

Cluster analysis for highway speed according to patterns and effects

Byungsoo Kim^{a,1} · Soyoung An^a · Jungmin Son^a · Hyemi Park^a

^aDepartment of Statistics, Inje University

(Received May 30, 2016; Revised July 25, 2016; Accepted August 8, 2016)

Abstract

This paper uses all sections of highway data (VDS) for two years (Jan. 2014–Dec. 2015), with 15 minute units. The first purpose of this study is to find clusters with similar patterns that appear repeatedly with time variables of month, week and hour. The cluster analysis results indicate a variety of patterns of average traffic speeds by time variables depending on the clusters; subsequently, these can be utilized to model for the forecast of the speed at a specific time. The second purpose is to do cluster analysis for grouping sections by effect nets that are closely related to each other. For the similarity measure we use cross-correlation functions calculated after pre-whitening the speed of each section. The cluster analysis gets 19 clusters, and sections within a cluster are geographically close. These results are expected to help to forecast a real-time speed.

Keywords: cluster analysis, cross-correlation function, pattern, sections of highway, speed

1. 서론

고속국도법 제2조(용어의 정의)에 의하면 고속도로는 중요한 도시를 연결하는 자동차 전용의 고속교통에 속하는 도로이며 1968년에 경인고속도로가 개통된 이래로 2015년 12월 기준으로 42개 노선에 1,066개 구간이 있다. 고속도로 통행량은 꾸준히 증가하고 있으며(2010년 13.77억대, 2014년 15.46억대; Korea Expressway Corporation, 2015), 이에 맞춰 고속도로도 신설 및 확장이 지속적으로 진행되고 있으나 늘어나는 고속도로 이용 수요를 충족시키기에는 역부족이기 때문에 정확하고 풍부한 교통정보의 필요성이 대두되고 있다. 이런 필요에 따라 TV, 라디오, 인터넷, 스마트폰, 도로전광표지(variable message sign; VMS) 등의 많은 도로관련 정보 매체들이 실시간으로 교통 정보를 제공하고 있으며, 컴퓨터 기술의 발달로 차량검지기시스템(vehicle detection system; VDS), 고속도로 통행료 수납 시스템(toll collection system; TCS), 단거리무선통신(dedicated short range communication; DSRC) 등을 이용하여 많은 교통 관련 데이터들이 수집되고 있다. 또한 차량의 내비게이션(navigation)과 스마트폰의 앱 등은 길을 안내하는 기능 뿐 아니라 속도를 예측하는 정보를 제공하고 있다.

Lee 등 (2003)은 검지기별 국도 3호선 상행 교통량 데이터를 이용하여 교통량 데이터에서 결측이 발생했을 때 원활한 교통정보 제공을 위한 실시간 보정 방법에 대해 연구하였으며, 결측이 발생했을 경우 주

This work was supported by the 2015 Inje University research grant.

¹corresponding author: Department of Statistics/Institute of Statistical Information, Inje University, 197 Inje-ro, Gimhae, Gyeongnam, 50834, Korea. E-mail: statkbs@inje.ac.kr

중, 토요일, 일요일로 구별하여 보정하는 방법을 제시하였고, 결측 검지기와 교통 패턴이 유사한 인근 검지기를 찾을 수 없다면 과거 프로파일을 이용하는데 이때는 직전 5주의 과거 데이터의 산술평균을 이용한 보정이 타당하다고 주장하였다. Kim과 Cho (2008)에서는 일반국도의 교통특성에 따른 도로를 분류하기 위해 계층적 군집방법을 비교 연구하였고 12개의 설명변수로부터 요인분석에 의해 얻은 4개의 요인점수로 군집분석을 실시하였으며, 그 결과 Ward 방법이 가장 우수하였고 총 5개의 군집이 적절하다고 결론 내렸다. Cho 등 (2011)에서는 경기도 안산시 3개 구간의 시계열 자료를 이용하여 거시적인 관점과 미시적인 관점으로 나누어 요일별로 패턴을 분류하여 유사한 통행패턴을 보이는 요일을 5분단위로 그룹핑하는 알고리즘을 개발하였다. Lee 등 (2014)은 경부고속도로 데이터를 활용하여 IC-IC구간 단위의 통행패턴을 의사결정나무 기법을 이용하여 요일을 기준으로 5개 그룹으로 분류하였고, ‘직전월 소통상황’의 추가 변수가 분류의 정확성을 높였다고 주장하였다. 또한 Choi 등 (2009)은 서울시내 교통자료와 1년간의 고속도로 구간 자료를 이용하여 통행속도 패턴을 분석하여 예측 모델을 구축하기 위해 도로 구간별 속도 자료만을 이용하여 푸리에 변환을 적용하고 삼각함수를 설명변수로 하는 시계열 회귀모형을 이용한 예측모형을 개발하여 구간별 및 시간대별 평균 속도를 예측하였다. Flaherty (1993)은 월별요인을 비계층적 군집분석인 k-평균 방법을 사용하여 연평균 일교통량(AADT)를 추정하였다. 기존의 연구들은 주로 일부 도로 또는 구간에 한정되어 있는 실험적인 연구이고 비슷한 패턴을 갖는 요일 또는 월을 묶는 연구가 다수였으며, 전국 단위로 패턴이 비슷한 구간들을 묶는 연구는 부족한 실정이다.

패턴이란 사전적 의미로는 “되풀이되는 사건이나 물체의 형태”이다. Lee 등 (2014)은 패턴을 “고속도로 구간별로 동일한 시간대에 반복적으로 발생하는 통행속도 추이”로 정의하였다. 고속도로에서 시간대별로 통행속도의 변화가 작은 구간들이 있고 변화가 심한 구간들이 있다. Choi 등 (2009)에서는 하루 24시간 사이에서의 속도 그룹핑을 통해 변화가 작은 시간대에 대해서는 하나의 값으로 예측하는 것을 제안하고 있다. 마찬가지로 구간에 따라 통행속도에 대한 패턴이 다양하기 때문에 모든 구간에 대해 하나의 모형으로 고속도로 통행시간을 예측하는 것은 무리가 있으며, 비슷한 패턴을 가지는 구간들을 군집으로 묶고 군집별로 통행시간을 예측한다면 정확도는 더욱 높아질 것으로 보인다. AASHTO (1994)의 정의에 따르면 통행속도는 정상적인 도로상황에서의 최고속도이지만, 실제 통행속도는 기상상황이나 사고 등에 의해 일시적인 속도 변화를 가질 수 있으며 이러한 변화는 주변의 구간들에 대해 영향을 미칠 수 있다. 따라서 이러한 영향들을 고려한다면 실시간으로 예측을 갱신하는데 도움이 될 것이다.

본 연구의 목적은 전국의 고속도로 자료를 이용하여 월·요일·시간대의 시간변수에 따라 반복적으로 나타나는 패턴이 비슷한 구간들의 군집을 구하는 것과, 통행속도에 있어서 직접 또는 간접적으로 영향을 주고받는 등 밀접한 관계가 있는 같은 영향권에 있는 구간들을 묶기 위해 군집분석을 하는 것이다. 본 연구의 성과는 미래의 특정한 일시 또는 실시간의 통행속도 예측을 위한 모형구축에 도움을 줄 것이다.

본 논문의 구성은 다음과 같다. 2절에서는 자료수집방법을 설명하고 통행속도의 기초분석을 통하여 자료의 특성을 파악하였다. 3절에서는 2절의 특성을 고려하여 유의성 분석 후 시간변수에 따른 패턴을 이용한 군집분석을 시행하고 각 군집들의 특징에 대해 알아보았다. 4절에서는 통행속도에 대해 구간들 사이의 교차상관함수를 이용한 군집분석을 시행한 후 각 군집들의 특징에 대해 살펴보았다. 마지막 5절에서 이에 대한 결론과 향후 과제에 대하여 논하였다.

2. 자료수집 및 기초분석

2.1. 자료수집

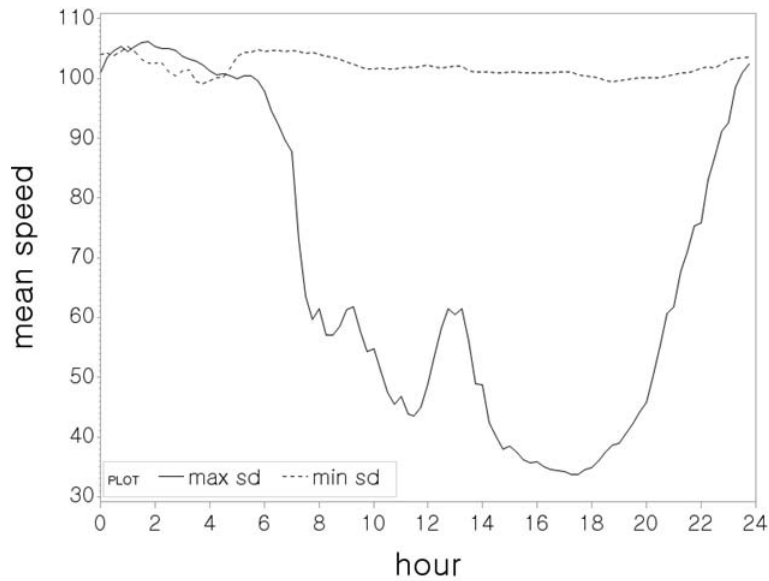
2015년 12월 기준으로 우리나라 고속도로는 총 42개의 노선에 1,066개 구간(conszone)이 있다. 본 논문에서는 한국도로공사 내 고속도로 공공데이터 포털에서 제공하는 통행속도(vehicle detection sys-

Table 2.1. Summary statistics of whole highway sections speed

N	Mean	Median	Standard deviation	Maximum	Minimum	Skewness
64,546,017	97.05	98.75	12.34	180.00	1.33	-2.08

Table 2.2. Summary statistics of mean speed by section

N	Mean	Median	Standard deviation	Maximum	Minimum	Skewness
929	97.05	98.23	7.14	111.19	57.67	-1.04

**Figure 2.1.** Mean speed of Seoul Ring highway(Jangsu IC-Songnae IC) and Incheon International Airport highway(Airport Nooji JC-New Airport TG) sections which have the maximum and minimum standard deviation respectively.

tem; VDS) 중 2014년 1월부터 2015년 12월까지 2년간의 모든 구간에 대해 15분 단위로 관측된 자료를 사용하였다. 자료의 단위는 시속(km/h)이다. 2년간의 연구기간 중에 신설되거나 없어진 구간 또는 기계결합 등으로 5%가 넘는 결측치를 가지는 구간은 연구대상에서 제외하였고 최종 929개 구간을 선정하였다. 15분 간격의 전체 통행속도 자료 수는 64,801,310개이고 이 중 0.39%가 결측이었다.

2.2. 통행속도의 기초분석

Table 2.1은 고려된 929개 구간의 전체 통행속도에 대한 기초통계량이다. 왜도는 -2.08로 음수이고 오른쪽으로 치우친 모양이었다. 최댓값은 180으로 매우 큰 값을 보이는데 제2경인선의 문학IC-학익JC 구간 등의 여러 구간에서 23시 30분과 4시 30분 사이의 밤 시간대에 관측되었으며, 최솟값은 1.33으로 대구부산선의 밀양IC-청도IC 구간의 2015년 3월 30일 2시 15분에 관측되었다. 최댓값과 최솟값을 보이는 시간대의 통행 차량은 대부분 한 대였다.

Table 2.2는 고려된 929개 구간의 평균속도에 대한 기초통계량으로 왜도는 -1.04로 음수였다. 최댓값

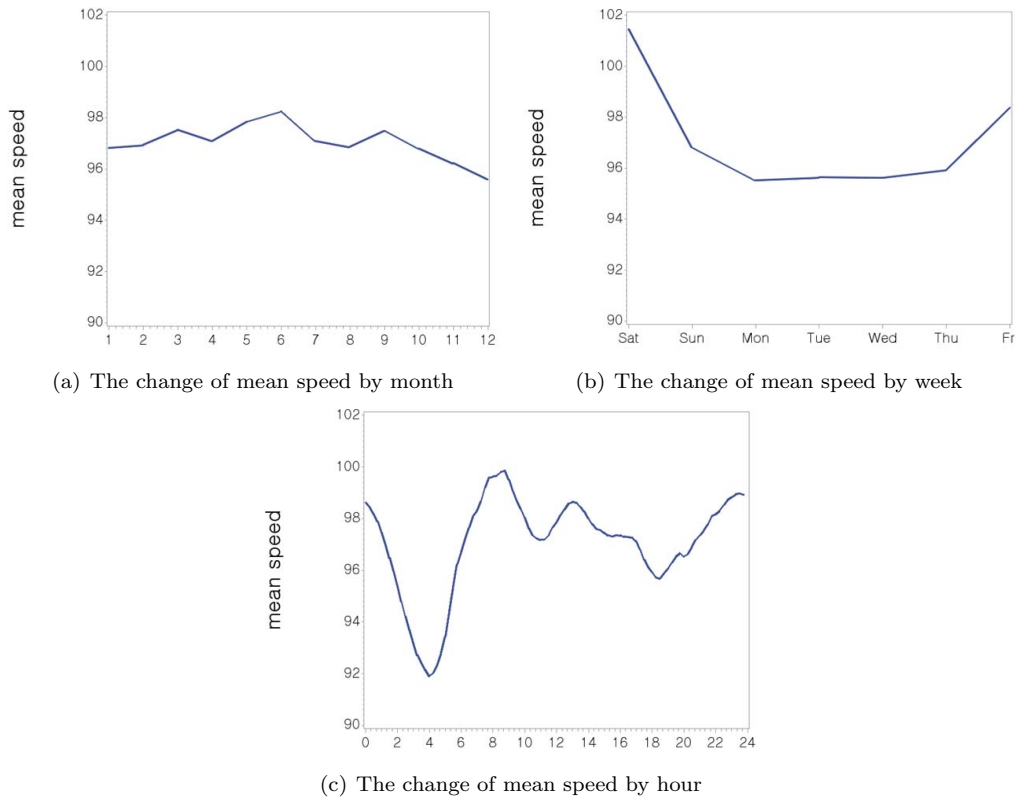


Figure 2.2. The change of mean speed of whole sections by (a) month, (b) week, and (c) hour.

은 111.19로 남해선(영암-순천)의 서영암IC-영암시점 구간이었고, 최솟값은 57.67으로 서해안선의 일직JC-금천IC 구간이었다.

구간별 통행속도의 변동을 비교하기 위해 각 구간별로 표준편차를 구하였다. 가장 큰 표준편차는 33.51로 서울외곽순환선의 장수IC-송내IC 구간이었고, 가장 작은 표준편차는 3.35로 인천국제공항선의 노오지JC-신공항TG 구간으로 표준편차에서 약 10배의 차이를 보였다. Figure 2.1은 두 구간에 대해 시간별 평균을 나타내는 시도표이며, 두 구간에서 변동의 차이가 크고 시간대별 패턴도 다르다는 것을 알 수 있다.

본 논문에서는 통행속도에 영향을 주는 주요 요인으로 월·요일·시간대를 고려하였다. Figure 2.2는 각 요인에 따른 통행속도의 전체 구간 평균을 나타내는 그림으로 월별 평균은 통행속도가 5월과 6월에 높았고 11월과 12월에 낮았으며, 요일별로는 토요일과 일요일에 높았고 주중에는 대체로 낮았다. 시간대별로 보면 4시경이 가장 낮았으며 11시15분경과 18시30분경에도 주변 시간대에 비해서 낮았다.

3. 시간변수에 대한 유의성 분석 및 군집분석

3.1. 시간변수에 대한 유의성 분석

각 구간의 통행속도는 구간마다 다른 통행속도의 수준, 월·요일·시간대와 같은 시간변수에 따라 반복적

Table 3.1. Three-way ANOVA table for speed

Source of variation	SS	df	MS	F	p-value
Month	2.75E+07	11	2,499,904	27,789	<.0001
Week	26.50E+07	6	44,186,646	491,177	<.0001
Time	23.10E+07	95	2,431,005	27,023	<.0001
Month × Week	1.71E+07	66	258,456	2,873	<.0001
Month × Time	7.64E+07	1,045	73,116	813	<.0001
Week × Time	8.29E+07	570	145,516	1,618	<.0001
Month × Week × Time	2.55E+07	6,270	4,061	45	<.0001
Error	581.00E+07	6.45E+07	90		
Total	654.00E+07	6.45E+07			

으로 나타나는 패턴, 휴가와 명절과 같이 특정한 일시에 나타나는 효과, 사고나 작업과 같이 실시간으로 일어나는 사건에 의한 효과 등 다양한 요인에 의해서 결정된다. 본 논문의 목적은 평상시의 통행속도를 분석하는 것이므로 시간변수에 따른 패턴의 효과만 고려하였다. 구간마다 다른 통행속도의 수준은 고려 대상이 아니며 구간 자료들을 통합하여 분석하기 위해 이후의 분석에서는 각 구간별로 평균속도를 뺀 자료를 사용하였다. 따라서 변환된 자료에서는 각 구간별 평균속도가 0이 된다. 변환된 데이터로 분산분석 한 결과는 Table 3.1과 같으며 월·요일·시간에 따라 속도의 차이가 유의하고 모든 2차와 3차 교호작용도 유의한 영향을 주는 것으로 나타났다. *F*-값으로 비교하여 보면 주효과와 비교하여 교호작용의 효과는 현저히 작았고, 주효과 중에서는 요일 효과가 가장 컸으며 월, 시간 효과는 비슷하였다.

3.2. 시간변수에 따른 패턴을 이용한 군집분석

월·요일·시간대의 시간변수에 따라 통행속도의 패턴이 비슷한 구간들을 묶기 위해 군집분석을 하였다. 원자료를 사용하는 경우 목적에 맞지 않게 통행속도의 수준에 따라 구간들이 묶일 가능성이 있기 때문에 구간별로 평균속도를 뺀 변환된 자료를 사용하였다. 개체는 각 구간으로 하였으며 사용한 변수는 각 구간별 월평균(12개), 요일평균(7개), 15분 간격의 시간대별 평균(96개)으로 총 115개 변수를 사용하였다. 변수의 수가 많고 변수들 간에 상관관계가 크기 때문에 주성분분석을 통하여 변수를 축약하였다. 주성분분석에서 공분산행렬을 사용하였으며 고유값이 115개 변수들의 분산의 평균(23.07) 이상인 6개의 주성분을 선택하였고 6개 주성분이 전체변동의 96.19%를 설명하였다.

6개의 주성분을 변수로 한 929개 구간에 대한 군집분석은 Ward 방법의 계층적 군집방법으로 군집 수를 정한 후 k-평균 방법을 이용하였다. 군집의 개수를 정하는 판정기준 중 R-square는 개체들 간의 총변동량 중에서 군집 간의 변동에 의하여 설명되는 비율이며, 그 값이 급격히 증가하다가 완만해지는 지점에서 군집의 수를 결정하게 된다. 본 연구에서는 R-square를 판정기준으로 하여 군집 수를 6개로 정하였고 그 때의 R-square는 0.726이었다.

Table 3.2는 군집분석 결과로 각 군집의 빈도는 군집1과 군집4에서 많았고 군집2와 군집6에서 매우 적었다. 평균속도는 원자료에서의 군집별 평균 통행속도이며 군집1이 가장 높았고 군집2가 가장 낮았다. 표준편차의 평균은 각 군집에 속하는 구간들의 속도의 표준편차에 대한 평균으로 군집2가 가장 컸으며 군집1이 가장 작았다. Table 3.2에서 고속도로 구간은 각 군집의 중심으로부터 가장 가까운 위치에 있어서 각 구간을 대표한다고 할 수 있는 구간이다.

Figure 3.1은 변환된 자료에 대해 시간대에 따른 군집별 평균속도를 나타내는 것으로 군집1은 새벽 시간대에 속도가 약간 낮았고 다른 시간대는 일정한 속도를 유지하고 있다. 군집2는 새벽 시간대에 속도가 높았다가 낮 시간대에 낮은 평균속도를 가지며 밤 시간대에 속도가 높아지는 변동이 큰 패턴을 보인다.

Table 3.2. Cluster analysis summary

Cluster	Freq	Mean speed	Mean of standard deviation	Highway	Section of highway
1	617	99.38	8.11	Jungang Jungang Namhae	S,Andong IC-Uiseong IC Sillim IC- S.wonju IC W.Suncheon IC-Suncheon IC
2	2	62.89	33.16	Seohaean Seoul Ring	Iljik JC-Geumcheon IC jangsu IC-Songnae IC
3	38	87.22	18.64	Seoul Ring Yeongdong Honam	Pangyo JC-Cheonggye TG Wolgot JC-Seochang JC E.Gwangju IC-Yongbong IC
4	233	94.92	9.24	Seoul Ring Yongin-Seoul Yongin-Seoul	Goyang IC-Tongillo IC Geumgwang Underway N.-Geumgwang Underway S. W.Pangyo IC N.-unjung Tunnel S.
5	36	83.57	16.83	Yeongdong 2nd Gyeongin Seoul Ring	Dundae JC-Gunpo IC Namdong IC-Seochang JC Jonam JC-Sanbon IC
6	3	90.16	23.35	Daegu-Pohang Daegu-Pohang Seoul-Yangyang	N.Yeongcheon IC-W.pohang IC W.pohang IC-N.Yeongcheon IC Misa IC-DeoksosampaeIC

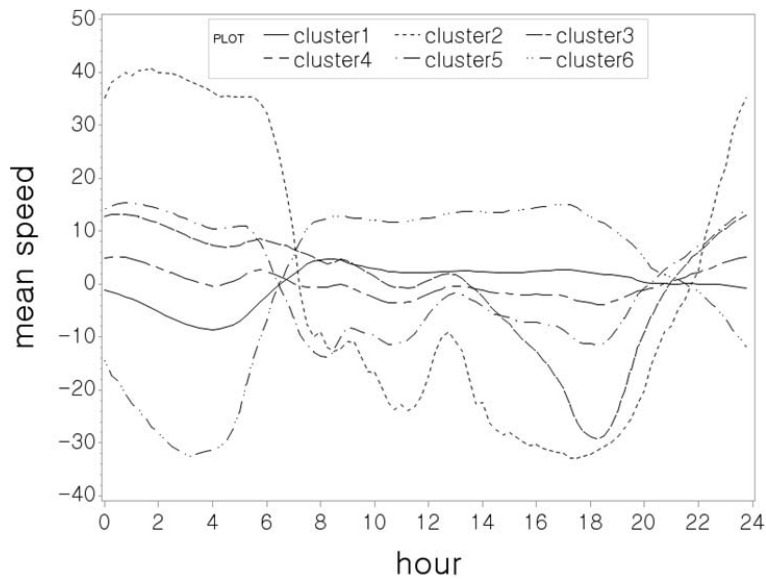


Figure 3.1. Mean speed of the centering data by the clusters.

군집3은 밤 시간대에만 속도가 낮아지고, 군집4는 일정한 평균속도를 유지하고 있으며 이는 군집1의 일정한 속도를 유지하는 특성과 유사하고 평범한 구간으로 보인다. 군집5는 밤과 새벽 시간대에 속도가 약간 높고 낮 시간대에 약간 낮아지는 패턴을 보인다. 군집6은 밤과 새벽 시간대에 낮은 평균속도를 가지고 낮 시간대에는 높은 평균속도를 가지는 구간들로 큰 변동을 보인다. Figure 3.2는 원자료에 대해

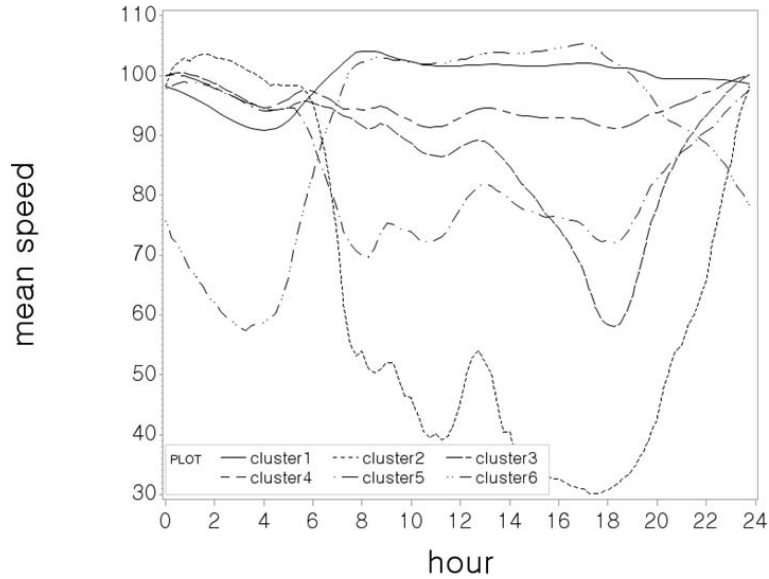


Figure 3.2. Mean speed of the non-centering data by the clusters.

시간대에 따른 군집별 평균속도를 나타내는 그림으로 Figure 3.1의 변환된 자료와 비슷한 패턴으로 묶여있음을 알 수 있다.

4. 통행속도에 대한 교차상관함수를 이용한 군집분석

한 구간의 정체가 근접한 하나의 구간 또는 여러 구간의 통행속도에 영향을 미칠 수 있다. 그 영향은 동 시간대일 수도 있고 시차를 두고 일어날 수도 있다. 본 절에서는 통행속도에 있어서 직접적으로 또는 간접적으로 영향을 주고받는 등 밀접한 관계가 있는, 즉 같은 영향권에 있는 구간들을 묶기 위해 통행속도에 대한 교차상관함수를 이용하여 군집분석을 하였다. 원자료를 이용하여 교차상관함수를 구하는 경우에는 일부 구간들이 같은 영향권에 있지 않지만 패턴이 비슷하여 높은 교차상관함수를 가지며 같은 군집에 묶이는 경우가 발생할 수 있다. 따라서 각 시계열에 대해 패턴을 제거 하는 등의 사전 백색화를 한 후 교차상관함수를 구함으로써 구간끼리의 순수한 영향을 살펴 볼 필요가 있다.

Figure 4.1의 (a)는 929개 구간에 대해 원 자료를 사용하여 자기상관함수를 구하고 시차별로 평균을 낸 그림으로 1차(15분에 해당) 자기상관계수는 0.874로 강한 자기상관을 보였고, 96차(1일에 해당)와 672차(7일에 해당)는 0.475와 0.535로 뚜렷한 주기성이 있었으며 하루 단위의 주기보다 일주일 단위의 주기가 좀 더 강했다. Figure 4.1의 (b)는 1차 차분한 자료에 대해 자기상관함수를 구하고 시차별로 평균을 낸 그림으로 1차 자기상관계수는 -0.208 로 음의 자기상관을 보였고, 96차와 672차는 0.065와 0.077로 약한 주기성을 보였다. 1차 차분하기 전과 비교하면 차분으로 주기가 거의 사라진 것으로 보이며 백색화가 되었다고 보아도 큰 무리가 없을 것으로 판단되므로 본 연구에서는 1차 차분한 시계열자료를 이용하여 교차상관함수를 구하였다.

정상인 두 시계열 X_t 와 Y_t 사이의 교차상관함수는 $\rho_{XY}(k) = \text{Corr}(X_t, Y_{t+k})$ 와 같이 정의되며 자기상관함수와 달리 시차 k 에 대해 대칭이 아니지만 $\rho_{XY}(k) = \rho_{YX}(-k)$ 의 관계가 있다 (Wei, 2006). 시차 k 가 양수이면 X_t 가 선행이고 Y_t 가 k 만큼 후행이 되는 것에 대한 교차상관함수이고, 시차 k 가 음

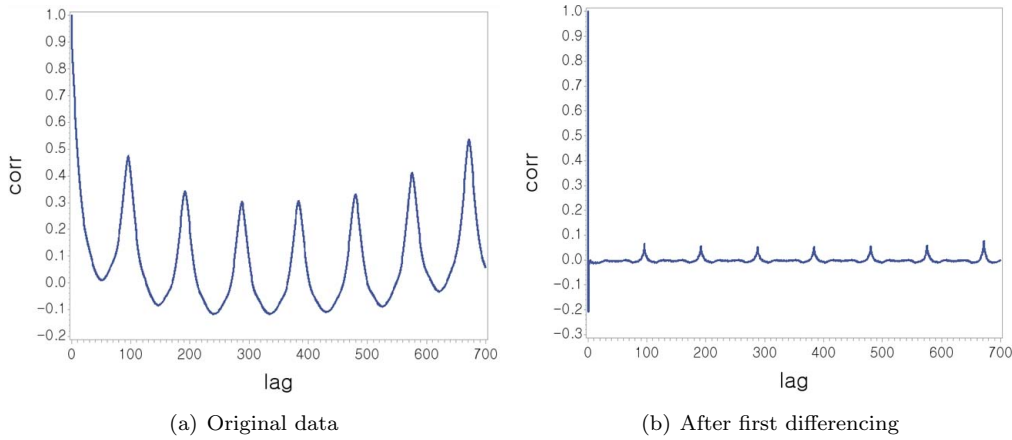


Figure 4.1. Mean of the autocorrelation functions.

Table 4.1. Frequency of the maximum lags(k^*)

Lag	-4	-3	-2	-1	0	1	2	3	4	Total
N	37,003	30,209	48,339	154,731	321,548	154,731	48,339	30,209	37,003	862,112
%	4.29	3.50	5.61	17.95	37.30	17.95	5.61	3.50	4.29	100

Table 4.2. Summary to the maximum value of the cross-correlation function($\hat{\rho}_{XY}(k^*)$)

N	Mean	Median	Standard deviation	Maximum	Minimum	Skewness
862,112	0.03	0.02	0.03	0.69	-0.06	6.34

수이면 그 반대가 된다. 본 연구에서는 929개 구간에서 2개씩 뽑은 조합 각각에 대해 1차 차분한 자료를 이용하여 시차 k 에 대해 -4 부터 4 까지의 범위 내에서 교차상관함수를 계산하였고 그 중에서 가장 큰 값을 가지는 교차상관함수의 최댓값($\hat{\rho}_{XY}(k^*)$)과 그 때의 시차인 최대시차(k^*)를 구하였다. 시차 k 의 범위를 $-4 \sim 4$ 로 정한 이유는 시차가 4 인 경우 실제 시간은 1 시간 차이가 되며 1 시간이 넘는 관계는 무시해도 좋다고 보았기 때문이다. 929개 구간에 대한 교차상관함수의 최댓값과 최대시차는 총 $862,112(929 \times 928)$ 개가 계산되었으며 많은 계산시간이 소요되었다.

최대시차에 대한 빈도표인 Table 4.1을 보면 0 에 대해 대칭이고 시차가 0 에서 멀어질수록 비율이 줄어들며 시차 ± 4 에서의 비율이 높지 않기 때문에 고려하는 시차 k 의 범위를 늘린다 하더라도 최댓값을 갖는 최대시차가 ± 4 를 넘어가는 경우는 많지 않을 것으로 판단된다. 교차상관함수의 최댓값에 대한 기초 통계량은 Table 4.2와 같으며 평균이 0.03 , 표준편차는 0.03 으로 0 근처에 많이 모여 있고, 경부선의 구서IC-영락IC 구간과 영락IC-부산IC 구간의 교차상관함수의 최댓값이 0.69 로 가장 높은 것으로 나타났다. 또한 왜도의 값은 0 을 기준으로 하였을 때 6.34 로 높은 값을 가지며, 왼쪽으로 치우친 모양이다.

특정한 구간에 대해 다른 구간들과 구한 교차상관함수의 최댓값들이 모두 작은 값을 가지면 그 구간은 다른 모든 구간들과 주고받는 영향이 작다고 볼 수 있다. 본 절의 목적은 상호 영향을 주고받는 구간들을 묶는 것이고 다른 구간들과의 영향 정도가 작은 구간들은 관심 대상이 아니기 때문에 교차상관함수의 최댓값들이 모두 작은 구간들을 제외하였다. 본 연구에서는 모두 0.55 이하인 구간들을 제외하였으며 남은 56 개의 구간들을 대상으로 군집분석을 하였다. 어떤 두 구간의 교차상관함수의 최댓값이 클수록 두 구간은 강하게 연결되어 있다고 볼 수 있고 교차상관함수의 최댓값을 유사성의 척도로 사용할 수 있으므로 ‘ $1 - \text{교차상관함수의 최댓값}$ ’을 두 구간 사이의 거리로 정의하여 군집분석을 하였다. 이렇게

Table 4.3. Clustering results based on the maximum cross-correlation function

Cluster	Freq	Highway	Section of highway	Range of CCF
1	3	Gyeongbu	Guseo IC-Youngnak IC Youngnak IC-Busan TG Busan TG-Nopo IC	0.47-0.69
2	4	Yongin-Seoul	W.Pangyo IC N.-Unjung Tunnel S. Unjung Tunnel N.-Geumto TG Godeung IC-Simgok Tunnel S. Simgok Tunnel S.-Simgok Tunnel N.	0.25-0.69
3	2	Muan-Gwangju	Donggwangsan TG-Seogwangsan IC Seogwangsan IC-Naju IC	0.67
4	2	Suncheon-Wanju	Namwon JC-N.Namwon IC N.Namwon IC-Namwon JC	0.67
5	4	Muan-Gwangju	E.hampyeong IC-Munpyeong IC Munpyeong IC-Naju IC Naju IC-W.Gwangsan IC W.Gwangsan IC-E.Gwangsan TG	0.25-0.65
6	4	Incheon International Airport	88 JC-Gimpo Airport IC Gimpo Airport IC-Nooji JC Gonghang Entrance IC-Gonghang New Town JC Gonghang New Town JC-Sinbul IC	0.17-0.63
7	2	Gochang-Damyang	Jangseong Logistics IC-S.Gochang IC S.Gochang IC-Gochang JC	0.62
8	5	Dangjin-Daejeon	W.Sejong IC-Gongju IC Gongju IC-Gongju JC Gongju JC-W.Gongju JC W.Gongju JC-Magoksa IC Magoksa IC-Yugu IC	0.21-0.58
9	2	Yongin-Seoul	Simgok Tunnel N.-Simgok Tunnel S. Segok Tunnel N.-Segok Tunnel S.	0.60
10	2	Incheon International Airport	Sinbul IC-Gonghang New Town JC Gonghang New Town JC-Gonghang Entrance IC	0.60
11	2	Namhae (Yeongam-Suncheon)	Dorong IC-Suncheonman IC Suncheonman IC-S.Suncheon TG	0.60
12	2	Honam	Daedeok JC-Changpyeong IC Changpyeong IC-Goseo JC	0.58
13	4	Dangjin-Daejeon	Yugu IC-Magoksa IC Magoksa IC-W.Gongju JC Gongju JC-Gongju IC Gongju IC-W.Sejong IC	0.20-0.58
14	2	Gochang-Damyang	Gochang JC-S.Gochang IC S.Gochang IC-Jangseong Logistics IC	0.58
15	4	Namhae	Sanin JC-Haman IC Haman IC-Jangji IC Jisu IC-Jinseong IC Jinseong IC-Munsan IC	0.31-0.57
16	4	Honam	Goseo JC-Changpyeong IC Changpyeong IC-Daedeok JC Okgwa IC-Gokseong IC Seokgok IC-Juam IC	0.14-0.57
17	4	Namhae	Sacheon IC-Jinju JC Jinju JC-Jinju IC Jinseong IC-Jisu IC Jisu IC-Gunbuk IC	0.28-0.57
18	2	Dangjin-Daejeon	Godeok IC-Myeoncheon IC Myeoncheon IC-Dangjin JC	0.56
19	2	88olympic	Goseo JC-Damyang JC Damyang JC-Damyang IC	0.55

정의한 거리는 0과 2 사이의 값을 갖는다. 군집분석은 계층적 군집방법 중에서 Ward 방법을 사용하였고, 군집 수는 R-square 통계량의 변화량을 고려하고 같은 군집 내에 있는 구간들 간의 거리는 모두 작은 값을 유지하도록 19개로 정하였다. 군집분석의 결과는 Table 4.3과 같으며 각 군집에 속한 구간들은 2-5개이다. 각 군집에 속한 구간들은 0.14에서 0.69 사이의 교차상관함수의 최댓값을 가지며 모두 같은 고속도로 노선에 속해 있고 서로 인접한 곳에 있는 구간들이다. 군집들의 지리적인 위치를 살펴보면 수도권, 대전서부, 전라도, 경남과 부산 등 남한의 서부 지역과 남부 지역에 걸쳐서 군집들이 형성되었다.

5. 결론

본 논문은 고속도로 통행속도 예측을 위한 사전작업으로 통행속도에 따라 구간들을 묶기 위해 군집분석을 하였다. 연구를 위해 사용한 자료는 한국도로공사에서 제공하는 15분 단위로 관측된 고속도로 통행속도(VDS) 자료 중에서 2년간(2014년 1월-2015년 12월)의 자료이다. 연구기간 중 5%가 넘는 결측치를 가지는 구간을 제외한 최종 929구간을 선정하였으며 15분 간격의 전체 통행속도 자료 수는 64,801,310개였다. 통행속도에 대한 기초분석 결과 구간에 따라 시간대별 변동의 차이가 크고 패턴도 다양함을 볼 수 있었다. 따라서 통행속도 예측의 정확도를 높이기 위해 구간들을 묶는 것은 의미가 있다고 보여진다.

첫 번째 목적은 월·요일·시간대의 시간변수에 따라 반복적으로 나타나는 패턴이 비슷한 구간들의 군집을 구하는 것이다. 따라서 시간변수에 따른 패턴의 효과만 고려해야 하므로 각 구간별로 평균속도를 빼 자료를 분석에 사용하였다. 시간변수를 요인으로 하는 분산분석의 결과에서 F -값을 보면 주효과와 비교하여 교호작용의 효과는 현저히 작으며, 주효과 중에서는 요일 효과가 가장 크고 월과 시간 효과는 비슷하였다. 월·요일·시간대의 115개 시간변수에 따라 통행속도 패턴이 비슷한 구간들을 묶기 위해 6개의 주성분분석으로 변수를 축약한 후 군집분석을 하였다. 이때 군집의 수는 6개로 정하였고 k-평균 방법을 이용하였다. 군집분석 결과 군집에 따라 시간변수별 평균 통행속도의 패턴이 다양한 모습을 보였으며, 이 결과는 군집별로 실시간이 아닌 미래의 특정한 일시의 통행속도 예측을 위한 모형을 만드는 데 활용될 수 있다. 이 예측모형으로 특정시간에 출발했을 때 언제 도착할 지에 대한 예측이 가능할 것이다.

두 번째 목적은 통행속도에 있어서 직접적으로 또는 간접적으로 영향을 주고받는 등 밀접한 관계가 있는 같은 영향권에 있는 구간들을 묶기 위해 군집분석을 하는 것이다. 구간끼리의 순수한 영향 정도를 보기 위해 통행속도에 대한 1차 차분으로 사전 백색화를 하였고 시차 k 에 대해 ± 4 의 범위 내에서 계산한 교차상관함수의 최댓값과 그 때의 시차인 최대시차를 도출하였다. 교차상관함수의 최댓값이 0.55 이상인 56개의 구간을 대상으로 '1 - 교차상관함수의 최댓값'을 거리로 정의하여 군집분석을 하였다. 군집분석 결과 각 군집에 속한 구간들은 모두 같은 고속도로에 속해 있고 지리적으로도 가까운 19개의 군집들이 도출되었다. 이러한 결과는 실시간 통행속도 예측모형 구축에 도움이 될 것으로 기대한다. 이 예측모형은 인근 구간에서의 사고 등의 상황 변화가 특정 구간에 언제 얼마나 영향을 줄 지를 가능하고 우회도로를 선정하는데 도움을 줄 수 있을 것이다.

대규모의 전국 고속도로 통행속도 자료를 이용하여 고속도로 통행속도 패턴과 영향 정도를 이용하여 군집분석을 한 것이 본 논문의 가장 큰 성과이다. 본 논문은 미래의 특정한 일시 또는 실시간의 통행시간 예측을 위한 기반을 마련하기 위한 것이고 예측 그 자체에 초점이 맞춰져 있지는 않다. 따라서 맹절이나 사고와 같은 특이한 상황 또는 이상치에 대한 고려를 하지 않았다. 향후의 연구 과제로는 보다 정확한 통행속도 예측을 위한 모형을 구축하는 것이고, 이를 위해서는 통행속도에 영향을 주는 시간변수에 대한 정보, 기상 정보, 사고 등의 고속도로 상황에 대한 정보 등과 함께 통행량이 고려되어야 할 것이다. 또한 본 연구의 결과를 이용하여 군집별로 모형구축을 한다면 보다 더 정확한 통행속도 예측이 가능할 것으로 사료된다.

References

- AASHTO (1994). *A Policy on Geometric Design of Highway and Streets*, Washington, D.C.
- Cho, J., Kim, B., Kim, S., and Kang, W. (2011). Development of a daily pattern clustering algorithm using historical profiles, *Journal of the Korea Institute of ITS*, **10**, 11–23.
- Choi, B., Kang, H., Lee, S., and Han, S. (2009). A study for traffic forecasting using traffic statistic information, *The Korean Journal of Applied Statistics*, **22**, 1177–1190.
- Flaherty, J. (1993). Cluster analysis of Arizona automatic traffic record data, *Transportation Research Record*, **1410**, 93–99.
- Kim, S. and Cho, J. (2008). A comparative study on a hierarchical clustering method for road classification by traffic characteristics, *Journal of Engineering & Technology*, **17**, 49–58.
- Korea Expressway Corporation (2015). 2014 highway traffic statistics.
- Lee, J., Do, M., Kim, S., and Ryu, S. (2003). Real-time adjustment of traffic volume-based on the national highway route 3, *The Korean Journal of Applied Statistics*, **16**, 203–215.
- Lee, M., Lee, S., Namkoong, S., and Choi, K. (2014). Study on the classification methodology for DSRC travel speed patterns using decision trees, *Journal of the Korea Institute of ITS*, **13**, 1–11.
- Wei, W. W. (2006). *Time Series Analysis* (2nd Ed.), Addison-Wesley, Redwood City, California.

고속도로 구간별 통행속도의 패턴과 영향에 따른 군집분석

김병수^{a,1} · 안소영^a · 손정민^a · 박혜미^a

^a인제대학교 통계학과

(2016년 5월 30일 접수, 2016년 7월 25일 수정, 2016년 8월 8일 채택)

요약

본 논문은 고속도로 전 구간에 대해 2년간(2014년 1월-2015년 12월) 15분 단위로 관측한 고속도로 통행속도 데이터(VDS)를 사용하였다. 본 연구의 첫 번째 목적은 월·요일·시간대의 시간변수에 따라 반복적으로 나타나는 패턴이 비슷한 구간들의 군집을 구하는 것이다. 115개의 시간변수를 주성분분석으로 축약한 후 군집분석을 한 결과 시간변수에 따라 다양한 패턴을 보이는 군집들을 얻었으며, 이 결과는 실시간이 아닌 미래의 통행속도 예측을 위한 모형을 만드는데 활용될 수 있다. 두 번째 목적은 통행속도에 있어서 직접적으로 또는 간접적으로 영향을 주고받는 등 밀접한 관계가 있는 같은 영향권에 있는 구간들을 묶기 위해 군집분석을 하는 것이다. 구간끼리의 순수한 영향 정도를 보기 위해 각 구간별로 통행속도에 대한 백색화한 후 계산한 교차상관함수를 이용하여 군집분석을 하였다. 군집분석 결과 지리적으로 가까운 군집들이 다수 도출되었으며, 이 결과는 실시간 예측에 도움이 될 것이다.

주요용어: 고속도로 구간, 통행속도, 패턴, 군집분석, 교차상관분석

본 논문은 2015학년도 인제대학교 학술연구조성비 보조에 의한 것임.

¹교신저자: (50834) 경남 김해시 인제로 197, 인제대학교 통계학과/통계정보연구소. E-mail: statkbs@inje.ac.kr