J. Korean Soc. Transp.
Vol.33, No.1, pp.14-28, February 2015
http://dx.doi.org/10.7470/jkst.2015.33.1.14

Journal of Korean Society of Transportation
pISSN : 1229-1366
eISSN : 2234-4217

Article

검색 트래픽 정보를 활용한 고속도로 교통지표 분석 연구

류인곤[1]* · 이재영[2] · 박경철[1] · 최기주[3] · 황준문[1]
[1] 경기개발연구원 휴먼교통연구실, [2] 중앙플로리다대학교 첨단교통시뮬레이션연구센터,
[3] 아주대학교 교통시스템공학과

Analysis of Highway Traffic Indices Using Internet Search Data

RYU, Ingon[1]* · LEE, Jaeyoung[2] · PARK, Gyeong Chul[1] ·
CHOI, Keechoo[3] · HWANG, Jun-Mun[1]
[1] Department of Transportation Policy, Gyeonggi Research Institute, Gyeonggi 440-290, Korea
[2] Center for Advanced Transportation Systems Simulation, University of Central Florida, Florida 32816, USA
[3] Department of Transportation Systems Engineering, Ajou University, Gyeonggi 443-749, Korea

Abstract

Numerous research has been conducted using internet search data since the mid-2000s. For example, Google Inc. developed a service predicting influenza patterns using the internet search data. The main objective of this study is to prove the hypothesis that highway traffic indices are similar to the internet search patterns. In order to achieve this objective, a model to predict the number of vehicles entering the expressway and space-mean speed was developed and the goodness-of-fit of the model was assessed. The results revealed several findings. First, it was shown that the Google search traffic was a good predictor for the TCS entering traffic volume model at sites with frequent commute trips, and it had a negative correlation with the TCS entering traffic volume. Second, the Naver search traffic was utilized for the TCS entering traffic volume model at sites with numerous recreational trips, and it was positively correlated with the TCS entering traffic volume. Third, it was uncovered that the VDS speed had a negative relationship with the search traffic on the time series diagram. Lastly, it was concluded that the transfer function noise time series model showed the better goodness-of-fit compared to the other time series model. It is expected that "Big Data" from the internet search data can be extensively applied in the transportation field if the sources of search traffic, time difference and aggregation units are explored in the follow-up studies.

2000년대 중반부터 인터넷 검색 트래픽을 활용한 다양한 연구가 진행되었다. 대표적으로 구글은 미국의 독감 발병 상황을 인터넷 유저의 검색 패턴을 통해 예측하는 서비스를 만들기도 하였다. 교통지표 역시 인터넷 검색 패턴과 유사할 수 있다는 가설을 확인하기 위하여, 검색 트래픽 데이터를 활용하여 고속도로의 진입 교통량과 구간 속도를 추정하는 모형을 구축하고 적합도 등을 확인하는 것이 본 연구의 목적이다. 그 결과, 첫째, 출퇴근의 상시적 통행이 이루어지는 지점의 TCS 진입 교통량 모형은 구글 검색 트래픽이 입력변수로 우수하였고, 검색 트래픽과는 음의 상관관계를 보였다. 둘째, 여가 통행이 집중적으로 나타났던 지점의 TCS 진입 교통량 모형은 네이버의 검색 트래픽이 입력변수로 선정되었으며, 검색 트래픽과는 양의 상관관계가 나타났다. 셋째, VDS 속도의 경우 시계열 도표상 검색 트래픽과 음의 상관관계를 보였다. 넷째, 검색 트래픽을 입력변수로 활용한 전이함수 잡음 시계열 모형은 그렇지 않은 시계열 모형에 비해 비교적 적합도가 우수하다는 결과를 도출하였다. 다만, VDS 속도 모형의 경우 다수의 입력변수가 포함되고 모형 계수의 부호가 상이함에 따른 한계가 존재하였다. 향후 검색 트래픽의 출처나 검색어, 혹은 시차 및 집계 단위에 대한 추가적 연구가 진행된다면, 교통 분야의 빅 데이터 연구시 활용 폭이 넓어질 것으로 판단된다.

Keywords

big date, google trends, naver trend, TCS traffic/VDS speed, transfer function-noise model
빅 데이터, 구글 트렌드, 네이버 트렌드, TCS 교통량/VDS 속도, 전이함수 잡음 시계열 모형

* : Corresponding Author
bhpark@chungbuk.ac.kr, Phone: +82-43-261-2496, Fax: +82-43-264-2496

Received 21 August 2014, Accepted 19 November 2014

서론

1. 연구의 배경 및 목적

인터넷 기술의 발달은 웹상의 정보와 사용자의 폭발적 증가를 야기하였다. 이는 자연스럽게 웹을 둘러싼 정보 생성과 정보 검색의 트렌드를 파악하여 활용하고자 하는 연구의 증가를 야기하게 된다.

특히 인터넷 포털 서비스 업체 중 국외에서는 구글, 국내에서는 네이버가 온라인 사용자들의 웹검색 트래픽 정보를 구글 트렌드, 네이버 트렌드 서비스를 통해 공개하고 있다. 이들이 제공하는 웹검색 트래픽 정보를 기반으로 온라인 사용자들의 정보 검색 행태에 대한 연구들이 학계·업계 등에서 주목받고 있다. 특히 웹검색 정보를 기반으로 사회 현상이나, 소비 동향, 정치 투표 결과 등을 예측해 볼 수 있음을 실증하는 연구가 활발히 이루어지고 있다(Jun and Park, 2013).

이와 같은 연구 결과가 교통 분야에서도 적용될 수 있다면, 교통과 관련된 연관어의 검색 트래픽이 활동(activity) 및 통행(travel) 선택에 대한 선행지표가 될 수도 있다. Figure 1은 일상생활 속에서 단계별 선택 직전에 어떠한 검색어를 사용하는지 예를 들어 본 것이다. 특히 "교통정보", "고속도로"와 같은 단어[1]는 특정 활동 계획을 특정 통행으로 전환하거나 선택된 통행의 경로를 전환할 때 사전적으로 웹 사용자에 의해 입력되는 검색어라고 예상해 볼 수 있다.

한편 고속도로의 운영 및 유지관리를 위한 전략 모색 측면에서 단기적 교통 상황 지표 추정이 필요한데, 첫째 특정 기간(예를 들어 명절 혹은 휴가철)에 앞서 간단한 설문조사 진행하여 출발일자, 경로, 시간 등을 확인하고 전수화하여 이용 수요를 추정하는 방법, 둘째 고속도로상 관측된 링크교통량을 이용하여 역으로 기종점 통행량 및 통행시간 등을 추정하는 방법, 셋째 링크의 과거 시점, 이전 구간의 교통류 특성을 활용하여 특정 지점 혹은 구간의 교통량이나 속도를 추정하는 방법 등이 활용된다.

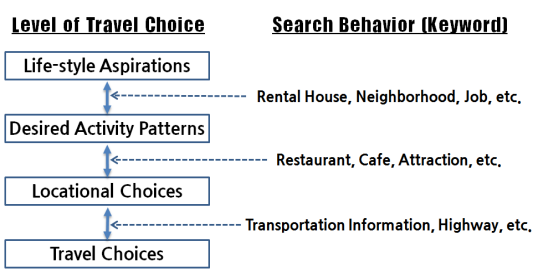

Figure 1. Levels of choice for an individual and search keywords (Manheim(1979), modified)

이 때 첫 번째 방법은 조사 시점과 통행 시점 사이의 통행자의 의사결정이 바뀔 수 있다는 점에서, 두 번째 방법은 관측 교통량이 일부 없을 경우 다중해의 기종점 통행량이 추정된다는 점에서, 세 번째 방법은 구간, 시점에 따라 교통량, 속도 모형이 모두 상이하게 추정된다는 측면에서 각각 한계점이 있다.

다만 현실적으로 세 번째 방법은 고속도로상 관측 자료만을 활용하여 교통량이나 속도가 추정되고, 그 값 자체로도 고속도로 운영에 직접 활용할 수 있다는 점 등에서 연구 활용도가 높다고 할 수 있다.

본 연구에서는 기본적으로 지점 교통량과 구간 속도라는 교통지표를 빅 데이터라 할 수 있는 검색 트래픽을 활용하여 추정해보고자 한다. 즉 교통지표와의 검색 트래픽의 연관성을 시계열 도표를 통해 확인하고, 검색 트래픽이 반영된 전이함수 시계열 모형을 구축하여 적합도 개선 여부를 확인하는 것을 그 목적으로 한다.

2. 연구의 내용 및 방법

본 연구에서의 다루고자하는 교통 관련 검색 트래픽은 높은 인터넷 이용 밀도가 위치한 대도시권을 중심으로 발생한다. 따라서 대도시권에서의 높은 교통수요가 교통 네트워크에 영향을 미칠 수 있다고 판단하여 수도권의 고속도로를 연구의 공간적 범위로 삼았다.

일차적으로는 검색 트래픽 데이터의 기초통계량 분석과 시계열적 특성을 확인한다. 이를 통해 포털 출처나 검색어에 따라 어떠한 차이가 있는지 검토한다. 이 때 검색 트래픽은 구글 및 네이버 트렌드의 자료를 활용하였으며, 검색어는 "교통정보"와 "고속도로" 각각을 사용하였다.

다음으로는 TCS(Toll Collection System) 교통량과 구간 통행속도를 출력 변수로 삼고 시계열 모형을 추정하였으며, 비교 모형 측면에서 검색 트래픽을 전이함수 입력 변수로 사용한 시계열 모형을 추정하여 그 개선

1) 구글 트렌드 서비스의 단어별 검색량 비교 결과, 운전자가 교통정보를 구득하기 위해 사용하는 "고속도로" 혹은 "교통정보" 단어의 검색량이 "교통상황", "고속도로상황", "고속도로교통정보", "실시간교통정보" 단어 검색량보다 상대적으로 많았으며, 패턴은 유사한 것으로 확인

정도를 파악하였다.

더불어 검색어 등에 따라 달라지는 지점별 최적화 전이함수 시계열 모형을 확인하고 검토하는 과정도 동반한다.

연구 대상은 TCS 지점 3개소의 진입교통량과 VDS (Vehicle Detection System) 지점 4개소(양방향 8개소)의 속도에 대한 모형으로 총 11개 모형이라 할 수 있으며, 이에 대해 우수한 모형을 선별적으로 제시할 수도 있으나 향후 연구자를 위해 분석 모형식 전체 결과를 제시하도록 한다.

기존 문헌 고찰 및 본 연구의 차별성

빅데이터 개념 이전부터 교통 분야에서는 첨단교통체계(ITS; Intelligent Transportation System)와 관련된 각종 센서 데이터, 교통카드 이용시 누적되는 통행패턴 데이터를 활용한 다양한 연구가 진행되고 있었다. 본 연구의 동기가 인터넷 이용간 추출된 검색 트래픽 데이터를 교통에 접목시키는데 있는 바, 빅데이터 중에서도 검색 트래픽 관련 연구를 중심으로 문헌 고찰을 수행한다.

1. 인터넷 검색과의 연관성 발견에 대한 연구[2)]

가장 초기에 Ettredge et al.(2005)는 실업과 관련된 인터넷 검색 수와 실제 미국 정부가 발표한 실업률과의 연관 관계를 연구하였다.

Cooper et al.(2005)은 여러 종류의 암에 대한 인터넷 검색 수와 실제 암 환자 발생 수와의 관계를 조사하였다. 암 관련 용어의 인터넷 검색 수는 암 발생건수와 상관성(상관계수 0.5)이 있으나 그 보다는 신문지 상에서 암과 관련된 뉴스의 빈도(상관계수 0.88)와 더 큰 연관성이 있다는 것을 발견하였다. 이를 통하여 검색빈도가 실제 암 발생에 대한 직접적인 연관성 보다 다른 요인(신문 뉴스 빈도)에 의해 영향 받을 수 있으므로 관련성 분석에 있어 여러 가지 면을 고려해야 한다는 점을 강조하였다.

Polgreen et al.(2008)과 Ginsberg et al.(2009)은 각각 Yahoo와 Google에서 독감(Influenza)과 연관된 검색 결과와 실제 독감발생건수와의 관계를 조사하고, 웹 검색 데이터를 이용하면 보건당국보다 먼저 독감의 유행을 예측할 수 있다는 것을 사례를 통하여 주장하였다. 실제로 Google에서는 이러한 개념을 이용하여 현재 Google Flu Trends를 통하여 국가별로 독감유행정보를 실시간으로 일반인에게 공개하고 있다.

Choi and Varian(2009b)은 구글 트렌드 데이터가 실업수당을 받기 위해 최초로 신청하는 사람들에 대한 사전 지표로 활용될 수 있음을 보였다. 의료건강 분야에서도 인터넷 검색 데이터를 활용한 연구가 활발히 연구되고 있다.

이 외에도 검색엔진에서의 검색 용어를 활용하여 노동과 주택시장(McLaren, 2011), 영화, 게임 및 음악 산업(Goel, 2010) 등 다양한 산업에 활용하려는 연구들이 시도되고 있다.

특히 주식금융산업에 웹 검색 결과를 활용하는 방안에 대한 연구가 최근 활발히 진행되고 있다. Preis et al.(2010)은 S&P500 기업의 주간 주식거래량이 대응되는 회사의 Google을 통한 인터넷 검색횟수와 연관관계가 있음을 발표하였다. Bollen et al.(2011)은 twitter.com에서의 감성(mood)과 관련된 용어의 빈도가 다우존스산업평균지수(DJIA; Dow Jones industrial average)에 영향을 미친다는 것을 발견하고, DJIA를 예측하는 데에 twitter의 감성관련 단어를 고려한다면 정확도를 높일 수 있다고 주장하였다. Moatet al.(2013)은 주식시장의 움직임이 있기 전에 인터넷 백과사전인 Wikipedia에서 관련된 용어에 대한 검색 빈도가 어떻게 변하는지에 대한 연구를 수행하였다.

교통과 관련된 연구는 폭 넓게 이루어지지 않았지만, 자동차 판매, 해외 여행을 주제로 인터넷 검색과의 관련성을 검토한 연구가 존재하였다. Choi and Varian(2009a)은 자동차 및 부품 판매액을 추정하는 계절 AR 시계열 모델에 구글 트렌드 검색 변수(검색어: Motorcycles, Auto Insurance, Trucks&SUVs)를 포함시킬 경우 평균절대오차의 18% 개선 효과가 나타났다고 분석하였다. 27개의 자동차 메이커에 대해서도 각각 분석을 하였는데 이때는 메이커별로 계절에 따른 출시 및 마케팅 효과가 상이함에 따라, 검색 입력변수를 포함시킨 모형이 그렇지 않은 모형에 비해 개선 효과가 나타나기도 하고 나타나지 않기도 했다. 9개 국가로부터 입국한 홍콩 방문객을 추정하기 위해 구글 인터넷 검색 횟수와 환율을

2) 교통 분야를 제외한 기존 문헌은 Kim, M. S., Koo, P. H.(2013), A Study on Big Data Based Investment Strategy Using Internet Search Trends, Journal of the Korean Operations Research and Management Science Society, 34(4)의 정리 일부 인용

시계열 입력변수로 설정하고 베이징 올림픽 기간을 의미하는 더미변수를 포함시켜 분석하였으며, 이를 통해 방문객은 1개월 전과 12개월 전의 방문객, 구글 검색량과 양의 상관관계가 있으며, 올림픽 기간은 음의 상관관계가 있다고 분석하였다.

Choi and Varian(2012)에서는 2009년 연구를 다소 변경하여 시계열 모형을 설정하였는데, 자동차 및 부품 판매액의 경우 입력변수로 사용한 검색어를 Trucks&SUVs와 Automotive Insurance를 사용하였으며, 홍콩 방문객의 경우 분석기간을 조정하고 환율변수 및 월드컵 더미변수를 제거하였으며, 일본을 제외한 분석을 수행하였다.

Jun et al.(2014)는 하이브리드 차량 구매에 대한 수용성을 분석하기 위해 통상적 변수인 신문기사, 특허, 유가, 국내총생산 증가율 변수와, 검색 데이터 변수를 비교하여 분석을 수행하였다. 특히 특정 기술 검색어와 차량 브랜드 검색어 등을 조합하여 분석을 수행하였는데, 그 결과 검색어 선택에 따라 신기술이 적용된 차량 판매량을 유용하게 분석할 수 있다는 결론을 얻었다. 또한 차량 브랜드 검색어를 활용할 경우 거시지표인 국내총생산 증가율이나 유가를 사용하는 것에 비해 우수한 추정력을 보일 수 있으며, 역시 서지정보라 할 수 있는 특허출원과 신문기사보다도 우수하다고 분석하였다.

2. 시계열을 활용한 교통지표 추정 연구

시계열 기법을 활용하여 각종 교통지표를 추정하는 연구는 국내외에서 다양하게 이루어졌다. 그 중 시계열 기법이 적용된 교통량 관련 국내 연구는 주로 철도, 항만, 항공 수요를 중심으로 연구가 진행 되었다. Kim and Kim(2011)에서는 개입 ARIMA 모형을 활용하여 KTX를 수요예측 할 때 경부고속철도 2단계 개통과 2008 금융위기의 개입효과를 검토하였으며, Lee and Kwon(2011)에서는 지역간 철도 수요를 추정함에 있어 계절형 ARIMA 모형을 제안하였고, Min et al.(2013)에서는 계절형 ARIMA 모형을 이용해서 인천국제공항발 유럽 항공노선의 화물 수요를 예측하는 모형을 제안하였다. 최근 Yoo et al.(2014)에서는 장래 해상 교통량을 예측함에 있어 기존의 회귀모형을 개선하고자 시계열 모형과 신경망이론을 적용하여 예측 모형을 작성하는 과정을 제시하였다. 지수평활과 인공신경망 결합 모형이 지수평활 모형, 인공 신경망 개별 모형보다 항상 좋은 추정력을 보일 것이라는 가설을 세웠으나 그렇지 못하였다고 밝히고 있다.

도로분야와 관련 Lee et. al.(2001)는 성분 분해 방법을 활용하였는데, 특히 일우량과 최저기온이라는 기상요인을 시계열상 나타나는 불규칙 변동으로 반영하여 AADT를 추정하는 방법론을 제안하였다. 그 결과 기상과 연결된 불규칙변동 요인을 사용하였을 경우 추정능력이 더욱 개선되었다는 결론을 내린다.

속도 예측 등에는 시계열 모형뿐만 아니라 회귀분석 모형, 패턴인식 기반 모형, 교통류 기반 모형 등 다양한 연구가 진행되고 있다. 이 중 시계열 관련 국내 연구로서는 Jang et al.(2005)은 도시부 도로내 단일 링크내 생성된 통행속도와 인접 링크내 생성된 통행속도를 이용하여 단기 통행시간을 예측하는 시계열 모형을 구축하였는데, 분석결과 다변량 모형의 검증력이 우수하다고 판단하기는 어렵다는 결과를 도출하였다. Kim(2009)에서는 연속류 도로의 종점부의 속도, 밀도 변화를 단기예측하기 위해 합류부 시점부의 밀도/속도/교통량을 변수를 사용하였으며, 60초 분석 단위로 속도변화를 예측하는 모형이 가장 우수하다는 결과를 도출하였다. Bin and Mun(2009)은 안산시와 수원시를 사례로 버스정보시스템의 속도 정보를 활용하여 도로상 속도를 추정하는 연구를 진행하였는데, 단순이동, 지수평활법, 이중이동, 이중지수평활법, ARIMA(p,d,q) 모형을 적용하였으며, 상호 비슷한 결과가 나타났다고 언급하고 있다. 따라서 버스정보를 활용하여 구간의 통행속도를 추정하는 것이 가능하다는 수준의 결과를 제시하고 있다.

3. 본 연구의 차별성

인터넷 검색 트래픽을 활용한 연구는 교통 분야에서 다양하게 적용되지는 않았다. 선행연구는 자동차 판매, 관광객 유입을 검토하는 등 여타 분야의 인터넷 검색 관련 연구를 교통과 관련된 소재에 접목시켜 보기 위한 연구로 판단된다.

본 연구는 구체적 교통지표라 할 수 있는 교통량이나 속도의 추정에 있어서도 인터넷 검색 트래픽이 이용될 수 있는지 검토하였다는 점에서 그 차이점이 있다. 또한 시계열 모형의 교통지표 추정 측면에서는 기존에 논의되던 국내 모형과 다르게 전이함수 모형을 활용함으로써 적합성이 개선될 수 있는지 검토했다는 점 역시 차이라 볼 수 있다.

모형 설정 및 분석

1. 변수 선정 및 가공

1) 검색 트래픽 변수

본 연구에서 적용한 검색 트래픽 정보는 현재 주간 단위로 무료로 트래픽 정보를 제공하고 있는 구글 트렌드와 네이버 트렌드의 자료를 활용하였으며, 검색어는 해당 사이트의 연관검색어 등을 고려하여 "교통정보"와 "고속도로" 두 가지로 선정하였다(Table 1).

구글 트렌드 검색 통계 데이터는 주간 단위로 구글상 실행된 총 검색수 대비 사용자가 입력한 용어의 검색수를 계산한 후, 어떤 사용자가 특정 시간대 특정 지역에서 특정 검색어를 검색할 확률을 나타내는 방식으로 계산된다. 검색된 비율이 최대치인 시점 혹은 지역을 100으로 최소치인 시점 혹은 지역을 0으로 환산하여 결과가 제시된다. 단, 검색 통계에서는 검색어에 대한 최소 트래픽 기준이 설정되므로 검색량이 적은 검색어는 통계로 표시되지 않으며, 짧은 기간 동안 특정 사용자가 반복적으로 검색한 검색어는 집계에서 제외되고, 각 지역별 총 트래픽을 기준으로 표준화 작업을 수행하여 검색량이 아닌 상대적 검색 확률을 제시한다.

구글 트렌드의 한글 단어 "교통정보"라는 검색어 추세는 2004년부터 1월부터 제시되며, 검색 확률 최상위 도시인 군포시를 100으로 가정했을 때 거제시 81, 의왕시 80, 광명시 69, 서울시 65, 과천시 62, 시흥시 51로 나타나고 있다[3]. 또한 관련 검색어[4]로 "교통", "교통정보고속도로", "고속도로정보", "고속도로", "고속도로교통정보", "교통정보실시간", "실시간교통정보"가 주로 검색되었다. "고속도로"라는 검색어의 경우 검색 확률 상위 도시인 군포시를 100으로 가정했을 때 거제시 90, 부산시

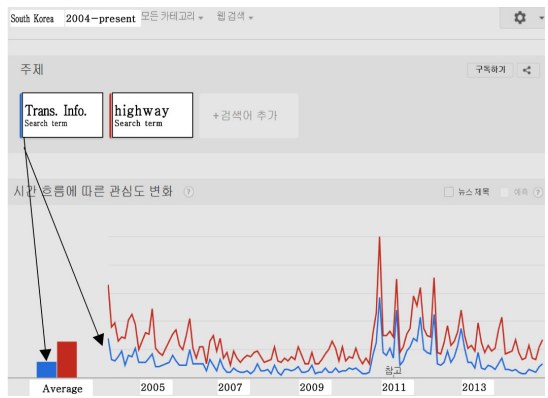

Figure 2. Google Trends by keywords-"Transportation information" and "Highway"

87, 의왕시 86, 서울시 85, 과천시 77, 광명시 68로 나타나고 있으며, 관련 검색어는 "교통정보"와 유사한 결과를 제시하고 있다. "교통정보"와 "고속도로"의 검색 확률을 동시에 비교할 경우 "교통정보"는 9, "고속도로"는 21의 평균값을 나타내고 있으며, 두 검색어 모두 2010년 하반기부터 검색량이 대폭 증가하기 시작한다(Figure 2).

네이버 트렌드는 특정 키워드가 통합검색에서 가장 많이 검색된 지점(주단위)을 기준(100)으로 하여 나머지 기간의 검색 횟수를 상대값으로 환산하여 보여준다. 예를 들어 "날씨"라는 키워드가 가장 많이 입력된 횟수가 1000회라면, 그 지점을 기준인 100으로 삼고 500회, 350회가 검색된 시점은 상대값인 50, 35로 환산하여 그래프로 제시한다. 단, 검색이 발생하는 하드웨어를 구분하여 PC의 경우 2007년 1월부터 현재까지, 모바일 검색의 경우 2010년 7월부터 검색 트렌드 데이터를 제공하고 있다.

네이버 트렌드에서 PC를 활용한 검색량은 2010년 하반기부터 월별 변동폭 및 절대 검색량이 감소하는 경향을 나타내며, 해당 기간 "교통정보"는 11, "고속도로"

Table 1. Data sources for each variable and index

Variables	Site	Explanation
Search traffic[g]	Google trends	Weekly search traffic in South Korea (2004-Present)
Search traffic[n]	Naver trend	Weekly search traffic by PC (2007-Present) Weekly search traffic by cell phone (2010-Present)
Traffic - Toll collection systems(TCS)	Korea Expressway Co.	Hourly traffic counts from toll(2003-Present)
Speed – Vehicle detection systems(VDS)	OASIS	Daily link travel speed by VDS(2010-Present)

3) 구글트렌드 웹사이트를 통해서 '04년 1월부터 '14년 9월까지를 기준으로 산출한 수치이며, 분석기간 설정에 따라 수치/순위 변경
4) 해당 용어를 검색한 사람이 자주 검색하는 검색어의 순위

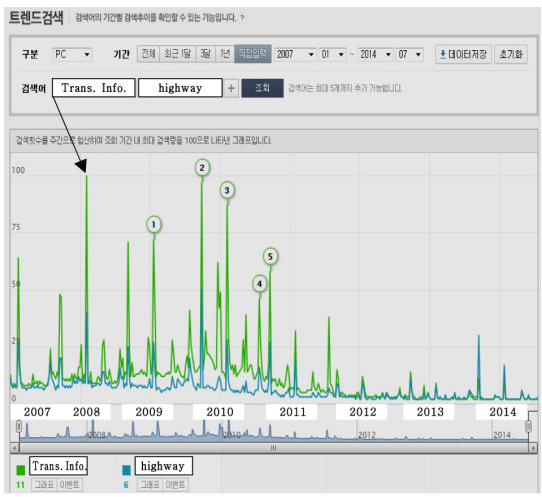

Figure 3. Naver Trend by keywords-"Transportation information" and "Highway"

는 6의 평균적 검색 정도를 나타내고 있다(Figure 3). "교통정보"의 평균적 검색이 더 많이 나타나고 있어, 구글에서의 검색 행태와 다른 양상을 보이고 있었다.

2) 교통지표 변수

교통관련 지표는 검색 트래픽 변수의 조사 기간 동안 비교적 균질한 상태로 정리되어 있는 자료를 활용하기 위하여 한국도로공사에서 제공하고 있는 오픈 오아시스 DB 사이트를 활용하여 조사하였다.

"교통정보" 혹은 "고속도로"라는 단어의 상대적 검색 확률이 높은 지역은 주로 수도권이므로, 교통지표 설정의 공간적 범위는 수도권 주요 진출입 고속도로 지점을 중심으로 선정하였다. 폐쇄식 형식으로 TCS 자료의 구득이 가능하고, 서울에 가장 인접한 톨게이트 3개소인 서울, 서서울, 남양주톨게이트와 수도권 경계에 인접한 안성IC를 포함하여 4개 지점의 진입 TCS 교통량을 교통지표 변수로 선정하였다.

또한 VDS를 통해 수집된 수도권 주요 구간의 통행속도 자료를 분석 대상으로 삼았다. 구간은 특송기간 주요 혼잡구간이라고 할 수 있는 경부 및 서해안 축의 서울 톨게이트↔신갈JC, 서평택JC↔서평택IC, 안성JC↔안성IC의 3개 구간을 선정하여 양방향 개별적 분석을 수행하였다.

3) 변수 가공

검색 트래픽 변수는 주간 단위로 집계된 자료를 바탕

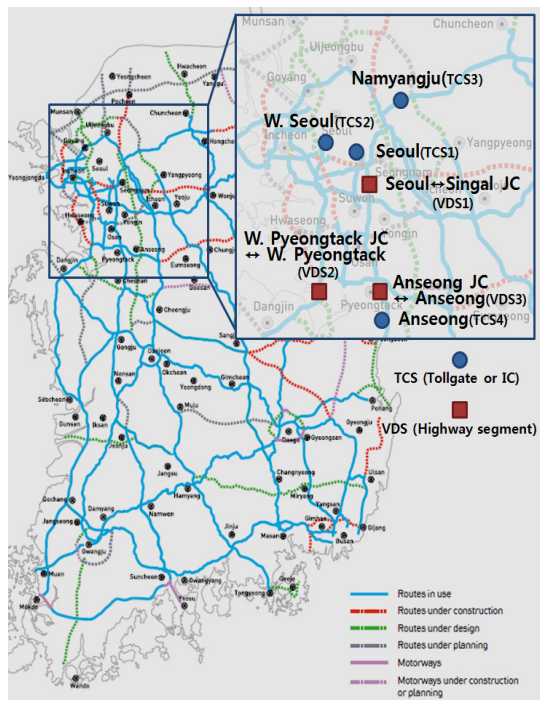

Figure 4. Study site (Highway traffic index)

으로 통계량이 제공되고 있어, 별도의 가공 없이 분석 모형의 변수로 이용하였다.

다만, 검색 트래픽 변수는 최고 검색 트래픽이 발생한 시점을 기준으로 값이 산정되므로, 어떠한 기간까지 추이를 검토하는가에 따라 그 값이 상이하게 된다. 본 연구에 활용한 도로공사의 오픈 오아시스 DB에서는 2013년 11월부터 12월 1주차까지의 TCS 교통량이 누락 제공됨에 따라 검색 트래픽 변수는 2010년 2월부터 2013년 10월까지를 기준으로 산출하였다(단, 네이버 트렌드의 모바일 검색이 2010년 7월부터로 데이터가 존재함에 따라 해당 기간의 데이터만을 최종 데이터로 사용).

교통지표 중 TCS 진입 교통량의 경우, 제공 시점은 2003년부터이며 집계주기는 1시간, 1일이고 제공주기는 1일, 1개월 단위로 제공된다. 1일 단위로 제공된 자료를 검색 트래픽 변수와 동일한 주기인 1주일 단위로 집계하여 주별 IC 진입 교통량으로 재가공하였다.

VDS 속도의 경우 제공 시점은 2010년부터로 집계주기는 5분, 15분, 1시간, 1일 단위이고, 제공주기상 1일, 1개월 단위의 평균 속도가 제공된다. 속도 산출의 공간적 범위로서 차로, 지점, 구간의 3가지 방식이 존재하는데 본 연구에서는 가장 포괄적 범위인 구간 기준의 통행속도를 사용하였다. 또한 1일 단위로 제공된 속도를 1주

일 기간 동안의 평균 속도로 재가공하여 분석에 사용하였다. 교통지표 변수는 검색 트래픽 변수와 동일하게 2010년 7월부터 2013년 10월까지를 기준으로 가공하였다. 통상 속도의 집계 주기로 1주일 단위는 적용되지 않는다. 그러나 혼잡류 상황에서 링크 교통량이나 TCS 진입 교통량은 잠재수요를 모두 대표하지 못한다. 따라서 주간 평균 속도라는 지표를 통해 인근 지역으로부터의 이용 수요까지 확인한다는 측면에서 결과를 검토하도록 한다.

2. 분석모형 구축

통상 시계열 모형 추정을 위해서는 Box-Jenkins 모형을 적용하는데, 식별 통계량을 이용하여 잠정적인 모형을 선택한 후, 선택된 시계열 모형의 모수를 추정하고, 해당 모형의 적합성을 진단하는 과정을 반복하여 예측모형을 만든다. 대부분의 시계열 자료는 일정기간 동안 평균이 같지 않거나, 분산이 같지 않은 경우가 많아 정상성(stationary)을 확보시켜주기 위하여 로그변환이나 차분(difference)을 사용하게 된다.

본 연구의 목적이 교통지표를 추정함에 있어 검색 트래픽 정보의 활용이 효과적일 수 있는지 확인하는 데 있으므로 일차적으로는 검색 트래픽 변수를 사용하지 않은 시계열 모형과 검색 트래픽 변수를 사용한 전이함수 잡음 시계열 모형(TFN; Transfer Function-Noise model)을 추정하여 상호 비교 분석을 수행하였다.

즉, 기본 모형(base model)으로서 지수평활 혹은 일변량 ARIMA 모형을 사용하였으며, 대안 모형(alternative model)으로서 전이함수 잡음 시계열 모형을 사용하였다.

일반적인 전이함수 잡음 시계열 모형은 Box-Jenkins(1976)에 의하면 다음과 같이 표현될 수 있다.

$$Y_t = \delta^{-1}(B)\omega(B)X_{t-b} + N_t \quad (1)$$

여기서, N_t는 잡음(noise)으로 백색 잡음 형태로 표현된다.

$$N_t = \theta^{-1}(B)\phi(B)a_t \quad (2)$$

위의 두 식을 결합하여 정리하면 다음과 같이 표현할 수 있다.

$$Y_t = \frac{\omega(B)}{\delta(B)}X_{t-b} + \frac{\theta(B)}{\phi(B)}a_t \quad (3)$$

위 식에서 b는 입력 시계열과 출력 시계열 사이의 지체시간을 의미하며 나머지 변수들은 다음과 같이 표현된다.

$\omega(B) = \omega_0 - \omega_1 B - \omega_2 B^2 - \cdots - \omega_s B^s$ TFN 모형의 추정 모수 분자

$\delta(B) = 1 - \delta_1 B - \delta_2 B^2 - \cdots - \delta_r B^r$ TFN 모형의 추정 모수 분모

$\phi(B) = 1 - \phi B - \phi B^2 - \cdots - \phi_p B^p$ AR 모형의 연산자

$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ MA 모형의 연산자

위 식에서 보면 일반적으로 ARIMA 모형에 가중함수를 추가하여 나타낸 것을 알 수 있다. 여기서 가중함수는 출

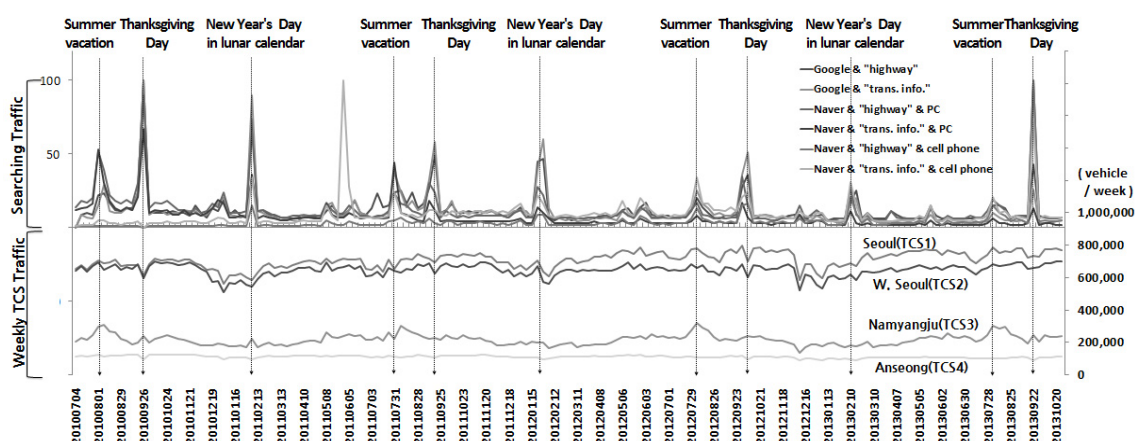

Figure 5. Comparison of weekly TCS traffic and search traffic

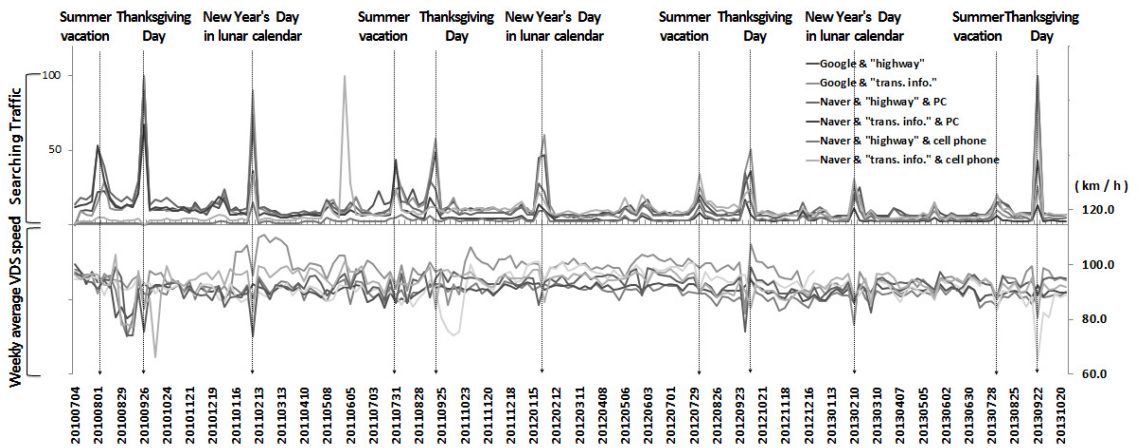

Figure 6. Comparison of weekly average VDS speed and search traffic

력 시계열에 작용하는 입력 시계열이라고 할 수 있다.[5)]

본 연구는 개별 지점의 최적화된 교통량/속도 모형 추정 보다는 포괄적 범위에서의 검색 트래픽 변수 적용 가능성을 검토하는 것이 그 목적이므로, 단계별 변수 선택 및 시차 선택 과정은 생략하고, SPSS 21의 Expert Modeler를 활용하여 각 지점별 최적화된 모형을 산출하고 이를 비교하는 방법을 사용하였다. Expert Modeler는 다수의 지수평활 모형과 ARIMA 모형을 적합한 후 그 중 가장 좋은 모형을 정규화 Bayesian Information Criterion(BIC) 기준을 적용하여 찾아내는 과정을 거치게 된다.

$$Normalized\,BIC = \ln(MSE) + k\frac{\ln(n)}{n} \qquad (4)$$

여기서 k는 모형 적합에 쓰인 파라미터의 수(모형 자유도)이며, n은 데이터 수를 의미한다. k가 증가할수록 더 우수한 값이 산출되는 다른 적합도 통계량을 극복한 통계량이라고 볼 수 있다.

3. 분석 결과

1) 시계열 도표 검토

2010년 7월부터 2013년 10월까지의 검색 트래픽과 TCS 교통량, VDS 주간 평균 속도를 시계열 도표를 통해 살펴보면 Figure 5, 6과 같다.

출처 및 검색 단어에 따라 총 6개의 검색 트래픽 유형을 검토한 결과 대부분 비정상적(non-stationary) 시계열의 형태를 보이고 있었다. 네이버의 PC를 통한 검색 트렌드나 구글 검색 트렌드는 2011년 1분기 정도를 기점으로 다소 감소하는 추세로 전환되었다.

또한 검색 트래픽은 계절변동(seasonal fluctuation)의 성격도 보이고 있는데, 이는 하계 휴가, 추석, 설날의 기간 동안 반복적으로 유사한 패턴이 나오고 있었다. 다만, 우리나라의 추석 혹은 설날이 음력에 의해 날짜가 정해짐에 따라 주기가 1년보다 다소 크거나 작거나 한 형태를 보이고 있다. 하지만 대부분의 집계 주기(1주 단위)간 증감 패턴은 유사한 형태를 보이고 있었다.

Figure 5에서는 검색 트래픽과 TCS 진입 교통량을 비교했는데, TCS 조사 지점마다 다른 특성을 나타내고 있었다. 서울(TCS1) 및 서서울 톨게이트(TCS2)의 경우 검색이 집중되는 주기에 교통량이 감소하는 경향을 나타냈다. 두 톨게이트는 상시적으로 수도권 지역의 출퇴근 교통을 처리하는 지점이다. 하계 휴가 혹은 명절 기간 동안은 주중 출근일이 휴일로 전환된다. 즉, 이 기간 동안 검색 트래픽이 증가했음에도 불구하고, 주간 단위로 집계된 TCS 교통량은 평소 출퇴근시보다 그 수요 크지 않았기 때문에 이와 같은 현상이 나타나게 된다. 하지만, 남양주 IC(TCS3)는 검색 트래픽 증가시 교통량이 증가하는 현상을 나타내고 있었다. 이는 여가 통행을 담당하고 있는 서울-춘천 고속도로 이용객이 검색 트래픽과 양의 상관관계를 맺고 변화하고 있음을 의미하게 된

5) Ju, J. W.(2013), Time series analysis and modeling for the investigation of seawater intrusion into fractured rock aquifer in Muan, Korea, Chonnam National University, A Master's Thesis, 인용

다. 안성 IC(TCS4)는 분석기간 동안 다른 지점에 비해 다소 낮은 11만1천대 수준의 주간 진입교통량을 보이고 있었다(서울 톨게이트의 경우 71만대). 서울 혹은 서서울 톨게이트처럼 상시적 교통량을 처리함에 따라 검색 트래픽이 증가했을 때 진입 교통량이 다소 감소하는 경향을 나타내었다.

Figure 6에서는 검색 트래픽과 VDS 속도를 비교했는데, 대부분 검색 트래픽 증가시 주간 평균 속도가 감소되는 경향을 나타냈다. 검색 트래픽 증가시 서울 혹은 서서울 톨게이트의 주간 진입 교통량이 감소하였음에도 불구하고, 인근 구간의 주간 평균 속도는 감소되는 현상이 나타났다. 통상 일자별 균등하게 출퇴근 하던 고속도로 통행자가 휴가 및 명절 기간에는 동일한 일자와 시간대에 집중되는 현상이 나타나게 된다(그 절대적 통행량은 다소 감소하였을 지라도). 이때 용량을 초과하거나 용량에 이른 상태에서 매우 낮은 속도를 경험하는 다수의 차량이 발생했기 때문에 교통량 감소에도 속도가 감소하는 현상이 나타난 것으로 추정된다.

2) 개별 모형 검토

기본 모형(base model)으로서 지수평활 혹은 일변량 ARIMA 모형을 선택하였는데, 그 결과는 Table 2와 같다. 각각의 지점별로 상이한 모형이 선택되었다. TCS1/2/4, VDS2-SB/NB, VDS3-SB/NB의 경우 지수평활 모형이 최적 모형으로 선정되었으며, TCS3, VDS1-SB/NB의 경우 ARIMA 모형이 최적 모형으로 선정되었다.

Table 2에서 확인할 수 있듯 각 모형의 추정 파라미터 p-value(significance)는 매우 낮아 지수평활 모형의 알파, ARIMA 모형의 상수, AR, MR 파라미터 모두 매우 유의함을 알 수 있다. 또한 Table 5에서 Box-Ljung Q 통계량의 p-value(significance)가 매우 크기 때문에 시차(lag)가 다른 잔차 간에 서로 상관관계가 없다고 할 수 있다. 상기 모형에 적합 시킨 것이 최소한 잘못된 것은 아니라는 판단이 가능하다.

대안 모형(alternative model)으로서 전이함수 잡음 시계열 모형이 최적 모형이 될 수 있게 분석을 수행하였다. TCS 진입교통량의 경우 Table 3에 최종 선택 모형이 제시되어 있는데, 시계열의 비정상성이 감안되어 TCS1, 2, 4 모형에서는 1차 차분이 적용된 모형이 도출되었다.

각 모형별로 입력 계열로 선택된 검색 트래픽 정보는 서울(TCS1)과 서서울(TCS2) 톨게이트는 구글에서 "교통정보"를 검색한 추세값이 입력 시계열로 선택되었으며, 남양주(TCS3) 톨게이트의 경우 네이버 모바일에서 "교통정보"를 검색한 추세값이 선택되었다. 안성(TCS4)

Table 2. Results of base model

Model			Estimate	SE	t	Sig.
TCS1	Exponential Smoothing		.579	.069	8.409	.000
TCS2	Exponential Smoothing		.513	.066	7.733	.000
TCS3	TCS traffic	Constant	217,480.818	6,099.879	35.653	.000
		AR 1	.688	.060	11.387	.000
		MA 2	-.279	.081	-3.459	.001
TCS4	Exponential Smoothing		.446	.064	7.016	.000
VDS1-SB*	VDS speed	Constant	90.059	.415	216.761	.000
		AR 1	.492	.066	7.421	.000
VDS1-NB	VDS speed	Constant	90.703	.340	267.122	.000
		MA 1	-.441	.069	-6.367	.000
		MA 2	-.421	.071	-5.889	.000
		MA 3	-.407	.072	-5.691	.000
VDS2-SB	Exponential Smoothing		.300	.054	5.539	.000
VDS2-NB	Exponential Smoothing		.429	.063	6.799	.000
VDS3-SB	Exponential Smoothing		.129	.035	3.636	.000
VDS3-NB	Exponential Smoothing		.213	.046	4.675	.000

Note* : SB=southbound, NB=northbound

IC의 경우 구글에서 "고속도로"를 검색한 추세값이 선택되었다. 출퇴근의 상시적 통행이 포함된 지점의 TCS 진입 교통량은 구글 검색 트래픽이 시계열 모형의 입력변수로 우수성을 나타냈으며, 검색 트래픽과 음의 상관관계를 보였다. 여가 통행이 집중적으로 나타났던 지점의 TCS 진입교통량에서는 네이버의 검색 트래픽이 입력변수로 선정되었으며, 양의 상관관계가 나타났다.

시계열 도표의 해석 과정에서 언급했듯이 하계 휴가 혹은 명절 기간 동안은 검색 트래픽이 증가했음에도 불구하고, 상시적 통행과 관련된 TCS 지점의 주간 단위 교통량이 평소 출퇴근시보다 크지 않았기 때문에 음의 상관관계가 나타났으며, 여가 통행과 관련된 TCS는 검색 트래픽 증가시 교통량이 증가함에 따라 양의 상관관계가 나타난 것이다.

상시적 통행과 여가 통행에 대해 검색 출처가 다르게 입력변수로 사용된 것은 포털 업체별 서비스와 이용자 특성이 다르기 때문으로 판단된다. 네이버의 경우 추석/설 명절 등에 더 특화된 서비스가 제공되고 있다. 반면 구글은 인터넷 서비스가 익숙한 오래된 이용자에 의해 상시적 검색이 이루어지다 보니 상시적 통행에 대한 묘사에 유리했던 것으로 판단된다.

참고로 대안 모형(alternative model)으로서 TCS 진입교통량 추정 모형을 수식으로 나타내면 아래와 같다.

$$\triangle TCS1_t = (-642)\triangle Google_t^{t.i.} + (1+0.402B)a_t \quad (5)$$

$$\triangle TCS2_t = (-642+337B^8)\triangle Google_t^{t.i.} + (1+0.455B)a_t \quad (6)$$

$$TCS3_t = 201,633 + \frac{(604+350B^2)}{(1-0.443B^2)} Naver_t^{Cell-t.i.} + \frac{(1+0.278B^2)}{(1-0.704B)}a_t \quad (7)$$

$$\triangle TCS4_t = (-294)\triangle Google_t^{highway} + (1+0.454B)a_t \quad (8)$$

VDS 속도의 경우 Table 4를 통해 최종 선택 모형을 확인할 수 있다. 각각의 모형별로 변수에 차분을 적용한

Table 3. Results of alternative model(TCS)

Model			Estimate	SE	t	Sig.
TCS1	TCS traffic	difference	1			
		MA 1	.402	.070	5.707	.000
	Google (Trans. Info.)[1)]	NUM[2)] 1	-642.059	160.672	-3.996	.000
		difference	1			
TCS2	TCS traffic	difference	1			
		MA 1	.455	.071	6.387	.000
	Google (Trans. Info.)	NUM 0	-662.371	149.158	-4.441	.000
		NUM 8	-337.410	147.834	-2.282	.024
		difference	1			
TCS3	TCS traffic	Constant	201,633.488	9,427.180	21.389	.000
		AR 1	.704	.061	11.616	.000
		MA 2	-.278	.083	-3.349	.001
	Naver[C*] (Trans. Info.)	NUM 0	604.811	144.789	4.177	.000
		NUM 2	-350.491	146.306	-2.396	.018
		DEN[2)] 1	.443	.220	2.010	.046
TCS4	TCS traffic	difference	1			
		MA 1	.454	.069	6.562	.000
	Google (Highway)	NUM 0	-294.591	26.350	-11.180	.000
		difference	1			

Note1 : ( ) is a keyword to search.
Note2 : NUM=numerator, DEN=denominator
Note* : C=cell phone, PC=personal computer

경우와 그렇지 않은 경우가 혼재되어 있다. 각 모형별로 입력 계열로 선택된 검색 트래픽 정보는 서울TG→신갈JCT(VDS1_SB)와 안성JCT→안성IC(VDS3_SB)는 네이버의 검색 추세값이 선택되었으며, 신갈JCT→서울TG(VDS1_NB)는 어떠한 정보도 선택되지 않았다. 서평택JC→서평택IC(VDS2_SB)는 구글과 네이버의 검색 추세값이 동시에 선택되었으며, 그 외 지점에서는 출처 및 단어가 다른 3개 이상의 검색 추세값이 혼용되어 모형의 입력변수로 선택되었다.

기본 모형과 마찬가지로 대안 모형의 추정 파라미터 p-value(significance)는 모두 매우 낮아, 모형의 상수, AR, MR 파라미터 대부분 유의하게 나타났다(Table 3, 4).

VDS 속도의 경우 구간별로 선택된 최종모형의 형태가 상이하고 일부 모형(VDS1_NB)은 검색 트래픽 변수

Table 4. Results of alternative model(VDS)

Model			Estimate	SE	t	Sig.
VDS1_SB	VDS speed	Constant	89.389	.397	225.003	.000
		AR 1	.443	.070	6.364	.000
	Naver^{PC*} (Trans. Info.)	Delay	1			
		NUM 0	.080	.023	3.544	.001
		DEN 1	.724	.115	6.285	.000
		DEN 2	-.755	.115	-6.583	.000
VDS1_NB	VDS speed	Constant	90.703	.340	267.122	.000
		MA 1	-.441	.069	-6.367	.000
		MA 2	-.421	.071	-5.889	.000
		MA 3	-.407	.072	-5.691	.000
VDS2_SB	VDS speed	difference	1			
		MA 1	.518	.066	7.838	.000
	Google (Highway)	NUM 0	.114	.041	2.747	.007
		difference	1			
	Naver^{PC} (Highway)	NUM 0	-.287	.039	-7.300	.000
		difference	1			
VDS2_NB	VDS speed	Constant	93.848	1.112	84.362	.000
		AR 1	.762	.051	14.981	.000
	Google (Trans. Info.)	NUM 0	-.131	.030	-4.306	.000
		NUM 1	-.128	.038	-3.359	.001
		NUM 2	.119	.029	4.034	.000
		DEN 1	1.032	.204	5.058	.000
		DEN 2	-.678	.202	-3.362	.001
	Naver^{PC} (Highway)	NUM 0	.130	.044	2.919	.004
	Naver^{C*} (Highway)	NUM 0	-.326	.043	-7.541	.000
VDS3_SB	VDS speed	difference	1			
		MA 1	.831	.045	18.356	.000
	Naver^{PC} (Highway)	NUM 0	-.170	.024	-7.032	.000
		difference	1			
VDS3_NB	VDS speed	difference	1			
		MA 1	.479	.066	7.249	.000
		MA 3	.353	.064	5.536	.000
	Google (Highway)	Delay	2			
		NUM 0	-.397	.082	-4.846	.000
		difference	1			
	Google (Trans. Info.)	NUM 0	-.290	.038	-7.566	.000
		NUM 2	-.269	.074	-3.635	.000
		difference	1			
	Naver^{PC} (Highway)	NUM 0	.261	.055	4.755	.000
		difference	1			
	Naver^{C*} (Highway)	NUM 0	-.319	.054	-5.916	.000
		difference	1			

가 전입 함수 잡음 시계열 모형의 입력 변수로 선택되지 않았음을 감안할 때, 지점별 혹은 검색어의 선택과 관련된 일반화된 결론을 내리기는 쉽지 않다. 또한 2개 이상의 검색어가 입력변수로 포함되고 모형의 계수 부호가 상이함에 따른 해석의 어려움 또한 존재하고 있다. 따라서 Figure 6에서 확인한 바와 같이 검색 트래픽과 속도 간 음의 상관관계를 가설로 수립하고 분석기간, 검색 출처, 시차를 개별적으로 통제한 상태에서 추가적인 연구가 수행되어야 할 것으로 판단된다.

참고로 대안 모형(alternative model)으로서 VDS 지점 중 VDS1, 2 지점에 대한 주간 평균속도 추정 모형을 수식으로 표현하면 다음과 같다.

$$VDS1SB_t = 89 + \frac{0.080}{(1-0.724B+0.755B^2)} Naver_{t-1}^{PC-t.i.} + \frac{1}{(1-0.443B)} a_t \quad (9)$$

$$VDS1NB_t = 91 + (1+0.441B+0.421B^2+0.407B^3)a_t \quad (10)$$

Table 5. Comparison of goodness-of-fit measures

Model			Model fit statistics					Ljung-Box Q(18)		
			Stationary R^2	R^2	RMSE	MAPE	Normalized BIC	Statics	Df	Significant
Base Model (A)	TCS1	Exponential Smoothing	.126	.592	30,203.410	3.221	20.661	13.658	17	.691
	TCS2	Exponential Smoothing	.186	.414	27,839.816	3.244	20.498	12.324	17	.780
	TCS3	ARIMA(1,0,2)	.615	.615	19,947.096	6.593	19.891	21.739	16	.152
	TCS4	Exponential Smoothing	.214	.492	5,473.032	3.576	17.245	20.236	17	.262
	VDS1 - SB	ARIMA(1,0,0)	.225	.225	2.818	2.361	2.132	14.989	17	.596
	VDS1 - NB	ARIMA(0,0,3)	.322	.322	2.036	1.553	1.541	15.227	15	.435
	VDS2 - SB	Exponential Smoothing	.348	.318	4.414	3.134	2.999	11.509	17	.829
	VDS2 - NB	Exponential Smoothing	.160	.406	3.981	2.966	2.793	31.454	17	.018
	VDS3 - SB	Exponential Smoothing	.442	.075	4.184	3.205	2.892	8.976	17	.941
	VDS3 - NB	Exponential Smoothing	.362	.101	4.376	3.218	2.982	23.604	17	.131
Alternative Model (Transfer function model) (B)	TCS1	ARIMA(0,1,1)	.200	.624	29,056.765	3.020	20.614	10.855	17	.864
	TCS2	ARIMA(0,1,1)	.290	.502	26,200.668	3.029	20.440	7.012	17	.983
	TCS3	ARIMA(1,0,2)	.655	.655	19,153.018	6.302	19.900	20.237	16	.210
	TCS4	ARIMA(0,1,1)	.544	.705	4,193.957	2.803	16.742	16.261	17	.505
	VDS1 - SB	ARIMA(1,0,0)	.265	.265	2.670	2.270	2.115	16.876	17	.463
	VDS1 - NB	ARIMA(0,0,3)	.322	.322	2.036	1.553	1.541	15.227	15	.435
	VDS2 - SB	ARIMA(0,1,1)	.551	.530	3.696	2.719	2.704	21.888	17	.189
	VDS2 - NB	ARIMA(1,0,0)	.665	.665	3.075	2.551	2.516	21.592	17	.201
	VDS3 - SB	ARIMA(0,1,1)	.562	.275	3.716	2.744	2.685	5.526	17	.996
	VDS3 - NB	ARIMA(0,1,3)	.576	.402	3.658	2.855	2.805	34.087	16	.005
Difference (B-A)	TCS1	-	.074	.032	-1,146.645	-0.201	-.047	-2.803	-	.173
	TCS2	-	.104	.088	-1,639.148	-0.215	-.058	-5.312	-	.203
	TCS3	-	.040	.040	-794.078	-0.291	.009	-1.502	-	.058
	TCS4	-	.330	.213	-1,279.075	-0.773	-.502	-3.975	-	.243
	VDS1 - SB	-	.040	.040	-0.148	-0.092	-.017	1.887	-	-.133
	VDS1 - NB	-	-	-	-	-	-	-	-	-
	VDS2 - SB	-	.203	.212	-0.719	-0.416	-.296	10.379	-	-.640
	VDS2 - NB	-	.506	.259	-0.907	-0.414	-.277	-9.863	-	.183
	VDS3 - SB	-	.120	.200	-0.468	-0.461	-.208	-3.450	-	.055
	VDS3 - NB	-	.214	.301	-0.718	-0.363	-.177	10.483	-	-.125

$$\triangle VDS2SB_t = (0.114)\triangle Google_t^{highway} + (-0.287)\triangle Naver_t^{PC-highway} + (1-0.518B)a_t \quad (11)$$

$$VDS2NB_t = 94 + \frac{(-0.131+0.128B-0.119B^2)}{(1-1.032B+0.678B^2)}Google_t + 0.130Naver_t^{PC} - 0.326Naver_t^{PC} + \frac{1}{(1-0.762B)}a_t \quad (12)$$

3) 전이함수 적용 전후 적합성 비교 검토

고속도로 진입 교통량과 이에 영향을 받는 고속도로 구간 속도를 추정함으로써 검색 트래픽 정보의 활용이 교통지표 추정 모형의 적합성을 높일 수 있는지 검토하는데 본 연구의 목적이 있으므로 기본 모형과 대안 모형에 대한 적합도 통계량을 검토하였다(Table 5).

시계열 자료의 적합도를 판단할 수 있는 통계량 중 본 연구에서는 큰 값을 가질수록 우수한 정상 R^2와 R^2, 정규화된 BIC(Normalized Baysian Information Criterion), 그리고 작은 값을 가질수록 우수한 평균제곱오차의 제곱근(RMSE; root mean square error), 평균절대백분위오차(MAPE; mean absolute percentage error)를 검토하였다.

대부분 모형에서 전이함수를 적용함에 따라 모형의 적합성이 높아졌다. 특히 VDS 2, 3의 주간 평균속도 모형은 정상 R^2와 R^2를 통해 확인된 적합도가 뚜렷하게 개선되었다. TCS 진입 교통량과 VDS 속도의 단위가 다름을 고려하여 MAPE를 통해 확인해 본 적합도는 전체적으로 0.092-0.773이 감소하였다. VDS 2, 3의 속도 추정 모형은 전이함수 적용으로 MAPE가 각각 0.4 정도 감소하는 등 안정적인 적합도 개선이 나타났다.

이는 단기 교통지표 추정시 검색 트래픽 정보가 시계열 모형의 입력변수로 사용되면 모형이 개선된다는 것을 의미한다. 회귀분석이 교통지표 추정시 입력계열의 미래값을 알아야 한다는 문제점이 있는 반면, 전이함수 모형은 현시점 혹은 과거 시점의 검색 트래픽 정보만으로 개선된 교통지표 추정 모형이 가능한 것이다. 또한 시계열 도표를 통해 확인한 출력계열 교통지표가 입력계열인 검색 트래픽을 통해 설명되고 있음을 확인한 것이다.

다만, Ljung-Box 통계량 확인 결과, 잔차간의 상관관계에 대한 검증통계량 일부가 약화되기도 하였다. 또한, SPSS의 Expert Modeler를 활용하여 최적화된 모형을 선택하다 보니, 변수의 가감 혹은 시차의 증감을 반복하는 과정에서 정규화된 BIC가 매우 낮은 수준이나마 증가하는 현상이 TCS3 지점에서 나타났다.

결론 및 향후 연구과제

2000년대 중반부터 구글 등 인터넷 포털 서비스 업체들은 온라인 사용자들의 웹검색 트래픽 정보를 서비스화하여 공개하고 있다.

각종 선행연구에서 검토된 바와 같이 검색 트래픽이 교통지표의 선행지표 기능을 가질 수 있다면 이는 일차적으로는 교통지표 추정의 불확실성을 보완하는 도구로 사용할 수 있게 된다. 이에 본 연구는 고속도로상에서의 교통지표인 진입 교통량과 이에 영향을 받는 고속도로 구간 속도를 추정함으로써 검색 트래픽 정보의 활용 가능성을 검토하였으며, 또한 검색 출처와 검색어의 형태를 구분하여 시계열 모형의 입력변수를 최적함으로써 추정 지점별로 그 모형을 비교 분석을 수행하였다.

1. 결론

2010년 7월부터 2013년 10월까지 교통 관련 단어의 검색 트래픽과 TCS 진입교통량 그리고 VDS 속도에 대해 시계열 도표를 통해 확인하고, 기본 모형으로서 지수평활 혹은 일변량 ARIMA 모형을 대안 모형으로서 전이함수 시계열 모형을 추정하였는데, 서울(TCS1)과 서서울(TCS2) 톨게이트는 구글에서의 "교통정보"를 검색한 추세값이 포함된 전이함수 잡음 시계열 모형이 최적모형으로 선택되었으며, 남양주(TCS3) 톨게이트의 경우 네이버 모바일에서의 "교통정보"를 검색한 추세값이 선택되었다. 안성(TCS4) IC의 경우 구글에서 "고속도로"를 검색한 추세값이 선정되었다. 출퇴근의 상시적 통행 비율이 높은 지점의 TCS 진입 교통량은 구글 검색 트래픽이 시계열 모형의 입력변수로 우수성을 나타냈으며, 검색 트래픽과 음의 상관관계를 보였다. 여가 통행이 집중적으로 나타났던 지점의 TCS 진입교통량에서는 네이버의 검색 트래픽이 입력변수로 선정되었으며, 양의 상관관계가 나타났다.

VDS 속도의 경우 구간별로 선택된 최종모형의 형태가 상이하고 일부 모형(VDS1_NB)은 검색 트래픽 변수가 전입 함수 잡음 시계열 모형의 입력 변수로 선택되지 않았음을 감안할 때, 지점별 혹은 검색어의 선택과 관련된 일반화된 결론을 내리기는 쉽지 않았다. 다만, 분석기간, 검색 출처, 시차를 개별적으로 통제한 상태에서 추가적인 연구가 수행되어야 할 것으로 판단되었다.

전이함수 적용 전후를 비교한 결과 적용 후 모형의 적합성이 높아졌다. 특히 VDS 2, 3의 주간 평균속도 모형은 정상 R^2와 R^2를 통해 확인된 적합도가 뚜렷하게 개선되었다. 또한 TCS 진입 교통량과 VDS 속도의 단위가 다름을 고려하여 MAPE를 통해 확인해 본 적합도의 경우, VDS 속도 추정 모형이 지점과 무관하게 안정적으로 전이함수의 적합성이 높다는 결론을 얻을 수 있었다.

이는 시계열 도표를 통해 확인된 교통지표와 검색 트래픽간의 관련성을 모형을 통해 재확인한 것으로, 단기 교통지표 추정시 검색 트래픽 정보가 시계열 모형의 입력변수로 사용됨으로써 모형이 개선될 수 있다는 것을 의미하는 것이다.

다만, 검색 트래픽의 대상 단어 선정 과정에 있어 본 연구에서는 다소 직관적으로 "교통정보", "고속도로"라는 용어를 사용한 측면과, 교통지표와 검색 트래픽의 집계주기를 자료 구득 가능성 등으로 일주일 단위로 설정함에 따른 활용의 한계점 또한 존재한다.

2. 향후 연구과제

본 연구에서 입력변수이었던 검색 트래픽과 출력 변수이었던 교통 관련 변수를 중심으로 다음 이슈에 대한 다양한 향후 연구가 필요할 것이다.

1) 검색 트래픽

검색 트래픽의 출처로 본 연구는 구글과 네이버를 그 대상으로 삼았다. 그러나 각각의 포털 사이트마다 이용자 특성이 상이함은 물론이며, 국내 검색 시장에서의 점유율 또한 각 업체별로 지속적으로 바뀌고 있다. 또한 포털 서비스 업체가 아니라 직접적으로 교통정보 제공 사이트의 접속 트래픽을 확인하거나 교통정보 관련 모바일 어플리케이션의 접속 트래픽을 확인하는 것도 한 가지 방법이 될 수 있다. 즉 교통과 검색 트래픽을 접목시킬 때, 어떤 출처의 트래픽을 활용하는 것이 적합한지에 대한 후속 연구가 필요할 것이다.

2) 교통 관련 변수

교통과 관련된 변수는 수단과 시차(time lag), 집계단위가 중요하다. 본 연구에서는 고속도로라는 공간에서 차량으로 기인되는 교통량과 속도를 분석 대상으로 검토하였다. 하지만 버스, 철도, 항공, 보행, 자전거 등 이용수요, 혹은 그 결과로 도출되는 각종 지표 역시 특정 검색어와 밀접한 관련을 맺고 변동 될 수 있다.

통상 검색 행위와 그에 기인한 통행 관련 행위가 나타남에 있어 시차가 발생한다. 항공과 도로 인프라를 활용해서 통행을 할 때 검색 트래픽이 선제적으로 발생되는 시차는 다를 것이다. 시계열 모형을 수립함에 있어서도 시차를 단계적으로 조정해 최적의 모형을 구축해야만, 교통과 관련된 수요 추정 혹은 운영계획 수립시 해당 시차를 감안하여 활용할 수 있다. 이 때 집계 단위를 1시간 혹은 일 단위 조정하여 자료 가공 및 분석을 하는 과정도 동반되어야 할 것이다.

ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea grant funded by the Korea Government(MSIP) (NRF-2010-0029446).

REFERENCES

Bin M. Y., Mun J. B. (2012), A Study on the Traffic Speed Estimation Using Bus Information, Gyeonggi Research Institute, Suwon, South Korea.

Bollen J., Mao H., Zeng X. J. (2011), Twitter Mood Predicts the Stock Market, Journal of Computational Science, 2(1), 1-8.

Choi H., Varian H. (2009a), Predicting the Present With Google Trends, Google Technical Report.

Choi H., Varian H. (2009b), Predicting Initial Claims for Unemployment Insurance Using Google Trends, Technical Report.

Choi H., Varian H. (2012), Predicting the Present With Google Trends, The Economic Record, 88, 2-9.

Cooper C., Mallon K., Leadbetter S., Pollack L., Peipins L. (2005), Cancer Internet Search Activity on a Major Search Engine, United States 2001-2003, Journal of Medical Internet Research, 7(3), e36.

Ettredge M., Gerdes J., Karuga G. (2005), Using Web-based Search Data to Predict Macroeconomic Statistics, Communications of the ACM, 48(11), 87-92.

Ginsberg J., Mohebbi M. H., Patel R. S., Brammer L., Smolinski M. S., Brilliant L. (2009), Detecting influenza

epidemics Using search engine query data, Nature, 457, 1012-1014.

Goel S., Hofman J. M., Lahaie S., Pennock D. M., Watts D. J. (2010), Predicting consumer behavior With Web search, Proceedings of the National Academy of Sciences, 7(41), 17486-17490.

Jang J. A., Chung W. H., Choi K. C. (2005), Bi-variate and Muli-variate Time Series-based Algorithm for Link Travel Speed Estimation, Proceedings of Korean Society of Civil Engineers, 3741-3745.

Ju J. W. (2013), Time Series Analysis and Modeling for the Investigation of Seawater Intrusion into Fractured Rock Aquifer in Muan, Korea, Chonnam National University, A Master's Thesis

Jun S. P., Park D. H. (2013), Intelligent Brand Positioning Visualization System Based on Web Search Traffic Information : Focusing on Tablet PC, J Intell Inform Syst, 19(3), 93-111.

Jun S. P., Yeom J. H., Son J. K. (2014), A Study of the Method Using Search Traffic to Analyze New Technology Adoption, Technol. Forecast. Soc. Change, 81, 82–95.

Kim H. S. (2009), A Study on Time and Space Variation of Traffic Flow Characteristics on a Merge Influence Section in Uninterrupted Facility, Hanyang University, Doctoral Dissertation.

Kim K. H., Kim H., S. (2011), KTX Passenger Demand Forecast With Intervention ARIMA Model, Journal of the Korean Society of Road Engineers, 14(5), 470-476.

Kim M. S., Koo P. H. (2013), A Study on Big Data Based Investment Strategy Using Internet Search Trends, Journal of the Korean Operations Research and Management Science Society, 34(4), 53-63.

Lee J. M., Kwon Y. J. (2011), A Study on Dynamic Change of Transportation Demand Using Seasonal ARIMA Model, J. Korean Soc. Transp, 29(5), Korean Society of Transportation, 139-155.

Lee S. J., Back N. C., Kwon H. J. Choi D. S., Do M. S. (2001), The AADT Estimation Through Time Series Analysis Using Irregular Factor Decomposition Method, J. Korean Soc. Transp, 19(6), Korean Society of Transportation, 65-73.

Manheim M. L. (1979), Fundamentals of Transportation Systems Analysis, MIT Press, Cambridge, Massachusetts, USA.

McLaren L., Shanbhogue R. (2011), Using Internet Search Data as Economic Indicator, Quarterly Bulletin, Q2, 134-140.

Min K. C., Jun Y. I., Ha H. K. (2013), Forecasting the Air Cargo Demand With Seasonal ARIMA Model: Focusing on ICN to EU Route, J. Korean Soc. Transp, 31(3), Korean Society of Transportation, 3-18.

Moat H. S., Curme C., Avakian A., Kenett D. Y., Stanley H. E., Preis T. (2013), Quantifying Wikipedia Usage Patterns Before Stock Market Moves, Scientific Report, 3, 1801, 1-5.

Polgreen P. M., Chen Y., Pennock D. M., Nelson F. D. (2008), Using Internet Searches for Influenza Surveillance, Healthcare Epidemiology, 47, 1443-1448.

Preis T., Reith D., Stanley H. E. (2010), Complex Dynamics of Our Economic Life on Different Scales : Insights From Search Engine Query Data, Philosophical Transactions of the Royal Society, 368, 5707-5719.

Yoo S. L., Kim J. S., Jeong J. S., Jeong J. Y. (2014), A Prediction of Marine Traffic Volume Using Artificial Neural Network and Time Series Analysis, Journal of the Korean Society of Marine Environment & Safety, 20(1), 33-041.

Google Trends: http://www.google.com/trends

Naver Trend: http://trend.naver.com

Korea Expressway Co. OASIS: http://data.ex.co.kr

♧ 주 작 성 자 : 류인곤
♧ 교 신 저 자 : 류인곤
♧ 논문투고일 : 2014. 8. 21
♧ 논문심사일 : 2014. 9. 17 (1차)
2014. 11. 2 (2차)
2014. 11. 19 (3차)
♧ 심사판정일 : 2014. 11. 19
♧ 반론접수기한 : 2015. 6. 30
♧ 3인 익명 심사필
♧ 1인 abstract 교정필