KIPS Tr. Comp. and Comm. Sys.
Vol.6, No.4 pp.189~196 pISSN: 2287-5891

https://doi.org/10.3745/KTCCS.2017.6.4.189

Arrival Time Estimation for Bus Information System Using Hidden Markov Model

Park Chul Young[†] · Kim Hong Geun[††] · Shin Chang Sun[†††] · Cho Yong Yun[†††] · Park Jang Woo[††††]

ABSTRACT

BIS(Bus Information System) provides the different information related to buses including predictions of arriving times at stations. BIS have been deployed almost all cities in our country and played active roles to improve the convenience of public transportation systems. Moving average filters, Kalman filter and regression models have been representative in forecasting the arriving times of buses in current BIS. The accuracy in prediction of arriving times depends largely on the forecasting algorithms and traffic conditions considered when forecasting in BIS. In present BIS, the simple prediction algorithms are used only considering the passage times and distances between stations. The forecasting of arrivals, however, have been influenced by the traffic conditions such as traffic signals, traffic accidents and pedestrians ets., and missing data. To improve the accuracy of bus arriving estimates, there are big troubles in building models including the above problems. Hidden Markov Models have been effective algorithms considering various restrictions above. So, we have built the HMM forecasting models for bus arriving times in the current BIS. When building models, the data collected from Sunchean City at 2015 have been utilized. There are about 2298 stations and 217 routes in Suncheon city. The models are developed differently week days and weekend. And then the models are conformed with the data from different districts and times. We find that our HMM models can provide more accurate forecasting than other existing methods like moving average filters, Kalmam filters, or regression models. In this paper, we propose Hidden Markov Model to obtain more precise and accurate model better than Moving Average Filter, Kalman Filter and regression model. With the help of Hidden Markov Model, two different sections were used to find the pattern and verified using Bootstrap process.

Keywords : Bus Information System, Hiddel Markov Model, Traffic Flow, Travel Speed Estimation

은닉 마르코프 모델을 이용한 버스 정보 시스템의 도착 시간 예측

박 철 영[†] · 김 홍 근[††] · 신 창 선[†††] · 조 용 윤[†††] · 박 장 우[††††]

요 약

버스정보시스템은 버스도착시간 예측과 같은 버스와 관련한 여러 정보를 제공한다. BIS는 우리나라 거의 모든 도시에 구축되어 있고 대중교통의 편의성 개선에 능동적인 역할을 하고 있다. 현재 BIS 시스템에서 버스 도착 예정시간을 예측하기 위하여 사용되는 대표적인 방법으로는 이동평균필터, Kalman Filter, 회귀 모형 등이 있다. 버스 도착 시간 예측의 정확성은 BIS 시스템에서 고려하고 있는 교통 상황이나 예측 알고리즘에 따라 차이가 크다. 현재 BIS에서 사용하는 예측 기법은 구간 통과 시간과 거리만을 이용한다. 그러나 도착시간 예측은 교통흐름, 신호주기, 이상 상황, 데이터 결측 등에 큰 영향을 받는다. 버스 도착 시간 예측의 정확도를 높이기 위해서는 위의 문제를 고려하여 모델링해야 하는 어려움이 있다. 은닉 마르코프 모델은 이와 같은 다양한 상황을 효과적으로 모델링 할 수 있다. 따라서 버스 도착 시간 예측의 정확도를 높이기 위해 도착시간에 대한 HMM 예측 모델을 구축했다. 이 모델에서는 순천시의 2015년 한 해 동안 수집한 데이터가 이용되었으며, 순천시에는 2298개의 정류장과 217개의 노선이 있다. 모델은 주중과 주말의 패턴을 다르게 적용하며, 다른 구간과 시간에 대해 모델이 적용된다. 본 논문에서는 버스정보시스템에 은닉 마르코프 모델 적용방법과 검증을 통해 버스정보시스템에서 사용 중인 이동평균필터, Kalman Filter, 회귀 모형을 사용한 예측 방법 보다 정밀한 정확도를 얻는 방법을 제안한다.

키워드 : 버스정보시스템, 은닉 마르코프 모델, 교통흐름, 운행 속도 예측

1. 서 론

버스정보시스템(Bus Information System, BIS)은 버스에 설치되어 운영되는 차내 장치(On Board Equipment, OBE)를 통해 현재 버스의 위치를 실시간으로 관제 센터로 전송되어 계산된 결과로 시민들에게 버스 도착 예정 시간 정보를 제공하고, 대중교통 이용의 편의성을 높이는 것이 주목적이다. 버스정보시스템은 사용되는 예측 알고리즘이나 교통 여건에 따라 정확도가 가변적으로 작용하며, 이는 예정시간 정보의 정확도에 큰 영향을 미친다[1].

버스 도착 예정시간을 예측하기 위하여 사용되는 대표적인

※ 이 논문은 2016년 순천대학교 학술연구비로 연구되었음.
† 비 회 원 : 순천대학교 전기·전자·정보통신공학과 박사과정
†† 준 회 원 : 순천대학교 전기·전자·정보통신공학과 박사과정
††† 정 회 원 : 순천대학교 정보통신공학과 부교수
†††† 정 회 원 : 순천대학교 정보통신공학과 교수
Manuscript Received : December 13, 2016
Accepted : January 25, 2017
* Corresponding Author : Park Jang Woo(jwpakr@sunchon.ac.kr)

방법으로는 이동평균필터, Kalman Filter, 회귀 모형 등의 기법을 사용한다[2]. 이러한 예측 기법들은 구간 통과 시간과 구간 거리를 예측 모형에 사용하지만 도로의 교통흐름, 신호 주기, 이상 상황, 데이터 결측 등의 상황을 고려하지 않았다.

본 논문에서는 교통흐름 분석 및 예측을 위한 기법으로 은닉 마르코프 모델(Hidden Markov Model, HMM)을 사용한다. 버스정보시스템에서 관찰되는 데이터는 정류장에 도착한 시간과 구간 거리이며 이에 의존하여 교통흐름이나 신호 주기 등의 모델링 데이터를 구해야 하는 어려움이 있다. 은닉 마르코프 모델은 이러한 제약조건에서 효과적인 모델링 방법으로 적용될 수 있다.

본 논문은 출/퇴근시간대, 신호 주기 등에 따른 과거의 교통흐름 패턴을 은닉 마르코프 모델을 이용하여 생성하고 이를 통해 버스도착 예측 시간 정보의 정확도를 높이고자 하는데 목적이 있다.

본 논문에서 제안하는 방법은 순천시의 버스정보시스템에서 사용 중인 이동 평균을 이용한 버스 도착 시간 예측 방법보다 정밀한 정확도를 얻을 수 있다.

2. 수집데이터 전처리

버스정보시스템에 구축된 데이터는 표준 노드.링크 체계를 이용하며 노드는 교차로, 도로 시/종점, 속성 변화점, 도로시설물 등의 속성을 의미하며, 링크는 노드와 노드 사이를 연결하는 역할을 한다.

버스정보 시스템의 차내 장치에서 수집되는 데이터 집합은 Table 1과 같은 형태를 갖는다.

Table 1의 정보를 시간과 버스 ID 집합으로 정렬하고, 노드 구분 필드의 구분자를 이용하여 시작 노드와 종료 노드 사이의 시간을 계산하는 전처리 과정을 수행하여 Table 2와 같은 정보를 생성했다.

Table 1. RAW Data Event Schema for Bus Information System

Field	Type	Comment
COLLECT_DT	DATETIME	Data generation time
BUS_ID	VARCHAR (9)	Area code (3)+ bus type (1) + sequence (5)
OBE_ID	VARCHAR (9)	OBE ID
ROUTE_ID	VARCHAR (9)	Route ID
LOCATION_TY	CHAR (1)	NODE/STATION
EVENT_CD	CHAR (2)	Accident/Breakdown
LOCATION_ID	VARCHAR (10)	Location ID
LOCATION_SEQ	SMALLINT (6)	Sequence
SYSTEM_DT	DATETIME	System time
GPS_X	DECIMAL (11,5)	GPS X
GPS_Y	DECIMAL (11,5)	GPS Y
MAP_X	DECIMAL (11,5)	MAP X
MAP_Y	DECIMAL (11,5)	MAP Y
SERVICE_TM	INT (11)	Service time
SATELLITE_CNT	BIGINT (20)	Satellite count
ERROR_CD	CHAR (2)	Error code
EVENT_TY	CHAR (1)	Event type
EVENT_SEQ	TINYINT (4)	Event sequence
COMM_SEQ	INT (11)	Communication sequence

Table 2. Regenerated Event Schema Based on Table 1

Field	Type	Comment
COLLECT_DT	DATETIME	Data generation time
BUS_ID	VARCHAR (9)	Area code (3)+ bus type (1)+sequence (5)
ROUTE_ID	VARCHAR (9)	Route ID
EVENT_CD	CHAR (2)	Accident/Breakdown
LOCATION_SEQ	SMALLINT (6)	Sequence
SERVICE_TM	INT (11)	Service time
TRAVEL_TM	BIGINT (12)	Travel time
SECTION_NM	VARCHAR (60)	Section name
SECTION_LEN	DECIMAL (6,1)	Section length
SPD_KMH	DECIMAL (5,9)	Bus speed (km/h)

본 논문에서는 노드와 노드 사이의 시간차를 이용하여 버스 도착 시간 예측 정보를 생성하며 데이터의 샘플링 주기는 5분으로 설정하고, 데이터의 수집 시간을 버스의 첫 운행 시작인 06시 기준 분으로 환산했다. 버스가 통과하지 않은 누락된 시간대의 데이터가 존재할 경우 시계열 자료의 흐름에 영향이 없도록 이에 대한 보정을 먼저 수행한다. 누락된 속도 데이터는 이전 구간 통행 속도와 다음 구간 통행 속도를 이용하여 평균값으로 산출했다.

3. 은닉 마르코프 모델

마르코프 연쇄(Markov Chain)는 시간에 따른 시스템 상태의 변화를 나타낸다. 이러한 상태의 변화를 전이(Transition)라 한다. 데이터는 상태를 바꾸거나 같은 상태를 유지하게 되며 마르코프 성질(Markov Property)은 과거와 현재 상태가 주어졌을 때의 미래 상태 조건부 확률 분포는 과거 상태와는 독립적으로 현재 상태에 의해서만 결정된다.

1차 마르코프 체인은 다음의 Equation (1)과 같이 표현한다[3-5].

$$P(q_t = S_j | q_{t-1} = S_i, t_{t-2} = S_k, \ldots) = P(q_t = S_j | q_{t-1} = S_i) \quad (1)$$

마르코프 모델은 위의 가정으로 확률적 모델을 정의한 것으로, 전이 확률(Transition probabilities)은 마르코프 연쇄에서 중요한 과정이다.

여기서,

$$a_{ij} \geq 0, \sum_{j=1}^{N} a_{ij} = 1 \quad (2)$$

Equation (2)에서 a_{ij}는 i상태 에서 j상태로 바뀌는 전이 확률이며, 0과 1사이의 값을 갖는다. 2가지의 상태를 갖는

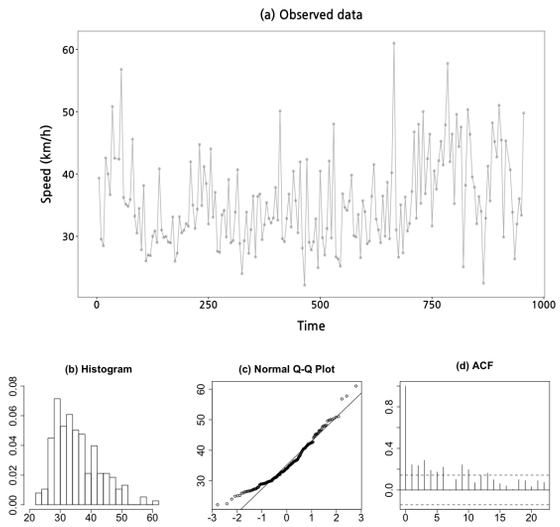

Fig. 1. (a) Observed Data, (b) Histogram, (c) Q-Q Plot and (d) ACF Fitted for the Bus Speed Series of "A" Section

마르코프 연쇄의 전이 상태 행렬은 다음과 같이 표현한다.

$$A = \{a_{ij}\}, A = \begin{pmatrix} a_{11} \cdots a_{1N} \\ \vdots \ \ddots \ \vdots \\ a_{N1} \ldots a_{NN} \end{pmatrix} \quad (3)$$

은닉 마르코프 모델(λ)은 다음과 같은 요소들로 정의된다.

$$\lambda = (A, B, \Pi) \quad (4)$$

- 상태들 간의 전이확률 행렬

$$A = \{a_{ij}\},\ a_{ij} = P(q_{t+1} = S_j | q_t = S_i), \quad 1 \le i, j \le N \quad (5)$$

- 관찰확률 행렬

$$B = \{b_{ik}\},\ b_{ik} = P(O_t = v_k | q_t = S_i), \quad 1 \le i \le N, 1 \le k \le M \quad (6)$$

- 상태들의 초기 상태 확률 벡터

$$\Pi = \{\pi_i\},\ \pi_i = P(q_1 = S_i), 1 \le i \le N \quad (7)$$

은닉 마르코프 모델을 이용하여 패턴을 찾기 위해서는 $\lambda = (A, B, \Pi)$와 $O = O_1, O_2, O_3, \ldots, O_{T-1}, O_T$의 주어진 관측열 사이에서 3가지의 문제를 해결해야 한다.

- 관측된 순서의 확률 $P = (O|\lambda)$을 시스템 부하 대비 효율을 고려하여 계산하는 문제
- 주어진 관측열로 부터 확률이 가장 높은 최적의 상태열 $Q = q_1, q_2, q_3, \ldots, q_{t-1}, q_t$을 찾는 문제
- 관측된 순서의 확률 $P = (O|\lambda)$을 최대로 하는 모델의 매개변수(parameter)를 결정하는 문제

Table 3. General Information on Analyzed Section

Section	A section	B section
Section distance (M)	593.2	263.0
Period	2015/06/16 06:00~21:00 (화)	
Number of lane	2	2
Limit speed	60km/h	60km/h
Number of intersection	1	1
Number of traffic signal	1	1
Number of route	69	62

위와 같은 문제에서 관측된 순서의 확률을 계산하는 방법으로는 전향 알고리즘(Forward algorithm)과 후향 알고리즘(Backward algorithm)을 사용하며, 주어진 관측열로 부터 최적의 상태열을 찾는 문제는 비터비 알고리즘(Viterbi algorithm)을 사용한다. 관측된 순서의 확률을 최대로 하는 매개변수를 결정하는 문제는 초기 모델과 관측열로 부터 구성된 모델 사이에 매개변수를 변경하며 매번 새로운 모델을 이용해 반복적으로 차이를 구하고 이를 통해 매개변수가 최대치를 갖는 모델을 찾는다.

4. 데이터 모델링 및 예측

사용된 자료는 순천시 버스정보 시스템에서 2015년 1월 1일 부터 2015년 12월 31일까지 수집된 데이터이다. 순천시의 버스노선은 대부분 시내권역 위주로 집중되어 있으며, 평균 운행 거리는 짧고 평균배차간격이 크다. 2015년 한해를 기준으로 1대당 평균 운행 거리는 29.59km이고 평균 운행 속도는 32.4 km/h이다. 데이터 분석을 위한 구간의 일반 현황은 다음과 같다.

Fig. 1의 구간 평균속도는 33.53 km/h이며 출근 시간대인 7시부터 9시까지의 속도 평균이 30.88 km/h로 교통 혼잡의 양상을 보인다. 반면 17시부터 19시까지의 속도 평균은 37.79 km/h로 구간의 전체 평균속도보다 높아 퇴근 시간대의 영향은 없다. 관측 데이터에서 이전데이터와 다음데이터 사이의 높낮이 차이를 보이고 있으며 이는 신호주기의 영향에 따른 것이다. Fig. 1의 속도 데이터의 분위수와 정규분포의 분위수를 산점도로 나타낸 분위수대조도(Quantile-Quantile Plot, Q-Q Plot)에서 보이는 바와 같이 관측 데이터는 정규분포를 따르고 있다고 보기 어렵다. 자기상관함수(Auto Corrrelation Function, ACF) 에서도 유의미한 모델을 찾을 수 없다.

버스정보시스템에서 수집된 데이터는 정류장에 도착한 시간과 구간 거리에 의존하여 교통흐름이나 신호 주기 등의 외부요인이 작용한 결과이기 때문이다. 추출한 데이터의 특성을 파악하는 것은 패턴을 분류함에 있어 매우 요소로 작용한다. 본 논문에서는 은닉 마르코프 모델을 이용하여 실측자료간의 전이 확률의 변동 양상 분석 및 버스 도착 예정시간을 예측하기 위해 가우시안 혼합 모델을 고려했다.

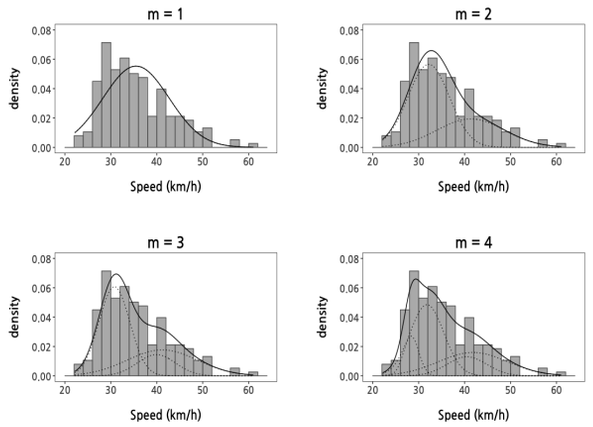

Fig. 2. Observed Speed Data on "A" Section : Histogram of Counts, Compared to Mixtures of One, Two, Three and Four Gaussian Distributions

A 구간의 속도 데이터는 Fig. 2와 같은 분포를 갖는다. 평균을 중심으로 양측에 동일한 분산을 가지는 가우시안분포로는 A 구간의 분포의 특성을 반영하기는 어렵다. 복수개의 가우시안 분포들의 합으로 구성된 가우시안 혼합 모델을 사용함으로서 데이터의 특성을 정밀하게 묘사한다. M은 사용된 확률밀도함수의 수이며, 사용된 모델의 수가 많을수록 정밀한 데이터를 추출할 수 있지만 사용되는 모델 수에 따라 시스템의 부하가 증가한다.

Table 4는 분석 구간에 적합한 모델을 선택하기 위해 AIC (Akaike Information Criterion), BIC(Bayesian Information Criterion) 그리고 우도를 계산한 결과이다.

정밀도가 높은 가우시안 혼합 모델을 선택하기 위해서는 AIC, BIC와 $-\log L$를 최소로 가진 모델을 선택하는 것이 좋은 방법이다.

그러나 모델의 수가 4개(M = 4)인 경우 $-\log L$ 값이 낮아지는 반면 AIC, BIC의 값이 높아짐을 알 수 있다. 모델의 수가 많아질수록 분포의 특성을 잘 반영하지만 이는 과대적합(Overfitting)의 문제로 분포의 오차를 더 많이 반영하는 결과를 보인다. M=2와 M=3의 $-\log L$ 차이가 다른 혼합 모델의 차이보다 많다. 이는 다른 혼합모델보다 M=3의 모델이 모수를 추정하는데 적합하다. 본 논문에서는 M=2 또는 M=3의 혼합 모델을 적용하여 구간의 속도를 예측한다.

Equation (3)의 표현에서 A 구간의 속도를 고려한 2가지와 3가지의 상태를 갖는 상태전이행렬(State transition matrix)은 다음의 Equation (8)과 같다.

Table 4. Gaussian Mixture Models Fitted for the Speed Data

Model	AIC	BIC	$-\log L$
M = 1	1286.921	1293.404	641.4603
M = 2	1247.395	1270.087	616.6975
M = 3	1242.667	1288.051	307.3334
M = 4	1255.955	1330.515	304.9775

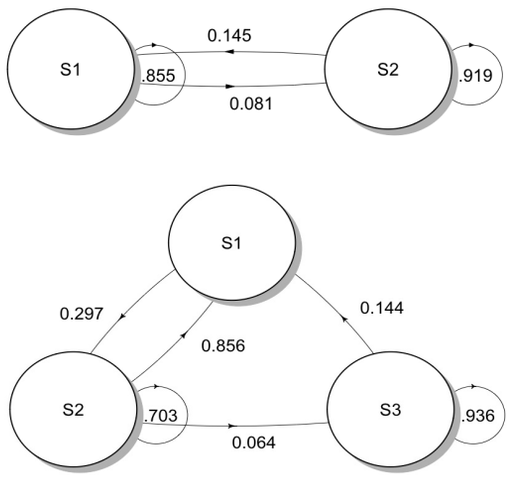

Fig. 3. State Diagram for 2-state and 3-state Model

$$A=\begin{pmatrix}0.855\ 0.145\\0.081\ 0.919\end{pmatrix}, A=\begin{pmatrix}0.000\ 0.856\ 0.144\\0.297\ 0.703\ 0.000\\0.000\ 0.064\ 0.936\end{pmatrix} \quad (8)$$

Equation (8)을 상태 전이 다이어그램으로 표현하면 다음 Fig. 3과 같다.

Fig. 3에서 구성된 2개의 상태를 갖는 전이 확률 모델의 경우를 보면 구간의 속도 변량이 크지 않아 각 상태에 머무를 확률이 크게 나타난다. 3개의 상태를 갖는 전이 확률 모델의 경우 S1에서 S3로 전이될 확률과 S3에서 S2로 전이될 확률은 없으며 S3상태에 머무를 확률과 S2에서 S1으로 전이될 확률이 크게 나타난다. Fig. 4는 상태 병합 과정을 통해 관측된 데이터와 함께 구성된 상태를 표현한 것이다.

본 논문에서 전처리 과정 중 데이터의 샘플링 주기는 5분으로 설정하고, 데이터의 수집 시간을 버스의 첫 운행 시작인 06시 기준 분으로 환산했다.

Fig. 4의 2개의 상태로 구성된 모델에서 S1은 41.338, S2는 32.182의 값을 가지며 3개의 상태로 구성된 모델에서 S1은 39.924, S2는 30.773, S3는 41.210의 값을 갖는다. A 구간은 교통흐름이 항상 혼잡한 구간이며 S1은 혼잡한 상태를 S2는 원활한 상태를 나타낸다. 3개의 상태로 구성된 모델에서 교통량이 많지 않은 새벽 06시와 07시 사이 19시와 21시 사이에서 S3는 교통흐름이 매우 원활한 상태를 나타낸다.

초기 모델과 관측열로부터 구성된 모델 사이에 매개변수를 변경하며 매번 새로운 모델을 이용해 반복적으로 차이를 구하고 이를 통해 매개변수가 최대치를 갖는 모델을 찾는 과정을 통해 생성된 2개의 상태를 갖는 패턴을 생성했다. Fig. 5는 관측된 데이터를 이용해 예측 모델을 구성한 결과이다.

은닉 마르코프 모델의 과정을 통해 3개의 상태를 갖는 예측 모델을 구성했다. 3개의 상태열로 구성된 모델의 경우 2개의 상태열로 구성된 모델을 이용한 예측보다 상태의 변화를 정밀하게 반영하고 있는 결과를 보인다.

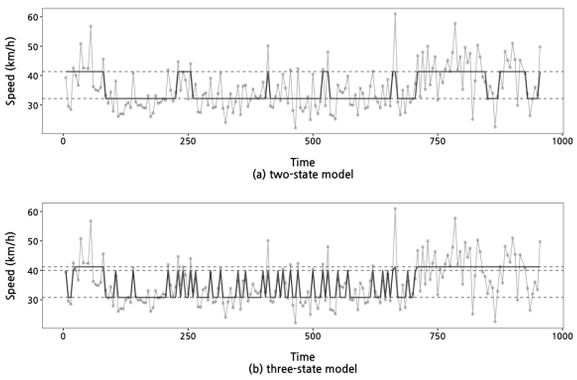

Fig. 4. State sequence of (a) two-state model and (b) three-state model for "A" section

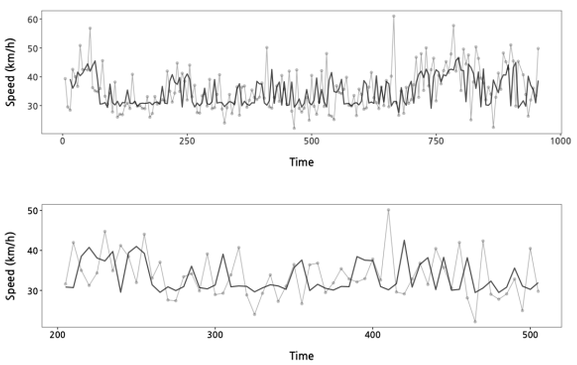

Fig. 5. Estimation of speed passing the section using two-state model for "A" section

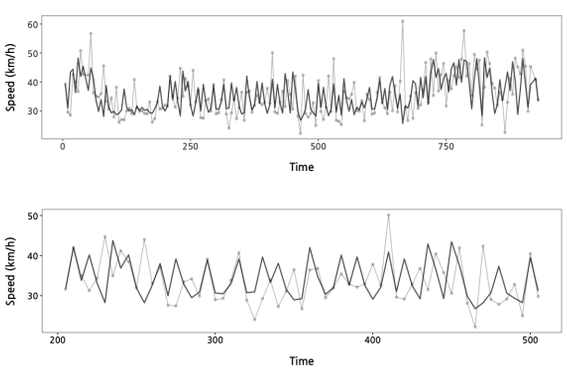

Fig. 6. Estimation of Speed Passing the Section Using Three-State Model for "A" Section

본 논문에서 생성된 모델의 정규성 검정을 위해 2개의 상태열로 구성된 모델의 잔차(Residuals)를 구했다. 잔차의 영역을 벗어나는 부분은 분위수대조도에 이상치(Specification)로 나타난다. 생성된 모델의 잔차는 이상치를 제외하고는 선형성(Linear)을 보이고 있으며 히스토그램(Histogram)에도 정규분포(Normal distribution)의 형태를 보인다.

3개의 상태열로 구성된 은닉 마르코프 모델을 이용한 예측

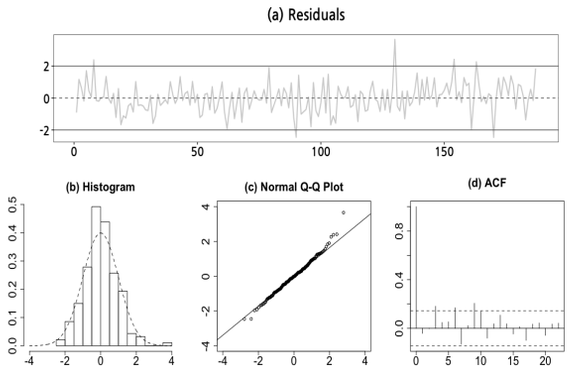

Fig. 7. Pseudo Residuals Between Observed Data and Two-State Model for "A" Section

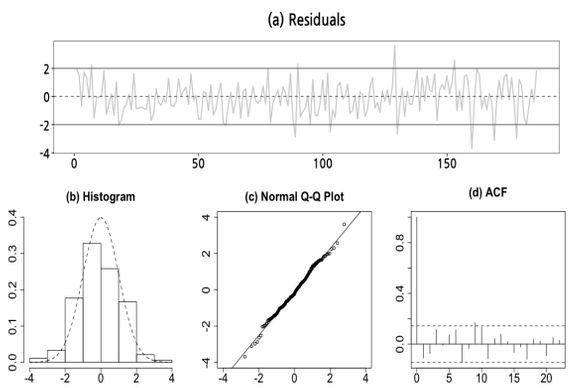

Fig. 8. Pseudo Residuals Between Observed Data and Three-State Model for "A" Section

모델의 잔차는 Fig. 8과 같다. 3개의 상태열로 구성된 모델은 2개의 상태열로 구성된 모델보다 정밀한 상태를 반영하고 있어 관측된 데이터의 특성을 반영한 결과를 나타낸다.

본 논문에서의 모델을 재차 검증하기 위해 A 구간보다 데이터의 변화폭이 큰 B 구간에 적용했다. Fig. 9의 구간의 평균속도는 35.21 km/h 이며 07시부터 09시까지의 속도 평균이 32.10 km/h로 출근 시간대 교통 혼잡의 양상을 보인다. 또한 퇴근 시간대인 17시부터 19시까지의 속도 평균은 31.5 km/h로 구간의 전체 평균 속도와 출근 시간대 평균속도보다 낮다. A 구간과 동일하게 관측 데이터에서 이전데이터와 다음데이터 사이의 높낮이 차이를 보이고 있으며 이는 신호주기의 영향에 따른 것이다.

Fig. 10의 2개의 상태로 구성된 모델에서 S1은 42.550, S2는31.241의 값을 가지며 3개의 상태로 구성된 모델에서 S1은 50.79313, S2는 30.910, S3는 41.450의 값을 갖는다. 2개의 상태로 구성된 모델에서 S1은 교통흐름이 원활하지 않은 구간을 나타내며 출/퇴근 시간대의 상태를 보인다. S2는 교통흐름이 원활한 구간을 나타내며 교통흐름이 일반적인 상황임을 보인다. 3개의 상태로 구성된 모델에서 교통량이 많지 않은 새벽 06시와 07시 사이에서 S3는 교통흐름이 매우 원활한 상태를 나타낸다.

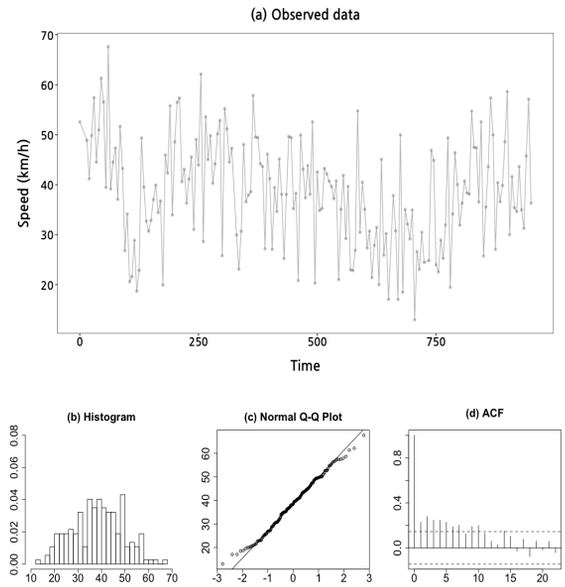

Fig. 9. (a) Observed Data, (b) Histogram, (c) Q-Q Plot and (d) ACF Fitted for the Bus Speed Series of "B" Section

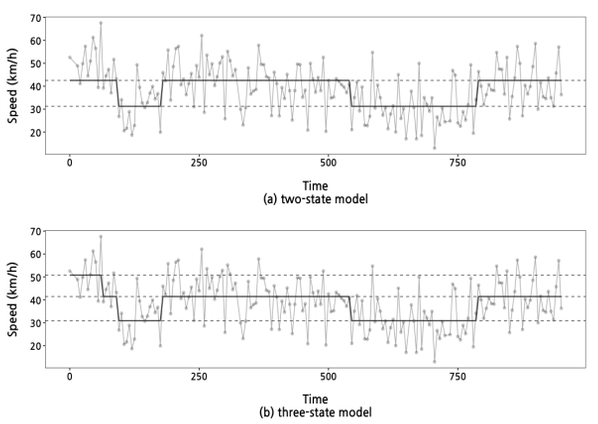

Fig. 10. State Sequence of (a) Two-State Model and (b) Three-State Model for "B" Section

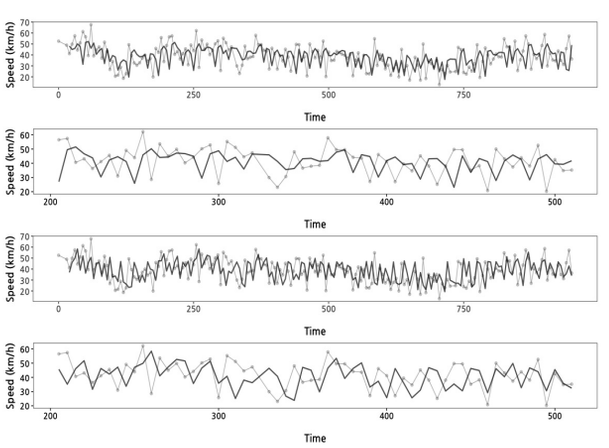

Fig. 11. Estimation of Speed Passing the Section Using Two-State Model and Three-State Model for "B" Section

2개의 상태로 구성된 모델과 3개의 상태를 갖는 은닉 마르코프 모델로 도출된 패턴을 이용한 예측 모델을 구성했다. 3개의 상태열로 구성된 모델의 경우 2개의 상태열로 구성된 모델을 이용한 예측보다 상태의 변화를 정밀하게 반영하고 있는 결과를 보이며 이는 A 구간과 동일한 양상으로 나타난다.

본 논문에서 생성된 모델의 통계적 검증을 위해 부트 스트랩(Bootstrap) 과정을 수행하였다. 1,000개의 은닉 마르코프 모델의 표본을 생성하고 A구간과 B구간 모델 데이터의 평균, 각 매개변수의 90% 신뢰 구간(Confidence intervals)의 분위수(Quantile)를 계산했다.

Fig. 12에서 A 구간 μ_1의 90% 신뢰구간은 32.869, μ_2는 43.666이고, B 구간 μ_1의 90% 신뢰구간은 36.276, μ_2는 44.933이다.

부트스트랩 과정을 수행한 결과 2개의 상태열로 구성된 모델은 정규 분포의 형태를 따르고 있다.

Fig. 13에서 A 구간 μ_1의 90% 신뢰구간은 31.793, μ_2는 41.491, μ_3는 44.867이고, B 구간 μ_1의 90% 신뢰구간은 34.614, μ_2는 42.261, μ_3는 52.457이다.

부트스트랩 과정을 수행한 결과 3개의 상태열로 구성된 모델은 2개의 상태열로 구성된 모델보다 관측 데이터의 정밀한 상태를 반영하고 있어 분포의 형태가 초기 관측데이터의 분포를 반영하고 있다.

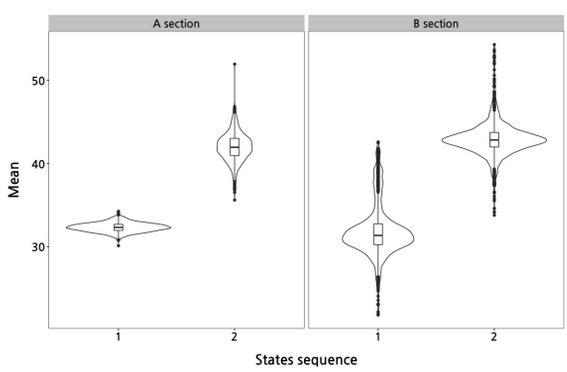

Fig. 12. Results of Bootstrap Process for Two-State Model by Section

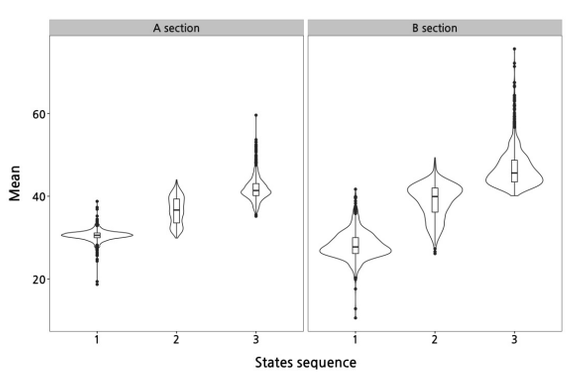

Fig. 13. Results of Bootstrap Process for Three-State Model by Section

Table 5. Maximum likelihood, AIC and BIC for model

	"A" section		"B" section	
	2 state	3 state	2 state	3 state
$-\log L$	549.858	536.206	650.188	631.435
AIC	1326.307	1351.131	1462.409	1440.289
BIC	1348.999	1396.516	1484.990	1485.449

Table 6. Root Mean Square Error

3-state HMM	Kalman Filter	Moving Aaverage
7.523	13.49	14.28

Table 5는 부트스트랩 과정을 통해 생성된 상태열의 최대 우도(Maximum likelihood), AIC, BIC를 나타낸다. A 구간에서 3개의 상태열로 구성된 모델은 $-\log L$ 값이 2개의 상태열로 구성된 모델보다 낮게 나타나며 AIC, BIC의 값은 높다. 2개의 상태로 구성된 모델보다 3개의 상태로 구성된 모델의 복잡도가 높지만 차이가 크기 않으며 예측 데이터에 대하여 모수를 추정하는 성능이 좋은 모델이다.

B 구간에서 3개의 상태열로 구성된 모델은 $-\log L$와 AIC가 낮은 값을 가진다. AIC에 비해 가혹한 벌점요소(penalty)를 부여하는 BIC의 경우 높게 나타나지만 차이가 크기 않다. 모델에 대한 AIC/BIC 값으로 복잡도를 고려하였을 때 3개의 상태로 구성된 모델의 복잡도가 높지만 그 차이는 크지 않으며 3개의 상태열로 구성된 모델이 2개의 상태열로 구성된 모델보다 출/퇴근시간대, 신호 주기 등에 따른 특성을 정밀하게 반영하고 있다.

Fig. 14는 관측데이터와 각 모델간의 비교를 위한 그래프이다. 모델의 성능을 평가하기 위해 관측데이터와 각 모델간의 잔차로 RMSE(Root Mean Square Error)를 계산하여 비교하였다.

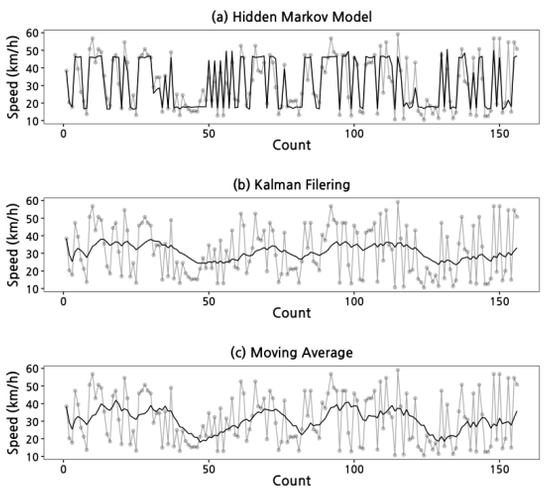

Fig. 14. Compare Observed Data with (a) 3-state HMM, (b) Kalman Filter and (c) Moving Average

Table 6은 본 논문에서 제안한 3-state HMM의 RMSE값은 7.523으로 오차가 가장 작다. 이는 3개의 상태열로 구성된 모델이 Kalman Filter/Moving Average 모델에 비해 출/퇴근시간대, 신호 주기 등에 따른 특성을 정밀하게 반영하고 있음을 보인다.

5. 결 론

기존의 버스정보 시스템에서 이용하는 칼만 필터링, 이동평균 방법과 같은 예측 모델과 달리 전처리 단계에서 추출된 상태를 기반으로 초기 모델을 구성하고 구간 운행 속도를 예측 하는 방법을 제안했다.

버스정보시스템에서 관찰되는 데이터는 정류장에 도착한 시간과 구간 거리이며 이에 의존하여 교통흐름이나 신호 주기 등의 모델을 구해야 하는 어려움이 있다. 기존의 칼만 필터링, 이동평균 방법과 같은 예측 모델에서는 반영되지 않는 데이터이다. 은닉 마르코프 모델은 이러한 제약조건에서 효과적인 모델링 방법으로 적용될 수 있다.

본 논문의 결과를 통해 은닉 마르코프 모델을 이용한 예측 모델을 생성하여 이를 활용하는 방법은 관측된 데이터의 특성을 충분히 반영하고 있음을 보여준다. 은닉 마르코프 모델은 관측된 순서열의 확률을 최대로 하는 매개변수를 결정하는 문제가 있다. 이는 초기 모델과 관측열로부터 구성된 모델 사이에 매개변수를 변경하며 매번 새로운 모델을 이용해 반복적으로 차이를 구하고, 이를 통해 매개변수가 최대치를 갖는 모델을 찾아야 하는 어려움이 있다. 버스정보 시스템과 같은 실시간 정보 제공 시스템에서의 반복 계산 과정은 정보 제공의 지연(delay)을 초래하는 문제의 소지가 될 수 있다.

향후 과제로는 일반화 될 수 있는 구간과 상황들에 따른 패턴을 사전에 생성하고, 이를 분류하여 각 구간에 따라 이용하는 방법을 고려하여 시스템을 설계하는 방향으로 위와 같은 문제를 해결해야 한다.

References

[1] H. G. Kim, C. Y. Park, D. C. Shin, C. S. Shin, Y. Y. Cho, and J. W. Park, "A Study on Traffic Analysis Using Bus Information System," *The KIPS Transactions on Computer and Communication Systems*, Vol.5. No.9, pp.261-267, 2016.

[2] S. H. Lee, B. S. Moon, and B. J. Park, "The Bus Arrival Time Prediction using Bus Delay Time," *Journal of Korean Society of Transportation*, Vol.28, No.1, pp. 125-134, Febrary, 2010.

[3] L. R. RABINER, 1989, "A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *IEEE,* Vol.77, No.2, Febrary.

[4] I. Visser., 2011, "Seven things to remember about hidden Markov models: A tutorial on Markovian models for time series", *Journal of Mathematical Psychology*, Vol.55, pp. 403-415, July.

[5] S. H. Lee, B. S. Moon, and B. J. Park, “The Bus Arrival Time Prediction using Markov Chain,” *The Journal of The Korea Institute of Intelligent Transport Systems*, Vol.8, No.3, pp.1-10, June, 2009.
[6] B. S. Choi, H. C. Kang, S, K, Lee, and S. T. Han, “A Study for Traffic Forecasting Using Traffic Statistic Information,” *The Korean Journal of Applied Statistics*, Vol.22, No.6, pp. 1177-1190, October, 2009.
[7] T. G. Kim, H. C. Ahn, and S. G. Kim, “Predictive Modeling of the Bus Arrival Time on the Arterial using Real-Time BIS Data,” *Journal of The Korean Society of Civil Engineers*, Vol.29, No.1, pp.1-9, January, 2009.

박 철 영
e-mail : naksu21@gmail.com
2010년 순천대학교 정보통신공학과(공학사)
2012년 순천대학교 정보통신공학과 (공학석사)
2012년～현 재 순천대학교 전기 · 전자 · 정보통신공학과 박사과정
관심분야 : 기계학습, 시계열 분석, IoT

김 홍 근
e-mail : khg_david@sunchon.ac.kr
2011년 순천대학교 정보통신공학과 (공학사)
2013년 순천대학교 정보통신공학과 (공학석사)
2013년～현 재 순천대학교 전기 · 전자 · 정보통신공학과 박사과정
관심분야 : 기계학습, 시계열분석, IoT

신 창 선
e-mail : csshin@sunchon.ac.kr
1996년 우석대학교 전산학과(학사)
1999년 한양대학교 컴퓨터교육과(석사)
2004년 원광대학교 컴퓨터공학과(공학박사)
2005년～현 재 순천대학교 정보통신공학과 부교수
2016년～현 재 순천대학교 정보전산원 원장
관심분야 : 분산컴퓨팅, 실시간 객체모델, 시계열분석

조 용 윤
e-mail : ycho@sunchon.ac.kr
1995년 인천대학교 전산학과(학사)
1998년 숭실대학교 컴퓨터학과(공학석사)
2006년 숭실대학교 컴퓨터학과(공학박사)
2009년～현 재 순천대학교 정보통신공학과 부교수
관심분야 : 시스템 소프트웨어, 유비쿼터스 컴퓨팅, 기계학습

박 장 우
e-mail : jwpark@sunchon.ac.kr
1989년 한양대학교 전자공학과(공학사)
1991년 한양대학교 전자공학과(공학석사)
1993년 한양대학교 전자공학과(공학박사)
1995년～현 재 순천대학교 정보통신공학과 교수
관심분야 : SoC, USN, 기계학습, 시계열 분석