KIPS Tr. Comp. and Comm. Sys.
Vol.6, No.7 pp.297~306 pISSN: 2287-5891

https://doi.org/10.3745/KTCCS.2017.6.7.297

Long-Term Arrival Time Estimation Model Based on Service Time

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

ABSTRACT

Citizens want more accurate forecast information using Bus Information System. However, most bus information systems that use an average based short-term prediction algorithm include many errors because they do not consider the effects of the traffic flow, signal period, and halting time. In this paper, we try to improve the precision of forecast information by analyzing the influencing factors of the error, thereby making the convenience of the citizens. We analyzed the influence factors of the error using BIS data. It is shown in the analyzed data that the effects of the time characteristics and geographical conditions are mixed, and that effects on halting time and passes speed is different. Therefore, the halt time is constructed using Generalized Additive Model with explanatory variable such as hour, GPS coordinate and number of routes, and we used Hidden Markov Model to construct a pattern considering the influence of traffic flow on the unit section. As a result of the pattern construction, accurate real-time forecasting and long-term prediction of route travel time were possible. Finally, it is shown that this model is suitable for travel time prediction through statistical test between observed data and predicted data. As a result of this paper, we can provide more precise forecast information to the citizens, and we think that long-term forecasting can play an important role in decision making such as route scheduling.

Keywords : Bus Information System, Arrival Time Estimation, Service Time Estimation, Hidden Markov Model, Generalized Additive Model

버스의 정차시간을 고려한 장기 도착시간 예측 모델

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

요 약

버스정보 시스템을 이용하는 시민들은 더 정확한 예측 정보를 원한다. 하지만 평균 기반 단기간 예측 알고리즘을 사용하는 대부분의 버스정보시스템에서는 교통흐름, 신호주기, 정차시간 등의 영향이 고려되지 않기 때문에 많은 오차를 포함하고 있는 실정이다. 따라서 본 논문에서는 오차의 영향요인 분석을 통해 예측정보의 정밀도를 향상시켜 시민들의 편의를 도모하고자 한다. 이에 현재 운영되고 있는 버스정보 시스템의 자료를 토대로 오차의 영향요인을 분석했다. 분석 데이터에서 시간대별 특성과 지리적 여건에 의한 영향이 복합적으로 나타나고, 정차시간과 단위구간속도에 미치는 영향도가 다름을 보였다. 이에 따라 정차시간은 일반화 가법 모형을 사용하여 시간, GPS 좌표, 통과노선수의 설명변수로 패턴을 구축하고, 단위구간에 대해 은닉 마르코프 모델을 사용하여 교통흐름에 따른 영향도를 고려한 패턴을 구축했다. 패턴 구축의 결과로 정밀한 실시간예측이 가능하고, 노선 통행속도의 장기간 예측이 가능했다. 마지막으로 관측 데이터와 예측 데이터의 통계적 검정 과정을 통해 전구간 예측에 적합한 모델임을 보였다. 본 논문의 결과로 시민들에게 더 정확한 예측 정보를 제공하고, 장기간 예측은 배차시간 등의 의사결정에 중요한 역할을 수행할 수 있으리라 생각한다.

키워드 : 버스정보시스템, 도착시간 예측, 정차시간 예측, 은닉 마르코프 모델, 일반화 가법 모형

1. 서 론

버스 정보 시스템(BIS)은 GPS가 포함된 차내 장치(OBE)를 통해 차량으로부터 수신된 정보를 노선, 통과구간, 정류장 등의 기반정보로 가공하여 시민들에게 버스 도착 예정시간과 버스의 현재위치 등의 정보를 제공한다. 버스 도착시간 예측 정보 제공은 대중교통의 편의성 개선에 능동적인 역할을 수행하고 있다[1-3]. 버스 도착시간 예측을 위한 대표적인 방법으로는 이동평균필터, 동적 선형 모형, 회귀 모형 등의 기법이 있으며 구간별로 계산된 데이터 테이블을 이용하는 방법이다[2, 3].

† 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 : February 27, 2017
Accepted : April 3, 2017
* Corresponding Author : Park Jang Woo(jwpark@sunchon.ac.kr)

일반적으로 버스정보 시스템에 사용되는 예측 모형은 도로의 교통흐름, 신호 주기, 이상 상황, 데이터 결측 등의 상황을 고려하지 않은 모델이다. 또한 시스템에서 수집되는 데이터는 시간정보와 위치, 거리 등의 기반정보에 의존하여 계산된 단순 데이터로 교통흐름과 신호 주기를 반영하여 모델링을 수행하기는 매우 어렵다. 은닉 마르코프 모델은 관측데이터에 은닉된 상태의 요소가 포함된다고 가정하고 데이터 순서의 확률을 계산하여 관찰된 결과로 은닉상태를 도출하는 모델이다. 은닉 마르코프 모델은 이러한 제약 조건에서 효과적인 모델로 적용될 수 있다[3, 4].

버스 노선의 운행시간은 구간 통행시간과 승객이 승/하차하는 정차시간이 포함된다. 그러나 현재의 시스템에서는 정차시간 예측을 위한 적합한 모델이 없으며, 정류장을 통과하는 이벤트가 생성되는 시점의 시간으로 예측 데이터를 보정하는 방법을 이용하고 있다. 이러한 방법은 시스템의 연산 부하를 줄여 실시간 정보를 제공하는데 이점이 있으나 시간이 보정된 정류장에서 가까운 정류장의 정확도가 높고, 통과하는 정류장이 많을수록 도착 시간 예측이 부정확해지는 문제가 있다. 운행거리가 짧고 통과하는 정류장의 수가 적은 소규모 버스정보 시스템에 적합한 예측 방법이며 정류장의 수가 많아질수록 이벤트 수가 증가한다.

도착 예정시간을 모델링하는 방법인 이동평균필터, 동적 선형모형 등의 방법으로 정류장의 특징을 반영한 정차시간을 예측하기 위해서는 정류장별로 데이터를 분류할 필요가 있다. 이러한 경우 정류장의 수 혹은 유사한 특징으로 묶은 그룹의 수에 따른 데이터 테이블이 필요하게 된다. 이는 시스템의 부하로 작용되어 실시간 연산을 요하는 버스정보 시스템에 적합하지 않다. 일반화 가법 모형은 자료의 특성을 반영한 연결함수와 반응변수에 영향을 미치는 설명변수를 사용하여 변수간의 관계를 분석하고 반응변수에 적합한 모형을 도출하는 방법이다. 관찰 데이터에서 도출되는 특징을 반영할 수 있는 모델로서 정차시간을 모델링하는데 효과적으로 적용될 수 있다.

본 논문에서는 2015년 한해 수집된 순천시 버스 정보 시스템의 데이터를 이용하여 구간 통행속도 예측 방법으로 은닉 마르코프 모델을 사용하고, 정차시간 예측을 위해 일반화 가법모형을 사용한 혼합 모델을 구축했다.

2. 관련 연구

일반화 가법 모형(Generalized Additive Model)은 일반화 선형 모형(Generalized Linear Model)의 속성에 가법 모형을 적용한 통계적 모형이다. 일반화 가법 모형은 선형(linear) 관계의 모형으로 적합하기 어려운 설명변수와의 관계를 평활(spline) 함수로서 모형을 적합한다[5, 6]. 일반화 가법모형은 지수 분포족의 특성을 갖는 반응변수에 대한 가법적 비선형(non-linear) 모형이다[6].

반응변수 $\mu = E(Y)$와 공변량 $X = (x_1, \cdots x_p)$는 다음과 같은 확률모델로 표현된다.

$$\mu = E(Y|x_1, \cdots x_p) \tag{1}$$

연결함수(link function, g)는 설명변수와의 연결구조를 나타내며, 예측치에 대한 연결함수와의 구조는 Equation (2)와 같이 표현된다.

$$\eta = g(\mu) = s_0 + \sum_{j=1}^{p} s_j(x_j) \tag{2}$$

여기서 $s_0, s_1(x_1), \cdots, s_p(x_p)$의 추정은 연결함수로 교정된 반응변수에 가중치를 부여한 값으로 계산된다. 일반화 가법모형은 비선형 관계를 표현하기 위해 설명변수에 비모수적인 평활함수(Smoothing spline fuction)가 사용된다. 일반화 가법모형에서는 x_j의 가법적 효과를 나타내는 평활함수 $s_j(x_j)$의 추정이 필요하다.

$$s_k(x_k) = E[Y - s_0 - \sum_{k \neq j} s_j(x_j)|x_k] \tag{3}$$

Equation (3)은 가법모형의 조건부 기댓값으로 다음과 같은 Backfitting 알고리즘을 통해 추정된다.

1) 초기화 : $s_j = s_j^{(0)}, j = 1, \cdots p$

2) 반복 : $j = 1, \cdots, p, 1, 2, \cdots, p, \cdots$

$$s_j = s_0 R_j(Y - s_0 - \sum_{k \neq j} s_k|x_j)$$

(R_j : Cubic smoothing spline)

3) $\hat{s}_j$의 값이 기준치 이하일 때 까지 2)의 과정을 반복한다.

3. 데이터 분석

3.1 정차시간 분석

본 논문에서 사용된 자료는 분석 대상 도시 버스정보시스템의 2015년 실제 운행 노선에서 수집된 데이터이다. 수집된 데이터의 정류장은 2298개소이며 데이터의 이상치(outlier)를 포함한 평균 정류장 정차시간은 26초이다. 하나의 노선은 최소 5개 이상의 정류장을 통과하며 순환노선의 경우 최대 149개의 정류장을 통과한다. 노선은 평균적으로 56개의 정류장을 통과한다. 정차시간의 데이터는 정류장에서 승객들의 승/하차가 없는 무정차통과, 버스의 시/종점, 순환노선의 대기지점의 데이터가 이상치로 반영된다.

Fig. 1은 2015년 4월 한달간 수집된 자료로서 이상치를 제외한 1일부터 30일까지의 시간대별 평균이다. 버스정보시스템의 노선 기반자료를 활용하여 시/종점, 순환노선의 대기 정류장을 제외했다. 또한 버스의 고장으로 인한 대기, 시간동기화 오류 등의 시스템 문제 등에 의한 이상치가 나타난

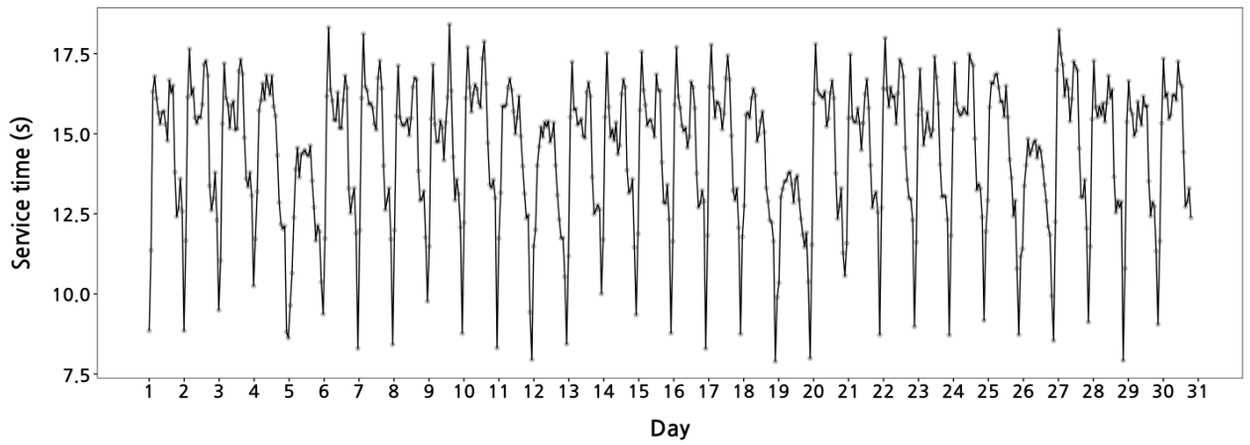

Fig. 1. Observed Data for Service Time of the Bus

다. 각 노드별 정차시간 데이터의 사분위 범위(inter quartile range; IQR)를 계산하고 3사분위수(75%) + 1.5 X IQR 보다 크면 이상치로 판단하여 제거했다[9-12].

Fig. 1에서 정차시간의 자료는 기본적으로 매우 유사한 패턴을 보인다. 그러나 2015년 4월의 주말인 4, 5, 11, 12, 18, 19, 25, 26일의 경우 주중의 형태(shape)와 다르게 나타난다. 정차시간 패턴은 주중 패턴과 공휴일을 포함한 주말 패턴의 두 가지 범주로 나뉠 수 있으며, 이에 따라 예측 모형은 두 가지로 고려된다.

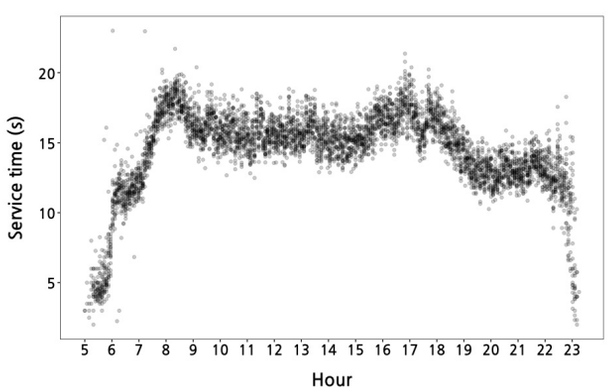

Fig. 2. Scatter Plot of Weekdays

Fig. 2는 평일의 시간대별 분포표이다. 8시경과 18시경의 출/퇴근 시간대의 영향으로 인해 정차시간이 높게 나타나며, 평일 시간대별 분포의 정차시간은 최저 7.89초 최고 18.41초 중앙값은 15.12초이다.

Fig. 3은 주말의 시간대별 분포표이다. 8시경과 18시경의 출/퇴근 시간대의 영향이 없으며, 8시부터 22시까지의 분포는 변량이 크지 않다. 주말 시간대별 분포의 정차시간은 최저 2초 최고 19.38초 중앙값은 13.99초이다. 5~8시 구간과 22시~23시 구간의 매우 짧은 정차시간이 존재하며 주중의 통계량에 비해 변량은 크지 않다.

주중 데이터에서 출/퇴근 시간대 특징이 나타난다. 정류장 이용패턴(정류장 위치 등)에 따른 일부 특정 구간이 주중/주말 패턴에 차이를 나타내는 중요한 특징으로 작용된다.

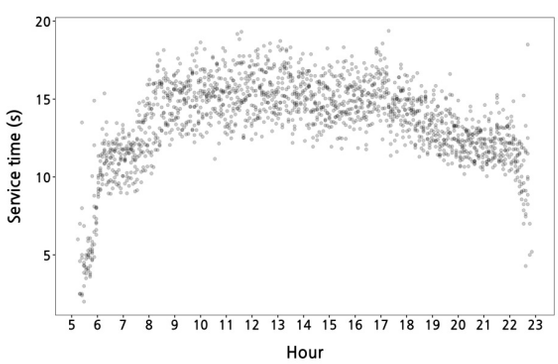

Fig. 3. Scatter Plot of Weekends

3.2 구간별 속도 분석

버스정보 시스템에 구축된 데이터는 노드. 링크를 이용하며 노드는 교차로, 도로 시/종점, 속성 변화점, 도로시설물(버스 정류장) 등의 속성을 의미하며, 링크는 노드와 노드 사이를 연결하는 거리를 포함하는 속성이다[2, 9-10]. 버스정보 시스템에서 수집되는 데이터는 정류장의 출발/도착 데이터이다. 분석 대상 도시의 정류장은 평균 8개의 노선이 통과한다. 링크의 거리 데이터는 모두 상이하다. 구간마다 공통의 속성을 적용하기 위하여 출발/도착 데이터의 시간차를 구하고 기반정보의 구간거리로 속도(km/h)로 변환했다.

Fig. 4는 수집된 데이터를 5분단위로 샘플링하고 하루단위로 같은 시간에 나타나는 그룹의 평균을 구한 평일의 시간대별 속도 분포표이다. 6시 이전과 22시 이후의 시간대에서 교통흐름이 매우 원활하여 이동속도가 높게 나타나며, 6시 이후와 22시 사이에는 평균속도를 보인다. 구간속도는 최저 21.9km/h, 최고 78.8km/h이며, 중앙값은 36.9km/h이다. 속도 데이터는 교통흐름에 영향이 있으며, 교통이 혼잡한 출/퇴근 시간의 영향이 구간마다 다르게 나타난다. 교통 혼잡은 구간에 따라 그 형태가 매우 상이하게 나타난다. 예측

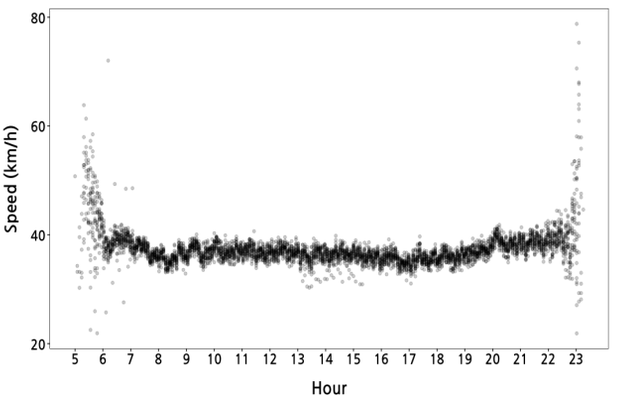

Fig. 4. Scatter Plot of Weekdays

데이터의 정확도를 높이기 위해서는 각 구간별로 패턴 테이블을 구축함으로서 구간별 영향이 고려되어야 한다.

Fig. 5는 데이터 분석을 위해 선택한 도심방향 구간의 시간대별 도표이다. 이 구간은 아파트단지가 밀집된 분석 대상 도시의 중심 주거지역이다. 점선은 평균을 나타내며 시간대별 박스 내부의 실선은 중위수를 표시했다. 8시와 18시에서 출/퇴근시간대의 영향이 나타난다. 8시에 중위수가 18시보다 낮게 나타나며 퇴근시간대보다 출근시간대의 영향이 더 크다.

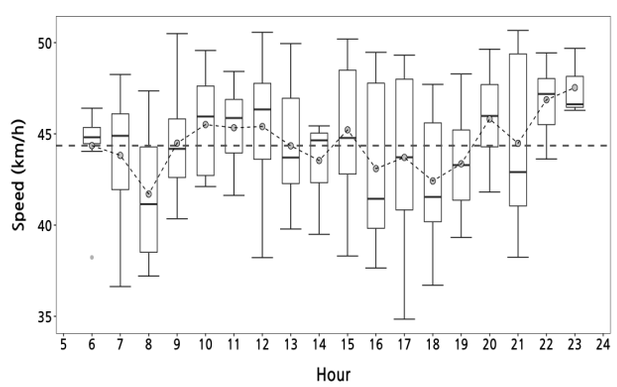

Fig. 5. Box-plot of the Section Toward the Urban Street Direction

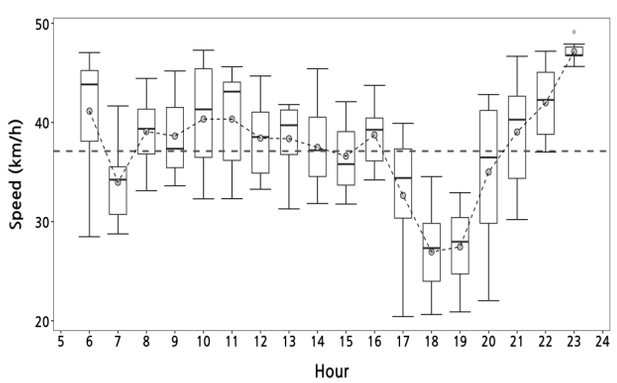

Fig. 6. Box-plot of the Section Toward the Residential District Direction

Fig. 6은 Fig. 5의 진행방향이 반대인 주거지역 방향 구간의 시간대별 도표이다. 7시, 17시, 18시, 19시에서 출/퇴근시간대의 영향이 나타난다. 17, 18, 19시인 퇴근시간대의 영향이 더 크다.

Fig. 7은 주거지역 방향 구간으로 데이터를 5분 간격 샘플링 속도를 나타낸 도표이다. 시간에 따른 영향을 보이고 있으나, 세부적으로 샘플링 시간대에서 변동 폭이 크게 나타나는 형태를 보인다. 이러한 변동 폭은 일시적인 교통흐름의 변화(신호주기 등)에 의해 나타난다. 은닉 마르코프 모델은 관측데이터에 은닉된 상태 요소의 포함을 가정하고 데이터 순서의 확률을 계산하여 관찰된 결과로 은닉상태를 도출하는 모델이다. 은닉 마르코프 모델은 이러한 제약 조건에서 효과적인 모델로 적용될 수 있다[3].

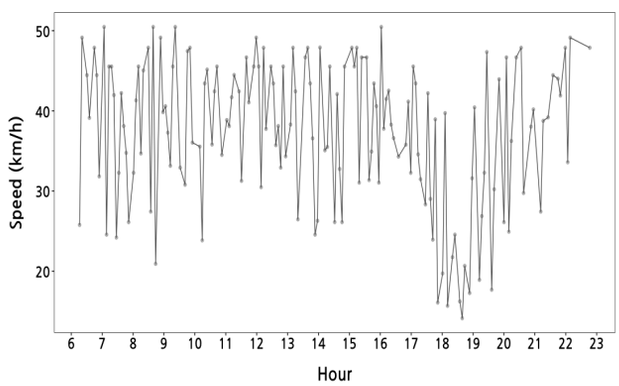

Fig. 7. The Section Toward the Residential District Direction

4. 데이터 모델링

4.1 정차시간 예측 : 일반화 가법모형

정차시간은 시민들이 많이 이용하는 시간대와 노선의 특성(운행간격, 시내/시외 통과 여부) 그리고 지역별 특성(주거지역 및 학교 등의 분포)에 따라 유동적으로 나타난다.

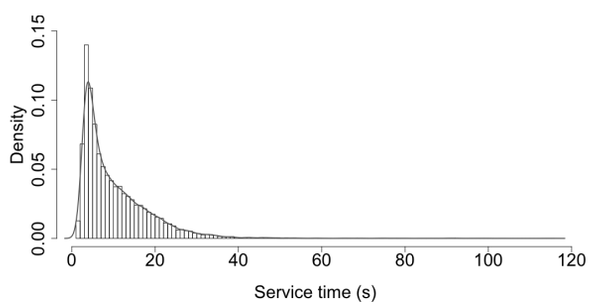

Fig. 8. Histogram for Observed Data

Fig. 8은 관측데이터의 분포를 보인다. 일반화 가법 모형은 반응변수의 분포와 연결 함수(link function)의 정의에 따라 다양한 형태의 모형을 적합할 수 있다. 비모수적 특징을 갖는 정차시간 데이터의 분포에 따라 로그함수를 연결함수

로 사용했으며 반응변수의 분포는 포아송 분포(Poisson distribution)로 모델을 적합(fitting)했다.

Fig. 9는 각 그룹별로 계산된 계수를 보인다. 분석 대상 도시의 정류장은 2298개소로 각 정류장별 특징 반영을 위해 요인변수로 정류장 고유번호를 이용하기에는 시스템의 부하에 미치는 영향이 매우 크다. 또한 추후 노선의 변경이나 노선의 추가삭제에 의해 요인변수가 변경되며 전체 변수들과 계수들의 재계산이 요구된다. 이러한 문제를 해결하기 위해 정류장별로 수집된 데이터의 3사분위수를 구하고 범주형 으로 변환 후 요인변수로 이용했다.

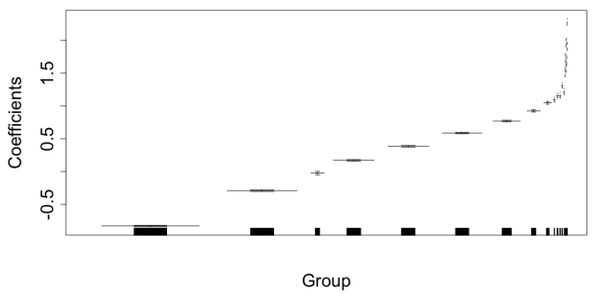

Fig. 9. Coefficients by Categorical Variable

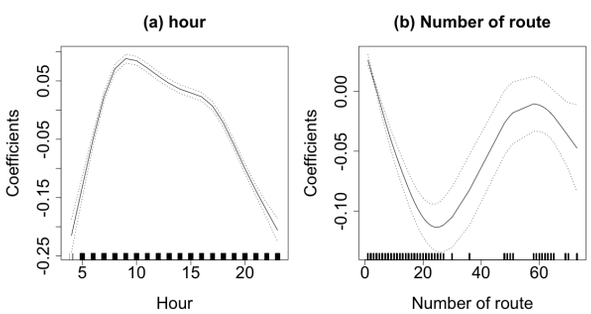

Fig. 10. Natural Smoothing Splines Fitted to (a) Hour and (b) Number of Route

Fig. 10에서 (a)는 시간에 따라 스플라인(spline)을 적합한 결과이다. 점선은 95%신뢰구간을 의미한다. 8시경의 출근 시간대에서 가장 높게 나타나며 8시 이전과 18시 이후에는 감소하는 경향을 보인다. (b)는 정류장에 통과하는 노선의 수에 따라 적합한 결과이다. 노선의 수 25에서 정차시간이 최소값을 가진다. 정차시간에 대한 노선의 수에 따른 영향을 반영하고 있다. 정류장을 통과하는 노선의 수는 주로 30개 이하로 분포되고 있으며, 30 이하에서 정류장을 통과하는 노선의 수가 증가할수록 정차시간이 감소하는 경향을 보인다. 노선의 수가 증가할수록 승객들이 분산되며 이로이한 효과로 정차시간이 감소함을 보인다. 30 이상 60 이하에서는 영향도가 상승하는 경향을 보이며, 60 이상에서는 다시 감소하는 형태를 보인다. 노선의 수 30~60 사이에서는 정차시간의 감소효과가 줄어든다.

Fig. 11은 위도와 경도에 따라 스플라인을 적합한 결과이다. 시내권의 주요 밀집지역에 나타나는 영향도가 크게 반영

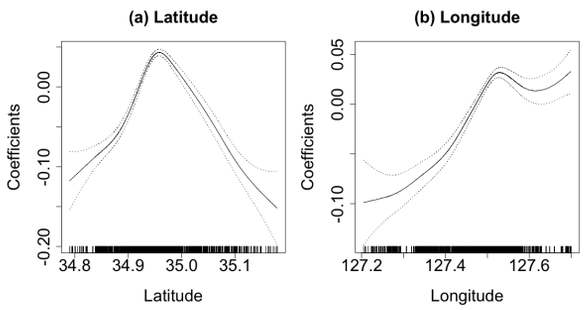

Fig. 11. Natural Smoothing Splines Fitted to (a) Latitude and (b) Longitude

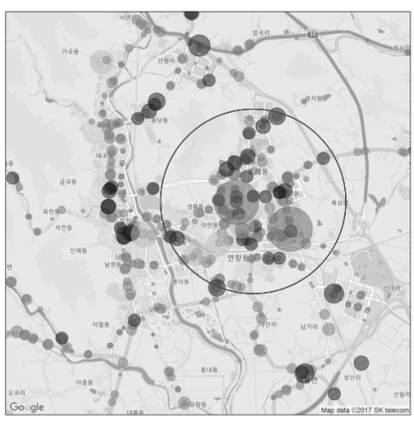

Fig. 12. Spatial Scatter Diagrams on 9 A.M.

되며 위도 34.9577, 경도 127.5331에서 최대치를 보인다.

Fig. 12는 GPS 좌표에서 최대 영향을 나타낸 지역의 오전 9시의 도표이다. 좌표의 지역은 반경 1km 내에 13개의 학교가 있는 아파트 단지가 밀집된 중심 주거지역이다. GPS 좌표의 설명변수는 정류장의 위치별 특성을 반영한다.

수집된 데이터와 기반정보를 통해 가공된 데이터에서 사용될 수 있는 반응변수는 정류장을 통과하는 노선의 수, 시간, GPS 좌표이며, 요인변수는 정류장 고유번호가 있다.

$$\log(\mu) = s_0 + s_1(NoR) + s_2(HOUR) + s_3(LON) + s_4(LAT) \quad (4)$$

(NoR ; Number of route, LON : Longitude, LAT : Latitude)

Equation (4)에서 NoR은 정류장을 지나는 노선의 수, LON은 GPS좌표의 경도 그리고 LAT는 위도를 나타낸다. Equation (4)의 모델에 데이터를 적합(fitting)한 결과는 다음과 같다.

Table 1은 분산분석 (ANOVA; Analysis of variance) 테스트 결과이다. 분산분석은 관측 데이터와 모델링 데이터 사이의 분산 비교를 통해 계산된 통계량으로 유의성을 검정하는 방법이다. 반응변수의 비모수적 특성에 따라 카이제곱 분포를 이용하여 계산된 통계량과 유의확률을 보인다. 유의

Table 1. ANOVA for Non-parametric Effects

Variable	Degrees of freedom	Chi-square statistics	P-value
NoR	3	193.19	< 2.2e-16
HOUR	3	1129.82	< 2.2e-16
LON	3	547.68	< 2.2e-16
LAT	3	107.82	< 2.2e-16

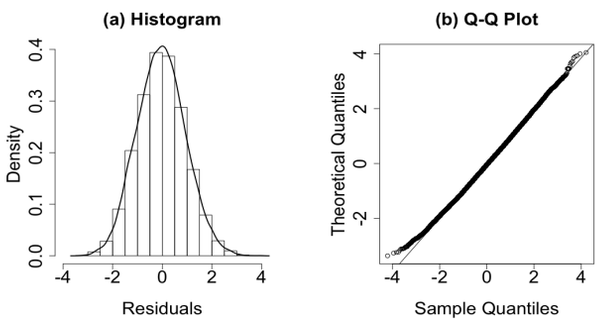

Fig. 13. Residual diagrams : (a) Histogram and (b) Q-Q Plot

수준 5%에서 모든 변수가 매우 작은 값으로 나타나고 있으며 사용된 변수 모두 매우 유의한 설명변수임을 보인다.

Fig. 13은 잔차의 정규성을 검정하기 위한 도표이다. (a)는 히스토그램을 나타내며, (b)는 분위수대조도(Quantile - Quantile Plot, Q-Q Plot)를 나타낸다. 버스의 승/하차가 없는 무정차 통과는 예측데이터와의 오차를 높이는 성분이다. 무정차 통과의 특성으로 인해 분위수 대조도(b)의 좌우측에 왜도(skewed)가 나타난다.

Table 2. GAM fitted Coefficients for Weekdays

Coefficients	Weekday	Weekend
Intercept	-23.1719981	-71.497272
GROUP	0.0756428	0.113214
NoR	-0.0008444	-0.001229
HOUR	-0.0111555	-0.010774
LAT	0.2316771	0.059453
LON	0.1377532	0.559848

Table 2는 일반화 가법 모형에 적합된 변수의 주말과 주중 모델의 계수이다. 정차시간은 시민들이 버스를 주로 이용하는 시간대와 노선의 특성(운행간격, 시내/시외 통과 여부) 그리고 지역별 특성(주거지역 및 학교 등의 분포)에 따라 유동적으로 나타난다. 버스정보 시스템에서 수집되는 데이터로 정차시간 모델링을 위한 설명변수가 매우 제한적이다.

본 논문에서 사용된 설명변수는 4개이다. 일반화 가법모형은 적합도가 높은 변수 선택을 위한 방법(stepwise, forward, backward)이 있으나, 정차시간의 설명변수가 매우 제한적이며 사용된 변수 모두 유의한 변수로 선택 방법은 고려되지 않았다.

4.2 통행시간 예측 : 은닉 마르코프 모델

버스정보시스템에 수집되는 구간별 통행속도는 진행방향에 따른 영향과 출/퇴근의 교통 혼잡 시간대에 따른 영향이 있다. 분석 구간의 일반현황은 다음과 같다.

Table 3. General Information on Analyzed Section

Section	Distance (Meter)	Lane	Traffic signal	Inter-section
A	518.7	3	1	3
B	478.5	3	1	1

Table 3은 분석구간의 일반현황이다. 교통흐름의 영향은 각 구간별로 상이한 패턴으로 나타나며 정확도가 높은 예측을 위해 각 구간별로 모델링이 필요하다. 교통흐름의 영향도가 다른 2개의 구간을 선택했다.

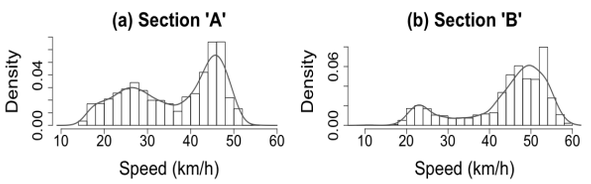

Fig. 14. Histogram by Section within Density Curve

Fig. 14는 분석 구간별 도수분포표와 확률밀도함수를 나타낸다. 각 구간에서 정규분포를 따르지 않는 비모수적인 형태를 보인다. 구간 속도 분포는 평균을 중심으로 양측에 동일한 분산을 가지는 가우시안 분포로 특징을 표현하기는 어렵다. 이러한 과정에서 복수개의 가우시안 분포들의 합으로 구성되는 혼합모델을 고려했다.

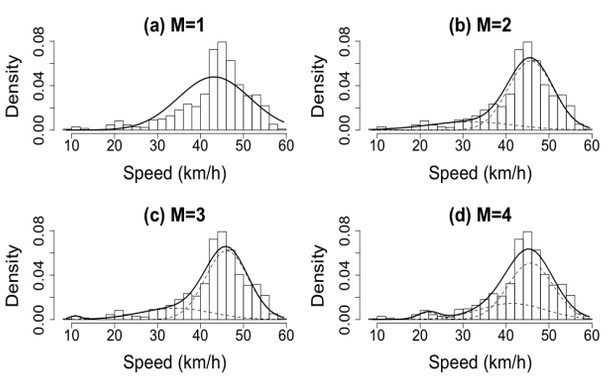

Fig. 15. Histogram of Counts, Compared to Mixtures of (a) One, (b) Two, (c) Three and (d) Four Gaussian Distributions

Fig. 15는 “B”구간의 하루 동안 수집된 속도 데이터이며 확률밀도함수의 모수(M)에 따른 혼합모델을 나타낸다. 사용된 확률밀도함수의 수에 따라 정밀하게 모분포의 특성을 반영할 수 있다. 은닉 마르코프 모델은 데이터가 확률적으로 어떤

분포로 부터 추정되는지에 대한 상태(state) 추정 모델이다[3, 7-9]. 즉, 혼합모델의 확률밀도함수의 모수(parameter)는 은닉 마르코프 모델의 상태 수이다.

Table 4. Model Selection Criterion by Parameter

Parameter	AIC	BIC	-logL
1	4697	4706	2347
2	4535	4566	2260
3	4492	4555	2232
4	4507	4611	2231

Table 4는 모형의 복잡도에 따른 과대 적합(Over fitting)의 문제를 고려하고 적합한 파라미터를 선택하기 위해 AIC (Akaike Information Criterion), BIC(Bayesian Information Criterion) 그리고 로그 우도(log-likelihood)를 계산한 결과이다. 모델의 복잡도가 낮고 정밀도가 높은 모수를 선택하기 위해서는 AIC, BIC와 -logL를 최소로 하는 모델을 선택해야한다. 모수가 3(M=3)인 경우 AIC와 BIC가 최소이며 모수가 4(M=4)인 경우와 비교해 -logL값의 차이가 크지 않다. 본 논문에서는 3개의 상태를 가진 은닉 마르코프 모델을 사용하여 구간의 속도를 예측한다.

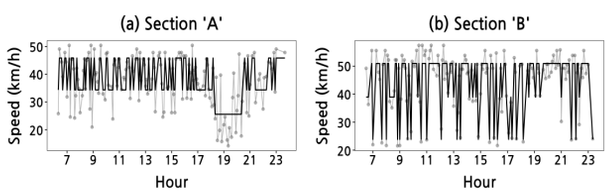

Fig. 16. State Sequence of Three-state Model for (a) Section 'A', (b) Section 'B'

Fig. 16은 모수를 3으로 상태병합과정을 통해 구성한 상태열 도표이다. 회색 실선은 관측데이터이며 검정색 실선은 상태열이다. 각 상태에 따라 교통흐름의 원활, 정체, 일반적인 상황을 나타낸다. 각 구간별 상태의 계수는 다음과 같다.

Table 5. HMM fitted Coefficients by Section (unit: km/h)

Section	State 1	State 2	State 3
A	34.37	[25.55]	45.89
B	38.93	50.95	[23.84]

Table 5는 은닉 마르코프모델의 상태별 계수를 나타낸다. 각 상태의 계수는 정규분포의 평균을 나타내며, 구간에 따라 가장 낮은 계수를 갖는 "A"구간 S2, "B"구간 S3는 교통흐름이 혼잡한 상태를 나타낸다.

Fig. 17은 은닉 마르코프 모델의 전이확률과 상태병합과정을 통해 생성된 상태열로 예측모델을 구성하고, 관측된 데이터에 따라 속도를 예측한 결과이다. 회색 실선은 관측데이터이며 검정색 실선은 예측 데이터이다.

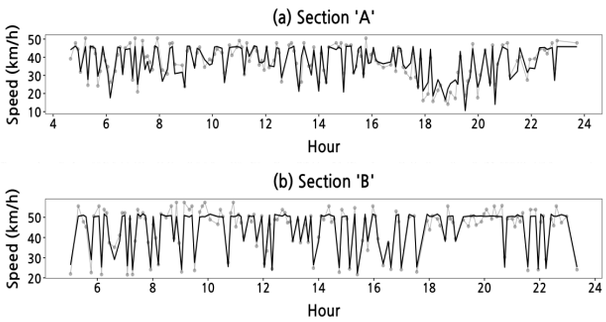

Fig. 17. Fitted Results of Speed Estimation with Transition Matrix and State Sequence

5. 결과 및 결론

본 논문에서 버스정보 시스템의 운행시간 예측은 두가지 모델로 적합(fitting) 되었다. 첫 번째로 정차시간 예측을 위해 일반화 가법 모형을 이용했다. 두 번째로 구간 통행속도 예측을 위해 은닉 마르코프 모델을 이용했다. 정차시간 모델은 정류장별로 별도의 테이블 구축 필요성이 없는 모델로 범주형 변수와 설명변수로 추정된 모델을 이용하여 주중과 주말의 정차시간을 예측한다. 은닉 마르코프 모델은 구간별로 상이하게 나타나는 영향을 반영하기 위해 구간별로 별도의 테이블을 구성했다.

Fig. 18. A Spatial Plot of the Route for Estimation

Fig. 18은 예측 데이터의 검증을 위해 선택한 노선의 좌표를 지도에 표시한 도표이다. 노선은 상업지역과 중심 주거지역을 경유하며 42개 구간을 통과한다. 노선의 총 운행거리는 19.3km이다.

Fig. 19는 누적 관측데이터와 누적 예측데이터의 비교 도표이다. 회색은 관측데이터이며, 운행일은 2015년 6월 16일 화요일이며 운행시간은 20시 30분부터 21시 20분까지 총 50분을

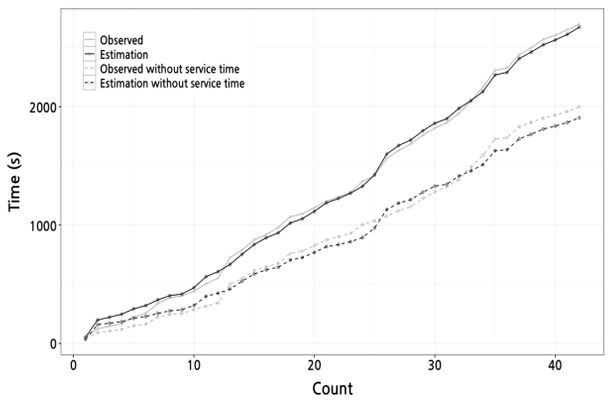

Fig. 19. Comparison of Observed Data and Travel Time Estimation on June 16

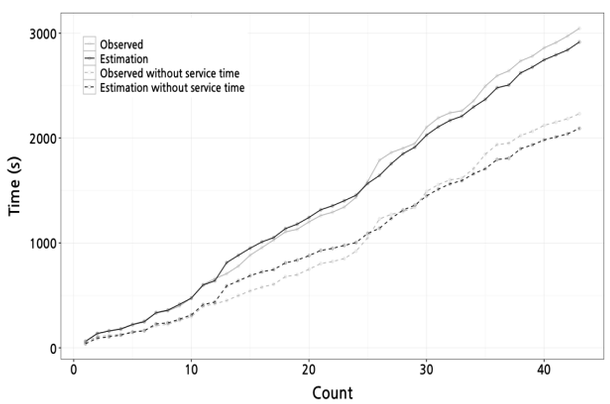

Fig. 20. Comparison of Observed Data and Travel Time Estimation on July 23

운행했다. 검정색은 출발시간을 기준으로 한 예측 데이터이다. 실선은 정차시간을 포함한 데이터이고, 점선은 정차시간을 포함하지 않은 데이터이다.

Fig. 20은 누적 관측데이터와 누적 예측데이터의 비교 도표이다. 회색은 관측데이터이며, 운행일은 2015년 7월 23일 목요일이며 운행시간은 17시 36분부터 18시 30분까지 총 54분을 운행했다. 관측 데이터와 예측 데이터사이 산포(dispersion)의 차이를 비교하기 위해 F-검정을 수행했다.

Table 6. Results of F-test

Route	DF	F-value	P-value
June 16	41	1.4	0.3
July 23	41	1.3	0.4

Table 6은 F-검정의 결과이다. 6월 16일의 결과에서 유의확률은 0.3이다. 7월 23일의 유의확률은 0.4로 유의수준 5%에서 두 결과 모두 등분산성을 가정한 귀무가설(null hyphothesis)을 만족한다.

Table 7. Standard Deviation and Mean by Route

Route	Observed data		Estimation data	
	Standard deviation	Mean	Standard deviation	Mean
June 16	35.02	64.24	35.07	63.17
July 23	39.37	70.84	33.07	68.07

Table 7은 관측데이터와 예측 데이터의 표준편차와 평균이다. 등분산성을 만족하는 관측 데이터와 예측 데이터 사이의 유의성을 확인하기 위해 T-검정을 이용했다.

Table 8. Results of T-test

Route	95% Confidence interval	T value	P value
June 16	-14.15 / 16.28	0.14	0.4
July 23	-15.21 / 16.72	0.094	0.9

Table 8은 T-검정의 결과이다. 6월 16일의 결과에서 유의확률은 0.4이다. 7월 23일의 결과에서 유의확률은 0.9 이다. 두 결과 모두 신뢰구간이 0을 포함하고 유의수준 5%에서 관측 데이터와 예측데이터 사이의 동질성 가설을 지지하는 매우 유의한 결과이다.

본 논문에서는 버스정보시스템에서 수집되는 데이터에 영향을 미치는 요인을 분석하고 제약사항을 극복하는 두 개의 모형을 제시했다. 또한, 모형의 정규성을 검증하고 최종모형의 유의성 검정을 통해 모델의 예측력을 보였다.

정차시간과 통행속도 예측의 정확도를 높이기 위해 정차시간 모델링에는 일반화 가법모형을 사용하고, 통행속도 예측은 은닉 마르코프 모델을 사용했다. 일반화 가법모형은 설명변수의 영향력을 고려하고, 정류장별 테이블의 구성이 필요하지 않은 정차시간 패턴에 적합한 모델임을 보였다.

은닉 마르코프 모델은 관측데이터의 은닉된 상태의 요소 즉, 교통흐름과 신호주기 요소를 포함성분으로 가정하고 데이터의 순서 확률을 계산하여 은닉상태를 도출하는 모델로 교통흐름과 신호주기에 따른 오차를 줄이는데 매우 효과적인 모델임을 보였다.

본 논문에서 사용된 데이터는 전라남도 순천시의 버스정보 시스템 데이터이다. 출/퇴근시간대의 영향력, 신호주기, 버스전용차선 운행여부 등의 조건이 비슷한 인근지역(광양, 여수 등)의 데이터에도 유의한 결과를 도출하는지에 대한 연구가 필요하다.

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] H. G. Kim, C. Y. Park, C. S. Shin, Y. Y. Cho, and J. W. Park, “Time Series Analysis for Traffic Flow Using Dynamic Linear Model,” *The KIPS Transactions on Computer and Communication Systems*, Vol.6, No.4, pp.179-188, 2017.

[3] C. Y. Park, H. G. Kim, C. S. Shin, Y. Y. Cho, and J. W. Park, “Arrival Time Estimation for Bus Information System Using Hidden Markov Model,” *The KIPS Transactions on Computer and Communication Systems*, Vol.6, No.4, pp. 189-196, 2017.

[4] 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, Feb., 2010.

[5] Liu. H, “Generalized Additive Model. Ph.D.” Thesis, University of Minnesota Duluth, Duluth, MN, USA, 2008.

[6] M. J. Choo, “A Study on the Application of Generalized Additive Model in predicting customer churn of the mobile phone company,” Master thesis, Ewha Womans Univ., Korea, 2011.

[7] L. R. RABINER, “A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” *Proceedings of the IEEE*, Vol.77, No.2, Feb., 1989.

[8] I. Visser. “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, Jul., 2011.

[9] 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, Jun., 2009.

[10] 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, Oct., 2009.

[11] 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, Jan., 2009.

[12] Y. Y. Lee, “A Study on Estimate to Link Travel Time Using Traveling Data of Bus Information System,” *Journal of Korean Society of Transportation,* Vol.30, No.3, pp.241-246, 2010.

[13] B. Portugais and M. Khanal, “State-Space Models With Kalman Filtering for Freeway Traffic Forecasting,” *International Journal of Modern Engineering*, Vol.15, No.1, pp.11-14, 2014.

[14] G. Petris, and S. Petrone, “State Space Models in R,” *Journal of Statistical Software*, Vol.41, No.4, pp.1-25, 2011,

[15] G. Petris, S. Petrone, and P. Campagnoli, “Dynamic Linear Models with R,” Springer Science Business Media, 2009.

[16] Indrabayu, R. Y. Bakti, I. S. Areni, and A. A. Prayogi, “Vehicle detection and tracking using Gaussian Mixture Model and Kalman Filter,” *2016 International Conference on Computational Intelligence and Cybernetics*, pp.115-119, Nov., 2016.

[17] C. Zeng and W. Li, “Application of Extended Kalman Filter for tracking high dynamic GPS signal,” *2016 IEEE International Conference on Signal and Image Processing (ICSIP)*, pp.503-507, Aug., 2016.

박 철 영

e-mail : naksu21@gmail.com

2010년 순천대학교 정보통신공학과(공학사)

2012년 순천대학교 정보통신공학과 (공학석사)

2012년~현 재 순천대학교 전기·전자·정보통신공학부 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·정보통신공학부 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 : yycho@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