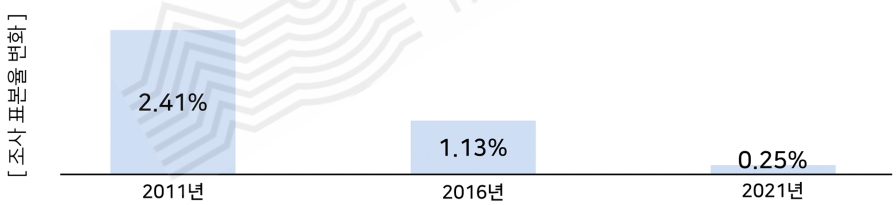
교통데이터의 통행목적·이용자특성 파악해 AI모형 활용하면 O/D의 신뢰성 향상 가능

가구통행실태조사, 직접 조사 줄고 교통카드데이터 등의 활용요구 증대

가구통행실태조사로 구축되는 여객 기종점통행량(Origin to Destination, 이하 O/D)은 교통계획 및 사회간접자본의 타당성 평가에 활용되며 각종 교통통계지표를 산출하기 위한 핵심 기초자료이다.

가구통행실태조사는 일부 표본으로 전체 모집단을 대표하는 자료를 만들기 때문에 '전수화' 과정을 거쳐야 한다. 하지만, 행정동 간 통행 표본이 부족하기 때문에 전수화 과정에서 필연적으로 오류가 발생하는데, 최근 조사비용이 축소되면서 표본율이 감소하여 이 문제가 더욱 심화되고 있다.



[그림 1] 가구통행실태조사 표본율 변화

표본율 감소에 따른 문제 해결을 위해 현장에서 수집되는 교통 빅데이터 활용이 논의되고 있다. 전수에 가까운 교통카드데이터와 택시데이터가 있음에도 불구하고, O/D 구축과정에서 이 데이터들의 구체적 활용방안은 여전히 미비한 실정이다.

교통데이터에 AI 방법론 적용해 통행목적·이용자특성 등 필요한 속성 추정

가구통행실태조사는 개인에 관한 풍부한 정보를 제공하지만, 극히 적은 표본이라는 단점이 있다. 반면, 교통카드데이터와 택시운행정보관리시스템(Taxi Operation Information Management System, 이하 TIMS) 데이터는 전수 통행데이터라는 엄청난 장점이 있지만 통행목적과 이용자특성에 대한 정보가 없다. 통신데이터인 생활이동데이터는 표본율이 가구통행실태조사 대비 높고 통행목적과 이용자특성에 대한 정보가 있지만, 교통수단이 구분되어 있지 않다.

이처럼 필요한 속성이 있는 표본 데이터와, 전수 데이터이지만 해당 속성이 없는 데이터가 존재하며, 각 데이터의 장점을 적절히 활용할 필요가 있다.

[표 1] 데이터 특성

데이터 구분	표본율	이용자특성	통행목적	수단 구분
가구통행실태조사	0.25%	성별, 연령, 소득 등	○	○
교통카드데이터	100%	아동/청소년/고령자 구분	X	○
TIMS 데이터	100%	X	X	○
통신데이터	23.97%	성별, 연령대	○	X

※ 교통카드 및 택시 데이터의 표본율은 100%에 가까우므로 100%라 표기함

※ 통신사 점유율 출처: http://www.smartchoice.or.kr/smc/smartreport/service_03_client.do

교통데이터의 통행목적과 이용자특성 추정에는 AI(Artificial Intelligence)모형 중 분류모형과 생성모형이 적용되고 있다. 분류모형과 생성모형은 가구통행실태조사에 포함된 통행정보, 통행목적, 이용자특성을 학습하여 교통카드데이터의 통행목적과 이용자특성을 추정한다. 기존 연구들에서도 유사한 방법론을 확인할 수 있었으며, 이 연구 또한 이를 참조하여 서울시 교통데이터 활용에 적합한 모형을 구축하였다.

AI모형의 정확도, 통행목적은 86%로 높고, 연령은 37%로 낮아 한계 있어

이 연구에서는 대중교통과 택시 각각의 AI모형을 구축하였다. 택시 AI모형의 경우, 표본이 매우 부족하여 신뢰성이 부족하다고 판단된다. 대중교통 AI모형은 입력 변수와 모형종류에 따라 총 15가지를 구축하였다. 구축된 모형별로 통행목적, 연령대, 성별에 대한 정답률과 세 속성의 결합분포 유사도를 평가하였다.

통행목적에 대한 정답률은 분류모형(XGBoost)에서 최대 86%의 정확도를 나타냈고, 성별 정답률은 최대 61%의 정확도를 보였다. 연령대에 대해서는 최대 37%로 낮은 정확도를 보여 한계가 있는 것으로 보였다.

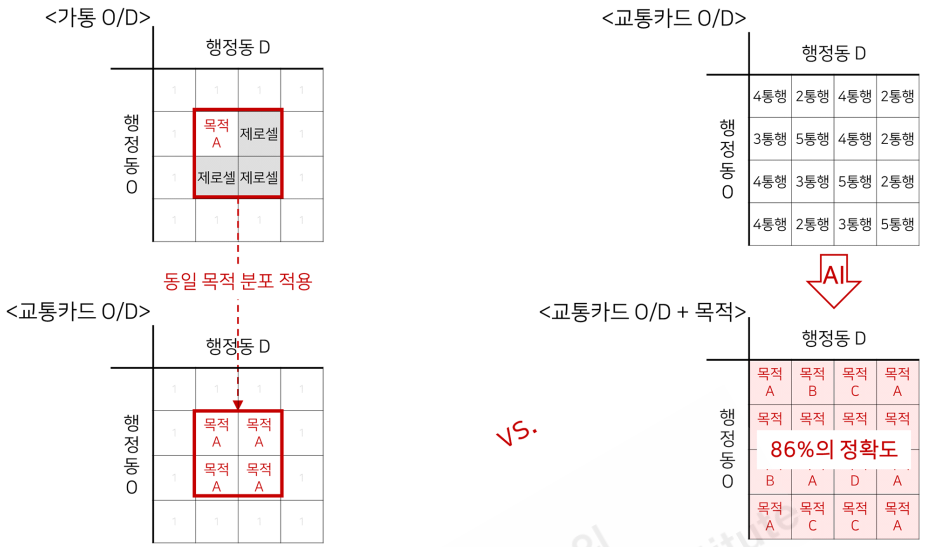
분포 유사도는 분포를 목적으로 학습하는 생성모형보다 오히려 분류모형에서 높게 나타났는데, 이는 표본의 부족과 속성의 개수가 적어 나타나는 현상으로 해석된다.

개별속성에 대한 정답률과 분포 유사도가 모두 전반적으로 양호한 모형은 분류모형과 생성모형에서 각각 2개 도출되었다. 분석 목적에 따라 모형을 선택하여 활용할 수 있다.

[표 2] 모형 구축 결과

모형 구분			개별속성 정답률			분포 유사도		모형 평가
			통행목적	연령대	성별	KLD	JSD	
분류 모형	XGBoost	가통	**0.86**	**0.36**	**0.61**	0.08	0.14	개별속성 추정 우수
		가통, 통신	**0.86**	**0.36**	**0.59**	0.06	0.12	
		가통, 통신, 토지	**0.86**	**0.37**	**0.59**	0.06	0.12	
	LSTM DNN	가통	0.79	0.31	0.55	0.04	0.07	전반적으로 우수
		가통, 통신	0.78	0.31	0.54	0.04	0.07	
		가통, 통신, 토지	0.57	0.16	0.50	**0.02**	**0.05**	유사분포 추정 우수
	Transformer DNN	가통	0.79	0.31	0.54	**0.02**	**0.05**	
		가통, 통신	0.79	0.31	0.53	**0.02**	**0.04**	
생성 모형	LSTM CVAE	가통	0.82	0.32	0.56	0.04	0.07	전반적으로 우수
	Transformer CVAE	가통	0.81	0.31	0.54	0.06	0.09	

위의 AI모형을 활용하면 다양한 통행정보를 바탕으로 최대 86%의 정확도로 통행목적을 추정할 수 있다. 기존 전수화 방법론에서 나타나는 표본의 과대대표와 주변 지역의 목적 분포를 그대로 사용하는 문제가 해결될 수 있다.



[그림 2] AI모형을 통한 대중교통 전수화 정확도 향상

AI 활용시 O/D 신뢰성 향상 위해 지도앱 등 활용, '가통조사' 대체 가능

전체 O/D와 대중교통 O/D의 차이는 곧 승용차 O/D와 유사하기 때문에, 대중교통 O/D의 정확도 향상은 곧 승용차 O/D의 정확도 향상으로 귀결된다. 승용차 통행에 대한 전수 데이터 수집이 쉽지 않은 만큼, 대중교통 O/D 정확도 향상을 통한 승용차 O/D의 정확도 향상은 효율적인 방법이 될 수 있다.

가구통행실태조사 표본율이 점차 축소되고 있는 실정에서 O/D의 신뢰성을 향상시키거나 유지하려면 이 연구에서 구축한 AI모형을 활용하는 것이 대안일 수 있다. 그러나, AI모형 또한 신뢰성을 확보하려면 모형 구축에 필요한 속성을 지닌 최소한의 표본 확보가 필요하다. 이를 위해 T맵, 네이버맵, 카카오택시 등 많은 이용자를 확보한 모바일앱상에서 가구통행실태조사를 대체할 수 있는 조사를 실시할 수 있다. 모바일앱의 데이터베이스에 자동으로 축적되는 통행정보를 기본적으로 확보할 수 있으므로, 연령대·성별 정보만 간단한 추가조사로 얻을 수 있다면, 예산과 시간 측면에서 효율적으로 표본을 확보할 수 있을 것으로 기대된다.