한국ITS학회논문지 J. Korea Inst. Intell. Transp. Syst.

버스정보시스템 데이터를 활용한 교통카드 정류장 정보 오류 보정 알고리즘

Algorithm for Correcting Error in Smart Card Data Using Bus Information System Data

송 혜 인* · 탁 화 정** · 신 강 원*** · 손 상 훈****

* 주저자 : ㈜솔루인 대표이사, 충북대학교 통계학과 박사과정
** 공저자 : 여수시청 교통과 주무관
*** 공저자 : 경성대학교 도시공학과 교수
**** 교신저자 : 제주연구원 환경도시연구부 연구위원

Hye Inn Song* · Hwa Jeong Tak** · Kang Won Shin*** · Sang Hoon Son****

* Soluin Corporation and Statistics, Chungbuk National University
** Transportation Department, Yeosu City Hall
*** School of Civil, Urban, and Environmental Engineering, Kyungsung University
**** Regional Planning and Environment Division, Jeju Research Institute

† Corresponding author : Sang Hoon Son, ssanghoon.son@gmail.com

Vol. 22 No.3(2023)
June, 2023
pp.131~146

pISSN 1738-0774
eISSN 2384-1729
https://doi.org/10.12815/kits.2023.22.3.131

Received 28 April 2023
Revised 19 May 2023
Accepted 22 June 2023

요 약

교통카드 데이터는 승하차 정류장과 시각 등 활용가능성 높은 정보들을 포함하고 있어 대중교통 분야에서 다양하게 활용되고 있다. 데이터 수집·저장 과정에서 물리적·환경적 요인에 의해 다양한 오류가 교통카드 데이터에 존재하지만, 오류 유형과 보정에 대한 연구는 부족한 상황이다. 본 논문에서는 교통카드 데이터의 승하차 정류장 정보 오류를 상세히 살펴보았다. 제주특별자치도에서 수행된 버스승하차조사 자료와 동일 기간을 대상으로 수집된 교통카드 데이터와 승차정류장을 중심으로 비교한 결과 교통카드 데이터의 승차정류장 정보 오류율이 6.2% 수준으로 보정이 필요함을 확인하였다. 6단계로 구성된 버스정보시스템 데이터 기반 교통카드 승하차 정류장 정보 오류 보정 알고리즘을 제시하였다. 버스승하차조사 자료와 버스정보시스템 데이터를 비교한 결과 승차정류장 정보 일치율은 98..3% 수준으로 버스정보시스템 데이터를 활용하여 정류장 오류 보정 가능성을 확인하였다. 본 논문에서 제시한 교통카드 승하차 정류장 정보 오류 보정 알고리즘의 성능을 승차정류장을 중심으로 누락을 제외하고 평가한 결과 교통카드 승차정류장 정보 오류율이 보정 전 6.2%에서 보정 후 1.0%로 5.2%p 감소한 것으로 나타났다. 정류장 정보 오류가 보정된 교통카드 데이터를 통해 버스 노선 조정과 대중교통 인프라 투자 정책의사 결정이 보다 합리적으로 수행될 수 있을 것으로 기대된다.

핵심어 : 교통카드, 대중교통, 정류장 오류, 버스정보시스템, 오류 보정

ABSTRACT

Smart card data is widely used in the public transportation field. Despite the inevitability of various errors occur during the data collection and storage; however, smart card data errors have not been extensively studied. This paper investigates inherent errors in boarding and alighting station information in smart card data. A comparison smart card data and bus boarding and alighting survey data for the same time frame shows that boarding station names differ by 6.2% between the two

data sets. This indicates that the error rate of smart card data is 6.2% in terms of boarding station information, given that bus boarding and alighting survey data can be considered as ground truth. This paper propose 6-step algorithm for correcting errors in smart card boarding station information, linking them to corresponding information in Bus Information System(BIS) Data. Comparing BIS data and bus boarding and alighting survey data for the same time frame reveals that boarding station names correspond by 98.3% between the two data sets, indicating that BIS data can be used as reliable reference for ground truth. To evaluate its performance, applying the 6-step algorithm proposed in this paper to smart card data set shows that the error rate of boarding station information is reduced from 6.2% to 1.0%, resulting in a 5.2%p improvement in the accuracy of smart card data. It is expected that the proposed algorithm will enhance the process of adjusting bus routes and making decisions related to public transportation infrastructure investments.

Key words : Smart Card Data, Bus Information System, Correcting Error algorithm, Jeju

Ⅰ. 서 론

대중교통 이용 시 교통카드가 널리 활용되고 있다. 휴대와 사용이 간편하고 환승 혜택을 받을 수 있다는 장점이 있기 때문이다. 운영자 측면에서는 요금을 보다 효율적으로 징수하거나 정산할 수 있어 이용이 장려되고 있다. 교통카드 이용 과정에서 대중교통 이용객의 승하차 정류장과 시각, 환승 위치 등 활용 가능성 높은 정보가 생성된다. 생성되는 정보는 데이터베이스에 저장되어 대중교통 운영관리에 필요한 기초이력 자료로써 활용되고 있다(Tak, 2016).

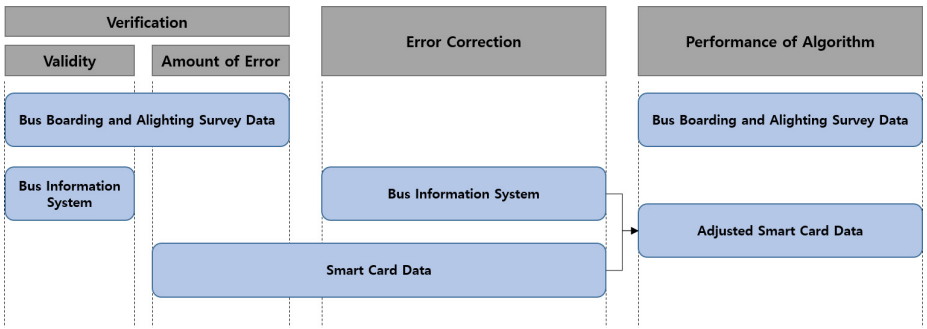
교통카드를 통해 통행행태 자료 수집에 요구되는 시간과 비용을 크게 절감할 수 있다. 그러나 교통카드 데이터 수집 특성상 이용자가 요금 단말기에 태그를 하는 과정에서 물리적·환경적 요인들로 인해 여러 가지 오류와 결측이 발생하는 것도 사실이다. 기존 연구들을 살펴보면 교통카드 데이터 내 기대치 오류, 논리적 오류, 결측이 발생하고 있으며, 이러한 오류와 결측은 그 특성에 따라 제거하거나 보정하고 있는 것으로 나타났다. 그러나 교통카드 데이터에 적재된 승하차 정류장 정보 자체가 오류일 가능성은 크게 부각되지 않았다. 승하차 정류장 정보는 교통카드 데이터가 포함하고 있는 가장 중요한 정보임에도 불구하고, 관련 연구가 제한적으로 수행되었다.

교통카드 데이터를 구축하고 활용하는 첫 번째 목적이 버스 요금의 정산이기 때문에 정산과 관련된 데이터 중심으로 관리가 이루어지고 있다는 측면과 버스 정류장의 신설과 조정이 빈번하게 일어남에도 불구하고 변경되는 정보가 즉각 혹은 정기적으로 반영되지 않는다는 현실을 고려할 때 교통카드 승하차 정류장 정보에 대한 다면적인 검증이 필요하다. 승하차 정류장 정보 자체에 오류(실제 승하차 정류장이 아닌 다른 정류장으로 기재)가 있을 경우 분석 방향에 따라 큰 문제를 야기할 수 있어 이에 대한 검증은 매우 중요하다. 교통카드 데이터에 기반하여 버스 정류장과 지하철역 이용 수요 분석 또는 O/D 분석 시 전혀 다른 결과가 도출될 수 있기 때문이며, 대중교통 노선 조정이 잘못 이루어지거나 인프라 투자에 대한 비합리적 정책의사 결정에 이를 수 있기 때문이다.

따라서 본 논문에서는 교통카드 데이터에 기재된 승하차 정류장 정보가 잘못 적재된 경우(실제로는 A정류장에 있었으나 교통카드 데이터 내에서는 B정류장으로 기재되는 경우)를 '승하차 정류장 정보 오류'로 정의하고 이에 대한 오류를 상세히 살펴보고자 한다. 아울러 교통카드 데이터의 승하차 정류장 정보에 오류가 존재할 경우 이를 보정할 수 있는 알고리즘을 제안하고자 한다. 제안하는 알고리즘의 핵심은 다수의 지자체

에서 구축하여 운영하고 있는 버스정보시스템을 활용한다는 점이다. 구체적으로는 버스정보시스템이 수집하는 여러 데이터 중에서 버스 운행 시각과 경유 정류장 정보를 활용한다는 것이다.

본 논문에서 제안하고자 하는 교통카드 데이터 승하차 정류장 정보 오류 보정 알고리즘의 구현은 제주특별자치도의 교통카드, 버스정보시스템 데이터를 활용하여 수행하고자 한다. 또한, 정류장 정보 오류 존재 여부와 규모, 버스정보시스템 데이터의 활용 가능성 검토 및 알고리즘 적용 결과 확인, 알고리즘 성능 평가는 제주특별자치도의 버스승하차조사 결과와의 비교를 통해 수행하고자 한다. 버스승하차 조사는 조사원이 직접 버스에 승차하여 버스의 운행, 승하차 정보를 조사지에 기입하므로 위와 같은 검토를 수행하기에 가장 알맞은 자료로 판단하였다. 본 논문에서 수행한 일련의 과정과 이에 따른 데이터의 활용여부를 <Fig. 1>에 제시하였다.



<Fig. 1> Data utilization process
(Bus Boarding And Alighting Survey Data, Bus Information System Data, Smart Card Data)

본 논문의 의의는 다음과 같다. 첫째 선행 연구에서 다뤄지지 않았던 교통카드 데이터 내 승하차 정류장 정보 오류를 제안하고 보정한다는 점, 둘째 교통카드 데이터, 버스정보시스템 데이터를 참값으로 판단할 수 있는 버스승하차 조사 결과로 검증한다는 점, 셋째 교통카드 데이터를 교통카드 데이터 자체만 활용하여 정제하는 것이 아닌 보다 정확도가 높다고 판단되는 타 데이터(버스정보시스템 데이터)를 활용하여 보정한다는 점, 넷째 서로 연관성이 높으나 각각 활용되었던 데이터를 통합하여 활용한다는 점이다.

본 논문은 다음과 같이 작성하였다. 2장에서 교통카드 데이터 오류 유형과 오류 보정에 대한 선행 연구들을 검토하여 기존 연구와의 차별성을 도출하고, 3장에서는 교통카드에 기재된 승차정류장 정보를 중심으로 교통카드 데이터 오류를 검증하고자 한다. 4장에서는 버스정보시스템을 활용한 교통카드 승하차 정류장 정보 오류 보정 알고리즘을 제시하였으며, 버스승하차조사 자료와 버스정보시스템 데이터를 비교하여 버스정보시스템의 활용 가능성을 타진하였다. 5장에서는 제안된 교통카드 승하차 정류장 정보 오류 보정 알고리즘을 적용 전후를 비교하여 오류 보정 정도와 오류 보정 상세 결과를 제시하였다. 6장에서는 결론, 연구의 한계, 향후 연구 과제를 논의하였다.

Ⅱ. 선행 연구 고찰

다양한 오류들이 교통카드 데이터에 존재한다. Ahn and Lee(2007)은 교통카드 데이터에 존재하는 오류를 승하차 정보가 없는 경우, 승차 정보만 없는 경우, 하차 정보만 없는 경우, 승하차 정류장이 동일한 경우로

구분하였다. 전체 자료의 13.13%가 오류를 포함하고 있는 것으로 제시하였다. Seoul Development Institute (2007)와 Park et al.(2008)은 교통카드 오류를 크게 기대치 오류와 논리적 오류로 구분하였다. 기대치 오류는 탑승객수가 '0'으로 입력된 탑승객수 오류와 첫 승차 기본요금이 250원 이하로 입력되거나 1,800원이 초과한 기본요금 오류이다. 논리적 오류는 하차시간이 승차시간보다 빠르거나 하차와 승차의 시간차가 3시간 이상인 승하차시간 관련 오류와 승차정류장과 하차정류장이 동일한 승하차정류장 동일 오류이다. 탑승객수 오류는 0.001% ~ 0.0001%, 기본요금 오류는 0.02% ~ 0.03%, 승하차시간 관련 오류는 0.004% ~ 0.03%, 승하차정류장 동일 오류 2.9% ~ 3.1%로 발생 비율이 크지 않은 것으로 나타났다. Korea Transport Institute(2009)의 연구에서는 교통카드 데이터 오류를 크게 2가지로 구분하였다. 첫 번째 오류는 교통카드 이용자의 최종 하차시간이 승차시간보다 빠른 오류이고, 두 번째 오류는 승하차 정류장 또는 지하철역이 동일하게 기록된 오류이다. 데이터 행수 기준 최종 하차시간이 승차시간보다 빠른 오류 발생 비율은 0.002%인 반면 승하차 정류장 또는 지하철역이 동일하게 기록된 오류는 52.77%에 이르는 것으로 나타났다. Kang et al.(2012)와 The Incheon Institute(2012)은 등록된 교통수단 이외의 교통수단 ID가 입력된 경우, 환승횟수가 '0'부터 '4'까지 값 이외의 값이 입력된 경우, 사용자 구분 코드가 1(일반)·2(초등학생)·4(청소년) 이외의 값이 입력된 경우, 탑승객수가 '0'인 경우, 첫 승차의 기본요금이 250원 이하 및 1,800원을 초과한 경우, 버스 회차 정보가 없는 경우를 기대치 오류로 구분하였다. 하차시간이 승차시간보다 빠른 경우, 하치시간과 승차시간 시간차가 3시간 이상인 경우, 승하차 정류장이 동일한 경우, 개인ID와 통행순서가 맞지 않은 경우, 승하차 거래일자가 불일치하는 경우를 논리적 오류로 구분하였다. Park et al.(2013)는 승하차 태그 오류, 승하차 정류장 동일 오류, 통행시간 오류를 제시하였으며, 발생빈도는 약 4%정도로 언급하였다. Jeon et al.(2014)은 오류와 결측을 구분하여 제시하였다. 오류는 기대치 오류(교통수단ID 오류, 환승 횟수 4번 초과 오류, 탑승객 수 '0'명 오류 등)와 논리 오류(하차시간이 승차시간보다 빠른 오류, 승하차 시간의 차이가 3시간 이상인 오류)로 구분하였고, 결측은 승차 결측과 하차 결측으로 별도로 구분하였다. 기대치 오류 발생빈도는 0.001% 미만, 논리 오류는 0.01%미만인 것으로 나타났다. Han(2016)은 9가지 오류(최초 승차 시 정류장 누락인 경우, 최초 하차 시 정류장이 누락인 경우, 대중교통 수단ID가 '0'인 경우, 승차인원이 '0'명인 경우, 통행시간이 모순인 경우, 환승하는 정류장이 누락인 경우, 환승시간이 '0'분 이하인 경우, 환승 횟수가 5인 경우, 총 통행거리가 '0'인 경우)를 구분하였다. Lee et al.(2018)은 하차태그가 누락된 승하차태그 오류, 대중교통 승하차 시간이 총 소요시간 보다 큰 통행시간 오류, 최초하차일시가 최초승차일시보다 빠르거나 승하차수집건수가 '0'인 논리적 오류 등을 식별하였다. Shin(2016)은 단순오류를 중복 태그, 통행자의 첫 통행이 환승 또는 하차인 경우, 통행의 승차ID가 누락된 경우, 통행 수단이 도시철도 또는 경전철임에도 하차ID이 누락된 경우를 제시하였고, 부산과 같이 단일요금제를 활용하는 경우는 하차태그의 필요성이 낮아 종점에 결측이 많음을 강조하였다. Lee et al.(2020)은 대중교통 이용자의 이용 수요를 파악하기 위해 교통카드 데이터가 활용되나 많은 지자체에서 단일요금제를 적용하고 있어 하차 태그에 결측이 많음을 강조하였다. Jung et al.(2022) 또한 교통카드 데이터 오류로 하차태그 결측을 제시하였다.

자료 정제(Data Cleaning) 과정에서 이러한 오류와 결측을 포함한 교통카드 데이터는 일반적으로 직관적인 값으로 보정하거나 제외한 후 분석이 수행되었다. Ahn and Lee(2007), Seoul Development Institute(2007), Park et al.(2008), Kang et al.(2012), The Incheon Institute(2012), Park et al.(2013), Jeon et al.(2014), Jung et al.(2022)은 필터링 기법을 사용하여 오류와 결측을 포함한 교통카드 데이터를 식별한 후 분석에서 제외한 것으로 나타났다. Han(2016)의 경우 오류 혹은 결측의 특성에 따라 교통카드 데이터를 분석에서 제거하거나 직관적인 값으로 보정하여 사용하였다. 예를 들어 승차인원이 '0'명인 경우 최소한 1명이 탑승하였다고 가정하고 1로 보정하는 것이다. Korea Transport Institute(2009)의 연구에서는 교통카드 이용자의 최종 하차시간이 승차시간보

다 빠른 오류는 교통존기반 기종점 통행량을 구축한 후 총량적으로 보정하였으며, 승하차 정류장 또는 지하철역이 동일하게 기록된 오류는 교통카드데이터의 정류장 기반 기종점 통행량 자료를 이용하여 정류장간 통행량 분포를 적용하여 보정하였다. Lee et al.(2018)은 하차태그 누락 오류(결측)의 경우 공간적 범위로 설정한 지역의 외부로 보정하고, 그 외 오류는 필터링을 통해 제외시켰다. Shin(2016)은 단순오류는 삭제, 2회 이상 통행한 경우의 하차태그의 결측에 초점을 두어 통행사슬 구조를 활용한 통행종점 추정 모형을 적용하여 결측을 보정하였다. Lee et al.(2020)은 하차태그 결측을 통행 유형(연속적 통행, 비연속적 반복 통행, 비연속적 비반복 통행)별로 통행사슬, 통행패턴, 통행기록을 활용하여 보정하였다.

선행 연구에서는 교통카드 데이터 내 기대치 오류, 논리적 오류, 결측을 식별하고 오류 혹은 결측의 특성에 따라 이를 제거하거나 보정하였다. 그러나 교통카드 데이터에 적재된 승하차 정류장 정보 자체가 오류일 가능성 및 이를 보정할 수 있는 방법이 제시되지 않았다.

교통카드 데이터의 정류장 정보는 버스가 운행하고 있는 노선 정보와 위치 정보를 통해 적재되는데, 노선 운행 전 버스 기사가 버스 내 단말기에 노선 정보를 입력하고 운행 시작 버튼을 누를 때 이를 잘못 입력하여 버스가 입력된 노선의 경로와 다른 경로로 운행될 경우, 동일한 경로인 경우만 정류장이 올바르게 적재되고 나머지 경우는 모두 이 전 정류장(동일한 경로의 마지막 정류장)으로만 적재될 수 있다. 유사한 이유로 정류장의 신설로 해당 정보가 노선 내 경유하는 정류장에 반영되지 않을 경우 신설된 정류장에 올바르게 정차하였어도 데이터상으로는 경로 이탈이 되어 이 전 정류장으로 적재될 수 있다.

이와 같은 이유로 발생하는 정류장 정보의 오류를 ‘승하차 정류장 정보 오류’로 칭하고, 본 논문에서는 선행 연구들과 달리 교통카드 데이터의 승하차 정류장 정보 오류의 존재 가능성을 제시하고자 한다. 이에 오류의 존재 여부 및 규모를 살펴보고, 더 나아가 이를 보정할 수 있는 알고리즘을 제안하고자 한다.

교통카드 데이터를 구축하고 활용하는 첫 번째 목적이 버스 요금의 정산이기 때문에 정산과 관련된 데이터 중심으로 관리가 이루어지고 있다는 측면, 버스 정류장의 신설과 조정이 빈번하게 일어남에도 불구하고 변경되는 정보가 즉각 혹은 정기적으로 반영되지 않는다는 측면을 고려할 때 승하차 정류장 정보 오류의 존재 가능성은 무시할 수 없다. 또한, 교통카드 데이터 내 적재된 승하차 정류장 정보 자체에 오류가 존재한다면 어떤 정류장이냐에 따라 대중교통 이용객의 승하차 패턴을 파악하거나 최다 승차, 최다 하차 지점을 파악하는 통행수요 분석 등의 과정과 결과에 큰 영향을 미칠 수 있어 정확한 분석을 위해서 필수적으로 보정될 필요가 있다.

Ⅲ. 교통카드 데이터 오류 검증

교통카드 데이터에 승하차 정류장 정보 오류가 존재할 경우 버스 수요, OD 등의 파악에 큰 영향을 미칠 수 있다. 제주특별자치도의 경우 2017년 8월 28일 대중교통개편과 함께 정류장 대규모 신설이 수행됐는데, 특히 제주공항에 신규 정류장이 들어서며 교통카드 데이터에 승하차 정류장 정보 오류가 발생하였다. 이에 제주공항의 버스 수요를 파악하고자 교통카드 데이터로 승하차인원을 집계하였을 때 제주공항으로 태그되어야 할 정류장이 모두 다호마을로 적재되는 현상이 발생하였다. 제주공항의 정확한 버스 수요를 파악하기 어려운 현실이었다. 2017년 9월 2일을 예로 들면 제주공항의 전체 승차인원은 2,590명이며 이에 해당되는 정류장은 ‘공항초소’, ‘제주국제공항’, ‘제주국제공항(대정, 화순, 일주서로)’, ‘제주국제공항(종점)’, ‘제주국제공항입구’ 6개로 ‘제주국제공항(대정, 화순, 일주서로’를 제외하고 개편 전 정류장이었으며, 다호마을은 1,067명으로 다호마을에 많은 승차건수가 적재되었고, 이는 2017년 7월(개편전) 다호마을 일평균 승차인원 약 232명

과 비교하여도 약 4.6배로 매우 많은 승차인원이었다. 따라서 제주공항의 승차인원이 다호마을로 적재되었다고 판단하기 충분하다. 이와 같이 승하차 정류장 정보 오류는 정류장에 따라 수요 분석 시 결과에 중대한 영향을 미칠 수 있으므로 보정이 필요하다.

제주특별자치도를 대상으로 교통카드 데이터의 승하차 정류장 정보 오류의 존재 여부와 규모를 분석하고자 한다. 이를 위해 2018년 1월 6일부터 2월 26일까지 조사원이 직접 버스에 탑승하여 실제 버스 운행 정보(경유 정류장명, 정류장 도착/출발 시각, 승하차 인원 등)를 작성한 버스승하차조사 결과를 활용하였다. 버스승하차조사 결과를 참값으로 전제하고, 같은 기간 수집되어 교통카드 데이터에 저장된 승하차 정류장 정보와 비교하였다.

<Fig. 2> The location of Jeju International Airport and Daho Village

교통카드 데이터의 승하차 정류장 정보 오류 존재 여부와 규모를 확인하기 전 선행 연구에서 제시된 교통카드 데이터 오류 규모를 검토하였다. 이때 각 오류의 기준은 환승 횟수처럼 지역별 차이가 있을 수 있는 경우 제주특별자치도의 기준에 맞게 변경하였다. 선행 연구에서 언급한 오류는 첫째, 데이터 결측(승하차 정보가 없는 경우), 둘째, 기대치 오류(탑승객수가 '0'으로 입력된 경우, 기본요금이 카드기준 최소 350원, 최대 1,150원[1]에 맞지 않는 경우, 환승횟수가 3회 이상인 경우, 사용자 구분 코드가 의도되지 않은 값이 기재된 경우, 환승시간이 0분 이하인 경우, 총 통행거리가 0인 경우), 셋째, 논리적 오류(승하차 정류장이 동일한 경우, 하차시간이 승차시간보다 빠른 경우, 하차와 승차의 시간차이가 3시간 이상인 경우, 승하차 거래일자가 불일치하는 경우)이다. 단, 제주특별자치도에서는 교통수단ID와 탑승시간 정보를 수집하고 있지 않으므로 '등록된 교통수단 이외의 교통수단ID' 오류와 '대중교통 탑승시간이 총소요시간보다 큰 경우'오류는 검토에서 제외하였다.

2018년 1월 6일부터 2월 26일까지의 교통카드 데이터의 오류를 검토한 결과를 <Table 1>에 제시하였다. 교통카드 데이터 수는 총 6,786,415건이며, 첫째, 데이터 결측은 승차 정보가 없는 경우 0%, 하차 정보가 없는 경우 34%(2,321,879건)로 나타났고, 둘째, 기대치 오류는 탑승객수가 '0'으로 입력된 경우 0%, 기본요금이 카드기준 최소 350원, 1,150원에 맞지 않는 경우가 0.0002%(15건)[2], 환승횟수가 3회 이상인 경우 0%, 사용자

1) 2018년 1월 기준 제주특별자치도 버스 요금
2) 교통복지카드, 환승인 경우를 제외하였고, 인당 요금(승차요금/인원)을 대상으로 집계하였으며, 1,150원 초과인 경우는 거리비례 요금을 적용하고 있는 급행, 공항리무진과 성인 5,000원, 어린이 3,000원 요금을 적용하고 있는 시티투어를 제외하였음

구분 코드가 의도되지 않은 값이 기재된 경우 0%, 환승시간이 0분 이하인 경우 0.0002%(14건), 총 통행거리가 0인 경우 0.01%(971건)[3]로 나타났다. 셋째, 논리적 오류는 승하차 정류장이 동일한 경우 3.3%(223,019건), 하차시간이 승차시간보다 빠른 경우 0.00003%(2건), 하차와 승차의 시간차이가 3시간 이상인 경우 0.03%(2,224건), 승하차 거래일자가 불일치하는 경우 0.000014%(1건)으로 나타났다. 제주특별자치도의 하차태그율이 약 70% 수준이기 때문에 하차 정보가 없는 비율이 높게 나타났지만, 그 외 승하차 정류장이 동일한 경우를 제외하고 매우 낮은 수준의 오류율을 보였다.

<Table 1> The error amount of smart card data in Jeju for each type of errors

Division	Type of Errors	Percent
Missing Data Error	Boarding Information	0%
	Alighting Infromation	34%
Expected Value Error	Count of Passengers = 0	0%
	350 < Basic Rate of Bus Fare < 1,150	0.0002%
	Count of Transfers ≥ 3	0%
	Unexpected Value in User Identification Code	0%
	Transfer Time ≤ 0	0.0002%
	Total trip distance = 0	0.01%
Logical Error	Boarding Station = Alighting Station	3.3%
	Alighting Time < Boarding Time	0.00003%
	|Alighting Time - Boarding Time| ≥ 3 hours	0.03%
	Boarding Date ≠ Alighting Date	0.000014%

교통카드 데이터의 승하차 정류장 정보 오류 존재 여부와 규모를 확인하고자 버스승하차조사 결과와 비교하였다. 버스승하차조사 전체 자료 중에서 첫째, 기점부터 종점까지 전 구간이 조사된 자료, 둘째, 충분한 승하차 인원이 존재하는 자료를 활용하였다. 그 결과 간선 5개 노선 10회 운행(365번 4회 운행, 351번 2회 운행, 342번 1회 운행, 312번 2회 운행, 311번 1회 운행), 지선 2개 노선 3회 운행(655번 1회 운행, 415번 2회 운행)이 선정되었다. 본 논문에서는 승하차 정류장 정보 중에서 승차정류장 정보를 중심으로 비교를 수행하였고, 그 결과를 제시하였다. 비교는 동일한 시간대에 버스승하차조사의 정류장 순서와 해당 정류장에서 수집된 승차인원이 교통카드 데이터에 기재된 승차정류장 순서와 정류장별 승차인원과 일치하는지를 확인하는 방법으로 수행되었다. 제주 전 지역이 고려된 총 승차인원 827명의 승차정류장 정보가 비교되었다.

비교 결과 일부인 노선번호 312번, 차량번호 3684번 버스가 운행한 시간대의 버스승하차조사 자료와 교통카드 데이터 승차정류장 정보 비교 결과를 <Table 2>에 제시하였다. 버스승하차조사를 통해 한라병원에 6명 승차, 그레이스호텔(현 정류장명 롯데시티호텔)에서 1명이 승차한 것으로 자료가 수집되어있으나, 동일한 시간대 교통카드 데이터에는 그레이스 호텔이 7명으로 적재되어 있어 6명의 승차정류장 정보는 일치하지 않는 것으로 나타났다. 해당 경우는 그레이스 호텔 6건이 한라병원으로 보정되어야 하는 오류로 판단된다.

3) 제주특별자치도에서 수집한 교통카드 데이터에는 급행인 경우만 총 통행거리(이용거리)가 적재되었으므로 급행인 경우를 대상으로 했으며, 하차태그가 되지 않는 경우 '0'으로 적재되어 해당 건은 하차 정보가 없는 경우에 해당되어 제외하였음, 또한 승차정류장 ID와 하차정류장ID가 동일하여 총 통행거리가 0으로 집계된 경우도 승하차 정류장 동일 오류에 해당하므로 제외하였음

<Fig. 3> The location of Halla Hospital and Grace Hotel

이 외에도 버스승하차조사 자료에서는 앞서 소개한 것과 같이 노선번호 365번, 차량번호 7002번 버스의 제주국제공항 정류장 승차인원이 23명, 제주국제공항의 다음 정류장인 다호마을 정류장 승차인원이 0명이나, 교통카드 데이터에는 다호마을 정류장 승차인원이 23명으로 적재되어있는 경우도 존재하였다. 교통카드 데이터의 다호마을 정류장이 제주국제공항 정류장으로 변경되어야 하는 오류 사례이다.

<Table 2> Comparison between bus boarding and alighting survey data and smart card data(Route 312)

Route	Veh. Reg. Num.	Bus Boarding and Alighting Survey Data		Smart Card Data		Consistent
		Time	Boarding Station Name	Time	Boarding Station Name	
312	3684	17:18	Halla Hospital	17:17:36	Grace Hotel	No
		17:18	Halla Hospital	17:17:39	Grace Hotel	No
		17:18	Halla Hospital	17:17:41	Grace Hotel	No
		17:18	Halla Hospital	17:17:43	Grace Hotel	No
		17:18	Halla Hospital	17:17:47	Grace Hotel	No
		17:18	Halla Hospital	17:17:48	Grace Hotel	No
		17:20	Grace Hotel	17:20:08	Grace Hotel	Yes

버스승하차조사 자료의 승차정류장 정보와 교통카드 데이터의 승차정류장 정보를 전체적으로 비교한 결과를 정리하면 <Table 3>과 같다. 총 승차인원 827명의 버스승하차조사 결과와 교통카드 데이터의 승차정류장 정보 중 10.7%에 해당하는 88명의 승차정류장 정보가 일치하지 않는 것으로 나타났다. 세부적으로는 6.2%에 해당하는 51명의 승차정류장 정보가 서로 다른 것으로 나타났다. 37명(4.5%)의 승차정류장 정보는 교통카드 데이터에 존재하지 않는 것으로 나타났다. 선행 연구에 따르면 교통카드 결제 과정, 데이터 수집・저장 과정에서의 물리적・환경적 요인과 카드사의 정산 시기 불일치 등에 기인하여 누락이 발생한 것으로 판

단된다. 본 논문에선 6.2%에 해당되는 정류장 정보 불일치를 교통카드 데이터의 승하차 정류장 정보 오류로 판단하고, 이를 버스승하차조사 결과와 같이 올바른 정류장으로 변경하는 알고리즘을 개발하고자 한다.

<Table 3> The result of comparison between bus boarding and alighting survey and smart card data

		Frequency		Percent	
Consistent		739		89.3%	
Inconsistent	Different Boarding Station Name	88	51	10.7%	6.2%
	No data in Smart Card Data		37		4.5%
Total		827		100%	

Ⅳ. 교통카드 정류장 정보 오류 보정 알고리즘

교통카드 데이터에 존재하는 승하차 정류장 정보 오류에 대한 보정 방안으로 다수의 지자체에서 구축하여 운영하고 있는 버스정보시스템을 활용하고자 한다. 버스정보시스템 데이터는 기본적으로 30초 간격으로 버스의 위치 정보가 수집되며 이벤트(정류장 도착, 출발, 경유 교차로 진입, 운행 시작, 운행 종료 등)가 발생할 시 추가로 데이터가 적재되고 있어 정보의 신뢰도가 높아 활용 가능성 또한 높을 것으로 판단되기 때문이다. 활용 가능성에 대한 검증은 버스승하차조사 결과와 버스정보시스템 데이터에 저장된 경유 정류장 정보를 비교하여 확인할 수 있다. 버스정보시스템의 활용 가능성 검토 이후 버스정보시스템 데이터의 버스 운행 시각과 경유 정류장 정보에 기반한 교통카드 승하차 정류장 정보 오류 보정 알고리즘을 제안하고자 한다.

1. 버스승하차조사 자료의 정류장과 버스정보시스템 데이터의 정류장 정보 비교

버스정보시스템의 활용 가능성을 파악하고자 버스승하차조사 자료와 버스정보시스템 데이터를 비교하였다. 구체적으로는 동일한 시간대에 동일 차량의 버스승하차조사 결과 내 승차정류장의 순서와 명칭이 버스정보시스템 내 경유 정류장 순서 및 명칭과 일치하는지를 확인하였고, 앞서 교통카드 데이터와 비교 시 활용된 버스승하차조사 자료를 동일하게 사용하였다. 즉, 간선 5개 노선 10회 운행(365번 4회 운행, 351번 2회 운행, 342번 1회 운행, 312번 2회 운행, 311번 1회 운행), 지선 2개 노선 3회 운행(655번 1회 운행, 415번 2회 운행)을 대상으로 하였고, 총 승차인원 827명의 승차정류장 정보에 대한 비교가 이루어졌다. 비교를 위해 버스정보시스템 데이터에서 정류장 도착 정보(EVENT_CD가 17인 경우)를 추출하여 비교하였다.

버스승하차조사 자료의 승차정류장 정보와 버스정보시스템 데이터의 승차정류장 정보를 전체적으로 비교한 결과를 정리하면 <Table 4>와 같다. 버스승하차조사를 통해 수집된 총 827명의 승차정류장 정보가 버스정보시스템 데이터 경유 정류장 정보와 일치하는 경우는 98.3%로 나타났다. 이는 앞서 확인한 교통카드 데이터의 승차정류장 정보 일치율인 89.3% 대비 매우 높은 수준이다. 따라서 버스정보시스템 데이터의 경유 정류장 정보는 오류 보정에 활용하기에 충분하다고 판단된다. 전체 비교 대상의 1.7%에 해당하는 승차인원 14명의 승차정류장 정보는 버스승하차조사에서는 수집되었지만, 버스정보시스템 데이터에는 존재하지 않는 것으로 나타났다. 기술적 문제로 버스정보시스템 데이터가 수집되지 않았거나 저장되지 않아 누락된 것으로 판단된다.

<Table 4> The result of comparison between bus boarding and alighting survey and bus information system data

		Frequency	Percent
Consistent		813	98.3%
Inconsistent	No data in Bus Information System data	14	1.7%
Total		827	100%

2. 버스정보시스템 데이터를 활용 교통카드 정류장 정보 오류 보정 알고리즘

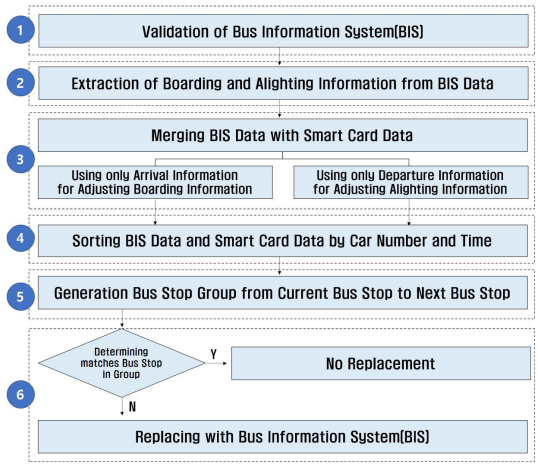
버스정보시스템 데이터를 기반으로 교통카드 승하차 정류장 정보 오류를 보정하는 알고리즘을 <Fig. 4>와 같이 제안한다. 알고리즘은 총 6단계로 구성된다.

1단계는 버스정보시스템 데이터의 정류장 정보(정류장 경유 순서와 정류장명)에 대한 정확도를 검증한다. 제안하는 교통카드 데이터 승하차 정류장 정보 오류 보정 알고리즘은 버스정보시스템 데이터의 정류장 경유 순서와 정류장명의 정확도에 영향을 크게 받기 때문인데, 본 논문에서는 버스승하차조사 결과를 활용하여 버스정보시스템 데이터의 정류장 경유 순서와 정류장명 정보에 대한 정확도를 살펴보았다. 이는 알고리즘을 본격적으로 적용하기 전 최소 1회 이상 수행해야하는 중요한 단계이다.

2단계는 버스정보시스템 데이터에서 정류장 출·도착 정보(차량번호, 노선ID, 노선번호, 정류장ID, 정류장명, 출·도착 시각, 이벤트 코드)만 추출한다. 버스정보시스템 데이터의 로우 데이터에는 차량번호, 노선ID, 정류장ID, 출·도착 시각, 이벤트 코드만 적재되어있으므로 노선번호와 정류장명은 버스정보시스템의 메타데이터와 병합을 통해 추가한다. 또한, 버스의 정차 전·후 움직임에 따라 출발·도착 정보가 중복되어 적재되는 경우가 있으므로 출발 정보는 가장 빠른 시각, 도착 정보는 가장 늦은 시각으로 추출하는 방법을 통해 정제한다.

3단계는 추출한 버스정보시스템 데이터를 교통카드 데이터와 병합한다. 교통카드 데이터의 승차정류장을 보정할 때는 버스정보시스템 데이터의 정류장 도착 정보만 활용하고, 교통카드 데이터의 하차정류장 정보를 보정할 때는 버스정보시스템 데이터의 정류장 출발 정보만 활용한다.

4단계는 병합된 버스정보시스템 데이터와 교통카드 데이터를 차량번호와 시간으로 정렬한다. 이때 시간은 위 3단계에 의해 승차정류장 보정 시에는 버스정보시스템 데이터의 도착시간, 교통카드 데이터의 승차시간



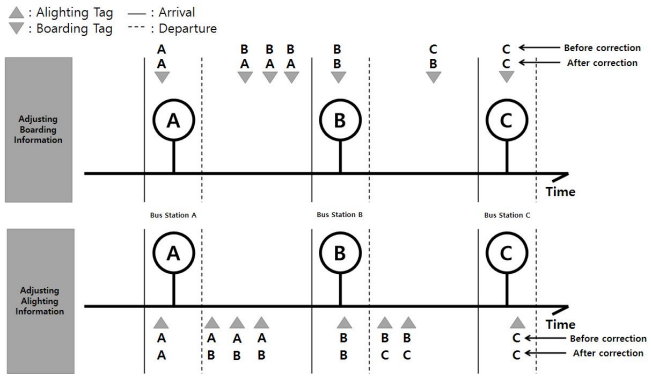
<Fig. 4> Error correction algorithm

이며, 하차정류장 보정 시에는 버스정보시스템 데이터의 출발시간, 교통카드 데이터의 하차시간이다.

5단계는 버스정보시스템 데이터의 출발(도착) 정류장과 다음 출발(도착)정류장까지를 그룹으로 지정한다. 승차 정류장 보정 시 버스 도착시간부터 다음 정류장 도착시간까지가 한 그룹, 하차정류장 보정 시 버스 출발시간부터 다음 정류장 출발시간까지가 한 그룹으로 선정된다.

6단계는 생성된 정류장 그룹 내 버스정보시스템 데이터의 정류장과 일치하지 않은 교통카드 데이터의 정류장을 오류로 판단하고 버스정보시스템 데이터의 정류장으로 대체한다. 단, 승차정류장의 경우 이전 정류장으로 대체, 하차정류장의 경우 다음 정류장으로 대체한다.

승차정류장과 하차정류장을 다른 정류장을 기준으로 하여 보정하는 이유는 승차 후 태그, 하차 전 태그가 존재하기 때문이다. 이를 직관적으로 나타낸 그림은 <Fig. 5>와 같다. 그림의 상단은 승차정류장 태그 상황을 보정 전과 후로 나누어 제시한 상황이고, 하단은 하차정류장 태그 상황을 보정 전과 후로 나누어 제시한 상황이다. 승차정류장 정보 보정의 경우 버스가 정류장(A)에 도착한 뒤 승객이 한번 승차태그를 하였고 해당 정보는 정류장(A)로 올바르게 수집되었으나, 버스가 출발한 이후(점선 이후) 태그한 경우는 모두 정류장(B)로 수집되었다. 하차정류장 정보 보정의 경우 버스가 정류장(A)에 도착한 뒤 승객이 한번 하차태그를 하였고 정류장(A)로 올바르게 수집되었으나 버스 출발 직후 하차태그를 할 경우 하차정류장이 정류장(A)로 잘못 수집되는 경우가 발생한다. 위 두 경우 승차정류장은 정류장(A)로 적재되어야하며 하차정류장은 정류장(B)로 적재되어야한다. 본 연구에서 제시한 알고리즘은 이 같은 늦은 승차태그, 빠른 하차태그를 보정한다.



<Fig. 5> Difference of adjusting algorithm between boarding and alighting information

Ⅴ. 교통카드 정류장 정보 오류 보정 알고리즘 검증

1. 교통카드 정류장 정보 오류 보정 알고리즘 적용

본 논문에서 제시한 교통카드 승하차 정류장 정보 오류 보정 알고리즘을 <Table 2>에서 검토한 교통카드 데이터에 적용한 후, 오류 보정 여부를 승차정류장을 중심으로 살펴보았다. <Table 5>는 노선번호 312번, 차량번호 3684번 버스의 버스승하차조사 결과와 교통카드 데이터의 승차정류장 정보의 일치 여부를 비교한 <Table 2>에 본 논문에서 제시한 오류 보정 알고리즘 적용 결과를 추가한 결과이다. 이를 통해 오류 보정 전 버스 승하차 조사 자료의 승차정류장 정보와 다르게 기재되었던 교통카드 데이터의 승차정류장 정보가 버스 승하차 조사 결과(참값)와 일치하도록 보정되었음을 확인할 수 있다.

<Table 5> The results of boarding station name error correction(Route 312)

Route	Car No.	Bus Boarding and Alighting Survey Data		Smart Card Data			Consistent	Error Correction
		Time	Boarding Station Name	Time	Boarding Station Name (Before Correction)	Boarding Station Name (After Correction)		
312	3684	17:18	Halla Hospital	17:17:36	Grace Hotel	Halla Hospital	No	Yes
		17:18	Halla Hospital	17:17:39	Grace Hotel	Halla Hospital	No	Yes
		17:18	Halla Hospital	17:17:41	Grace Hotel	Halla Hospital	No	Yes
		17:18	Halla Hospital	17:17:43	Grace Hotel	Halla Hospital	No	Yes
		17:18	Halla Hospital	17:17:47	Grace Hotel	Halla Hospital	No	Yes
		17:18	Halla Hospital	17:17:48	Grace Hotel	Halla Hospital	No	Yes
		17:20	Grace Hotel	17:20:08	Grace Hotel	Grace Hotel	Yes	-

구체적으로는 노선번호 312번, 차량번호 3684번 버스승하차조사 자료에서는 한라병원 정류장에서 6명이 승차한 것으로 나타났지만, 교통카드 데이터에서는 해당 승차인원이 그레이스 호텔 정류장에서 승차한 것으로 기재되어 있었다. 승차정류장을 버스정보시스템 데이터의 동일한 시간대에 동일 차량의 경유 정류장 정보를 통해 한라병원 정류장인 것으로 확인하였고, 교통카드 데이터의 승차정류장 정보(<Table 5>의 오류보정 전)를 오류로 판정한 후 버스정보시스템 데이터의 경유 정류장 정보(<Table 5>의 오류보정 후)로 변경하였다. 즉, <Table 5>에 오류 보정 전 그레이스 호텔이 보정 후 한라병원으로 변경된 부분은 버스정보시스템 데이터에 해당 시간의 동일한 차량의 정류장 정보이다.

앞서 교통카드 데이터 정류장 오류 사례로서 제시한 제주국제공항 정류장이 다음 정류장인 다호마을 정류장으로 기재된 경우 역시 본 논문에서 제시한 교통카드 승하차 정류장 정보 오류 보정 알고리즘을 적용한 결과 오류가 알맞게 보정된 것으로 나타났다. 노선번호 365번, 차량번호 7002번 버스승하차조사 자료에서는 제주국제공항 정류장 승차인원이 23명, 제주국제공항의 다음 정류장인 다호마을 정류장 승차인원이 0명이나, 오류 보정 결과 승차인원 23명의 승차정류장이 다호마을 정류장에서 제주국제공항 정류장으로 보정되었고, 버스승하차조사 자료의 승차정류장 정보와 일치하는 것으로 나타났다.

2. 교통카드 정류장 정보 오류 보정 전후 오류율 비교

본 논문에서 제시한 교통카드 승하차 정류장 정보 오류 보정 알고리즘의 성능을 평가하기 위해 오류 보정 전후에 대한 오류율을 승차정류장을 중심으로 비교하여 <Table 6>과 같이 제시하였다. 오류 보정 전은 <Table 3>에서 검토한 승차인원 827명의 오류율이고 오류 보정 후는 동일한 데이터의 오류 보정 알고리즘 적용 후 오류율이다.

비교 결과, 교통카드 승차 정류장 정보 오류를 포함한 전체 오류율이 보정 전 10.7%에서 보정 후 5.5%로 감소한 것으로 나타났다. 교통카드 데이터 수집・저장 과정에서의 물리적・환경적 요인과 카드사의 정산 시기 불일치 등에 기인하여 발생한 데이터 누락을 제외하고, 교통카드 승하차 정류장 정보 오류 보정 알고리즘을 통해 승차정류장 정보 오류는 6.2%에서 1.0%로 5.2%p 감소한 것을 확인하였다.

데이터 누락 규모는 알고리즘 보정 전과 후가 동일하므로 전체 오류를 대상으로 보정 전, 후 오류 발생 확률이 같은지를 파악하기 위해 카이제곱 검정을 수행하였으며 검정 결과 검정통계량 값이 15.02이고 p-value

가 0.00012로 0.001보다 매우 적어 충분히 유의하게 오류율이 감소하였다고 할 수 있다.

교통카드 데이터의 승차정류장 정보 오류를 본 논문에서 제시한 알고리즘을 적용하여 보정하였음에도 불구하고 1.0%의 오류는 여전히 존재하는 것으로 나타났다. 버스정보시스템 데이터의 누락, 교통카드 데이터 내 승차시각과 버스정보시스템 내 승차시각 정보의 불일치 등이 주요 이유일 것으로 판단된다. <Table 4>에서 버스승하차조사 자료와 버스정보시스템 데이터를 비교한 결과 버스정보시스템 데이터 누락률이 1.7% 수준임을 확인하였다. 교통카드 데이터 내 승차시각과 버스정보시스템 내 승차시각 정보의 미일치는 교통카드 단말기와 버스정보시스템 단말기의 시간 설정 체계가 달라 발생할 수 있다.

이밖에 교통카드 데이터가 처음부터 누락되는 경우에 대해서는 보정할 수 없는 한계가 존재한다. 본 논문에서는 교통카드 데이터의 누락률은 4.5%로 나타났다. 미미할 것으로 판단되지만, 승하차 태그 시 사용하는 카드에 따라 데이터에 포함되는 기간이 다르다는 측면에서 교통카드 데이터 누락이 발생할 수 있고, 이는 교통카드 데이터의 승차정류장 정보 오류 보정에도 불구하고 여전히 오류로 존재하는 원인이 될 수 있다.

또한, 참값으로 선정하였으나 버스승하차조사 자체의 오류가 교통카드 데이터 누락률에도 영향을 줄 가능성이 존재한다. 이는 조사원이 버스 도착 전, 도착 시, 정차 중, 출발 시, 출발 후 중 어느 시점에서 조사표에 시각과 승차인원을 기입하였는지 정확히 알 수 없다는 한계가 있기 때문이다.

본 논문에서는 승차정류장을 중심으로 교통카드 정류장 정보 오류 보정 알고리즘이 적용되고 검증이 수행되었지만, 제시된 알고리즘의 내용대로 하차정류장의 경우에도 알고리즘을 적용하여 오류를 보정할 수 있다. 버스정보시스템이 생성하는 버스 운행 시각과 경유 정류장 정보를 교통카드 데이터의 하차 시각 및 하차 정류장과 비교하고 다를 경우 버스정보시스템 데이터의 정류장 정보를 활용하는 오류 보정 과정이 동일하기 때문이다. 이때 승차정류장 보정의 경우 버스정보시스템의 도착 정보를, 하차정류장 보정의 경우 버스정보시스템의 출발 정보를 사용하여야 한다.

<Table 6> The results of comparison between before and after error correction

Division		Comparison between bus boarding/ alighting survey and smart card data (before correction)		Comparison between bus boarding/ alighting survey and smart card data (after correction)	
Consistent		739(89.3%)		782(94.5%)	
Inconsistent	Different station name	88(10.7%)	51(6.2%)	45(5.5%)	8(1.0%)
	No data in smart card data		37(4.5%)		37(4.5%)
Total		827(100%)		827(100%)	
χ^2(p-value)				15.02(0.00012)	

Ⅵ. 결 론

교통카드 데이터는 대중교통 운영관리에 필요한 기초이력 자료로 다양하게 활용되고 있다. 그러나 교통카드 데이터는 물리적· 환경적 요인들로 인해 오류를 포함하고 있는 것도 사실이다. 교통카드 오류와 관련한 다수의 연구가 수행된 바 있으며, 이러한 오류들은 그 특성에 따라 삭제 혹은 제한적으로 보정되어 버스 정류장의 신설과 조정 등 대중교통 분석과 정책 의사 결정에 사용되고 있다.

본 논문은 기존 연구에서 제시된 교통카드 데이터 오류 외에도 교통카드 데이터의 승하차 정류장 정보 자

체에 오류가 존재함을 제주지역에서 수집된 실측자료를 통해 제시하였다. 2018년 1월 6일부터 2월 26일까지 조사원이 직접 버스에 탑승하여 실제 버스 운행 정보(경유 정류장명, 정류장 도착/출발 시각, 승하차 인원 등)를 작성하는 방식으로 버스승하차조사가 수행되었고, 버스승하차조사에서 수집된 자료 중 승차인원 827명의 승차정류장 정보와 교통카드 데이터의 승차정류장 정보를 비교하였다. 비교 결과, 전체 승차인원의 10.7%에 해당하는 88명의 승차정류장 정보가 일치하지 않는 것으로 나타났다. 세부적으로는 6.2%에 해당하는 51명의 승차정류장 정보가 서로 다르고, 버스승하차조사에서 수집된 37명(4.5%)의 승차정류장 정보는 교통카드 데이터에 존재하지 않는 것으로 나타났다.

교통카드 데이터 승하차 정류장 정보 오류를 보정하는 알고리즘을 제안하였다. 제안하는 알고리즘의 핵심은 다수의 지자체에서 구축하여 운영하고 있는 버스정보시스템을 활용한다는 점이다. 버스정보시스템의 활용 가능성을 파악하고자 버스승하차조사 결과와 버스정보시스템 데이터를 비교하였고, 버스승하차조사를 통해 수집된 총 827명의 승차정류장 정보가 버스정보시스템 데이터 경유 정류장 정보와 일치하는 경우는 98.3%로 나타나 버스정보시스템 데이터의 활용 가능성이 높은 것으로 판단하였다.

버스정보시스템이 수집하는 여러 데이터 중에서 버스 운행 시각과 경유 정류장 정보를 연계하여 교통카드 데이터의 승하차 정류장 오류를 보정할 수 있음을 제시하였다. 버스정보시스템 데이터에서 정류장 출·도착 정보(차량번호, 노선ID, 노선번호, 정류장ID, 정류장명, 출·도착 시각, 이벤트 코드)만 추출하고, 추출한 버스정보시스템 데이터를 교통카드 데이터와 병합한 후, 버스정보시스템 데이터의 출발(도착) 정류장과 다음 정류장 출발(도착)까지를 그룹으로 지정하는데, 생성된 정류장 그룹 내 버스정보시스템 데이터의 정류장과 일치하지 않은 교통카드 데이터의 정류장을 오류로 판단하고 버스정보시스템 데이터의 정류장으로 대체하는 알고리즘이다. 본 논문에서 제시한 교통카드 정류장 정보 오류 보정 알고리즘의 성능을 평가하고자 교통카드 데이터의 오류 보정 전후를 비교한 결과 교통카드 승차정류장 정보 오류율이 보정 전 10.7%에서 보정 후 5.5%로 감소한 것으로 나타났다. 교통카드 데이터 수집・저장 과정에서 발생한 데이터 누락을 제외하고, 교통카드 승하차 정류장 정보 오류 보정 알고리즘을 통해 승차정류장 정보 오류는 6.2%에서 1.0%로 5.2%p 감소하였고 카이제곱 검정 결과 p-value가 0.00012로 유의수준 0.001에서 유의한 것으로 확인하였다.

본 논문은 다음과 같은 한계가 존재한다. 제주특별자치도의 사례를 중심으로 교통카드 데이터의 오류 존재 여부 및 규모를 파악하였고, 버스정보시스템 데이터에 기반한 교통카드 데이터 정류장 오류 보정 알고리즘을 평가하였다. 따라서 본 논문에서 교통카드 데이터 중 정류장 정보 오류율을 6.5%로 제시하였으나 교통카드 데이터 관리 정도가 지역별로 상이하다는 측면에서 이를 모든 지자체에 일반화하기에는 무리가 있다. 그 밖에도 본 논문에서 제시한 정류장 오류 개선 정도가 지역에 따라 다를 수 있다. 각 지자체마다 버스정보시스템의 자료 수집 방식, 버스기사의 키패드 조작 특징 등이 상이하기 때문이다.

또한, 본 논문에서 제시한 알고리즘을 적용하여 보정하였음에도 불구하고 1.0%의 수준의 오류가 여전히 존재하는 것으로 나타났다. 실제 승차가 이루어졌으나 교통카드 데이터에 해당 승차 정보가 존재하지 않는 경우, 실제 버스가 운행되었으나, 버스정보시스템 데이터에 해당 버스가 경유하는 정류장이 누락되는 경우, 버스정보시스템 데이터와 교통카드 데이터의 시간 설정 체계가 상이한 경우가 원인으로 판단된다.

이에 향후 교통카드 정류장 정보 오류 알고리즘을 고도화하는 연구가 추진될 필요가 있다. 위 언급한 승하차 또는 경유 데이터의 누락의 원인과 감소 방안에 대한 연구도 필요하다. 더불어 버스정보시스템 데이터와 교통카드 데이터의 시간 설정 체계에 대한 조사와 버스정보시스템 관리 업체, 교통카드사, 지자체가 협업하여 시간 설정을 공유하거나 이를 지원하는 시스템을 개발할 필요가 있다. 시간 설정과 더불어 정류장 신설, 이동, 폐지 등과 같은 정보도 실시간으로 공유할 수 있다면 보다 정확한 데이터 비교가 가능해질 것이다. 이와 같은 공유는 지자체에서 결정한 노선 또는 정류장 등의 변경사항 및 시간 설정을 버스정보시스템 관리

업체와 교통카드사가 따로 적용하기보다 시간 설정값과 변경사항 데이터를 하나의 시스템에 저장하고 이를 각 관리 업체가 병합하여 활용하는 방향이 효과적일 것이다.

본 논문에서 제시한 교통카드 데이터 승하차 정류장 정보 오류와 보정 알고리즘은 제주특별자치도에서 해당 오류로 인해 제주공항의 수요를 부정확하게 수집했던 것을 바로잡은 것과 같이 정류장을 기준으로 수요분석(최다 승차인원 정류장 집계 등), OD 산출 등의 분석을 할 때 선수행하면 보다 정확한 결과를 도출할 수 있을 것으로 기대된다. 지역단위, 동단위 등 범위가 클 경우에 승하차 정류장 정보 오류는 큰 영향을 미치지 않지만, 특정 지점(예 공항, 병원, 학교 등)을 대상으로 범위를 한정할 때에는 영향이 매우 커지기 때문이다. 지자체 또는 연구기관에서는 교통카드 데이터를 수집할 때 버스승하차조사 결과 및 버스정보시스템 데이터를 함께 수집하여 본 논문에서 제시한 승하차 정류장 정보 오류가 존재하는지를 검토할 필요가 있고, 존재한다면 교통카드 데이터를 버스정보시스템 데이터로 보정하여 보다 정확하게 정책 의사 결정에 도움이 되는 집계 결과를 도출할 수 있을 것으로 판단된다.

승하차 정류장 정보 오류 보정 알고리즘은 네 가지 의의가 있다. 첫째 선행 연구에서 다뤄지지 않았던 교통카드 데이터 내 승하차 정류장 정보 오류를 제안하고 보정한다는 점, 둘째 교통카드 데이터, 버스정보시스템 데이터를 참값으로 판단할 수 있는 버스승하차 조사 결과로 검증한다는 점, 셋째 교통카드 데이터를 교통카드 데이터 자체만 활용하여 정제하는 것이 아닌 보다 정확도가 높다고 판단되는 타 데이터(버스정보시스템 데이터)를 활용하여 보정한다는 점, 넷째 서로 연관성이 높으나 각각 활용되었던 데이터를 통합하여 활용한다는 점이다.

ACKNOWLEDGEMENTS

본 논문은 2018년도 한국 ITS학회 추계학술대회에서 우수논문상으로 선정되었던 논문을 수정·보완하여 작성된 것입니다.

REFERENCES

An, H. J. and Lee, Y. I.(2007), “A Study on the Establishment and Utilization of Bus/Subway Station-based OD and Network using Seoul City's Smart Card Data”, *Transportation Technology and Policy*, vol. 4, no. 4, pp.31-58.

Han, A. R.(2016), “Estimating Transit Origin-Destination Trip Table Using Smart Card Data”, *The Conference of the Korea Institute of Intelligent Transportation Systems*, pp.372-380.

Jeon, S. W., Lee, J. W. and Jun, C. M.(2014), “Development of an Algorithm for Minimization of Passengers' Waiting Time Using Smart Card Data”, *Journal of Korea Spatial Information Society*, vol. 22, no. 5, pp.65-75.

Jung, Y. M., Lee, T. Y., Choi, S. H. and Do, M. S.(2022), “Evaluation of Accessibility to Public Transportation Using Traffic Card-based Data-Daejeon Case Study-”, *The 86th Conference of Korean Society of Transportation*, pp.295-296.

Kang, M. H., Kim, J. H. and Son, J. E.(2012), “The Analysis of Public Transport Travel Behavior

in Incheon by Card Data", *The Conference of Korean Society of Transportation*, vol. 67, pp.683-688.

Korea Transport Institute(2009), *A Study on the Collection and Utilization Methods of Advanced Transportation Survey Data, such as Transportation Cards*, pp.49-50.

Lee, J. W., Lee, S. B., Kim, G. W. and Cheon, S. H.(2020), "Estimating Destination of Bus Trips Using Public Transit Transaction Data", *The 83th Conference of Korean Society of Transportation*, pp.299-300.

Lee, Y. M., Oh, S. J. and Lee, S. J.(2018), "A study on Prediction of Road Freezing in Jeju", *Journal of Environmental Science International*, vol. 27, no. 7, pp.531-541.

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

Park, M. S., Eom, J. K. and Heo, T. Y.(2013), "The Spatial Correlation of Mode Choice Behavior based on Smart Card Transit Data in Seoul", *The Korean Journal of Applied Statistics*, vol. 26, no. 4, pp.623-634.

Seoul Development Institute(2007), *The estimation and application of origin-destination tables by using smart card data*, pp.56-64.

Shin, K. W.(2016), "Inferring the Transit Trip Destination Zone of Smart Card User Using Trip Chain Structure", *Journal of Korean Society of Transportation*, vol. 34, no. 5, pp.437-448.

Tak, H. J.(2016), "Estimation of passenger gate delay in urban metro system using smart card data", *Journal of Korean Society for Urban Railway*, vol. 4, no. 4, pp.705-717.

The Incheon Institute(2012), *The Analysis of transit Transfer in Incheon by Card Data*, pp.36-37.