율성을 감안하여 스플라인 모형의 넛의 수를 5-10정도로 하는 스플라인 모형을 제안한다.

다음으로 본 논문에서는, 24시간 동안의 교통속도 패턴을 알아보기 위해 k-평균 군집화를 통해 비슷한 패턴을 갖는 군집을 구축하였다. 이 때 이 분석에 사용한 자료는 올림픽대로 전구간인 152구간으로 5분 간격 평균속도 자료이다. 와드(Ward)의 방법으로 나온 결과에 시행착오를 통하여 군집수를 9로 정하고, k-평균 군집분석을 실시하였다. 군집분석의 결과 군집들의 속도 패턴의 특징에 따라 80km 속도 구간, 70-80km 속도 구간, 미약한 퇴근정체구간, 출퇴근정체구간, 미약한 출근정체구간, 30-40km 속도 구간, 50km 속도 구간, 60-70km 속도 구간, 극심한 퇴근 정체구간이 군집을 나뉘어진다. 이와 같이 24시간 속도에 대한 패턴이 각 구간마다 다

른 것을 이용하여 도로이용자는 효율적으로 도로를 활용할 것으로 기대된다.

향후 연구과제로서 먼저, 전체 속도자료를 이용하는데 있어서 구간별 5분 간격 평균 속도자료를 이용하여 분석을 하였는데 평균속도 대신 중위수로 대체하여 사용한다면, 특이치도 함께 고려되므로 좀 더 효율적인 분석 자료가 생성될 것으로 기대된다. 또한, 9가지 요일변수를 생성한 것을 단순히 기초적인 자료 분석에만 그치지 않고 속도예측이나 군집 분석 시에도 요일별로 세분화하여 분석할 수 있을 것이다.

속도예측을 위한 퓨리에 급수를 이용한 회귀모형이나 스플라인 회귀 모형을 비교하는 방법으로 결정계수를 이용하였으나, 이외에 다른 방법으로 모의실험을 통한 적중률을 통해 두 모형을 비교해 볼 수 있다. 또한 특이치를 제외한 평균속도 자료를 이용하여 순수한 속도 효과만을 분석에 이용하고자 하였으나, 실제 도로상황에서는 돌발 상황 즉 공사, 사고, 기후 등에 수시로 상황이 변하므로 도로에서의 여러 가지 교통상황을 포함하여 속도를 예측할 수 있는 모형을 개발하는 것이 실제 도로를 이용하는 도로이용자들에게 실질적으로 도움이 되는 모형이 될 것이다.

마지막으로, 군집분석에서는 올림픽대로만을 분석대상으로 하였는데 분석대상을 좀 더 확장시켜 한강을 중심으로 한 주변도로의 속도패턴을 군집화 하기 위해서 강변북로 자료 또한 함께 군집분석을 실시할 수 있을 것이다. 나아가 분석 자료는 패턴에 관한 함수 데이터(functional data)로 볼 수 있으므로 함수 데이터에 적합한 함수데이터 군집(functional data clustering)방법을 적용하여 군집화를 할 수 있을 것이다.

참 고 문 헌

[1] 권성진(1999). *요일별 가구통행행태를 고려한 통행발생예측*, 연세대학교 대학원, 도시공학과.

[2] 김기영, 전명식(1990). *SAS 군집분석*, 자유아카데미, 서울.

[3] 김종태, 이성호, 김경무(1997). 퓨리에 급수기법에 의한 밀도함수 추정의 최적화 고찰, *Journal of Statistical Theory & Methods*, Vol.8, No.1, 9-20.

[3] 박전수(2007). *MATLAB 수치해석입문*, 아진출판사, 서울.

[4] 이성규, 홍성언, 박수홍(2006). 평균연결법과 k-means 혼합클러스터링 기법을 이용한 공시지가 유사가격권역의 설정, *대한지리학회지*, 제 41권, 제 1호, 121-135.

[5] 이해기(2009). *공업수학*, 태영문화사, 서울.

[6] 한상태, 강현철, 이성건, 최보승(2007). *통계교통정보분석 최종보고서*, 팅크 웨어.

[7] 허명회(2005). *사회과학을 위한 다변량 자료분석*, 자유아카데미, 서울.

[8] John A. Hartigan(1975). *Clustering Algorithms*, John Wiley & Sons, New York.

[9] Lawlence C. Marsh(2002). *Spline Regression Models*, Sage Publication, California.

[10] Qingzhi Guo & Ralph E. White(2004). *Cubic Spline Regression for*

the Open-Circuit Potential Curves of a Lithium-Ion Battery, University of South Carolina, Department of Chemical Engineering.

ABSTRACT

A study on the traffic velocity patterns using statistical modeling

Aeran Park
Department of Statistics
The Graduate School
Sungshin Women's University

As the number of cars increases, the road traffic conditions becomes complex. The aim of this study is to discriminate roads using statistical methods as well as to identify traffic characteristics of roads. To accomplish the objectives, regression analysis and cluster analysis are performed for forecasting mean velocity and classifying road using traffic data that are collected in the Olympic express way and the Gangbyeon express way during the years 2005-2007.

For forecasting traffic mean velocity, firstly we present that spline regression model with 5-10 knots. This model is more efficient than regression model using Fourier series with 30 trigonometric function, in

that the estimated parameter of our model is the smaller than the other model.

Secondly, we propose a daily velocity pattern determination method using k-means cluster analysis. The optimal number of clusters through the Ward's method is yielded to nine clusters, and each cluster shows a different velocity pattern. K-means clustering on the traffic velocity patterns enables us to examine overall traffic patterns.

감사의 글

늦게 대학을 진학한 지가 엊그제 같은데 벌써 대학원까지 마치고 졸업을 준비하니 감개가 무량합니다. 두려움과 기대를 가지고 시작한 대학원 생활동안 기쁘고 슬프고 힘든 일들이 있었습니다. 이렇게 졸업을 한다니 한편으로는 아쉬운 점이 많지만 제 앞날의 좋은 영양분이 될 것임을 믿어 의심치 않습니다.

저의 미흡한 논문을 완성하기 까지 많은 분들의 은혜를 입었습니다. 조용히 진리를 깨우쳐 주시는 송일성 교수님, 큰 꿈을 갖게 해주신 이해용 교수님, 따뜻한 관심을 보여 주시는 이우선 교수님, 열의를 가지고 꾸짖어 주시는 이종협 교수님께 고개 숙여 감사의 말씀을 전하고 싶습니다. 마지막으로 항상 논문의 목표를 잃지 않도록 가르침을 주신 지도교수님인 이성건 교수님께 뭐라고 감사의 말씀을 드려야 할지 모르겠습니다.

특히 일을 하는 동안에 남은 석사생활을 물심양면으로 도와주신 연세대 의과대학 연구부 성지민 교수님과 송혜령 선생님, 김 하나양에게 감사의 말씀을 전하고 싶습니다. 자료 정리를 도와준 이든, 대희, 소연에게도 감사의 말을 전합니다.

옆에서 많은 조언을 주었던 영은, 주현, 희라, 가영과 2년간 같이 고생하였던 희원, 인경에게도 고마움을 전합니다. 우리 대학원의 귀여운 후배들 경혜, 정윤, 보미, 하얀에게는 남은 대학원생활 열심히 하여 좋은 결실 있기를 기원합니다. 또한 항상 마음의 위안이 되었던 혜란, 윤주, 효정에게 감사의 마음을 전합니다.

어려울 때나 즐거울 때나 내편이 되어주고 옆에 있어준 윤정, 정민, 명하, 수연, 성윤, 10년 만에 연락하였는데도 반갑게 도움의 손길을 준 상훈에게 진심으로 고마운 마음을 전합니다. 바쁘다며 제대로 만나지 못한 수경, 은하, 기원, 지원에게 미안

한 마음을 전합니다. 멀리 타지에서 고생하는 지현에게도 격려의 말을 전하고 싶습니다.

마지막으로 오늘의 제가 있도록 도와주신 부모님과 동생, 그리고 P에게 작은 보답이 되길 바라며 이 논문을 바칩니다.