의사결정 학습 모델 기반 교통카드 데이터 하차 정류장 추정 모델 연구

A Study of Estimating the Alighting Stop on the Decision Tree Learning Model Using Smart Card Data

유 봉 석* · 추 상 호**

* 주저자 및 교신저자 : 홍익대학교 도시공학 박사과정
** 공저자 : 홍익대학교 건설도시공학부 교수

Bongseok Yoo* · Sangho Choo**

* Dept. of Urban Planning, Univ. of Hongik
** Dept. of Urban Planning, Univ. of Hongik
† Corresponding author : Bongseok Yoo, yoobs77@gmail.com

Vol.18 No.6(2019)
December, 2019
pp.11~30

pISSN 1738-0774
eISSN 2384-1729
https://doi.org/10.12815/kits.2019.18.6.11

Received 27 January 2019
Revised 4 March 2019
Accepted 20 December 2019

요 약

교통카드 데이터는 다양한 대중교통 통계 지표 산출, 정책 및 평가를 위한 자료로 활용되어 그 활용범위가 상당히 높다. 그러나 교통카드 데이터 내 주요 문제점은 하차 정류장에서 태그를 안 하고 하차하는 경우가 대부분으로 이는 교통카드 이용자의 불완전한 OD 통행 자료로 활용범위에 있어 한계가 있다. 본 연구는 의사결정 모델 기반 교통카드 데이터 하차 정류장 추정 방법을 적용한 결과 오차 범위 2개 정류장 이하에서 하차 정류장 추정 정확도는 89.7%으로 분석되었다. 이를 통하여 교통카드 데이터의 불완전성을 해소함으로써 다양한 대중교통 분석 및 평가 등에 대한 기초 자료로 활용 될 수 있을 것으로 판단된다.

핵심어 : 대중교통, 교통카드 데이터, 트립체인, 의사결정, 학습 모델

ABSTRACT

Smartcards are used as the basic data for utilizing the various transportation policies and evaluations, etc. and provided the transportation basic statistics index. However, the main problem of the smartcard data is that the most of users do not take the alighting tag at the stop, so there is a limit to the scope of use for the total O-D trip data because incomplete O-D traffic data of transportation card users. In this study, a decision tree of learning model is estimated for the alighting stop of smartcard users. The model estimation accuracy in range less than 2 stops interval was 89.7% on average. By eliminating the incompleteness alighting stop of smartcard data through this model, it is expected to be used as the basic data for various transportation analyses and evaluations.

Key words : Public transportation, Smartcard data, Trip chain, Decision tree, Learning model

Ⅰ. 서 론

교통카드 데이터는 현재 대중교통 현황분석, 대중교통 노선개편 및 이용수요 추정 등 다양한 대중교통 정책 및 평가를 위한 자료로 활용되고 있으며, 공공기관 및 지자체에서는 이러한 교통카드 데이터를 활용한 다양한 대중교통 통계 분석 지표를 제공하고 있다. 특히, 교통카드 데이터는 대중교통 이용자의 통행분석과 대중교통 노선체계 개편 등에서 활용도가 높으며, 대중교통 이용자의 개인별 통행경로를 파악할 수 있는 장점이 있다. 일반적으로 교통카드 데이터는 이용자의 완벽한 승차 및 하차 정류장 데이터가 수집되었을 때 데이터로서의 활용성이 높을 수 있지만, 현재 수도권을 제외한 대부분의 광역시 및 시/군에서는 하차 정류장 데이터가 누락된 불완전한 데이터가 수집되고 있다. 이러한 하차 정류장 누락 문제는 대중교통 이용자가 하차 시 교통카드를 버스 단말기에 태그를 안 하고 하차하는 경우에 발생된다. 따라서 불완전한 교통카드 데이터는 활용성이 낮아지며 그에 따른 데이터 샘플 수 또한 적어짐에 따라 개별 이용자의 정확한 통행경로를 분석하는데 한계가 있다.

Shin et al.(2016)은 부산광역시 교통카드 데이터를 활용하여 대중교통 이용자의 통행경로를 시공간적으로 연결하여 트립체인(Trip Chain, 통행경로 또는 통행사슬)을 구성하고, 결측된 하차 정류장에 대하여 통행의 승차지점 또는 최초 승차지점이 속한 교통존으로 하차 정류장을 추정하였다. 추정된 하차정류장 데이터 기반 대중교통 기종점 분석을 수행 하여 대중교통 이용자의 기종점 통행량을 추정하였다.

본 연구에서는 교통카드 데이터 상의 하차 정류장 결측 유형별 하차 정류장 추정 방법과 오차 검증 결과를 제시하고자 한다. 이를 위하여 교통카드 데이터 상의 하차 정류장 결측 유형을 정의하였으며, 과거 일주일 이력데이터와 분석 당일 데이터 기반 의사결정 트리 학습 방법을 구현하여 교통카드 데이터 상의 하차 정류장을 추정하였다. 추정된 하차 정류장은 실 하차 정류장 데이터와 비교 분석을 통한 하차 정류장 추정 오차 검증을 수행하였다.

본 연구의 수행 절차는 Ⅱ장 교통카드 데이터 관련 기존 연구 검토 및 데이터 구조 분석, Ⅲ 장 분석 대상 지역 및 교통카드 데이터 하차 정류장 결측 유형, 교통카드 하차 정류장 추정 모델 및 정확도 분석 결과, Ⅳ 연구 내용 요약 및 향후 계획 순으로 제시하였다.

Ⅱ. 선행연구

대중교통 분석 및 정책 수립에 있어서 교통카드 데이터에 대한 중요성이 높아짐에 따라 교통카드 데이터에 대한 국내외 다양한 연구가 진행되었다. 교통카드 데이터를 활용한 다양한 분석을 위해서는 승하차 정보가 모두 포함된 데이터가 필요하기 때문에 하차 정류장 결측 데이터에 대한 보정 및 추정 방법에 대한 연구의 중요성이 높아졌으며, 이를 위하여 국내외적으로 교통카드 이용자의 시・공간 및 통행 패턴을 활용한 연구가 수행되었다.

Barry et al.(2002)은 미국 뉴욕시 교통카드 데이터의 승차 시각 및 위치 자료를 활용하였다. 이용자별 1일 통행 경로 상 통행의 하차 지점은 이전 직전 승차 통행의 종점 부근이며, 1일 마지막 통행의 종점은 당일 최초 통행의 기점 인근이라는 가정을 기반으로 교통카드 상의 트립체인 통행 데이터를 활용한 연구를 수행 하였다. 이 연구에서는 교통카드 데이터 상의 개별 트립체인 통행에 대한 하차 정류장 추정을 통하여 지하철 역간 추정 O/D 통행과 통행조사(travel diary survey)의 O/D와 비교 분석 결과 약 90%가 일치하는 것으로 연구되었다.

Zhao et al.(2007)는 자동요금징수시스템을 통해 수집된 교통카드 이용자의 개별 통행 자료 중 하차 정류장 결측이 발생한 통행 데이터를 보정하여 철도 이용자의 총 O/D를 추정하였다. 이 연구에서는 철도 이용객의 하차 역사에 대한 보정 데이터로 개별 트립체인 통행 자료를 활용하였으나, 기존 연구와 다르게 철도와 버스 간 환승 통행 시 발생한 하차 결측 지점에 대한 추정 연구를 수행하였다. 개별 통행에서 환승 및 하차 정류장 추정을 위하여 환승거리, 승차 및 하차 정류장 간 연계거리 범위를 도보 이동 시간 기준인 5분 이내 도달 가능한 0.4km로 가정하였다. 이를 통하여 미국 시카고의 2004년 1월에 수집된 교통카드 데이터에 대한 분석 결과, 전체 통행의 71.2%의 하차 정류장 추정이 가능하였으며, 하차 정류장 추정을 못 한 28.8% 데이터의 대부분(18.6%)은 일일 1회 단일 통행인 것으로 분석되었다.

Tréanier et al.(2007)은 하나의 통행은 이전 통행의 목적지와 연관 있으며, 하루의 마지막 통행은 첫 통행의 승차 지점으로 회귀한다는 가정을 기초로 통행 O/D 모형인 TOOM(Transportation Object-Oriented Modeling) 방식을 활용하였다. 이 연구에서는 Network Object, Operations Object, Administrative Object 및 Demand Object 등을 주요 구성요소로 설정하였으며, 개인별 통행 경로인 통행사슬 구조에서 연속된 대중교통 통행에서 이전 통행의 하차 지점과 다음 통행 승차 지점 간 연계는 통행 종점 2km 이내에서 발생한다는 전제를 기초로 Sociétéde transport de l'Outaouais(STO) 지역을 대상으로 버스 하차 정류장을 추정하였다. 2003년 7월 및 10월 교통카드 데이터 상의 하차 정류장을 추정한 결과 66%의 추정 결과를 제시하였으며, 하차 정류장을 추정하지 못한 34%의 결과는 대부분 단일 및 불규칙적인 통행이며 대부분 비첨두시에 발생하는 것으로 분석되었다.

Barry et al.(2009)는 대중교통(지하철, 시내 및 시외버스, 여객선, 트램 등)에 대하여 하차 정류장 추정 모형을 제시하였으며, 교통카드 데이터 상의 각 수단별 운행 정보, 이용자 승차 수단 및 위치 등을 기반으로 모형을 추정하였다. 대부분의 이용자 하차 정류장은 다음 승차 정류장과 가까운 정류장에서 발생하며, 마지막 목적지 하차 정류장 또한 첫 통행의 출발지와 가장 가까운 정류장에서 발생한다는 가정을 기반으로 뉴욕시의 2004년 4월 19부터 5월 2일 데이터를 구축하였으며, 하차 정류장 추정을 통한 존 기반의 O/D 추정이 가능한 프로그램을 구현하였다.

Munizaga et al.(2012)는 대중교통 통행 O/D 산정을 위하여 교통카드 데이터의 하차 정류장 추정 모형을 제시하였으며 하차 정류장 추정을 위하여 정류장 거리뿐만 아니라 시간 요소를 고려하였다. 잠재적 하차 정류장 선택 거리는 1km이며, 도보 이동 시간이 최소가 되는 정류장을 선택하는 추정 모델을 활용하였다. 2009년 3월과 2010년 6월의 칠레 산티아고시 교통카드 데이터 상의 하차 정류장을 각 분석 기간별 80.8%와 83.0% 정도 추정하였으며, 추정된 하차 정류장 결과를 이용하여 통행 O/D를 비교 분석한 결과 각 분석 기간별 통행 O/D가 하차 정류장 추정을 통하여 산출된 통행 O/D간 분포와 유사성이 있음을 확인하였다.

He et al.(2015)는 결측된 하차 정류장은 정류장 순서와 교통카드의 정류장 지점(위도 및 경도)과의 연관성을 기반으로 이용자의 목적지는 통행의 연속성을 가지며, 마지막 목적지는 첫 번째 통행이 발생한 곳으로 회귀한다는 통행 목적지 결정에 대한 가정을 수립하였다. 잠재적 하차 정류장 선택 및 추정을 위하여 Kernel 밀집 방법을 사용하였으며, 오스트리아 브리스본 교통카드 데이터를 활용하여 승차 정류장과 잠재적 하차 정류장간 거리에 따른 하차 정류장 추정 결과에 대한 신뢰성 분석을 수행하였다. 분석결과 하차 정류장 선택의 추정 공간적 범위를 0~400m로 설정 시 하차 정류장 추정 정확도가 79.17%로 가장 높게 나타나 하차 정류장 추정을 위한 공간적 범위를 400m로 제안하였다.

He and Tréanier(2015)는 캐나다 퀘벡 주 가티뉴시의 2009년 10월 교통카드 데이터를 활용하여 하차 정류장 추정을 위한 비연결 통행거리 기준을 2km로 정의하였다. 또한 교통카드 이용자의 이력자료에서 하차 정류장 리스트를 추출하고 추출된 정류장 리스트 중에서 하차가 가능한 정류장에 대하여 Kernel 밀도함수를 산출 후 최종 하차 정류장을 추정하였다. 분석결과 과거 이력 자료 기반의 Kernel 밀도함수를 적용한 하차

추정률의 기존의 하차 정류장 추정률인 약 80% 수준을 넘어 약 91%까지 증가한 것으로 나타났다.

Alsger et al.(2016)는 대중교통 통행 O/D 추정에 대한 방법 및 검증을 위한 프로그램을 구현하였다. 이 연구에서는 교통카드 데이터와 대중교통 운행데이터를 활용하여 하차 정류장을 추정하는 연구 방법을 제시하였다. 기존의 하차 정류장 추정 시 활용했던 승차 정류장 기준 잠재적 하차 정류장간 거리에 실제 승차 시간 및 대중교통 운행 시간을 고려하였다. 이는 기존 모델 개선을 위하여 경험적 조사 및 분석을 수행한 것으로 잠재적 하차 정류장 추정 시 발생할 수 있는 하차 정류장 역전 현상 및 방향성 문제를 해결하였다.

Park et al.(2008)은 2007년 3월에 수집된 서울특별시 교통카드 자료를 분석하여 전체 데이터 중 하차 정류장 결측이 6.2-6.7% 수준임을 분석하였으며, 결측 보정을 위한 방법으로 개인별, 노선별, 총량적 보정 방법 등 총 3가지를 제시하였다. 이 중 개인별 보정방법과 노선별 보정방법에 대한 하차 정류장 보정 결과를 제시하였다. 개인별 보정방법을 통한 하차 정류장 보정 정확도에 대한 RMSE 및 MAPE 오차율은 각각 1.06 및 6.44, 노선별 보정방법의 RMSE 및 MAPE 오차율은 각각 0.75 및 1.96으로 분석되었다. 그러나 개인별 통행특성에 대한 패턴 분석이 용이하지 않았으며, 서울시 1개의 샘플 노선에 대한 분석으로 신뢰성 및 적용성에 있어서 연구의 한계가 있음을 제시하였다.

Cho et al.(2015)는 청주시 2013년 9월 교통카드 데이터를 활용하여 텍스트마이닝 기법 중 잠재 디리클레 할당법(latent Dirichlet allocation, LDA)을 이용하여 시내버스 이용자들의 이동패턴 분석을 통한 하차위치를 추정하였다. 여기서 시내버스 이용자들의 하차위치는 교통카드 원시자료와 버스정보시스템의 DB를 비교하여 시간차가 적은 노선에 정류장 매칭 후 통행사슬(트립체인)을 생성하여 각 통행자의 하차위치를 추정하였다. 연구 결과 전체 교통카드 데이터의 약 68.4%의 하차 위치를 추정하였으나, 교통카드 이용자의 탑승 노선 정보에 대한 정확성 및 이용자 통행에 대한 공간적 연계성을 고려하지 않은 한계가 있다.

Shin et al.(2016)은 대중교통 이용자가 하루 동안 발생시킨 n개의 대중교통 통행들이 서로 연계되어 있다면 그 통행자의 i번째 통행의 종점과 i+1번째 통행의 기점은 매우 근접해있을 것이며, 통행자의 최종 통행(n번째 통행)의 종점과 최초 통행의 기점은 매우 근접하다는 기준을 설정하였다. 교통카드 개별 이용자의 통행사슬 구조를 이용하여 통행 간 승차 및 하차 지점의 연계 거리가 수락할만한 거리 내 존재하는 위치를 기반으로 통행의 하차 지점을 추정하였다. 2014년 10월 주중 부산시 교통카드 데이터를 활용하여 연구 분석한 결과, 개인별 통행사슬 구조를 이용한 통행 하차 지점 결측 비율은 적용 전 71.40%(718,915통행)에서 21.74%(218,907통행)로 감소하였으나, 통행횟수가 1회인 통행에 대해서는 모형 적용이 불가능한 한계가 있다.

Yoo et al.(2017)은 교통카드 데이터의 기반정보 중 교통카드 이용자의 승차 및 하차 정보를 이용하여 버스 노선별 정류장 정보 생성 방법을 제시하였다. 교통카드 데이터 상에 노선별 정류장 정보가 없는 경우 또는 교통카드 데이터와 노선별 정류장 정보가 상이한 경우에 대하여 교통카드 데이터만을 활용하여 노선별 정류장을 생성하는 방법이다. 대전시 2017년 10월 주중 교통카드 데이터 상의 3개 노선에 대하여 노선별 정류장 정보를 생성하였으며, 생성된 교통카드 상의 노선별 정류장 정보와 BIS 정보를 비교 분석한 결과 81%의 노선 정류장 매칭 정확도를 도출하였다.

기존 선행연구에서는 교통카드 데이터를 이용한 다양한 하차 정류장 추정 방법 및 노선별 정류장 정보 생성 방법을 제시하였다. 그러나 교통카드 데이터 하차 정류장 추정 시 1일 통행 횟수(환승포함)가 2회 이상인 통행에 대해서만 하차 정류장 추정 연구 결과만을 제시하고 있어, 전체 통행(1일 단일 통행 및 1일 2회 이상 통행)에 대한 연구 결과를 제시하시 못한 한계를 보여주고 있다. 본 연구에서는 이러한 한계점을 보완하기 위하여 1일 단일 통행을 포함한 전체 통행 수로 연구범위를 확장하였다. 이를 위하여 교통카드 데이터 상에서 발생한 다양한 하차 정류장 결측 유형을 분석하였으며, 하차 정류장 결측 유형에 따른 하차 정류장 추정 방법과 연구 분석 결과를 제시하였다.

Ⅲ. 데이터 분석 및 방법론 선정

1. 교통카드 데이터 분석

본 연구에서는 대전광역시 교통카드 데이터를 활용하였다. 대전광역시 교통카드 데이터 구성은 교통카드 가상ID, 교통수단 코드 노선ID 및 노선명, 승차 및 하차 일시, 승차 및 하차 정류장 등 총 23개로 구성되어 있다. 테이블 구성은 <Table 1>과 같으며, 노선별 정류장 정보는 적용일시, 노선 ID 및 명, 정류장 ID 및 명, 정류장 순서, 정류장 경위도 좌표 및 정류장간 거리로 구성되어 있으며 <Table 2>와 같다.

<Table 1> Smart Card Data Structure of Daejeon City

Division	Contents
Card No.	2Aehppf+k4Z9Vm
Transportation code	105
Transaction ID	65
Line ID	11110609
Line name	No. 13
Transportation operator ID	111520010
Transportation operator name	Yeongseong Co., Ltd.
Vehicle ID	111753634
Vehicle registration number	Seoul 75Sa3634
Vehicle Departure time	20150427171220
Vehicle End time	20150427180809
Date of boarding	20150427171816
Boarding stop ID	9031004
Boarding stop name	POSCO The Shop Apartment
Date of getting off	2.01504E+13
Alighting stop ID	9009427
Alighting stop name	daejeon station
Transfer number	0
User code	1
Passenger number	1
Total travel fee	750
Total travel distance	359
Total travel time	66

<Table 2> Base Information

Division	Contents
Date and Time	20171001000000
Line ID	30300001
Line name	1
Stop ID	8002941
Stop Name	Hanbit Apt
Stop sequence	2
Latitude	127.353
longitude	36.36286
Distance between Stops	353

2. 대중교통 환승체계 및 통행 유형

대전광역시 대중교통 환승체계는 총 3회 및 30분이며, 일부 노선은 60분까지 환승 시간이 적용된다. 특히 대전광역시는 환승요금이 다음 승차 시 정산되는 형태이며, 대중교통 환승체계는 <Table 3>과 같다.

<Table 3> Transfer system feature of Daejeon City

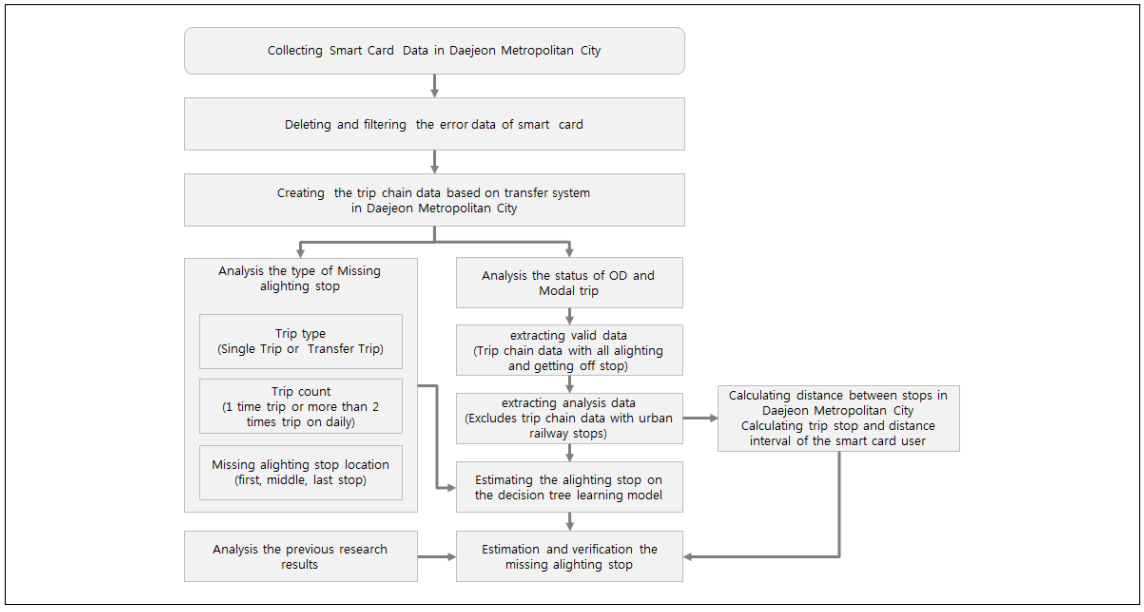
Division	Contents
Rate system	• If the charge for each public transport is different, the difference is subtracted from the higher fare minus the lower fare. • Free (discounted) transfers between the same routes are not available.
Transfer Number	• 3 times (total 4 times in Boarding)
Transfer available time	• Within 30 minutes of alighting : Weekday average dispatch interval is 15 minutes, 36 routes • Within 60 minutes of alighting :From the average interval of 16 minutes on weekdays, 61 routes

대중교통 이용자의 통행은 일반적으로 1일 통행 횟수에 따라 단일통행과 복수통행으로, 환승 유무에 따라 단일통행과 환승통행으로 분류 할 수 있다. 특히 대전광역시는 균일요금제(버스 탈 때 기본요금에서 추가 요금 없이 이용) 및 환승 시 추가 환승 요금이 적용되며, 환승 승차 시 환승 요금을 지불하기 때문에 대부분의 이용자들이 최종 하차 정류장에서 교통카드 태그를 안 찍는 상황으로 인하여 하차 정류장에 대한 결측이 발생한다. 하차 정류장 결측이 발생된 유형이 1일 단일통행인 경우에는 이용자의 통행을 참조 할 수 있는 데이터가 없기 때문에 과거 이력 자료를 활용 할 수 밖에 없으며, 복수 통행인 경우에는 1일 통행 횟수가 2회 이상이기 때문에 당일 발생된 통행 데이터를 활용 할 수 있다. 따라서 교통카드 이용자의 1일 통행을 단일통행과 복수통행으로 구분할 필요성이 있다.

3. 교통카드 하차 정류장 추정 방법 및 모델

본 연구는 대전시 교통카드 데이터 상의 결측된 하차 정류장에 대하여 보정 및 추정을 통하여 이용자별 통행사슬 구축하는데 목적이 있다. 이를 통하여 이용자별 OD 통행 구축을 통하여 다양한 대중교통 분석에 활용한다.

이를 위하여 본 연구에서는 대전시 교통카드 데이터를 수집하고 이를 트립체인 데이터로 변환 및 오류 데이터 정제를 통한 유효 데이터를 구축한다. 또한 노선별 정류장 정보 등 기초 데이터를 구축한다. 구축된 트립체인 데이터를 활용하여 하차 정류장 결측 유형, 대전시 교통카드 기반 OD 및 Modal(수단) 통행 분석을 수행한다. 결측된 하차 정류장 추정을 위하여 의사결정 트리 학습 모델을 활용하고 추정된 결과를 검증한다. 본 연구의 데이터 분석, 하차 정류장 추정 및 검증에 대한 연구 방법은 <Fig. 1>과 같다.

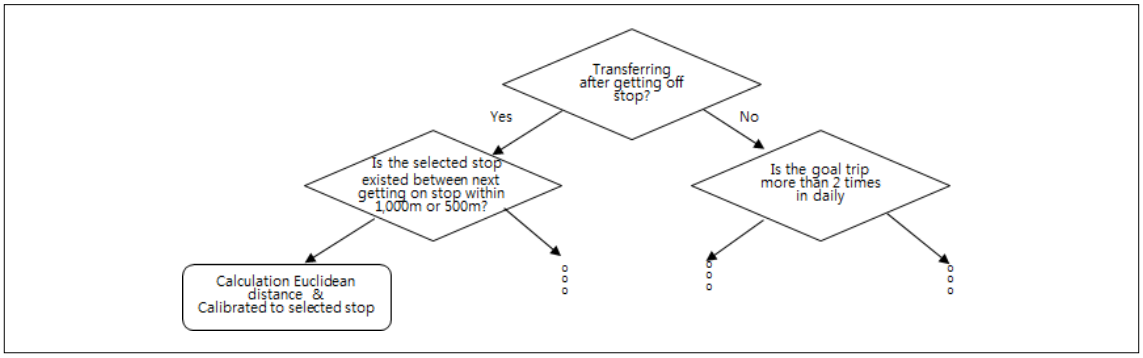


<Fig. 1> Procedure for the estimation of the missing alighting stop based on smart card data

본 연구에서 교통카드 데이터 하차 정류장 추정 모델은 의사결정 트리 학습(Decision Tree Learning) 모델을 적용하였다. 의사결정 트리 학습 모델은 데이터 마이닝에서 일반적으로 사용되는 방법론으로 입력 변수

를 기반으로 목표 변수의 값을 추정하는 모델이다. 특히 지도 분류 학습에서 가장 유용하게 사용되고 있는 기법 중 하나로 모든 속성들이 유한한 이산값들로 구성된 정의역을 가지고 있으며, 분류를 단일 대상 속성으로 지니고 있다고 간주한다. 분류별 정의역의 각 원소들은 클래스라고 불리며, 결정 트리 또는 분류 트리의 모든 내부 노드들은 입력 속성이 일대일로 대응된다.

의사결정 트리 학습 모델은 다른 머신러닝 알고리즘에 비해 직관적으로 분류 기준을 파악할 수 있다는 장점이 있으며, 이력자료가 늘어나면 정확도 또한 향상되는 장점이 있다. 이러한 모델은 교통소통정보 예측 등 다양한 입력 변수 기반 이력 및 패턴 데이터를 활용한 통행시간 예측 정보, 교통사고 위험요소 기반 교통사고 위험성 예측 분석 등 미래 상황 추정 및 예측에 활용되고 있다.



<Fig. 2> Process of the decision tree learning model for estimating the alighting stop

하차 정류장 추정 방법인 의사결정 트리 학습 모델에 사용되는 속성정보는 누락된 하차정보의 통행유형 즉 단일통행(환승통행이 없는 1회 통행) 및 환승 통행(첫 출발지에서 목적지까지 2회 이상 통행이 발생된 통행)으로 구분되며 누락된 하차정보 이후 발생된 통행에 대한 환승 유무를 판단하며, 이용자의 출발 및 도착지 간 통행 기준 하차정보 누락이 발생한 통행 위치 즉 n번의 통행(환승 횟수는 n-1회) 시 첫 번째 통행 정류장, 중간 통행 정류장(2~n-1번째), 마지막 통행 정류장인지를 구분한다. 또한 교통카드 이용자의 일일 통행유형 즉 단일통행 및 환승통행과 이에 대한 각 통행에 대한 횟수 등을 포함한다. 모델의 파라미터 변수로는 정류장 정보(위도 및 경도 포함), 노선별 기종점 방향 및 정류장 순서, 정류장간 직선거리, 노선별 정류장 승차 인원 이력 정보로 구성된다.

본 연구에서는 교통카드 데이터를 하나의 통행(Trip)으로 연결하기 위하여 트립체인(통행 기종점) 데이터를 구축하였다. 이를 통하여 하차 정류장의 결측이 환승 중간에 발생한 것인지, 마지막 목적지의 하차 정류장에서 발생한 것인지를 분석 할 수 있다.

교통카드 데이터 내 1일 단일 및 2회 이상 복수통행, 출발지에서 목적지까지 도착하는 통행에 대한 단일 및 환승통행을 기준으로 하차 정류장 결측 유형을 아래와 같이 9가지 Case로 구분하였다.

- CASE 1 : 1일 1회 단일통행, 결측 하차 정류장 위치가 정류장 0,1,2,3~N(환승 포함)번째, 마지막 통행인 경우 및 환승이 없는 통행
- CASE 2 : 1일 1회 단일통행, 결측 하차 정류장 위치가 정류장 0,1,2,3~N(환승 포함)번째, 마지막 통행인 경우 및 환승이 있는 통행(환승 내 하차 정류장 결측 발생)
- CASE 3 : 1일 1회 단일통행, 결측 하차 정류장 위치가 정류장 N번째(환승 N번째 및 환승 발생 시 마지막 환승 노선에서 발생 한 경우), 마지막 통행인 경우 및 환승이 없는 통행

- CASE 4 : 1일 2회 이상 통행, 결측 하차 정류장 위치가 정류장 0,1,2,3~N(환승 포함)번째, 마지막 통행인 경우 및 환승이 없는 통행
- CASE 5 : 1일 2회 이상 통행, 결측 하차 정류장 위치가 정류장 0,1,2,3~N(환승 포함)번째, 마지막 통행인 경우 및 환승이 있는 통행(환승 내 하차 정류장 결측 발생)
- CASE 6 : 1일 2회 이상 통행, 결측 하차 정류장 위치가 정류장 0,1,2,3~N(환승 포함)번째, 첫번째 또는 중간 통행인 경우(1~N-1)이며 환승이 없는 통행
- CASE 7 : 1일 2회 이상 통행, 결측 하차 정류장 위치가 정류장 0,1,2,3~N(환승 포함)번째, 첫번째 또는 중간 통행인 경우(1~N-1)이며 환승이 있는 통행
- CASE 8 : 1일 2회 이상 통행, 결측 하차 정류장 위치가 정류장 N번째(환승 N번째 및 환승 발생 시 마지막 환승 노선에서 발생 한 경우), 마지막 통행인 경우 및 환승이 없는 통행
- CASE 9 : 1일 2회 이상 통행, 결측 하차 정류장 위치가 정류장 N번째(환승 N번째 및 환승 발생 시 마지막 환승 노선에서 발생 한 경우), 첫 번째 또는 중간 통행인 경우(1~N-1)및 이후 환승이 없는 통행

결측 하차 정류장 추정 시 하차 정류장 결측이 발생된 통행 행태를 우선 분류한다. 1일 기반 단일 통행 또는 복수 통행인지를 판단하며, 1일 단일 통행인 경우에는 과거 이력 데이터 상에서 노선 및 승차 위치의 유사성을 기반으로 이력 통행 정보 상에 하차 정류장을 선정한다. 만약 이력 통행에서도 하차 정류장이 없는 경우에는 당일 이용 노선의 정류장 인원에 대한 비율을 산정하여 최빈 하차 정류장을 선정한다. 당일 복수 통행인 경우에는 우선 하차 정류장 결측 위치가 환승 시 발생한 것인지, 마지막 정류장(최종 목적지 하차 정류장)에서 발생한 것인지로 분류한다. 이는 환승통행 중간에 발생한 경우 다음 승차 정류장을 기준으로 하차 가능 정류장을 판단하지만 마지막 정류장인 경우는 다음 승차가 없기 때문에 당일 첫 승차가 발생한 지점으로 회귀한다는 가정을 기반으로 첫 승차 정류장을 중심으로 하차 가능 정류장을 판단할 수 있다.

일반적으로 하차 정류장 결측이 마지막 통행이 아닌 1~n+1회 통행에서 발생한 경우에는 하차 가능 정류장에서 최종 하차 정류장 선택은 다음 승차 정류장을 기준으로 하차 가능 정류장간 거리와 승차 인원수를 고려한 유클리드 거리 계산을 통하여 최소 유클리드 거리 값의 정류장을 선택한다.

단 환승 중간에 발생한 하차 정류장은 다음 승차 정류장과 하차 가능 정류장간 최소 거리의 정류장을 선택하며, 1일 마지막 통행의 하차 정류장 결측의 경우는 우선적으로 첫 승차 정류장을 기준으로 하차 가능 정류장을 선택 후 최소 거리의 정류장을 하차 정류장으로 선택하며, 하차 가능 정류장이 일정 범위 내 없을 경우에는 이용 노선의 1일 정류장별 이용량을 기준으로 가장 많이 이용한 정류장을 선택한다. 하차 정류장 선택 시 방향성과 회차 지점에 대해서는 모두 고려한다는 가정을 포함하고 있다.

하차 정류장 결측 Case별 세부 추정 방법은 다음과 같다.

- CASE 1 & 3 : 이력데이터에서 통행 이력을 검토하며 1차 이력데이터에서 승차정류장 매칭 후 하차 정류장 추출 이후 유클리드 거리 계산, 만약 이력 하차 정류장이 없을 경우 분석 데이터 노선에서 승차 인원수 가중치를 설정하여 유클리드 거리 계산
- CASE 2 : 교통카드 정산사업자에서 일부 환승 내 하차 정류장이 결측된 형태로 트립체인 내 다음 승차 정류장 기반 분석 데이터 노선의 정류장 유클리드 거리 계산
- CASE 4 & 8 : 마지막 통행이지만 분석 대상 날짜에서 통행이 존재하며, 1일 2통행 이상 발행한 데이터로 마지막 통행의 경우 가정으로 귀가하는 기준을 1차 적용, 그러나 이벤트 및 예외적인통행(다른 노선 승차)이 발행하는 경우 이력 데이터 마지막 정류장과 인접한 정류장 추정(1차 분석날짜에서 분석 이후 2차 이력 데이터 분석 순으로 프로세스 적용)

- CASE 5 : 1일 통행 중 이전 통행이 있는 경우로 교통카드 정산사업자에서 일부 환승 내 하차 정류장이 결측된 형태로 트립체인 내 다음 승차 정류장 기반 분석 데이터 노선의 정류장 유클리드 거리 계산
- CASE 6 & 9 : 하차 정류장 결측 데이터에 대하여 환승 유무로 구분할 수 있으며, 이전 통행과 다음 통행간 시간 차이가 많이 발생하거나 하차 및 승차 정류장의 거리가 긴 경우로 당일 분석데이터 내 다음 승차정류장과 분석데이터 노선의 정류장과의 유클리드 거리 분석 또는 이력 데이터에서 유사한 노선의 승차 정류장 매칭 후 하차 정류장 추출 이후 유클리드 거리 계산
- CASE 7 : 하차 정류장 결측 데이터에 대하여 환승 유무로 구분할 수 있으며, 각 통행의 마지막 하차 정류장이 결측된 경우로 이전 통행과 다음 통행 간 시간차이가 많이 발생하거나, 하차 및 승차 정류장간 거리가 긴 경우로 환승 데이터 분석, 선택할 정류장 반경 결정, 각 단계별 유클리드 거리 계산 후 노선 정류장 방향성을 고려하여 정류장 순서가 역전이 안 되는 정류장 선택

하차 정류장 추정 중 유클리드 거리 계산 방법은 두 점 사이의 거리를 계산할 때 쓰는 방식으로 평면 및 공간에 대한 유클리드 공간으로 정의할 수 있다. 본 연구에서는 이용자의 다음 노선의 승차 정류장과 하차 가능 정류장과의 거리, 이전 이용 노선의 정류장별 이용량에 대한 가중치를 사용한다. 정류장 유클리드 거리 계산은 정류장의 위도(X좌표)와 경도(Y좌표)를 기반으로 각 정류장간 거리를 계산하며, 통행 순서를 기준으로 다음 통행의 첫 승차 정류장을 기반으로 이전 통행의 하차 가능 정류장간 거리와 이전 마지막 이용 노선의 1일 승차 이용량에 대한 가중치를 적용하여 하차 가능 정류장별 유클리드 거리를 계산하여 최소값의 정류장을 하차 정류장으로 선정한다. 유클리드 거리 계산은 식 (1), 유클리드 거리 계산 개념도는 <Fig. 3>과 같다.

$$d = dict(d_i),\ d_i = S_a, S_{bi} = \sqrt{(P_{ax} - P_{bix}) + (P_{ay} - P_{biy})} \quad \cdots\cdots \text{식 (1)}$$

$$u = \min(d_i \times \alpha_i)$$

S_a : i번째 통행 승차 정류장

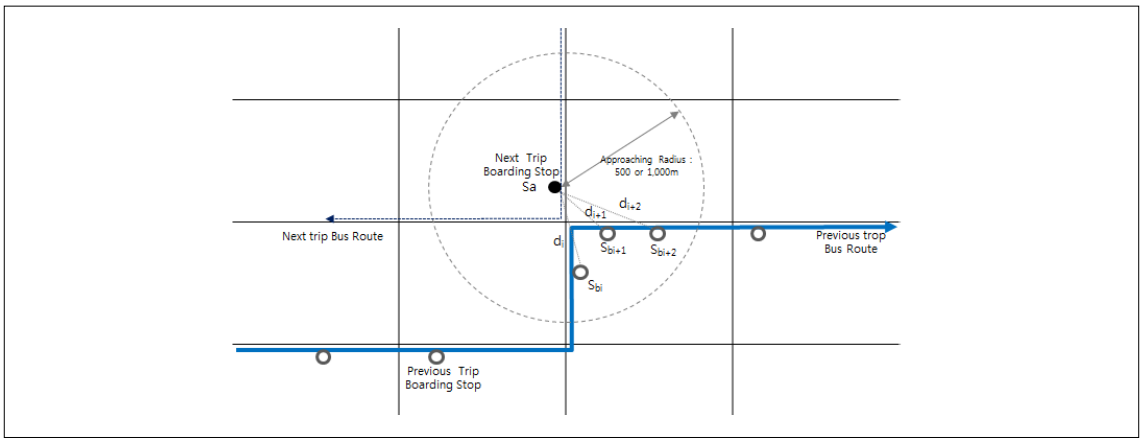
S_{bi} : 이전 노선에서 i번째 통행의 승차 정류장 반경 내 하차 가능 정류장

$S_a = (P_{ax}, P_{ay})$, $S_{bi} = (P_{bix}, P_{biy})$ 여기서 P_{bx} : 정류장 x좌표, P_{by} : 정류장 y좌표

d : 각 정류장별 거리, d_i : i번째 승차 정류장 S_a 와 하차 가능 정류장 S_{bi} 간 정류장 거리

a_i : 이용노선의 하차 가능 정류장별 승차 인원에 대한 가중치

u : 정류장별 유클리드 거리 값



<Fig. 3> Euclidean distance calculation method

본 연구는 교통카드 데이터 상에서 발생하는 다양한 결측 유형과 다음 승차 정류장이 이전 승차 정류장의 위치보다 앞에 있는 경우, 마지막 통행에서 첫 통행의 출발 정류장과 마지막 이용 노선에서 출발 정류장의 일정 반경에 정류장이 없는 경우, 1일 1회 단일 통행의 경우 등에 대하여 이력 데이터 기반 유사 노선 및 정류장 이용 데이터 및 이용 노선의 정류장별 승차 인원을 고려한 정류장 선택 확률을 적용하는 등 기존 연구와의 차별성을 두었다.

4. 하차 정류장 유효 거리 산정 방법

교통카드 데이터 내 결측 하차 정류장 추정을 위하여 다음 승차 정류장을 중심으로 이전 이용 노선 및 승차 정류장 기반 하차 정류장 추정 그룹을 선정하기 위한 유효 거리 산정 방법에 대하여 국내외 연구 사례를 검토하였다. He et al.(2015)는 호주 브리즈번시에서 수집된 40,431건의 대중교통 교통카드 기종점 자료를 이용하여 실제 하차지점과 통행사슬 구조를 이용한 추정 하차지점간 거리가 1km 이상이면 종점 추정률에 큰 변화가 없음을 밝혔다. He and Tréanier(2015)는 캐나다 퀘벡 주 가티뉴시 2009년 10월 교통카드 데이터를 Kernel 밀집 방법을 통하여 승차 정류장과 잠재적 하차 정류장간 거리를 0~400m로 제시하였다.

Alsger et al.(2016)은 호주 퀸즐랜드에서 2013년 3월에 수집된 버스, 철도, 페리 기종점 교통카드 자료를 이용하여 대중교통 통행 연계거리(allowable walking distance)를 0.4km, 0.8km,1.0km, 1.1km로 구분하여 통행종점을 추정한 후 이를 실제 통행 기종점과 비교하였으며, 통행연계거리가 0.8km 이상일 때 추정 통행량에 큰 변화가 없음을 제시하였다. 대전광역시 버스 정류장 간 거리는 평균 584m이며, 정류장간 거리는 200~500m사이가 50.99%로 가장 높으며 정류장간 거리 비율은 <Table 4>와 같다.

<Table 4> Distance between bus stops on Daejeon City

Division	Number of stop intervals	Ratio(%)
Under 200m	156	4.07
Under 200m ~500m	1,955	50.99
Under 500~700m	858	22.38
Under 700~1,000m	516	13.46
More than 1,000m	349	9.10
Total	3,834	100.00

분석된 대전광역시 정류장간 거리가 500m를 초과하는 구간이 발생할 수도 있다. 따라서 500m 이내 유효 정류장을 찾을 수 없는 제약사항을 고려하기 위하여 정류장간 간격을 1,000m까지 차등적으로 확대 적용하였다. 따라서 이전 하차 정류장 추정 그룹은 다음 승차 정류장 기준 500m 및 1,000m이며, 정류장 반경 기준은 우선 500m내 정류장을 선택 후 500m 내 정류장이 없는 경우 1km 내 정류장을 선택하도록 선정하였다. 또한 하차 정류장 추정 시 노선에 대한 기점 및 종점, 회차 등 운행 방향을 고려해야 하며, 이를 위하여 기존 교통카드 기반정보에 노선 기종점, 회차 지점 및 방향에 대한 코드를 추가하였으며 데이터 구조는 <Table 5>와 같다.

<Table 5> Start and end direction codes of the base information

Line ID	Line name	Stop ID	Stop name	Stop sequence	Latitude	longitude	Start direction	End direction	Start direction Shortest distance	End direction Shortest distance	Cumulative distance
30300001	No.1	8002736	Chongdae Agricultural College	1	127.351	36.366	1	0	10200	0	0
30300001	No.1	8002941	Hanbat APT	2	127.353	36.362	1	0	10200	0	353
⋮											
30300001	No.1	8007226	JoEun JEONGMIL	52	127.406	36.416	1	0	10200	0	23470
30300001	No.1	8007228	Cheongbuksan PARK	53	127.407	36.413	1	1	10200	11000	23787
30300001	No.1	8007227	JoEun JEONGMIL	54	127.407	36.415	0	1	0	11000	24002

5. 분석 대상 및 하차 정류장 추정 결과

1) 분석 대상 및 범위

본 연구는 2017.10.16~2017.10.22(7일) 대전광역시 교통카드 데이터를 활용하였으며, 분석 대상은 2017.10.19(목요일) 데이터(전체 데이터 수 : 458,431통행)이며, 승차 및 하차 정류장이 같은 통행, 승차 및 하차 시간이 역전된 통행, 노선별 정류장상에 승차 및 하차 정류장이 없는 통행 등도 포함되어 있다. 이들 분석자료 중 시간 및 공간상 연속성이 확보된 완전한 트립체인 데이터(O-D trips)는 213,954통행으로 나타났다. 본 연구는 버스의 하차정류장 정보를 추정하는 것이므로 이 트립체인 데이터 중 지하철 역사가 포함된 트립체인(도시철도를 한 번이라도 이용한 통행이 포함된 O-D trips, 91,736통행)은 분석에서 제외하였으며, 버스 통행으로만 구성된 122,218통행을 검증자료로 활용하였다. 검증을 위해 이들 자료의 하차정류장 정보를 제거한 후 본 연구의 하차정류장 추정 모델을 적용하여 그 결과를 실제 하차정보와 비교 하였다. 또한 실제 운행 노선에 대한 신뢰성 검증을 위하여 시내버스 4개 노선 및 광역버스 1개 노선에 대하여 버스 내 설치된 CCTV 영상 자료를 비교 분석하여 정확도 검증을 수행하였으며, <Table 6>은 본 연구에 활용된 데이터 및 검증 범위이다.

<Table 6> Data analysis object and range

Division	Contents	Note
Analysis object	Daejeon City	
Analysis date	2017.10.19(Thursday)	History data : 2017.10.16.~2017.10.22.(7일)
Total data Number	458,431	Total number of trips per day
Valid trip chain data Number	213,954	
Analysis & Verification data Number	122,218	Excludes trip chain data with urban railway stops Only bus modal OD trip data(trip chain data)
Accuracy verification number base on real bus	5 routes including city and metropolitan buses	Verification person of CCTV in Bus

대전광역시 교통카드 원시 거래 데이터를 대전광역시 환승체계 기준 이용자별 통행 트립체인 데이터로 변환하였다. 트립체인 데이터는 이용자가 출발지에서 목적지까지 이동하기 위하여 이용한 환승을 포함한 수단별로 연결한 데이터이다. 이를 Shin et al.(2016) 연구에서는 트립체인을 이용자의 통행 사슬이라 명칭 하였으며, 본 연구에서는 대전광역시의 기초 통계 분석으로 트립체인 즉 통행사슬 기반 트립 기종점(Trip Origin-Destination)

통행수와 통행사슬 내 첫 승차 수단부터 마지막 하차 수단 통행까지 환승을 포함한 수단별로 구분하여 산정한 수단 통행수를 산정하였다. 그 결과 트립 기종점 통행수는 458,431통행, 수단통행수는 579,698통행으로 분석되었으며 수단통행 중 버스 통행이 463,888 (80.0%)통행으로 가장 높았으며, 다음으로 도시철도(109,804통행) 통행이 높은 것으로 분석되었다. <Table 7>은 이용자 유형 및 통행수단별 트립 기종점 통행과 수단통행 분석결과이다.

<Table 7> Trips by mode with respect to user type

Division	O-D trips				Modal trips			
	shuttle bus	intra-city bus	BRT	urban railway	shuttle bus	intra-city bus	BRT	urban railway
senior	109	5,361	25	16,418	140	7,456	34	16,650
men of national merit	1	993	-	229	1	1,104	-	234
multi-child parent	1	288	1	1,087	1	370	1	1,101
child	118	7,303	49	352	125	8,174	52	530
adult	4512	292,088	4,623	66,649	5,407	379,298	5,501	81,041
Disabled	9	922	9	2,648	13	1,301	14	2,702
Youth	706	48,178	312	5,440	848	59,650	404	7,546
Sub Total	5,456	355,133	5,019	92,823	6,535	457,353	6,006	109,804
Total	458,431				579,698			

트립 기종점 통행은 단일 통행이 350,113통행(76.37%), 환승 통행이 108,318(23.63%)통행이며 대부분 1회 환승 통행이 가장 많은 비율을 차지하고 있다. 교통카드 데이터 내 하차 정류장 결측 비율은 전체 458,431통행 중 243,294통행(53.07%)이며, 1일 단일통행 하차 정류장 결측 비율은 단일 통행이 55.37% 및 환승통행이 56.12%, 1일 2회 이상 통행 중 하차 정류장 결측 비율은 단일통행이 51.35% 및 환승통행 55.05%이다. <Table 8>은 단일 및 환승통행별 기종점 통행 수(O-D trips)이며, <Table 9>는 1일 통행 수 기준 1회 통행과 2회 이상 기종점 통행별 하차 정류장 결측 통행 수 및 분포 결과이다.

<Table 8> O-D trips volume by the trip type in one trip

Division		O-D trips	Ratio(%)
Single Trip		350,113	76.37
Transfer Trip	Transfer 1 time	96,814	21.12
	Transfer 2 times	10,059	2.19
	Transfer 3 times	1,445	0.32
	Sub Total	108,318	23.63
Total		458,431	100.00

<Table 9> Missing stop volumes of O-D trips on daily

Division		Total trips	Missing alighting stop volumes	Ratio(%)
1 time trip on daily	Single Trip	90,394	50,054	55.37
	Transfer Trip	23,912	13,419	56.12
more than 2 times trip on daily	Single Trip	259,719	133,358	51.35
	Transfer Trip	84,406	46,463	55.05
Total		458,431	243,294	53.07

2) 교통카드 데이터 기반 이용자 이동 정류장 및 거리 분포

대전시 노선별 정류장 기반 평균 정류장간 거리는 584m이며, 대전시 교통카드 데이터 중 승하차 정류장 결측이 없는 트립체인 데이터 중에서 도시철도 수단통행(도시철도를 이용한 기종점 통행, 91,736통행)이 포함된 트립체인 데이터를 제외한 순수 버스통행만 포함된 122,218건(유효 O-D통행 수 213,954통행에서 도시철도 O-D통행 수 91,736통행을 제외한 통행 수)의 기종점 통행(O-D trips)과 161,404건의 수단통행(Modal trips)에 대하여 각각 이동 정류장 수를 분석 한 출발지에서 목적지까지 전체 통행 정류장 수는 10~20구간의 정류장 수가 가장 많았으며, 이용자의 출발지 및 목적지까지 이동하기 위한 각 버스수단(버스 노선)에 대한 평균 이동 정류장 수는 5~10구간의 정류장 수가 가장 많았다. 한 번의 버스 이용 시(교통카드 이용자가 한 번의 버스를 이용하여 이동한 통행, Modal trips)서 각 이동 정류장 구간은 10정류장 이하 약 59.6%로 분석되었으며, 교통카드 이용자의 O-D통행 및 각 버스 수단별 통행에 대한 이동 정류장 구간 통행 수 및 분포는 <Table 10>과 같다.

<Table 10> Distribution of bus stops by bus users

Division	O-D trips(using bus trip)		Modal trips(only using one bus trip)	
	Stop Interval Volume	Ratio(%)	Stop Interval Volume	Ratio(%)
Under 5 stops	19,292	15.78	41,034	25.42
Under 5~10 stops	36,449	29.82	55,182	34.19
Under 10~20 stops	43,047	35.22	47,505	29.43
Under 20~30 stops	14,630	11.97	11,552	7.16
Under 30~50 stops	7,480	6.12	5,364	3.32
More than 50 stops	1,320	1.08	767	0.48
Total	122,218	100.00	161,404	100.00

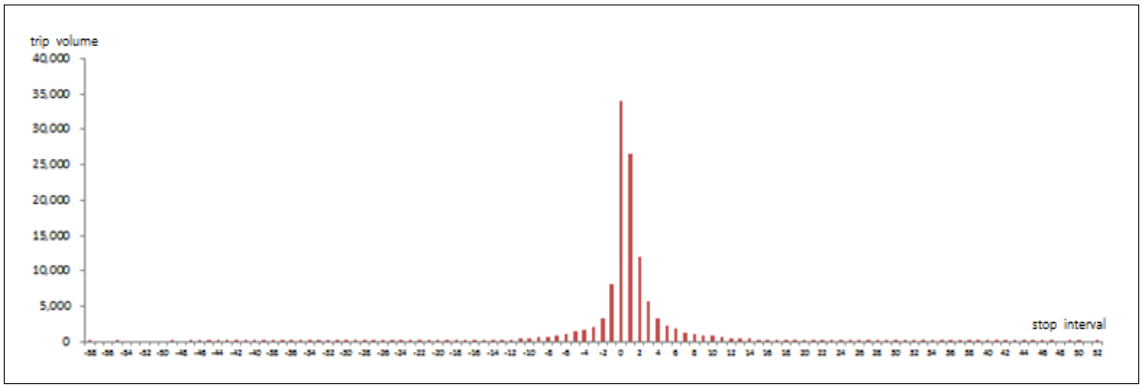
또한 교통카드 이용자의 이동 거리를 분석한 결과 출발지에서 도착지까지 이동 거리는 1,000~5,000m이하에서 가장 많은 비중을 차지하는 것으로 분석되었으며, 전체 O-D 통행에서 각 버스 수단별로 이동 거리를 분석한 결과도 1,000~5,000m이하가 가장 높았다. 이는 교통카드 이용자의 단일 및 환승 통행 구간의 거리가 유사하다고 볼 수 있다. <Table 11>은 교통카드 이용자의 통행 거리 분포이다.

<Table 11> Distribution of bus stop distance by bus users

Division	O-D trips		Modal trips	
	Distance Interval Volume	Ratio(%)	Distance Interval Volume	Ratio(%)
Under 1,000m	3,139	2.57	9,567	5.93
Under 1,000~5,000m	51,606	42.22	84,115	52.11
Under 5,000~10,000m	37,915	31.02	42,678	26.44
Under 10,000~20,0000m	21,229	17.37	18,556	11.5
Under 20,000~30,0000m	6,062	4.96	4,740	2.94
More than 30,000m	2,267	1.85	1,748	1.08
Total	122,218	100.00	161,404	100.00

3) 하차 정류장 추정 결과

하차 정류장 추정 시 트립체인 데이터 중 승차 및 하차 정류장이 단일 회차 구간 이외에 존재한 경우 즉 회차 지점을 초과하여 하차한 경우, 승차 정류장이 종점 정류장 번호로 입력된 경우 등 데이터 내 승차 정류장과 하차 정류장의 방향성이 상이한 데이터 등 교통카드 데이터 내 이상치 데이터를 제거한 유효 데이터(213,954 통행)에서 도시철도 수단통행(도시철도를 이용한 기종점 통행, 91,736통행)이 포함된 트립체인 데이터를 제외한 순수 버스통행만 포함된 122,218건(유효 O-D통행 수 213,954통행에서 도시철도 O-D통행 수 91,736통행을 제외한 통행 수)의 기종점 통행(O-D trips)에 대하여 마지막 하차 정류장을 임의로 삭제하고 이를 본 연구의 하차 정류장 결측 모형에 적용 및 R 프로그램을 활용하여 하차 정류장을 추정하였다. 임의 결측된 하차 정류장은 검증 데이터(122,218건)의 실제 하차 정류장과 비교 분석한 하차 정류장 추정 결과, 하차 정류장 추정 오차구간은 평균 2.82 정류장 수로 분석되었으며, 추정 값과 실제 정류장 간 오차 분포는 <Fig. 4>와 같다.



<Fig. 4> The stop interval distribution between the actual alighting stop and estimated alighting stop

선행 연구 결과에서 Shin et al.(2016)은 부산광역시 교통카드 자료를 활용하여 통행횟수가 2회 이상인 통행을 기준으로 하차 정류장 추정 모형 결과 75.23%의 추정 성공률을 도출하였으며, Kim et al.(2017)은 서울시와 광주광역시에 대한 교통카드 데이터 상 결측된 하차 정류장을 추정한 결과 추정 성공률은 서울시 및 광주광역시 각각 78.2%, 81.6%이며, 추정 정확도는 각각 54.25% 및 33.4%로 검증 결과를 제시하였다. 본 연구의 대전광역시 교통카드 데이터 기반 하차 정류장 결측 보정은 100%이며, 추정 정확도는 하차 정류장이 정확히 일치하는 비율은 27%, 정류장 오차 범위가 2개소 이하는 67.2%로 분석 되었다. 선행연구 결과(Shin et al., 2016; Kim et al., 2017)와 비교 시 광주광역시 보다 추정 정확도는 6.4% 감소하는 것으로 분석되었다. 본 연구에서는 통행별 제약조건을 최소화하여 하차 정류장 추정 모델에 대한 현실성을 적용하기 위하여 단일통행을 포함한 전체 통행에 대하여 결측된 하차 정류장을 추정하였다. 본 연구에서 추정된 하차 정류장 중 1일 1회 단일통행에 대한 하차 정류장 추정 오차율이 높았으며, 이는 1일 기반 교통카드 데이터 상에서 동일 이용자의 과거 통행을 참조하여 하차 정보를 추정하기 때문에 추정 오차가 상대적으로 높게 나타난 것으로 분석된다. 하차 정류장 추정 오차 분석 결과는 아래 <Table 12>와 같다.

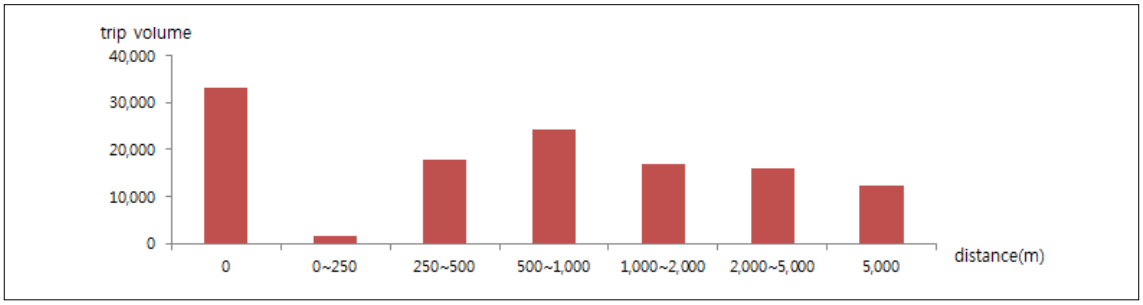
<Table 12> Stop interval between the estimated and actual volumes of alighting stop

Division		Missing trip Volume	Stop interval between the estimated and actual alighting stop						
			Trip volume /Ratio(%)	Absolute stop interval(± stop interval)					
				0 stop	1 stop	2 stop	3 stop	4 stop	5 stop	more than 5 stop
1 time trip on daily	Single Trip	41,942	Trip volume	9,525	8,924	4,606	3,196	2,454	2,042	11,195
			Ratio	22.71%	21.28%	10.98%	7.62%	5.85%	4.87%	26.69%
	Transfer Trip	16,167	Trip volume	3,637	3,833	2,091	1,178	921	740	3,767
			Ratio	22.5%	23.71%	12.93%	7.29%	5.7%	4.58%	23.3%
more than 2 times trip on daily	Single Trip	46,302	Trip volume	14,529	15,197	6,119	2,763	1,479	1,097	5,118
			Ratio	31.38%	32.82%	13.22%	5.97%	3.19%	2.37%	11.05%
	Transfer Trip	17,807	Trip volume	5,372	5,719	2,594	1,105	618	391	2,008
			Ratio	30.17%	32.12%	14.57%	6.21%	3.47%	2.2%	11.28%
Total		122,218	Trip volume	33,063	33,673	15,410	8,242	5,472	4,270	22,088
			Ratio	27.05%	27.55%	12.61%	6.74%	4.48%	3.49%	18.07%

<Table 13> Stop interval between the estimated and actual volumes of alighting stop by real moving bus stop

Real moving bus stop	missing trip Volume	stop interval error(number & interval)	volume	Ratio(%)
Under 5 stops	19,292	interval ±0 stop	5,840	4.78
		under interval ±1~±2 stop	5,642	4.62
		under interval ±2~±5 stop	4,527	3.7
		under interval ±5~±10 stop	1,853	1.52
		more interval ±10 stop	1,430	1.17
Under 5~10 stops	36,449	interval ±0 stop	10,499	8.59
		under interval ±1~±2 stop	10,700	8.75
		under interval ±2~±5 stop	9,577	7.84
		under interval ±5~±10 stop	4,422	3.62
		more interval ±10 stop	1,251	1.02
Under 10~20 stops	43,047	interval ±0 stop	11,172	9.14
		interval ±1~±2 stop	11,890	9.73
		interval ±2~±5 stop	10,229	8.37
		interval ±5~±10 stop	5,738	4.69
		more interval ±10 stop	4,018	3.29
Under 20~30 stops	14,630	interval ±0 stop	3,527	2.89
		interval ±1~±2 stop	3,637	2.98
		interval ±2~±5 stop	3,188	2.61
		interval ±5~±10 stop	1,527	1.25
		more interval ±10 stop	2,751	2.25
Under 30~50 stops	7,480	interval ±0 stop	1,787	1.46
		interval ±1~±2 stop	1,645	1.35
		interval ±2~±5 stop	1,420	1.16
		interval ±5~±10 stop	778	0.64
		more interval ±10 stop	1,850	1.51
More than 50 stops	1,320	interval ±0 stop	238	0.19
		interval ±1~±2 stop	159	0.13
		interval ±2~±5 stop	183	0.15
		interval ±5~±10 stop	79	0.06
		more interval ±10 stop	661	0.54
Total	122,218	-	122,218	100.00

또한 추정된 하차 정류장을 거리 구간으로 분류하여 분석한 결과 오차 정류장 거리가 500m미만의 추정 정확도는 42.9%로 분석되었으며, 추정 정류장 오차 범위가 정류장 구간 2개(대전시 평균 정류장 간격 584m 반영 시 1,000m는 2개 정류장을 포함한다는 기준 적용)이하 즉 추정 오차범위 1,000m이하는 62.9%로 분석되었다. <Fig. 5>와 <Table 14>는 하차 정류장 추정 거리별 오차 분포 결과이다.



<Fig. 5> The distance interval distribution between the actual alighting stop and estimated alighting stop

<Table 14> Difference between the estimated and actual volumes of getting off stop by stop interval

Division		missing trip Volume	Difference between the estimated and actual getting off stop by distance(unit :m)						
			distance 0	under distance 0~250	under distance 250~500	under distance 500~1,000	under distance 1,000~2,000	under distance 2,000~5,000	more distance 5,000
1 time trip on daily	Single Trip	41,942	9,525	443	4,630	6,795	6,155	8,049	6,345
			22.71%	1.06%	11.04%	16.2%	14.68%	19.19%	15.13%
	Transfer Trip	16,167	3,637	150	1,968	3,051	2,559	2,745	2,057
			22.5%	0.93%	12.17%	18.87%	15.83%	16.98%	12.72%
more than 2 times trip on daily	Single Trip	46,302	14,529	749	8,035	10,499	5,860	3,863	2,767
			31.38%	1.62%	17.35%	22.68%	12.66%	8.34%	5.98%
	Transfer Trip	17,807	5,372	258	3,177	4,064	2,385	1,475	1,076
			30.17%	1.45%	17.84%	22.82%	13.39%	8.28%	6.04%
Total		122,218	33,063	1,600	17,810	24,409	16,959	16,132	12,245
			27.05%	1.31%	14.57%	19.97%	13.88%	13.2%	10.02%

또한 본 연구에서는 추정된 하차 정류장 결과에 대하여 실제 버스 승하차 인원과 비교분석을 통한 정확도 검증을 수행하였다. 이는 하차 정류장 추정에 대한 실제 이용통행(승차 및 하차 통행량)과 비교를 위하여 버스에 설치된 CCTV자료를 활용하여 정류장별 승차 및 하차 인원을 조사하였다. 실제 교통카드 데이터 내 하차 정류장이 결측된 이용자에 대하여 개인별 승차 및 하차 정류장을 매칭하기 어렵기 때문에 분석 대상 노선의 정류장별 승차 및 하차 인원수, 재차인원을 비교분석 하였다. 우선 CCTV 상의 정류장별 승차 및 하차인원 조사 후 현금 이용자는 제외하고, 조사된 노선별 정류장 승하차 인원 기반 정류장간 재차인원을 산정 하였다. 교통카드 데이터 상에서는 본 연구에서 적용된 동일한 하차 정류장 추정 방법 적용하여 대전광역시 1일 전체 교통카드 데이터 중 하차 정류장 결측이 발행한 통행에 대하여 하차 정류장 추정을 통한 하차 정류장을 보정한 트립체인 데이터를 산출하였다. CCTV를 활용한 교통카드 하차 정류장 비교 분석 대상 노선은 시내버스와 급행버스 등 5개 노선을 선정하였으며, 분석 대상 노선에 대한 주요 정보는 <Table 15>와 같다.

<Table 15> The line running information about the analyzed bus line

Line Number	Line type	Start and terminal point direction	service interval (min)	running distance (km)	ridding passengers on 1daily	transferring passengers on 1daily
No.201	urban bus	wonnae-dong~birae-dong	8~9	39.4	10,260	2,883
No.211	urban bus	cargo terminal~government offices	11~14	33.9	5,830	1,417
No.301	urban bus	bongsan-dong~ocean world terminal point	8~12	46.6	10,234	3,176
No.311	urban bus	sindae garage~ocean world terminal point	7~8	49.5	11,911	3,574
No.1001	intercity bus	Daejeon Station~Osong Station	15	103.5	3,719	1,037

교통카드 데이터 상의 노선 및 차량별, 정류장별 하차인원과 버스 CCTV 상의 하차인원을 계산하여 각 정류장별 하차인원 차이값을 산정한다. 교통카드 상에서 정류장별 재차인원을 산정 한 후 이를 하차인원 차이값별로 나눈 후 절대 오차율(MAPE)을 산정하였다. 노선별 오전 및 오후 첨두시, 비첨두시로 구분하여 하차 정류장 추정 결과에 대한 오차율을 분석한 결과, 오차율은 10.3%이며 유효 정류장(오차 정류장 2개소 이하) 내 하차 정류장 추정 정확도는 평균 89.7%으로 분석되었다. 이는 하차 정류장 추정 시 광역버스 등 환승통행과 이용자의 복수통행이 많을수록 하차 정류장 추정 오차율이 감소하는 것으로 분석 되었으며, 각 노선별 하차 정류장 추정 오차율은 아래 <Table 16>과 같다.

<Table 16> The analyzed result of this estimation model of the alighting stop comparison of CCTV data

Division	Start and terminal point direction	Am peak period error rate(%)	Off-peak period error rate(%)	Pm peak period error rate(%)
No.201	wonnaedong~biraedong	10.09	6.47	7.13
	biraedong~wonnaedong	18.28	13.48	10.43
No.211	cargo terminal~government offices	12.64	17.00	8.98
	government offices~cargo terminal	12.95	15.65	5.12
No.301	bongsandong~ocean world terminal point	12.10	13.61	12.62
	ocean world~bongsandong	8.65	17.86	5.77
No.311	sindae garage~ocean world terminal point	15.59	13.03	9.52
	ocean world~신sindae garage	12.61	7.38	22.10
No.1001	Daejeon Station~Osong Station	1.26	3.26	5.17
	Osong Station~Daejeon Station	1.09	5.76	2.37
Average		10.53	11.35	8.92

Ⅳ. 결론 및 향후 연구과제

교통카드 데이터는 대중교통, 시내버스 노선개편, 대중교통 이용 수요 등 다양한 공공부분의 정책 기초자료로 활용되고 있으며, 교통안전공단, 국토교통부 및 각 지자체에서는 이러한 교통카드 정보를 활용한 다양한 대중교통 통계 지표를 제공하고 있다. 그러나 하차 정류장이 누락된 불완전한 교통카드 데이터는 활용성이 낮아져 그에 따른 데이터 샘플 수 또한 낮아지는 문제가 있다.

이에 본 연구에서는 하차 정류장 추정 방법을 통하여 대중교통에 활용되는 교통카드 데이터의 사장(또는 trash data)을 최소화하여 교통카드 데이터의 재활용과 대중교통 분석에 대한 활용성을 높이고자 하였다.

이를 위해 교통카드 데이터 상에서 발생하는 다양한 하차 정류장 결측 유형을 분석하였다. 예를 들면, 다음 승차 정류장이 이전 승차 정류장의 위치보다 앞에 있는 경우, 마지막 통행에서 첫 통행의 출발 정류장과 마지막 이용 노선의 정류장이 출발 정류장의 일정 반경에 없는 경우, 1일 1회 단일 통행의 경우 등이다. 또한, 교통카드 데이터 내 하차 정류장 누락 및 결측으로 인한 불완전한 데이터를 교통카드 이력데이터 및 승/하차 정류장 데이터 기반 의사결정 트리 학습 모델을 통하여 교통카드 데이터 하차 정류장 추정 모형을 제시하였다. 또한 기존 교통카드 데이터 상에서 승차 및 하차 정류장이 완벽한 트립체인 데이터를 활용하여 마지막 하차 정류장 데이터를 임의 결측 후 본 연구의 하차 정류장 추정 모델 결과를 비교하였다. 분석결과 유효 데이터(122,218통행)를 기준으로 하차 정류장 오차구간은 평균 2.82로 분석되었으며, 하차 정류장이 정확히 일치하는 비율은 27%, 정류장 오차 범위가 0~2개소 이하는 67.2%(MAPE : 평균 절대 백분율 오차)로 분석되었다. 또한 실차 기준의 하차 정류장과의 정확도 검증을 위하여 버스 내 설치 된 CCTV 영상 데이터를 통해 하차 정류장 추정 결과를 시내버스와 급행버스 등 5개 노선에 대하여 비교 검증하였다. 노선별 오전 및 오후 첨두시, 비첨두시로 구분하여 하차 정류장 추정에 결과에 대한 오차율 분석 결과, 오차율은 10.3%이며 오차 범위 정류장(오차 정류장 2개소 이하)이하에서 하차 정류장 추정 정확도는 평균 89.7%로 분석되었다. 이는 환승통행 및 복수통행 비율이 높을수록 하차 정류장 추정 정확도가 향상되는 것으로 판단된다.

따라서 본 연구의 결과는 교통카드를 활용하는 지자체의 하차 정류장 미태그로 인한 교통카드 데이터의 불완전성을 해소할 수 있으며, 또한 완전한 교통카드 데이터의 트립체인을 생성함으로써 다양한 대중교통 현황 및 문제점 분석, 대중교통 정책 분석 및 평가를 위한 기초 자료로 활용될 수 있을 것으로 판단된다.

다만, 교통카드 데이터 하차 정류장 추정 모델 시 1일 단일통행에 대한 오차율이 높은 것으로 분석되었으며, <Table 17>과 같이 실제 교통카드 이용자의 첫 승차 정류장과 마지막 통행 하차 정류장과의 거리를 분석한 결과, 첫 승차 정류장과 동일 날짜의 마지막 통행의 하차 정류장 거리가 1,000m이상인 통행자는 5,763인(19.5%)으로 분석되었다.

<Table 17> Distance between the first boarding stop and the final alighting stop of the last trip for a day

Division	Distance					
	0~500m	500~1,000m	1,000~2,000m	2,000~5,000m	more 5,000m	Total
User number	15,331	8,458	1,224	2,294	2,245	29,552
Rate(%)	51.88	28.62	4.14	7.76	7.6	100

본 연구는 대전광역시를 대상으로 하차 정류장 추정 모델을 적용하여 결측된 하차 정류장 추정에 거리 및 정류장 간격에 따른 추정 오차 검증(추정 정확도)을 수행하였다. 그러나 지자체별 대중교통 운영 및 환승체계 등이 상이하다. 특히 일부 지역의 경우는 하차 정류장 태그율이 상당히 저조한 지역도 분명 존재한다. 특히 단일통행에 대한 하차 정류장 오차율이 높은 원인 중 하나는 단일 통행 이용자의 불규칙한 통행과 이력통행 패턴의 부재일 가능이 높다. 본 연구에서는 교통카드 데이터만을 활용하여 보정한 연구 방법으로 한계가 있으므로 단일 통행에 대한 추가적인 통행 패턴과 공간적 밀접성 등에 대한 분석이 필요할 것이다. 또한 일반적으로 마지막 통행의 하차 정류장은 첫 승차 정류장과 인접할 것이라는 가정에 있어서 일부 모순이 발생하였다. 이는 마지막 통행의 하차 정류장 추정 시 예외적인 상황을 같이 검토할 필요성이 있다.

따라서 향후 연구에서는 1일 단일통행 시 출발 및 도착 정류장에 대한 시공간적 분석과 하차 정류장 결측

추정 시 교통특성을 고려한 반복통행 및 비 반복통행, 평일 및 주말, 시간대 등 통행특성을 고려한 추가적이 연구가 필요하며, 단일통행에서 이용자의 목적지 즉 하차가 발생되는 다양한 영향요소 분석을 통한 추가 모형 연구가 필요하다. 또한, 1일 마지막 통행에서 발생된 최종 하차 정류장 추정 시 1일 첫 승차 정류장과 마지막 통행의 최종 노선 및 정류장과의 연결성과 하차 정류장 추정 시 참조하는 데이터를 기존 이력 데이터 상의 통행 패턴 뿐만 아니라 분석 대상 다음 날의 첫 승차 정류장까지 참조하는 추가적인 연구 방법에 대한 분석이 필요할 것이다.

ACKNOWLEDGEMENTS

본 연구는 2017년도 홍익대학교 학술연구진흥비의 지원을 받아 수행하였습니다.

REFERENCES

Alsger A., Assemi B., Mesbah M. and Ferreira L.(2016), “Validating and improving public transport origin – destination estimation algorithm using smart card fare data,” *Transportation Research Part C: Emerging Technologies*, vol. 68, pp.490-506.

Barry J. J., Freimer R. and Slavin H.(2009), “Using Entry-Only Automatic Fare Collection Data to Estimate Linked Transit Trips in New York City,” *Transportation Research Record*, vol. 2112, pp.873-882.

Barry J., Newhouser R., Rahbee A. and Sayeda S.(2002), “Origin and Destination Estimation in New York City With Automated Fare System Data,” *Transportation Research Record*, vol. 1817, pp.183-187.

Cho A. H., Lee K. H. and Cho W. S.(2015), “Latent mobility pattern analysis of bus passengers with LDA,” *Journal of the Korean Data Information Science Society*, vol. 26, no. 5, pp.1061-1069.

He L. and Tréanie M.(2015), “Estimating the Destination of Unlinked Trips in Public Transportation Smart Card Fare Collection Systems,” *Transportation Research Record*, vol. 2535, pp.97-104.

He L., Nassir N., Trépanier M. and Hickman M.(2015), “Validating and calibrating a destination estimation algorithm for transport smart card fare collection systems,” *13th Conference on Advanced Systems in Public Transport, Rotterdam*, Pays-Bas, pp.19-23.

Kim K. T. and Lee I. M.(2017), “Public Transportation Alighting Estimation Method Using Smart Card Data,” *J of the Korean Society for Railway*, vol. 20, no. 5, pp.692-702.

Marcela M. A. and Palma C.(2012), “Estimation of a disaggregate multimodal public transport origin : Destination matrix from passive smart card data from Santiago, Chile,” *Transportation Research Part C*, vol. 24, pp.9-18.

Park J. H., Kim S. G., Cho C. S. and Heo M. W.(2008), “The study on error, missing data and impulation of the the smart card data for the transit OD construction,” *Journal of Korean Society of Transportation*, vol. 26, no. 2, pp.109-119.

Shin K. W.(2016) “Inferring the Transit Trip Destination Zone of Smart Card User Using Trip Chain

Structure," *J. of Korean Society of Transportation*, vol. 34, no. 5, pp.437-448.

Tréanier M., Chapleau R. and Tranchant N.(2007), "Individual Trip Destination Estimation in a Transit Smart Card Automated FareCollection System," *Journal of Intelligent Transportation Systems Technology Planning and Operations*, vol. 12, no. 5, pp.873-882.

Yoo B. S. and Choo S. H.(2017), "A research on the generation of Bus Stop Information by the line using passenger's boarding and getting off information on Smart cards," *J. of the Korea Institute of Electronic Communication Science*, vol. 12, no. 05, pp.873-882.

Zhao J., Rahbee A. and Wilson N.(2007), "Estimating a Rail Passenger Trip Origin-destination Matrix Using Automatic Data Collection Systems," *Computer-Aided Civil and Infrastructure Engineering*, vol. 22, pp.376-387.