J. Korean Soc. Transp.
Vol.33, No.5, pp.497-507, October 2015
http://dx.doi.org/10.7470/jkst.2015.33.5.497

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

Article

사고등급별 고속도로 교통사고 처리시간 예측모형 개발

이숭봉[1] · 한동희[2] · 이영인[1]*
[1]서울대학교 환경대학원, [2]한국도로공사 도로교통연구원

Development of Freeway Traffic Incident Clearance Time Prediction Model by Accident Level

LEE, Soong-bong[1] · HAN, Dong Hee[2] · LEE, Young-Ihn[1]*
[1]Graduate School of Environmental Studies, Seoul National University, Seoul 151-742, Korea
[2]Transportation Research Division, Korea Expressway Corporation, Gyeonggi 445-812, Korea

Abstract

Nonrecurrent congestion of freeway was primarily caused by incident. The main cause of incident was known as a traffic accident. Therefore, accurate prediction of traffic incident clearance time is very important in accident management. Traffic accident data on freeway during year 2008 to year 2014 period were analyzed for this study. KNN(K-Nearest Neighbor) algorithm was hired for developing incident clearance time prediction model with the historical traffic accident data. Analysis result of accident data explains the level of accident significantly affect on the incident clearance time. For this reason, incident clearance time was categorized by accident level. Data were sorted by classification of traffic volume, number of lanes and time periods to consider traffic conditions and roadway geometry. Factors affecting incident clearance time were analyzed from the extracted data for identifying similar types of accident. Lastly, weight of detail factors was calculated in order to measure distance metric. Weight was calculated with applying standard method of normal distribution, then incident clearance time was predicted. Prediction result of model showed a lower prediction error(MAPE) than models of previous studies. The improve model developed in this study is expected to contribute to the efficient highway operation management when incident occurs.

고속도로의 비반복 혼잡은 주로 돌발상황에 의해 발생된다. 돌발상황의 주요 원인은 교통사고로 알려져 있다. 따라서 교통사고 시 사고처리시간을 정확하게 예측하는 것은 돌발상황 관리에서 매우 중요하다. 본 연구에서는 전국고속도로의 2008-2014년 총 7년치(60,473건)의 사고 자료를 이용하였다. 사고처리시간 예측모형은 과거의 교통사고 이력자료를 바탕으로 비모수모형인 KNN (K-Nearest Neighbor) 알고리즘을 활용하였다. 사고자료 현황 분석결과 사고등급별로 사고처리시간에 미치는 영향이 매우 큰 것으로 분석되었다. 따라서 사고처리시간은 사고등급별로 분류하여 모형을 구축하였다. 그리고 현재 발생한 사고의 교통상황과 도로 기하구조를 반영하기 위하여 교통량, 차로수, 시간대를 구분하여 데이터를 추출하였다. 추출된 데이터 중 현재 교통사고와 유사한 사고를 검색하기 위하여 사고처리시간에 영향을 미치는 요인들을 분석하였다. 마지막으로, 상태간 거리 산정을 위해서 세부항목별 가중치를 산정하였다. 가중치산정은 정규분포 표준화방법을 적용하였고, 이를 통해 사고처리시간을 예측하였다. 본 연구에서 개발된 모형의 예측결과는 기존의 연구들의 결과에 비해 낮은 예측오차(MAPE)를 보여 모형의 우수성을 입증할 수 있다고 판단된다. 본 연구를 통해 고속도로의 돌발상황 발생 시 효율적인 고속도로의 운영관리에 기여할 수 있고, 기존의 모형들이 갖고 있던 한계를 개선 및 보완할 수 있을 것으로 판단된다.

Keywords

accident level, freeway incident, Incident clearance time, KNN(K-Nearest Neighbor) model, weight
사고등급, 고속도로 돌발상황, 사고처리시간, KNN모형, 가중치

* : Corresponding Author
yilee@snu.ac.kr, Phone: +82-2-880-1430, Fax: +82-2-871-8847

Received 3 April 2015, Accepted 18 August 2015

서론

교통혼잡은 반복적인 혼잡(recurrent congestion)과 비반복적 혼잡(nonrecurrent congestion)으로 분류 할 수 있다. 반복적인 혼잡은 교통수요가 용량을 초과하였을 때 발생되는 지정체를 의미하여, 비반복적 혼잡은 공사, 행사 등과 같이 예측 가능한 사건들과 사고 및 이상기후 등 예측 불가능한 사건들에 의해 발생되는 혼잡을 말한다. Friedrich(2012)의 연구에 따르면 고속도로 구간의 정체원인 중 사고(복수요인 포함)와 관련된 정체의 비율은 44%로 매우 높은 것으로 나타났다. 따라서 교통사고에 대해 즉각적으로 정보를 제공하고, 신속히 교통류를 관리하는 정도는 교통류관리시스템의 수준을 의미한다고 할 수 있다. 또한 교통사고 시 사고처리시간을 정확히 예측하는 것은 돌발 상황관리에서 매우 중요하다.

교통사고 처리시간(Incident Clearance Time)은 교통사고가 발생한 후 이를 검지하고 작업차량 및 응급차량이 도착한 이후부터 교통사고 상황을 종료시킬 때 까지 소요되는 시간을 의미한다. 검지시간은 사고가 발생한 후 이를 인지 확인할 때 까지 소요되는 시간을 의미한다. 대응시간은 사고를 접수한 후 현장에 대응반이 도착할 때까지의 소요시간이며, 이미 시스템화 되어있어 교통사고별로 시간의 편차가 크지 않다. 고속도로 사고자료에서는 검지시간과 대응시간을 합산하여 현장도착시간으로 표기하고 있다. 이는 사고가 발생하면 다양한 방법으로 사고가 접수되고, 전파되며, 여러 유관기관들이 각자 현장에 도착하므로 검지시간과 대응시간으로 구분하는 것이 현실적으로 불가능하기 때문이다. 현장도착시간은 사고가 발생한 지점의 사고발생요인, 도로 및 교통요인들의 영향에 의해 결정된다고 보기 어렵다. 따라서 본 연구에서는 교통사고지속시간 중에서 현장 도착시간을 제외한 사고를 처리하는데 소요되는 시간인 교통사고 처리시간만을 대상으로 분석을 수행하였다.

교통사고처리시간은 사고점유 공간(차로수), 날씨, 사고차량수, 사고심각도(A-D), 도로 및 기하구조, 교통요인 등 다양한 변수들의 영향으로 결정될 것이다. 하지만, 사고가 발생하였을 때 사고처리시간이 어떻게 될지 예측하는 것은 어려운 일이다. 사고가 발생하였을 때 교통사고 처리시간 예측은 교통영향권 내의 대기행렬, 지체, 정상류 회복시간 등을 산정하기 위해 필수적인 자료로서, 돌발상황에 따른 효율적인 고속도로의 운영관리 및 비반복정체의 효과를 최소화하기 위하여 필요하다. 또한, 실시간으로 교통사고에 따른 처리시간 예측을 위해서는 현장에서 바로 확인이 가능한 변수로 구성되어질 필요가 있다.

기존의 교통사고 처리시간과 관련된 연구는 국내에서는 활발하게 연구되고 있지 않은 실정이며, 국외에서는 수학적 모형(회귀모형) 및 decision tree, 비모수모형(NPR, Non Parametric Regression)등을 이용한 연구들이 진행되었다. 교통사고 처리시간의 경우 다양한 원인에 의해 복합적으로 작용하므로 교통사고 처리시간 예측 시 비모수모형이 유용하게 사용되어 질 수 있다. Smith(2002)는 고속도로 관리시스템에서 비모수회귀식은 긍정적인 결과를 보였고, 교통상황 예측기술에 유용하게 적용될 수 있음을 확인하였다.

본 연구에서는 과거의 교통사고 이력자료를 바탕으로 NPR모형의 일종인 KNN(K Nearest Neighbor) 알고리즘을 이용하여 교통사고 처리시간 예측모형을 개발하였다. NPR모형은 설명변수를 고려하는 모수 회귀식(Parametric Regression) 과는 달리 설명변수를 고려하지 않는 특징을 갖는 기법이다. 모수회귀식의 경우에는 설명변수에 의한 영향, 즉 파라미터(Parameter)가 장래에도 동일하게 종속변수에 영향을 미친다는 가정을 전제로 하는 문제점이 존재한다. 본 연구에서 사용한 모형은 현재 조건과 유사한 과거의 관측치를 탐색하여 장래의 상태를 예측하는데 적용이 용이하다고 알려져 있다.

기존연구 고찰

1. 고속도로 교통사고등급

고속도로에서 발생하는 교통사고 등급분류기준은 인명피해, 관련차량, 교통차단, 직원관련 사고, 기타 요인들에 의해서 A-D등급으로 분류하여 사고자료를 구축하고 있다. 사고등급 D는 경미한 사고로 해당기준에 포함되지 않는 사고들이 모두 포함된다. 교통사고 등급은 사고의 심각한 정도를 분류하는 기준으로 사고처리시간과 가장 밀접하게 관련되어 있다. Ha(2010)의 연구에서도 사고등급이 사고처리시간에 가장 큰 영향을 미치는 변수임을 밝히고, 사고등급별로 사고처리시간 예측모형을 개발하였다.

실제 고속도로 상의 사고자료를 분석한 결과 Figure 1과 같이 사고등급별 사고처리시간 분포현황은 차이를

Table 1. Incident clearance time by incident level

Index	A	B	C	D	total
Clearance time(m)	128	67	43	36	39
Number of incidents	57	1,773	13,270	45,373	60,473

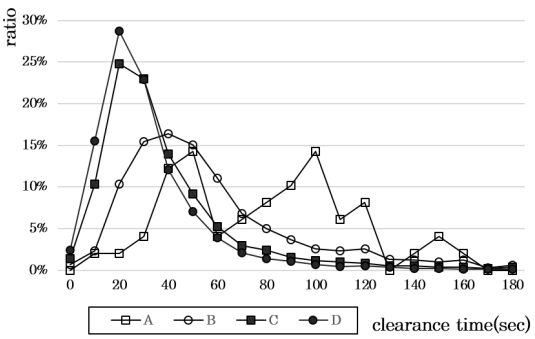

Figure 1. Incident clearance time distribution by accident level

보이고 있다. 경미한 사고인 C, D등급의 사고는 사고처리시간이 60분 이내인 경우가 거의 90%에 달하며, B등급은 1시간 이상이 약 40%, 가장 심각한 사고등급인 A는 65%가 넘는 것으로 분석되었다. Table 1은 등급별 평균 사고처리시간이다. 등급별로 큰 차이를 보이므로 사고처리사간 예측 시 분류가 필요하다.

2. 교통사고처리(지속)시간 관련연구

Shin(2002)의 연구에서는 돌발상황 시 지속시간 예측모형을 개발하기 위하여 독립변수는 돌발상황 확인시점에서 수집 가능한 변수이어야 하며, 모형의 현장적용을 위해서는 해당도로의 교통관제시스템의 수준과 함께 정보원(CCTV, 순찰반 등)의 특성을 고려해야 한다고 설명하고 있다. 지속시간 예측모형에 사용된 자료는 교통사고발생 속보 21개월 자료를 정리하여 본선에서 발생한 각종 돌발상황 168건을 사용 자료로 추출하고 이를 DB화하여 통계분석을 수행하였으며, 다중회귀모형을 기반으로 모형을 구축하였다.

Chung(2007)은 AFT(accelerated Failure Time) metric 모형을 적용하여 사고지속시간을 분석하였다. AFT모형은 예측모형으로도 널리 사용되었기 때문에 추정된 모형은 사고 발생 시 사고 관련 기본정보 접수 즉시 고속도로에서의 사고 지속시간 예측에 사용될 수 있다고 설명하고 있다. 결과적으로 예측된 사고 지속시간 정보는 사고를 처리하기 위한 제반 의사결정에 도움을 줄 뿐 아니라 교통 혼잡의 감소 및 사상자의 감소로 그 효과가 이어질 것으로 예상하고 있다.

Ha(2010)은 경부선의 교통사고 과거이력자료를 바탕으로 돌방상황 처리시간 예측모형 개발하였다. 종속변수인 사고처리시간을 사고등급 A, B, C등급으로 구분하였으며, 처리시간에 영향을 미치는 주요변수로는 사고등급, 교통량, 주·야 구분, 중차량 포함여부, 사고차량수 등이 영향을 미치는 것으로 분석하였다. 사고등급은 돌발상황 처리시간에 큰 영향을 미치는 변수로서 본 연구에서는 사고등급별로 처리시간 예측모형을 도출하였다. 또한, 처리시간에 영향을 미치는 주요변수를 이용하여, 고속도로 운영관리자가 돌발상황 발생 시 신속하게 적용할 수 있는 의사결정나무(decision tree)를 구축하였다.

Lee(2012)은 고속도로 교통사고를 대상으로 돌발상황 지속시간에 영향을 주는 요인들을 찾아내기 위한 모형을 개발하였다. 모형은 모든 대상을 포함한 통합모형과 일반구간, 교량, 터널 등 교통사고 장소별로 구분하여 분석한 세부모형 등 모두 4개의 모형을 구축하였다. 분석결과 교통사고가 발생한 장소에 따라 돌발상황 지속시간에 영향을 주는 것으로 나타났으며, 현장 처리를 위한 작업차량 도착시간이 가장 민감한 요인으로 분석되었다. 또한, 차대차 사고, 화물차에 의한 사고, 야간사고, 주말사고 등의 시사점 있는 요인을 찾아냈다.

Wang(2015)은 NPR(non-parametric regression) 모형인 KNN(k-Nearest Neighbor) 알고리즘을 이용하여 사고지속시간을 예측하였다. 사용된 주요 사고유형은 경미한 충돌(sideswipe)이며 평균 사고지속시간은 32.69분이였다. 이상치 제거를 위해서는 Vivatrat method를 사용하였다. 사고지속시간에 영향을 미치는 변수는 Kruskal-Wallis test를 통해 설정하였다. 주요요인으로는 6가지 요인이 설정되었다. 6가지 요인은 day first shift, weekday, incident type, congestion, incident grade, distance 이다. 본 연구에서는 사고지속시간 분포의 특징을 기반으로 원시자료의 log변형을 수행하였다. 그리고 모형성능의 개선을 증명하기 위하여 다른 상태간거리와 예측알고리즘을 검토하여 비교분석하였다. 그 결과 KNN모형은 가중된 예측알고리즘과 의사결정나무(decision tree) 기반의 가중된 상태거리를 사용할 때 예측의 정확도가 더 높은 것으로 나타났다. 또한 동일한 자료를 기반으로 다른 모형들과 비교한 결과 매우 짧거나 긴 사고지속시간을 제외하고는 예측

Table 2. Evaluation of predictive error with different method (Qing. H(2011))

Index	KNN	CART	URP	Quantile reg.
MAPE1	58.5%	57.4%	54.1%	49.3%
MAPE2	43.1%	42.8%	40.2%	34.8%

note: 1) MAPE1 : Mean Absolute Percentage Error
2) MAPE2 : Median Absolute Percentage Error

Table 3. Evaluation of predictive error with different method (Gaetano. V(2008))

Index	MLR	DT	ANN	RVM	KNN
MAPE	34%	43%	44%	36%	36%
RMSE	20.04	23.07	19.80	17.29	20.29

력이 우수한 것으로 나타났다.

Chang(2013)은 기존 연구들이 갖고 있는 이론적인 한계를 보완하기 위하여 분류트리(classification tree) 기반의 사고지속시간 예측모형을 개발하였다. 본 연구에서는 대상변수(사고지속시간)가 큰 범위(range)를 가지므로 연속형 변수를 범주형 변수로 변형하기 위해 클러스터 분석(short-duration, medium-duration, long-duration)을 사용하였다. 클러스터의 수는 K-means방법을 사용하고, 이를 이용하여 분류트리를 구축하였다. 트리분석은 CART 알고리즘을 이용하였고, 분류기준은 Gini값을 사용하였다. 관측값과 예측값을 비교한 결과 단기(5-41분)는 96.7%, 중기(42-118분)는 17.2%, 장기(119-391분)는 11.4%의 정확도를 보였다.

Qing. H(2011)는 비반복정체의 영향 최소화를 목표로 사고 지속시간 예측을 위해 Hybrid tree 기반의 분위회귀분석(Quantile Regression) 방법론을 제안하였다. 또한, 동일한 자료를 기반으로 KNN, CART (Classification and Regression Tree), URP(Unbiased Recursive Partitioning) 등의 기존의 방법론과 비교분석하였다. 분석결과는 Table 2와 같다.

Gaetano. V(2010)는 교통사고 발생 시 교통운영자에게 실시간으로 교통사고지속시간 예측정보 제공을 위하여 5가지 예측모델(MLR : Multiple Linear Regression, DT : Decision Tree, ANN : Artificial Neural Network, RVM : Relevance Vector Machine, KNN)을 비교분석하였다. 분석에 사용된 자료는 동일한 자료를 이용하였다. 분석결과 사고지속시간이 90분 미만의 경우 예측의 정밀도 가 높은 것으로 나타났다. 이러한 원인은 심각도가 높은 사고의 발생빈도가 상대적으로 낮기 때문인 것으로 분석하였다. 모형별 분석결과는 Table 3과 같다.

Y. Qi(2004)는 교통관리 시스템에 의해 축적된 빅데이터를 효과적으로 이해하기 위해서는 현재의 교통흐름과 유사한 과거이벤트를 검색해야하며, 이러한 유사이벤트는 현재 사건에 대한 최인접이웃(Nearest Neighbor)이라고 말하였다. 데이터의 인접 이웃을 식별하기 위해서 거리 매트릭스는 case 사이에서의 유사성 측정을 위해서 필요하다. 일반적으로 교통 이벤트 데이터를 구성하는 범주형 변수는 유클리드 거리가 유효하지 않다. 이러한 문제를 해결하기 위하여, 본 연구에서는 범주형 데이터를 사용할 수 있는 거리 매트릭스를 개발하였다. 매트릭스 값은 가장 인접한 이웃을 선택하는데 객관적인 지료로 사용된다. 매트릭스는 사고 지속시간 예측을 목적으로 현재 사고와 유사한 과거 사건을 식별하기 위하여 개발되었다. 이모형은 비모수 회귀예측모형과 통합되었을 때 모수예측모형을 능가하는 것으로 입증되었다.

교통사고 처리(지속)시간 예측모형 개발과 관련하여 국내연구는 국외와 비교하였을 때 많이 진행되지 않은 것으로 나타났다. 이는 그동안 교통사고의 세부자료의 구득이 쉽지 않았기 때문으로 판단된다. 국내의 대부분의 연구에서 교통사고 처리시간 예측 모형은 여러 가지 변수들을 통해 회귀모형을 이용한 분석이 대부분 이였고, 분석에 사용된 샘플 수는 매우 적어 신뢰도 측면에서 우수성을 입증하기 어렵다. 또한, 연구의 공간적인 범위가 일부 노선으로 한정되어 있어 지역별, 노선별, 기하구조 특성 및 교통특성을 고려한 분석의 한계가 있다. 또한, 교통사고 지속시간은 대응시간과 처리시간으로 구분되는데, 대부분의 연구에서 이에 대한 명확한 구분이 없이 연구를 진행하였다. 마지막으로, 국외 논문의 교통사고 지속시간 예측결과 정확도가 높지 않아 현장에 적용하기에는 한계를 보였다. 이는 교통사고 지속시간에 미치는 요인이 다양하지만, 이러한 요인들을 충분히 반영하지 못한 결과라고 판단된다.

본 연구에서는 교통사고 지속시간 중 대응시간을 제외한 교통사고처리사간에 초점을 맞추어 사고이력자료를 바탕으로 교통사고 처리시간에 영향을 미치는 요인을 분석하고, KNN (K-Nearest Neighbor) 알고리즘을 이용한 교통사고 처리시간 예측모형을 개발하고자 한다.

3. KNN 알고리즘

데이터 기반의 교통상황 예측방법은 과거의 교통데이

터와 실시간의 교통데이터를 이용하여, 실시간의 교통데이터와 가장 유사한 패턴을 가진 과거자료를 검색하여 교통상황을 예측하는 방법이다. 현재 한국도로공사에서는 KNN(K-Nearest Neighbor), ANN(Artificial Neural Network) 방법들을 이용하여 통행시간을 예측하는데 사용되고 있다.

Devijver(1982)는 KNN기법은 풍부한 데이터가 이용 가능할 때 모수기법과 유사한 또는 능가하는 결과를 낳을 수 있다고 주장하였다. Yakowiz(1987)는 시계열 자료의 예측문제를 해결하기 위하여 KNN모형을 이용하였고, 예측치는 최소평균제곱오차(Minimum Mean Squared Error, MMSE)로 수렴하고, 수렴율(Rate of Convergence)은 최적임을 주장하였다.

KNN 알고리즘은 비모수 기반의 예측모형으로 과거 이력데이터로부터 현행 상태백터(Current state vector)와 유사한 K개의 과거 이력자료(Input state vector) 즉, K개의 최인접한 이웃(Nearest Neighbor)으로 구성되는 군집으로 독립변수의 집합을 구축하고, 구성된 군집을 구성하는 각각의 이웃에 해당하는 과거의 장래 상태인 출력 상태벡터(Output state vector)를 종속변수로 이용하여 장래 상태를 예측하는 기법이다.

KNN 알고리즘의 장점은 데이터에 대한 특별한 가정이 필요 없고 간단하다. 즉, 모수적 접근법들에서의 엄격한 가정인 독립변수간의 엄격한 통계적 독립성(Independence)과 각 독립변수 분포(Distribution)의 정규성을 가정하지 않는다. 회귀식이나 알고리즘 기반의 대부분의 모형들은 이론적으로 복잡하여 현장에서 적용 시 이에 대한 깊은 지식이 없으면 잘못 이해되거나 잘못 적용될 수 있다. 반면, KNN 기법은 간단하면서 이용자의 깊은 지식을 요구하지 않기 때문에 현장에서 용이하게 이용이 가능한 장점이 있다.

또한, KNN 모형은 다른 모형들과 달리 사용에 있어서 매우 탄력적이다. 기존의 모형들은 어떠한 상황이 변화되면 기존의 수학적 모형의 경우 새로이 추가되는 입력값과 변수를 수정해야하는 번거로움이 있지만, KNN 기법은 입·출력값과 모형의 재구성이 매우 용이한 장점이 있다. 반면에 많은 데이터가 필요하며 탐색시간이 많이 소요될 수 도 있는 단점 있다고 알려져 있다. 하지만, 탐색소요시간과 관련해서는 첨단 컴퓨터 연산기능과 DB 검색기술의 발달로 검색에 소요되는 모형의 연산시간은 더 이상의 단점이 될 수 없다.

분석방법론

고속도로에서 사고발생 시 사고처리시간 예측을 위하여 본 연구에서는 사고등급별로 사고처리시간을 예측하고자 한다. 사고등급별 현황분석결과 평균사고처리시간과 분포의 형태가 큰 차이를 보이므로 유사자료를 추출 시 데이터를 분류하여 작업을 수행하는 것이 필요하다 또한 사고구간의 기하구조, 교통소통상황을 고려하기 위하여 차로수, 교통량, 시간대를 기준으로 하여 1차적으로 그룹을 분류하였고, 분류된 자료를 대상으로 현재 발상한 사고의 조건과 가장 유사한 과거의 사고자료를 검색하여 사고처리시간을 예측할 수 있는 알고리즘을 개발하고자 한다.

사고처리시간예측 방법론의 경우 노선별 적용성과 해당 요인에 대한 추가 및 제거 등을 고려할 때 비모수 기반의 KNN모형이 분석방법론이 적합하다고 판단된다. KNN모형의 구축을 위해서는 현재의 사고자료와 과거의 사고자료의 유사성을 판단하기 위해서는 상태간 거리산정이 필요하다. 또한 사고처리시간의 경우 개별요인들이 미치는 영향의 크기가 다르고, 범주형 변수의 형태로 구축된 자료들이 존재하므로 이를 해결하기 위하여 가중치를 산정하였다. 즉, 세부요인별 가중치를 이용하여 상태간 거리를 산정하게 된다. 또한 선정된 유사자료를 통해 장래상태(사고처리시간)를 예측할 수 있는 모형의 개발이 필요하다. 마지막으로 KNN모형은 데이터 기반의 모형으로 검증에 사용되는 데이터에 따라 예측력의 차이를 보일 수 있으므로, 2014년 6월, 7월 각각에 대하여 비교 분석하였다. 이와 관련된 세부적인 내용은 Figure 2와 같다.

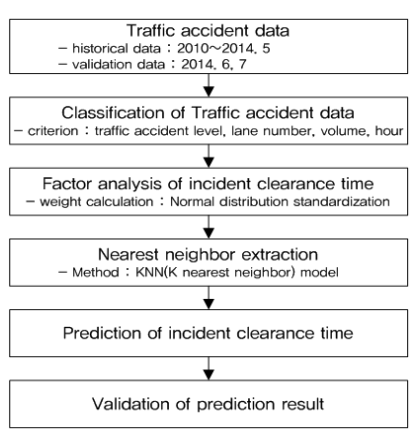

Figure 2. Research procedure

1) 상태간 거리산정

KNN모형은 최인접(Nearest) K개의 이웃(Neighbor)으로 구성되는 군집(Neighbor-hood)을 구축하기 위해서는 현행 상태벡터를 이용하여 과거 상태벡터를 탐색하는 과정을 수행하게 되며, 이때 두 상태간의 유사성을 판단하기 위한 방법으로 상태거리(Distance metric)가 이용된다. KNN과정에서 상태 간 유사성을 결정하는 Distance 방법으로 L_m거리를 이용하게 되며, $m=1$일 경우 맨하튼(Manhattan) distance, $m=2$일 경우 유클리디안(Euclidean) distance, $m=3$일 경우 Max distance 방법들이 사용되고 있다. 본 연구에서는 L_m 거리 중 가장 널리 사용되는 유클리디안 거리를 이용하였다.

$$L_m = [\sum_{i=1}^{k}(|x_i - y_i|)^m]^{1/m} \quad (1)$$

$$Euclidean = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (2)$$

$$U_d^i = \left[\sum_{a=1}^{11}\sum_{b=1}^{m}|x_a^b - y_a^b|^2\right]^{1/2} \quad (3)$$

여기서,

U_d^i : 상태간거리

x_a^b : 현행벡터(현재 사고자료)의 가중치

y_a^b : 입력벡터(과거 사고자료)의 가중치

a : 사고요인(ex: 작업장구분)

b : 사고요인별 세부항목(예 : 작업구간, 비작업구간)

2) 가중치 산정

본 연구에서는 교통사고 처리시간에 영향을 미칠 것으로 판단되는 요인으로 총 9가지를 검토하였다. 9가지 요인들의 세부항목들에 따른 현황 분석결과 요인별 각 세부항목별로 평균사고처리시간은 차이를 보이는 것으로 분석되었다. 일반적으로 교통사고 이력자료를 구성하는 범주형 변수는 유클리드안 거리가 유효하지 않다. Y. Qi(2004)는 교통사고 이벤트 데이트를 구성하는 범주형 변수는 유클리드 거리가 유효하지 않으므로, 이러한 문제를 해결하기 위해서는 범주형 데이터를 사용할 수 있도록 평균과 표준편차를 이용한 거리매트릭스를 이용하였고, Wang(2015)는 명목척도 분석에 적용되는 overlap metric을 적용하였다. 본 연구에서는 범주형 데이터를 사용할 수 있도록 세부요인별로 가중치의 개념을 도입하였다. 선택된 9가지 요인을 이용하여 현행상태벡터와 유사한 자료를 찾기 위해서는 하나의 상태간 거리(Distance metric)의 산정이 필요하다. 그러나 각 요인의 세부항목별로 사고처리시간의 크기가 다르므로, 요인별로 사고처리시간에 미치는 영향의 크기도 다를 것이므로 이를 분석하기 위해서는 해당 요인의 세부항목별 가중치 산정이 필요하다.

본 연구에서 적용한 가중치의 산정방법은 평균과 표준편차를 이용한 정규분포의 표준화 방법을 사용하였다. 표준화된 값은 확률분포의 면적값 즉, 누적분포값으로 0부터 1사이의 값을 갖는다. 정규분포의 표준화 식은 다음과 같다.

$$Z = \frac{X-\mu}{\frac{\sigma}{\sqrt{n}}} \quad (4)$$

여기서,

σ : 주어진 분포의 표준편차

μ : 주어진 분포의 평균

X : 측정 값

3) 장래상태(사고처리시간) 예측모형

현재의 사고 자료와 유사한 과거의 사고 자료가 추출되면 장래의 상태는 예측모형을 통하여 평가된다. 예측모형은 가장 널리 이용되는 산술평균방법과 현행상태와 입력상태 간 상태간 거리의 역수로 가중 평균한 방법으로 구분할 수 있다. 산술평균은 상태간의 유사성을 고려할 수 없는 단점이 있으므로 본 연구에서는 상태간 거리의 역수를 이용한 가중평균 방법을 적용하였다. 기존의 연구에서 산술평균에 비하여 우수한 예측결과를 보였다(Smith et al., 2002; Chang et al., 2010).

상태간 거리의 역수로 가중 평균한 방법론을 이용하여 사고처리시간을 예측하였으며, 식(5)와 같다. $\hat{t}$는 예측된 사고처리시간이며, d는 상태간 거리를 나타낸다.

$$\hat{t} = \frac{\sum_{i=1}^{k} t_i \times d_i^{-1}}{\sum_{i=1}^{k} d_i^{-1}} \quad (5)$$

개발모형 적용 및 평가

1. 분석데이터

본 연구에서는 고속도로 전국노선을 대상으로 하였고, 시간적 범위는 2008-2014년 7월까지 총 7년치(60,473)의 교통사고이력자료를 분석데이터로 사용하였다. 분석데이터 중 2008-2014년 5월의 자료는 과거이력자료로

Table 4. Incident clearance time by factors

Factor	Detail	Clearance time(avg)	Std
Number of casualties	0	37.6	34.5
	1	46.8	37.8
	2	51.5	40.8
	3	50.5	36.4
	4	54.9	46.0
	5-9	66.3	57.7
	10+	87.2	42.9
Accident type	veh-facility	36.2	32.4
	etc	43.3	39.0
Number of accident cars	1	37.3	35.2
	2	42.2	34.5
	3	46.7	36.6
	4	47.0	36.8
	5	49.2	31.1
	6	53.4	33.8
	7	59.8	39.8
	8	66.4	66.6
	9	63.6	31.8
	10+	78.8	62.1
Works	non-work	38.8	35.2
	work	45.4	39.3
Heavy vehicle	non-include	34.7	29.8
	include	48.6	43.9
Vehicle damage (partial+full destruction)	0	36.8	33.3
	1	49.6	40.4
	2	70.6	64.6
	3+	103.6	103.3
Road occupying vehicles	0	38.0	35.5
	1	38.9	33.6
	2	45.4	36.8
	3	47.1	37.1
	4	48.8	47.1
	5-9	51.3	34.1
	10+	96.2	57.4
Vehicle state (overturn)	0	37.4	34.2
	1+	46.5	39.3
Vehicle state (fire+jackknife)	0	38.7	34.9
	1	56.4	47.5
	2+	123.3	115.9

의 형태로 구축하였고, 이를 바탕으로 2014년 6, 7월의 각각의 자료를 대상으로 교통사고처리시간 예측모형의 검증에 사용하였다. 분석에 사용된 사고 자료들은 고속도로 상에서 교통사고 발생 시 현황파악을 목적으로 입력된 자료이다. 자료의 정보는 노선명, 사고시간, 발생지점, 차로, 요일, 날씨, 사고유형, 사고차량수, 평면선형, 종단선형, 원인자/피해자 차종, 사고처리 시간 등 개별 사고에 대한 세부 정보로 구성되어 있다.

현재 고속도로 교통사고 현장에서 수집되는 사고자료는 정보의 정확성과 신뢰도에 문제가 발생하고 있다. 현재 고속도로 사고현장에서는 조사원이 사고조사서와 수첩에 조사항목을 수기로 작성하고 있다. 이 과정에서 생기는 문제로는 현장에서 기록되지 않은 조사항목이 존재하고, 이 조사항목을 모든 사고처리 후 조사원의 기억에 의존하여 작성되어 사고자료의 정확성이 떨어지는 문제가 발생한다. 이러한 이유로 일부자료는 입력과정에서의 오류로 인해 완전한 형식을 갖추지 못하거나, 오기입이 발생할 가능성이 있으므로 이러한 자료는 분석에서 제거하였다. 또한, 사고처리시간의 분포범위는 0-596분으로 다양하게 나타났다. 교통정체에 큰 영향을 주지 않는 0분의 자료는 분석에서 제외하였다. 분석에 사용된 전국고속도로의 연도별 사고자료 현황은 Table 4와 같으며, 연간 사고건수는 점차 감소추세에 있는 것으로 나타났다.

2. 요인별 현황분석

교통사고 처리시간에 영향을 주는 요인들은 사상자수, 사고유형, 사고차량수, 도로요인, 교통요인, 차량피해정도, 본선도로 점유 차로수 등 다양한 원인에 의하여 영향을 받게 될 것이다. 각 요인에 따른 세부적인 항목의 분류는 Table 5와 같다. 각 요인별 수준에 따라 평균처리시간은 차이를 보이고 있으며, 표준편차의 경우 동일한 조건상에서도 매우 큰 것으로 나타났다. 이것으로 보아 교통사고 처리시간은 하나의 요인에 의해서 설명하기 어렵고, 복합적인 요인에 의해서 영향을 받는다고 판단할 수 있을 것이다.

Table 5. Accident data

Index	2008	2009	2010	2011	2012	2013	-2014. 7	Total
Number of incidents	9,805	10,116	10,042	9,223	8,902	8,061	4,324	60,473
Variation	-	3%	-1%	-8%	-3%	-9%	-	-

3. 이상치 제거

고속도로 교통사고 현장에서 수집되는 사고 자료는 정보의 정확성과 신뢰도문제 발생의 가능성에 대해서 앞서 제기하였다. 따라서 이러한 자료들은 제거될 필요가 있다. 본 연구에서는 개별 사고이력자료의 세부요인의 가중치를 바탕으로 사고처리시간을 예측할 것이다. 따라서 개별요인의 가중치 합과 사고처리시간 간의 상관성은 모형의 정확도를 결정하는 중요한 요인이다. 그러므로 가중치의 합과 사고처리시간 간의 관계를 이용하여 이상치를 제거하는 작업을 수행하였다. 즉, 가중치의 합이 큰 경우 사고처리시간은 커지는 경향을 보이고, 가중치의 합이 작은 경우 사고처리시간은 대체적으로 작을 것이라는 것은 앞서 현황분석 부분을 통해 확인할 수 있었다. 따라서 이러한 가정에 근거하여 사고등급별로 구분하여 회귀분석 시 이상치 진단에 사용되는 표준잔차를 이용하여 이상치를 제거해 주었다. 하지만 잔차의 크기가 상대적으로 크다고 해서 무조건 이례적인 사례로 간주하여 제거할 경우 표본의 크기가 줄어들게 되어 정보를 손실하게 될 우려도 있으므로 어떤 기준에 비추어서 이상치인가의 여부를 판정해야한다. 일반적으로 표준잔차(ZRESID)가 ±3보다 클 경우 일단 이상치라고 의심할 수 있다. 또한 표준잔차가 ±2보다 클 경우 측정치와 예측치가 5% 유의수준에서 다르다고 볼 수 있다. 본 연구에서는 총 60,473건의 사고자료 중 표준잔차가 ±2 보다 큰 경우인 2320건(약3.8%)을 제거하였다.

4. 이력자료 추출

교통사고처리시간에 큰 영향을 미치는 요인은 사고등급, 교통량, 차로수, 사고발생지점 등이 있다. 사고등급은 앞에서 살펴본바와 같이 사고처리시간에 가장 큰 영향을 미치므로 우선적으로 고려할 필요가 있다. 사고등급의 경우 A등급의 사고건수는 총 57건으로 매우 적으므로 데이터 기반의 모형에서는 예측의 한계를 보이므로 B등급에 포함시켰다. 또한, 교통량의 경우도 교통량수준이 낮을 경우 교통사고가 발생하여도 차로가 완전히 차단되지 않는 한 차로변경을 시도하기가 용이하여 교통정체에 미치는 영향이 낮고, 또한 사고를 처리하는데 있어서도 영향이 낮을 것이다. 그밖에 해당구간의 기하구조(차로수)의 영향도 클 것으로 판단된다.

현재 한국도로공사에서 수집하고 있는 사고 자료에는 해당구간의 교통량수준과 기하구조가 명시되어 있지 않은 한계가 있다. 이러한 한계를 극복하기 위하여 본 연구에서는 2013년 기준의 고속도로 전구간에 대한 차로수와 AADT 자료를 DB로 구축하였고, 사고자료 중 노선과 이정정보를 활용하여 매칭하는 작업을 수행하였다. 즉, 노선이 다르더라도 동일한 차로수, 유사한 교통량 수준을 갖는 구간은 사고처리시간에 미치는 영향의 정도가 유사할 것이라는 가정 하에 자료를 구축하였다. 또한, 유사한 교통량 수준이더라도 시간대에 따라 교통량 수준이 다를 수 있으므로 시간대를 분류하는 작업을 수행하였다. 시간대 분류는 영동고속도로 양방향의 5년치 VDS자료를 시간대별로 평균 교통량과 속도를 분석한 결과 Figure 3와 같이 통행의 특성이 4가지의 시간대로 분류되는 것으로 나타났다. 본 연구에서도 시간대의 경우 4가지(① 00-06시, ② 06-13시, ③ 13-19시, ④ 19-24시)로 분류하여 분석하였다. 단, 실시간 교통량 대신에 AADT 자료를 활용하였기 때문에 명절 및 휴일 등의 교통량이 평소에 비하여 많은 특송 기간에 대해서는 본 연구에서는 고려하지 못하였다.

교통량의 경우는 분포가 매우 다양하므로 사고발생일 교통량을 기준으로 일정한 범위 내에 포함되는 자료를 추출하도록 설계하였다. 교통량을 기준으로 자료를 추출할 경우 사고등급이 높은 경우 사고건수가 적어 실제로 추출 가능한 자료가 제약되므로 모형의 정확도에 영향을 미칠 수 있으며, 사고등급이 낮은 경우에는 추출 가능한 사고자료가 많으므로 최대한 교통량이 유사한 자료를 추출하는 것이 사고처리시간 예측 모형의 정확도를 높이는데 영향을 줄 것이다. 즉, 사고등급에 따라 추출 가능한 이력자료의 양이 다르므로 교통량 자료 검색을 위한 범위는 오차(MAPE)가 가장 낮은 최적값을 찾아 적용하였다.

Step 1) 사고등급 검색(B(A포함), C, D등급)

Step 2) 차로수 검색 (2, 3, 4, 5차로 이상)

Step 3) 유사한 교통량(AADT/차로수) 검색
- 범위 설정
  (예 : 사고발생시 교통량 × (100% ± 30%)
- 사고등급별로 최적의 범위 설정

Step 4) 동일한 시간대자료 검색

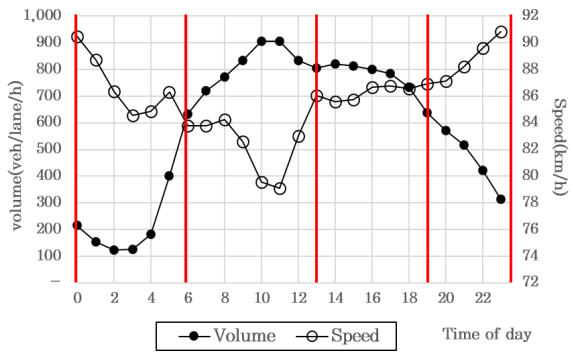

Figure 3. Traffic status by hour

- 통행특성(속도, 교통량)이 유사한 시간대로 분류
- ① 00-06시, ② 6-13시, ③ 13-19시, ④ 19-24시

5. 개발모형의 평가지표

NPR 모형은 대용량의 이력자료를 바탕으로 현행상태 벡터와 유사한 과거자료를 검색하여 장래의 상태를 예측하는 방법이다. 본 연구에서는 2008-2014년 5월까지의 교통사고 이력자료를 구축하였고, 개발모형의 평가를 위하여 2014년 6, 7월의 사고자료가 이용되었다.

개발된 모형의 평가를 위한 지표로는 ① 평균절대퍼센트오차(Mean Absolute Percentage Error(%), MAPE), ② 평균절대오차(Mean Absolute Error(분), MAE)를 이용하여 검증을 수행하였다.

$$MAPE(\%) = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{y}_i - y_i|}{y_i} \times 100,\ y_i > 0 \quad (6)$$

$$MAE(\text{분}) = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \quad (7)$$

여기서,

N : 분석 자료의 개수
y_i : i번째 관측값
$\hat{y}_i$: i번째 예측값

6. 사고처리시간 예측결과

KNN 알고리즘의 예측 정확도는 군집을 구성하는 최인접 이웃의 개수(K)에 의하여 결정된다. 따라서 최적(Best) 또는 적정(Optimal)의 K값을 설정해야 한다. 본 연구에서는 2014년 6, 7월의 검증자료를 바탕으로 실제 사고처리시간과 예측된 통행시간과의 예측오차를 최소화 할 수 있는 K값을 분석하였다.

사고등급별 평균사고처리시간 분석결과 등급별 차이가 큰 것으로 분석되어 사고처리시간 예측 시 등급을 분류하여 적용할 필요가 있다. 따라서 본 연구에서는 사고등급(B, C, D) 및 과거이력자료 추출기준(차로수, 교통량, 시간대)에 따라 각각의 Optimal K값을 산정하였다.

이력자료 추출시 적용한 교통량 범위는 사고등급별로 차이를 보였으나, 일반적인 경향성은 보이는 것으로 나타났다. 사고등급 B의 경우는 사고건수가 많지 않아 자료수집을 위한 교통량의 범위는 다른 등급의 사고와 비교하였을 때 큰 것으로 나타났으며, 사고등급 C, D는 B에 비하여 충분한 사고자료가 확보되어 교통상황을 좀 더 정확하게 반영할수록 예측의 정확도가 높은 것으로 나타났다. Optimal K값은 사고등급별로 경향성을 보이지는 않았으나 평균적으로 약 10개의 이웃을 추출할 경우에 모형의 정확도가 높은 것으로 나타났다.

오차율(MAPE, MAE) 분석 결과는 Table 6과 같다. 6월의 경우 B등급은 오차율 33.7%, C등급은 29.8%, D등급은 35.0%로 평균 34.2%의 오차를 보이는 것으로 나타났고, MAE는 평균 12.1분으로 나타났다. 7월의 경우는 B등급은 오차율 24.2%, C등급은 32.4%, D등급은 31.1%로 평균 30.5%의 오차를 보였고, MAE는 평균 11.4분으로 분석되었다.

Figure 4, 5는 2014년 6, 7월의 사고자료를 바탕으로 관측값과 예측값을 나타낸 그림이다. 분석결과 사고처리시간이 긴 경우에는 평균적으로 과소 예측된 결과를 보였고, 특히 60분 이상의 사고건수에 대해서는 그 정도가 더 큰 것으로 나타났다. 이는 과거이력자료가 60분 이상의 자료가 충분치 않아 유사자료로 추출이 되지 못한 것으로 판단된다.

Table 6. Prediction result of incident clearance time(June, 2014)

Month	Level	Optimal k	Vol Range	MAPE		MAE	
June	B	6	0.3	33.7	34.2	16.0	12.1
	C	13	0.1	29.8		11.1	
	D	10	0.1	35.0		12.1	
July	B	11	0.3	24.2	30.5	14.4	11.4
	C	9	0.1	32.4		14.4	
	D	10	0.1	31.1		11.4	

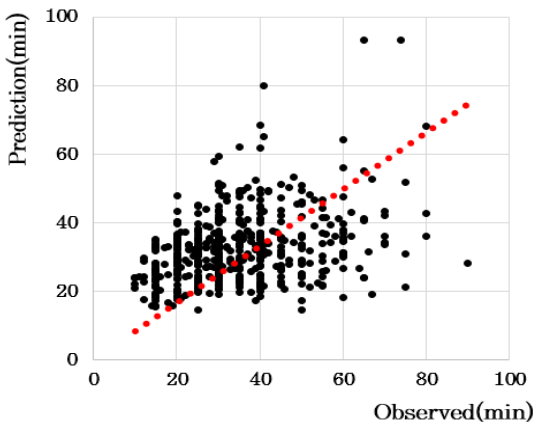

Figure 4. Prediction result of incident clearance time (June, 2014)

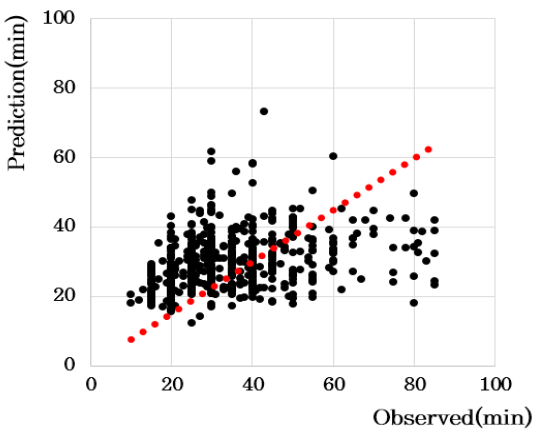

Figure 5. Prediction result of incident clearance time (July, 2014)

KNN모형은 데이터기반의 모형으로 검증에 사용되는 데이터의 형태에 따라 차이를 보였으나, 평균적인 검증력은 큰 차이를 보이지는 않은 것으로 나타났다.

결론

고속도로 구간의 교통 혼잡은 반복적 혼잡과 비반복적 혼잡으로 구분된다. 비반복 혼잡의 경우는 주로 돌발적인 상황에 의해 발생되며, 가장 큰 원인은 교통사고로 인한 도로의 용량감소 영향으로 혼잡이 발생하게 된다. 따라서 교통사고 시 사고처리시간을 정확히 예측하는 것은 돌발상황 관리에서 매우 중요하다. 사고가 발생하였을 때 교통사고 처리시간 예측은 교통영향권 내의 대기행렬, 지체, 정상류 회복시간 등을 산정하기 위한 필수적인 자료로서, 돌발상황에 따른 효율적인 고속도로의 운영관리 및 비반복정체의 효과를 최소화하기 위하여 필요하다.

본 연구에서는 전국 고속도로 노선의 2008-2014년 총 7년치(60,473건)의 사고자료를 이용하여 분석을 수행하였다. 교통사고 처리시간의 경우 다양한 원인에 의해 복합적으로 작용하므로 교통사고 처리시간 예측 시 비모수모형이 유용하게 사용되어 질 수 있다. 본 연구에서는 사고등급별 예측모형을 개발을 위하여 KNN 알고리즘을 이용하였다.

본 연구에서 개발된 모형의 예측결과는 기존의 연구들의 결과에 비해 예측의 오차를 나타내는 MAPE(%) 값이 낮은 것으로 나타나 모형의 우수성을 입증할 수 있다고 판단되며, 고속도로의 돌발상황 발생 시 효율적인 고속도로의 운영관리와 도로 이용자들에게 신속하고, 정확한 교통정보를 제공하는데 있어 기여할 수 있고, 기존의 모형들이 갖고 있던 한계를 개선 및 보완할 수 있을 것으로 판단된다. 하지만 본 연구에서 개발된 모형이 현장에 적용하기에는 아직은 오차값이 높아 한계가 있으므로 정확도 향상을 위해서는 향후 추가적인 연구가 필요할 것이다.

첫째, 가중치산정 방법론에 대한 추가적인 고찰이 필요하다. 본 연구에서는 정규분포의 표준화를 이용하여 가중치를 산정하였는데, 이는 개별 요인들에 대한 변동성을 제대로 반영하기 어려운 단점이 있다. 둘째, 교통사고자료는 조사원의 기억에 의존하여 사고발생 후 작성되므로 이상치가 발생할 가능성이 높다. 본 연구에서는 이상치제거를 위하여 표준잔차를 적용하였지만, 추가적인 이상치제거 방법론에 대한 검토가 필요하다. 셋째, 본 연구에서는 교통사고 시 교통소통상황을 반영하기 위하여 AADT값을 적용하였다. 하지만 AADT 값은 10월 셋째 주 수요일의 대푯값으로 사고발생시의 교통상황을 제대로 반영하지 못하므로 실제 사고발생시의 교통량자료를 반영할 필요가 있다. 넷째, 사고처리시간 예측결과 사고처리시간이 큰 경우에 대부분 과소 예측되는 결과를 보이는 것으로 나타났다. 이러한 문제를 해결할 수 있는 방법론의 개발이 필요할 것으로 판단된다. 마지막으로 기존에서 개발된 사고처리시간 예측방법론과 본 연구에서 개발된 방법론을 융합한 방법론의 개발이 필요할 것이다. 이러한 연구들이 향후에 연구되어진다면 돌발상황 시 정확도 높은 사고처리시간 예측과 통행시간 예측분야에 기여할 수 있을 것으로 판단된다.

ACKNOWLEDGEMENT

This work was supported by the BK 21 Plus program(5281-20130100) of the National Research Foundation of Korea.

REFERENCES

Chang H. L., Chang T. P. (2013), Prediction of Freeway Incident Duration based on Classification Tree Analysis, Eastern Asia Society for Transportation Studies, 9, 1964-1977.

Chang H., Park D., Lee S., Lee H. Baek S. (2010), Dynamic Multi-interval Bus Travel Time Prediction Using Bus Transit Data, Transportmetrica, 6(1), 19-36.

Chung Y. S., Song S. K., Choi K. C. (2007), A Prediction Model on Freeway Accident Duration Using AFT Survival Analysis, J. Korean Soc. Transp., 25(5), Korean Society of Transportation, 135-148.

Devijver P. (1982), Statistical Pattern Recognition, Applications of Pattern Recognition, K. S. Fu, ed., CRC Press, Boca Raton, Fla., 15-36.

Friedrich M., Lohmiller J. (2012), Factors Influencing the Travel Time Reliability of Motorway Section, Proceedings of the 6th International Symposium Networks for Mobility, Stuttgart.

Gaetano V., Maria L., Domenico C. (2010), A Comparative Study of Models for the Incident Duration Prediction, Eur. Transp. Res. Rev. 2, 103-111.

Ha O. K., Park D. J., Won J. M., Jung C. H. (2010), The prediction Models for Clearance Times for the unexpected Incidences According to Traffic Accident Classification in Highway, The Journal of The Korea Institute of Intelligent Transport Systems, 9(1), 101-110.

Lee K. Y., Seo I. K., Park M. S., Chang M. S. (2012), A Study on the Influencing Factors for Incident Duration Time by Expressway Accident, International Journal of Highway Engineering, 14(1), 85-94.

Qi Y., Smith B. L. (2004), Identifying Nearest Neighbors in a Large-Scale Incident Data Archive, Journal of the Transportation Research Board, 1879, 89-98.

Qing H., Yiannis K., Klayut J, Laura W. (2011), A Hybrid Tree and Quantile Regression Method for Incident Duration Prediction, TRB 91th Annual Meeting, Washington, D.C.

Shin C. H., Kim J. H. (2002), Development of Freeway Incident Duration Prediction Models, J. Korean Soc. Transp., 20(3), Korean Society of Transportation, 17-30.

Smith B., Williams B., Oswald R. (2002), Comparison of Parametric and Nonparametric Models for Traffic Flow Forecasting, Transportation Research Part C, 10, 303-321.

Wang S., Li R., Guo M. (2015), Application of Nonparametric Regression in Predicting Traffic Incident Duration, TRANSPORT, in press.

Yakowitz S. (1987), Nearest-neighbor Methods for Time-series Analysis, Journal of Time Series Analysis, 8(2), 235-247.

♧ 주 작 성 자 : 이숭봉
♧ 교 신 저 자 : 이영인
♧ 논문투고일 : 2015. 4. 3
♧ 논문심사일 : 2015. 5. 7 (1차)
2015. 8. 18 (2차)
♧ 심사판정일 : 2015. 8. 18
♧ 반론접수기한 : 2016. 2. 29
♧ 3인 익명 심사필
♧ 1인 abstract 교정필