2023년
공공부문 데이터 분석·활용 우수사례집



행정은 스마트하게,



국민은 편리하게





행정안전부 NIA 한국지능정보사회진흥원

Contents

PART 1 국민안전

1. 잡았다! '그놈 목소리' 세계 최초 범죄조직 잡는 한국형 음성분석모델 06
   행정안전부 통합데이터분석센터/국립과학수사연구원
2. 데이터 분석으로 주민의 안전을 찾다! 울주형 재난관리시스템 18
   울산광역시 울주군 에너지정책과
3. 노인을 위한 나라는 있다. 교통 빅데이터로 지정한 '노인보호구역' 28
   한국교통안전공단 디지털혁신처
4. 고속도로 낙하물 사고, AI 기술로 예방하겠습니다 38
   한국도로공사 ITS처
5. 안심하고 먹을 수 있게, 빅데이터로 식품 제조공정을 관리하다 48
   한국식품안전관리인증원 IT전략팀
6. 지역사랑상품권 부정 사용 NO! 숨어봤자 빅데이터 손바닥 위 60
   한국조폐공사 기술처 기술개발부

PART 2 공공행정

1. CCTV 영상을 업로드하면 차종별 교통량이 보인다 74
   행정안전부 통합데이터분석센터
2. 빅데이터 효자손으로 긁어드립니다. 지역이 균형있게 발전하도록 86
   충청남도 예산군 전산정보팀
3. 차곡차곡 쌓인 소방 활동 데이터, 과학 소방정책 이끈다 98
   경남소방본부 소방행정과
4. 중학교 배정으로 인한 고충민원, 민원 데이터 분석으로 개선책을 찾다 108
   국민권익위원회 민원정보분석과

PART 3 산업경제

1. 흔들림 없는 천연가스 공급, AI 수요예측 모형으로 가능해졌습니다 120
   한국가스공사 스마트기술부
2. 정전을 막아라! 기상 데이터와 전력 데이터의 만남 130
   기상청 기상융합서비스과
3. 우리 경제, 지금 어떤 상태지? 가장 빠른 답은 통계속보에 있다 142
   통계청 경제동향통계심의관 서비스업동향과

PART 4 관광활성화

1. 지역관광을 부흥시킬 지름길, 여기 있습니다! 154
   한국관광공사 관광컨설팅팀
2. 내 마음에 쏙 든 철도여행, 알고 보니 빅데이터 덕분이었다 164
   한국철도공사 미래기술처

PART 1. 국민안전

PART 1
국민안전

1. 잡았다! '그놈 목소리' 세계 최초 범죄조직 잡는 한국형 음성분석모델
   행정안전부 통합데이터분석센터/국립과학수사연구원
2. 데이터 분석으로 주민의 안전을 찾다! 울주형 재난관리시스템
   울산광역시 울주군 에너지정책과
3. 노인을 위한 나라는 있다. 교통 빅데이터로 지정한 '노인보호구역'
   한국교통안전공단 디지털혁신처
4. 고속도로 낙하물 사고, AI 기술로 예방하겠습니다
   한국도로공사 ITS처
5. 안심하고 먹을 수 있게, 빅데이터로 식품 제조공정을 관리하다
   한국식품안전관리인증원 IT전략팀
6. 지역사랑상품권 부정 사용 NO! 숨어봤자 빅데이터 손바닥 위
   한국조폐공사 기술처 기술개발부

잡았다! '그놈 목소리' 세계 최초 범죄조직 잡는 한국형 음성분석모델

보이스피싱 사기범 식별 등
과학수사를 위한 음성데이터 분석

행정안전부 통합데이터분석센터/국립과학수사연구원



이렇게 바뀌었습니다!

AS-IS	TO-BE
• 한국어 기반 분석모델 부재에 따른 외국산 분석모델 사용	• 한국어 기반 딥러닝 분석모델 개발 및 활용
• 범죄 유사도에 따른 군집화가 어려워, 여죄 수사 및 입증 곤란	• 범죄 유사도에 따른 군집화 가능으로 여죄 수사 가능
• 인력/기술 한계로 대량의 음성데이터 분석 불가능	• 보유한 누적 음성데이터의 절반가량 분석 완료

추진 배경

최근 경찰청의 발표에 따르면, 지난 5년간(2018~2022년) 국내에서 발생한 보이스피싱 피해 건수는 총 15만 6,249건으로, 피해액이 3조 원을 넘어서는 등 국민 생활에 심각한 영향을 미치는 것으로 나타났다.

그동안 국내에서는 국립과학수사연구원(이하 국과수)이 러시아·영국에서 개발한 음성분석모델을 활용해 보이스피싱 수사에 필요한 음성 감정을 진행하였으나, 외국어로 학습된 분석모델 특성상 한국어 사용자의 동일 화자 여부 판단이 어려워 음성 감정 불가 사례가 지속해서 발생하였다.

특히, 보이스피싱 범죄조직은 통상적으로 그룹(수사관·검사 역할 등)을 이루어 활동하지만, 외국산 모델에서는 그룹별 군집화 기능이 없었고 오직 화자(話者) 구분만 가능했다. 러시아에서 군집화 유사 분석모델을 개발한 바 있지만, 사건별 연루자들을 그룹화한 기능은 없었다.

또한, 분석모델의 구매 및 유지 관리에 큰 비용이 소요되어 사용자 확장성이 부족하였으며, 경찰청 및 각 시·도 경찰서에서 다량의 보이스피싱 범죄자 음성데이터를 확보하고 있었지만, 이를 분석할 수 있는 시스템이 없어 수사의 어려움을 겪고 있었다.

이러한 문제점을 해결하고자, 화자 구분의 정확도를 높이고 보이스피싱 범좌자 식별과 군집화까지 가능한 한국어 기반 음성분석모델 개발을

목표로 설정하였으며, 행정안전부 통합데이터분석센터를 중심으로 국과수와 협업체계를 구축, 분석모델 개발을 추진하였다.

활용데이터

연번	데이터명	기준연도	형태	내 용	보유기관
1	보이스피싱 신고데이터 1	2015~2019	m4a, wav, mp3등	국민이 신고한 보이스피싱 피해 통화녹음 파일	금융감독원
2	보이스피싱 신고데이터 2	2020~2023	m4a, wav, mp3등	국민이 신고한 보이스피싱 피해 통화녹음 파일	충청남도경찰청, 서울 경찰청
3	검증데이터 1	2017~2021	wav	실제 사건 관련자 음성	국립과학수사연구원
4	검증데이터 2	2015	wav	다양한 환경에 놓인 발화자의 음성파일	국립과학수사연구원

데이터 비식별화 작업 및 발화자 인식 모듈 개발

먼저, 보이스피싱 신고 음성데이터에 있는 피해자들의 정보 노출을 방지하고자 데이터 보유기관과 협의하여 비식별화된 데이터를 확보하였다.

아울러, 음성의 변화 폭은 발화자(發話者)의 신체적, 감정적, 환경적 요인에 따라 다르므로, 음성데이터 추출·분석 전문기술을 보유한 서울대학교 산학협력단과 함께 발화자 인식 모듈을 개발하였다.

한국어 기반의 세계 최초 보이스피싱 예방 음성분석모델 개발

음성분석모델을 개발하기 위해서는 발화자의 동일인 검증이 최우선이었으며, 이를 위한 데이터 및 모델링 과정을 추진, 국내·외 발화자 6천여 명으로부터 추출한 100만 개의 음성데이터를 활용했다.

특히, 한국어의 경우, 10만여 개의 일반인 음성데이터와 국과수가 보유 중인 보이스피싱 사기범의 음성데이터를 함께 사용하였다.

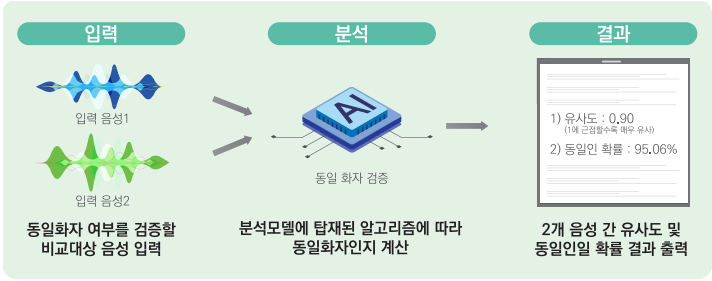
이후, 다양한 학습 과정 및 반복과정을 거쳐 최적의 알고리즘을 도출하였으며, 한국어 기반의 보이스피싱 예방 음성분석모델을 세계 최초로 개발하게 되었다.

동일 화자 여부 정확도 향상, 동일 범죄 가담자 그룹화

개발한 음성분석모델은 크게 두 가지 기능이 있다.

첫째, '동일 화자 여부 검증'으로, 두 개의 음성파일을 입력하면 두 음성이 동일인에 의해 발성되었을 확률값(유사도)을 확인할 수 있다. 보이스피싱 범죄자 검거 시 분석모델을 활용해 기존 범죄자 음성 데이터들과 비교한 후 일치 여부에 따라 여죄를 확인하는 데 사용할 수 있다.

〈그림 1. 보이스피싱 수사를 위한 음성데이터 분석 과정〉



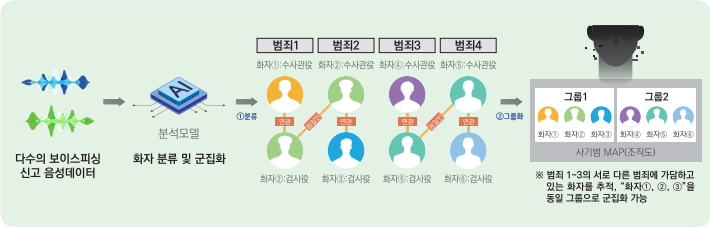
둘째, 보이스피싱 사기범에 대한 '범죄 가담자 그룹화'이다.

규모가 큰 사기는 통상 여러 명이 그룹을 이루며 각자 맡은 임무를 수행한다는 점에 착안하여, 사건별 범죄자들의 목소리를 연쇄 비교하는

과정을 거쳐 발화 특징이 유사하고 연관성 있는 음성들을 군집화하여 사기범 조직을 특정할 수 있다.

이를 이용해 조직원 한 명을 검거하면, 함께 활동한 구성원들을 체포할 가능성을 높일 수 있으므로 해당 기능을 중점적으로 연구하였다. 아래 그림과 같이 사건(범죄 1~4)별로 범죄자 목소리를 연쇄적으로 비교한 후 동일인 여부를 파악하고, 연루자들을 차례차례 군집화하면서 사기 집단의 윤곽을 확인하는 식이다. 연관 음성을 군집화할 수 있는 기술은 사실상 세계 최초로 구현했다고 볼 수 있다.

〈그림 2. 보이스피싱 범죄 가담자 그룹화 〉



외국산 모델 대비 음성 판독률 약 77% 향상

음성분석모델의 성능 검증을 위해 1차 150명의 660여 개, 2차 200명의 12,000여 개 음성데이터를 각각 사용하여 평소 발성/위장 발성, 일반녹음/무선 통화녹음/유선 통화녹음 등 다양한 환경과 상황을 가정하였다. 그 결과, 범죄자의 음성을 정확하게 판별하는 판독률이 기존 외국산 모델 대비 약 77% 정도 향상된 것으로 나타났다. 그리고 동일화자 및 군집화에 대한 분석모델의 정확도는 약 96.6% 수준으로 확인되었다. 여기서 의미하는 '정확도'는 분석모델이 '범죄자를 범죄자로 판별'한 경우의 신뢰도 수준에 해당한다고 보면 된다.

개발된 모델을 활용하여 12,323개의 범죄자 음성데이터를 분석한 결과에 따르면, 2회 이상 범죄 가담자가 2,741명이었고, 한 사람이 최대 34회까지 범죄를 저지른 경우가 있다는 것이 확인됐다. 또한, 총 235개 범죄조직에서 633명이 2,860건 이상의 범죄에 가담한 사실도 밝혀냈다.

〈표 1. 범죄조직 규모 및 가담 범죄 수〉

조직 규모	인원	2명	3명	4명	5명	6명	8명	9명	10명	18명
	조직수 (235건)	160개	47개	11개	7개	5개	1개	1개	1개	2개
가담 범죄수 (2,866건)		1,188	614	232	255	184	47	52	70	224

음성분석모델 특허 출원

한편, 행정안전부는 개발한 '보이스피싱 예방 음성분석모델'과 관련, "군집화 기법을 기반으로 하는 음성파일 관리 장치 및 그 방법"{대한민국 특허 출원번호 10-2023-0044590(2023.4.5.)}으로 특허 출원을 마쳤다.

범죄 수사에 적극 활용 및 해외 확산 추진

행정안전부는 '보이스피싱 예방 음성분석모델'이 성공적으로 개발됨에 따라 국과수, 경찰청과 협력해 보이스피싱 조직 수사와 검거 과정의 적극적인 활용을 지원하고 있다.

올해 2월 말부터 국과수는 보이스피싱범 목소리 감정에 개발 모델을 활용하고 있으며, 경찰청은 10월 중순부터 '전화사기 수사지원시스템'에 탑재하여 전국 경찰 수사 현장에 개발 분석모델을 사용하고 있다.

아울러, 몽골 등 개도국 수사기관을 대상으로 연수 프로그램을 통해 한국형 음성분석모델을 공유하였으며, 국제적인 전문학회 등을 통해 개발 성과를 발표할 예정이다.

기관소개

• 행정안전부 통합데이터분석센터



행정안전부 통합데이터분석센터는 데이터기반행정법 제정(2020)에 따라 중앙부처, 지방자치단체, 공공기관을 아우르는 범정부적 관점에서 데이터 분석과 활용을 지원하기 위해 2021년에 신설되었다.

데이터 분석·활용 및 데이터 분석·활용 역량강화 지원, 범정부 데이터 분석시스템 구축·운영 등을 통해 각 기관의 객관적·과학적 의사결정과 이를 통해 좋은 정책이 만들어질 수 있도록 지원하고 있다.

• 국립과학수사연구원



국민의 기본권 보장, 국민의 생명과 안전을 지키기 위해 1955년에 설립된 국립과학수사연구원은 범죄 수사 증거물에 대한 과학적 감정 등을 통해 사건을 해결하고 범인을 검거할 수 있도록 지원하고 있다.

국립과학수사연구원은 감정 장비 현대화로 과학수사 역량을 제고하고 감정기법 표준화에 의한 감정의 신뢰성과 투명성 확립을 통해 독립적이고 공정한 감정기관으로 거듭나고 있다.

Coffee Time (인터뷰)



현재 하시는 일을 포함해 자기소개 부탁드립니다.

(통합데이터분석센터) 저는 행정안전부 통합데이터분석센터에서 빅데이터 분석 및 활용 업무를 맡고 있는 김현우 주무관입니다. 컴퓨터공학(IT)을 전공했고 직장에서 정보보호 관련 업무를 계속 담당했는데 데이터를 다루는 일은 익숙하지 않았습니다.

(국과수) 저는 국립과학수사연구원 디지털과 오디오 포렌식실에서 근무하는 박남인 연구사입니다. 정보통신공학을 전공했고 음성/오디오 신호처리로 석·박사 학위를 취득했습니다. 2016년 알파고-이세돌 대국 이후 딥러닝 기법이 인기를 끌었을 때 음성 감정 업무에도 해당 기술을 접목하려고 노력했는데 이번에 결실을 본 것 같습니다.

이번 과제는 어떻게 참여하게 되었고, 기존 일과 어느 정도 관련이 있었나요?

국과수는 2015년부터 금융감독원과 '전화금융사기 근절을 위한 업무협약'을 체결해 실제 보이스피싱 사기범의 통화 녹음파일을 계속 가공 처리하고 있었습니다. 피해자 음성을 제거하고, 사기범 음성만 추출해 아카이브 형태로 데이터베이스를 구축하다가 하나의 음성파일에 다수의 조직원이 등장하는 것을 확인했는데요. 이걸 잘 활용하면

보이스피싱 사기범 조직을 군집화할 수 있겠다는 확신이 들어 행정안전부 통합데이터분석센터와 NIA의 공공 빅데이터 분석사업에 아이디어를 제출했습니다. 이후 범부처 과제로 선정되어 이번 모델을 성공적으로 개발할 수 있었습니다.



과제 추진 시 참여했던 관계자는 누구였나요?

본 과제 참여기관은 행정안전부 통합데이터분석센터, 국립과학수사연구원, 서울대학교이며, 행정안전부 통합데이터분석센터는 총괄기관으로 범죄 수사 현장의 요구를 충분히 반영하여 분석 과제를 발굴하고 상세 기획부터 활용 방안까지 과제 전반을 관리했습니다.



과제를 수행하면서 어려운 점이 있었다면 말씀해주세요.

이 분석 과제는 음성을 전문적으로 분석해야 했기에, 성공적으로 모델을 구현할 수 있는 기관을 찾는 게 가장 어려웠습니다. 또한, 약 6개월 만에 과제를 완료해야 했기에 시간과 사투를 벌이듯 진행해야 하는 것도 힘들었어요.



데이터 분석 시 어떤 알고리즘을 사용했나요? 그 과정에서 생긴 시행착오를 어떻게 극복하셨는지 노하우가 있다면 알려주세요.

하나의 음성파일에 사기범 2인 이상이 등장하면 '음성파일명-01', '음성파일명-02'와 같이 구분합니다. 기존 화자인식 모델은 음질이 좋은 데이터로 훈련된 모델인데, 실제 해외에서 걸려 온 저음질의 실험데이터와 도메인이 불일치 하는 문제가 생겨서 화자인식 모델을 저음질 데이터로 다시 훈련해야 했어요. 이 문제를 해결하는데 시간이 다소 부족했습니다. 분석 노하우라고 하면... 글쎄요. 검증 데이터세트를 구체화하고 단계적으로 접근할 필요가 있다고 말씀드리고 싶네요.



개발 모델을 공유, 확산하기 위해 어떻게 노력하셨는지 듣고 싶습니다.

보이스피싱 음성분석모델을 수사에 적극적으로 활용하기 위해 국과수, 경찰청 등과 긴밀하게 협업을 진행하고 있습니다. 모델 개발 이후, 국과수의 음성 감정 및 경찰청 '전화사기 수사지원시스템'에 활용되고 있으며, 보이스피싱범 음성 즉시 판별 및 비교·분석이 가능해져서 범죄자 특정, 영장 신청 및 검거 등 보다 빠른 수사가 이루어질 전망입니다.

아울러, 몽골, 인도네시아 등 개도국 대상 연수 프로그램을 운영하여 최신 음성 감정기법을 공유하였으며, 전문학회 참석을 통한 모델 개발 성과를 발표하였습니다.



데이터를 활용해 업무를 개선하고 싶으나 선뜻 나서기가 망설여지는 분들께 도움이 될 만한 말씀 부탁드립니다.

먼저, 국과수에서는 이번 음성데이터를 확보하는 데만 8년이 걸렸습니다. 정기적으로 양질의 데이터를 확보하고 실패를 두려워하지 말며, 문제를 잘 정의하고 명확한 목표를 설정하는 게 중요하다고 생각해요. 또, 협업하면 광범위한 관점과 아이디어를 얻을 수 있고, 데이터 기반의 의사결정과 업무 개선은 꾸준한 학습과 연습을 통해 이뤄진다고 생각합니다.

그리고, 이번에 제작되는 데이터 분석·활용 우수사례집의 내용이나 범정부 데이터 분석시스템(insight.go.kr)의 사례들을 참고해 데이터기반행정에 대한 이해를 넓힌다면, 데이터기반행정이 어렵게 느껴지지 않을 것입니다. 또한, 소속기관이나 동일 분야에서 추진한 과제 등을 하나, 둘 살펴보면, 현장에 적용할 만한 아이디어를 찾을 수 있을 거예요.

PART 1-2. 국민안전

데이터 분석으로 주민의 안전을 찾다! 울주형 재난관리시스템

데이터 분석 활용을 통한
재난 업무의 디지털 전환

울산광역시 울주군 에너지정책과



이렇게 바뀌었습니다!

AS-IS	TO-BE
• 안전취약계층에 대한 불분명한 정의로 구체적인 재난대응 미비 • 재난대응체계를 텍스트 위주로 명시하여 현실과 동떨어진 대응 다수 발생	• 안전취약계층을 세분화하고 거주지 통계를 반영한 데이터 기반 재난대응 모델 수립 • 데이터 융합분석을 통해 재난 관련 정보를 현행화하여 실효성 있는 재난 대응 가능

추진 배경

울산광역시는 2016년 태풍 '차바'로 인명, 재산 피해가 컸고 크고 작은 규모의 지진도 계속 발생해 재난 대비·대응이 중요해졌다. 특히 울주군은 원자력발전소 12기가 소재하거나 인접해 있어, 방사능재난에 대한 효율적 대응 및 주민 안전을 위한 신속·정확한 맞춤형 정보 제공의 필요성이 대두되었다. 그러나 문서에 의존하는 텍스트 위주의 재난대응 매뉴얼 체계로는 직관적인 상황판단이나 대응에 한계가 있었다.

만일 방사능재난이 발생한다면, 방사선비상계획구역(30km) 내 거주자 90여만 명(새울·고리 기준)이 대피해야 하며, 이는 울산 인구의 약 90%에 해당한다. 따라서 대규모 인원의 대피와 이재민 관리를 위해서는 더욱 더 체계적이고 과학적인 계획이 필요했다. 특히, 고령자, 장애인 등 이동이 어렵거나 디지털 정보에 취약한 사람들을 위한 세심한 재난 대응 내용도 포함되어야 했다.

이에 울주군은 지금이 재난관리의 패러다임을 완전히 전환해야 할 시점이라고 판단했다. 과거 경험에 의존하기보다는 명확한 데이터에 근거해 재난관리계획을 세워야 할 때였다. 그래서 울주군은 몇 가지 내·외부 데이터를 분석하고 융합해서 이전과는 다른 재난관리시스템을 만들어보기로 했다.

활용데이터

연번	데이터명	기준연도	형태	내 용	보유기관
1	주민전산데이터	2022	csv	취약인구 분석	새올시스템(내부)
2	이재민구호소	2022	xlsx	수용인원파악	울주군 에너지정책과
3	도로 (링크/노드)데이터	2022	shp	차량 속도 변화 발생점	공공포털(외부)
4	GIS (기초행정구역) 정보	2022	shp	지역 공간정보	자체구축(공간다듬이)

분석과정

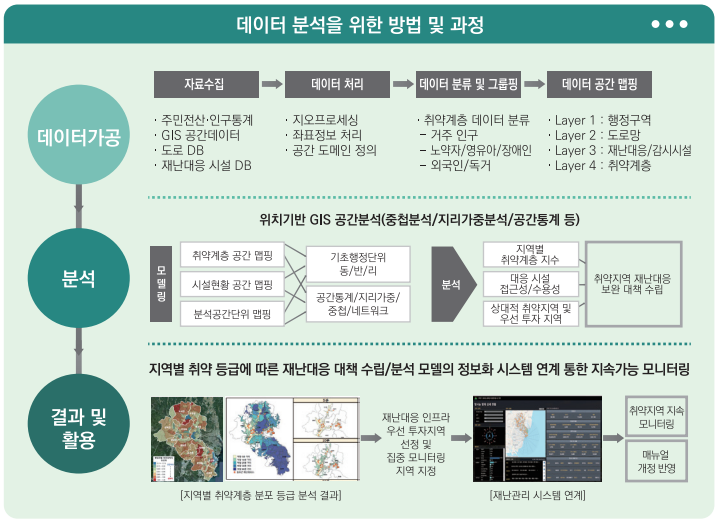
일단 울주군에서는 재난에 대비해 다양한 데이터를 모았다. 이미 보유하고 있는 내부 데이터 즉, 집결지 및 구호소 정보 180건, 안전취약계층 확인을 위한 주민 전산 데이터 23만 건, 수송로 확인을 위한 3,375㎞의 도로(링크/노드) 데이터, 390통의 GIS(기초행정구역)[1] 정보를 기본으로 수집했다. 또한, 각각 다른 공간정보를 가지고 있는 행정리·법정리 기반의 지도와 지리 데이터를 다듬어(공간다듬이 사업)[2]최신의 공간 데이터를 활용할 수 있도록 했다.

아울러, 공공데이터포털, 빅데이터포털, 국립재난안전연구원, 한국국토정보공사에서 제공하는 각종 데이터를 모았으며, 공간정보 오픈 플랫폼 및 지도, 보건의료 빅데이터 개방시스템 등도 적극적으로 활용했다.

그리고 안전취약계층(고령자, 장애인 등)을 위한 실질적인 재난대응 대책 수립을 위해 울주군은 안전취약계층에 관한 지표를 산정하기로 했다. 이를 위해 데이터를 지오코딩[3]을 이용하여 그룹 분류, 공간 정보화 및 데이터 중첩 등의 전처리 작업을 함께 수행했다.

1) 지리정보체계. 지리공간 데이터를 분석, 가공하여 활용할 수 있는 시스템
2) 한국국토정보공사가 울주군과 함께 2015년부터 추진하고 있는 사업으로, 공간정보를 다듬어 과학행정을 지원함
3) 주소 등의 고유명칭으로 위도와 경도의 좌푯값을 얻는 것

〈그림 3. 데이터 분석 과정 〉



다른 재난과 달리, 방사능재난이 발생하면 주민들은 먼 거리(30km 바깥)에 있는 구호소로 가야 한다. 하지만 노인, 미취학아동, 장애인 등 안전취약계층은 이동과 각종 시설 이용에 제약이 많기에, 이를 해결하기 위해서는 정확한 공간분석이 필요했다. 그래서 울주군은 GIS 공간분석(중첩분석, 지리가중분석, 공간통계 등)을 실시했다. GIS에 안전취약계층 공간정보, 시설현황 정보, 분석 공간 단위를 대응시켜 모델링하고, 이를 단순히 일회성으로 분석하여 활용하지 않도록 정보화 시스템과 연계하여 모니터링을 지속할 수 있도록 했다.

또한, 광범위한 방사능재난의 특성을 반영하기 위해 이재민구호소의 수용률을 포함한 경우와 제외한 경우 두 가지 유형으로 분석하여 재난대응 사각지대를 도출했다. 구호소가 원전 30km 바깥에 있으므로 이동에 영향을 미치는 불량도로 지표 또한 반영하여 분석했다.

정확한 통계가 정확한 재난대응 체계를 세운다

이러한 분석과정을 통해 울주군의 안전취약계층 지수와 이재민구호소의 접근성, 수용률을 파악할 수 있었으며, 아울러 지역마다 재난대응을 해야 할 중점이 달라야 함을 알게 되었다.

울주군은 12개의 행정구역(6읍, 6면)[4]이 있으며, 전체 인구는 약 22만 명이다. 그중 노인은 약 3.6만 명(16.3%), 미취학아동은 약 6,600명(3.0%)으로 대다수가 아파트, 주택 등에 살고 있는 것으로 나타났다. 장애인은 약 1.2만 명(5.3%)으로 그중 절반 정도가 65세 이상 노인층이며, 주로 인구수가 적은 읍·면이나 장애인 복지시설 등에 거주하고 있었다.

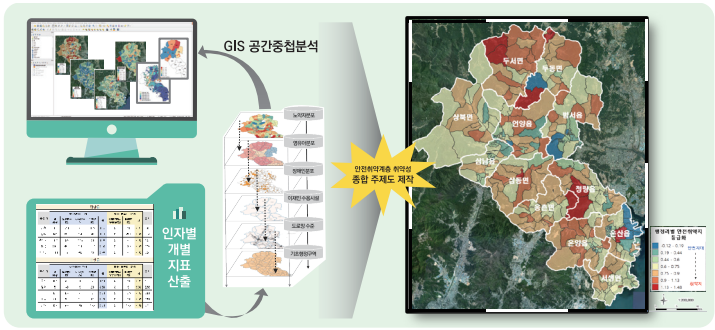
불량도로 비율은 약 19.9%이며 두동면(26.8%), 서생면(24.2%) 순으로 높았다. 울주군 내 이재민구호소는 총 35개소이며 장기 수용할 수 있는 인원은 4만여 명이었다. 이재민구호소 수용률은 상북면(44.4%), 언양읍(43.6%), 범서읍(25.9%) 순이고, 수용인원은 범서읍(1.8만 명), 언양읍(1.2만 명), 삼남읍(0.5만 명) 순으로 나타났다.

방사능재난 시 이재민구호소는 원전으로부터 30km 떨어진 곳(언양읍, 범서읍, 삼남읍 이북 지역)만 이용할 수 있으므로, 해당 지역은 다른 곳보다 구호 물품과 자원봉사자를 더 많이 확보해야 했다. 반면 원전 인근 지역은 신속하고 효율적인 대피가 가능하도록 장기간 차량 이동을 최소화하고 폭 4m 미만 불량도로의 개선 등의 계획 마련이 필요했다.

이러한 데이터 분석 결과는 재난 대응체계에 관한 정확한 통계를 제시하고 태풍, 홍수, 지진, 방사능 등 다양한 재난에 맞는 정책을 수립할 수 있는 근거자료가 되었으며, 이를 바탕으로 울주형 재난관리모델을 만들 수 있었다.

4) 범서읍, 삼남읍, 온양읍, 온산읍, 언양읍, 청량읍, 두동면, 두서면, 삼동면, 상북면, 서생면, 웅촌면

〈그림 4. 지역별 안전취약계층의 안전 취약성 종합분석 〉



〈그림 5. 빅데이터 기반 재난대응 체계 수립 〉



정책 활용 및 성과

울주형 재난관리모델은 안전취약계층을 하나로 묶지 않고 노령층, 미취학아동, 장애인 등으로 세부 분류하여 지역별 취약계층 분포도를 분석한 후, 재난 대응 대책을 좀 더 체계적이고 지역 실정에 맞게 수립했다는 데 의의가 있다.

울주군은 향후 재난관리모델을 통해 드러난 재난 대응 사각지대 및 인프라 우선 구축지역, 집중 관리지역을 지속해서 모니터링할 계획이다. 또한, 재난관리모델을 고도화하여 스마트 방사능 방재 시스템에 탑재하여 다양하게 활용하면서 대민 서비스를 더욱 강화할 예정이다.

기관소개(울산광역시 울주군청)

울산광역시 울주군은 1997년 울산광역시 승격에 따라 울주구에서 울주군으로 개편되었으며, 한반도에서 해가 가장 먼저 뜨는 간절곶이 있는 아름다운 지역이다. 울산광역시 전체의 70%가 넘는 넓은 면적과 1차 산업에서부터 미래 첨단 산업에 이르기까지 모든 산업 분야를 두루 갖추고 있는 산업 중심지이며, KTX 울산역이 개통되면서 교통·물류의 중심지, 동남권 개발의 중심으로 거듭나고 있다.



Coffee Time (인터뷰)



현재 하시는 일을 포함해 자기소개 부탁드립니다.

저는 울산광역시 울주군 에너지정책과에서 원전방재업무(원전사고 대응)를 맡고 있는 공해열 주무관입니다. 순수화학 전공자로 저는 원래 '데이터'라고 하면 휴대전화 요금제에 나오는 '데이터양(?)' 밖에 모르는 사람이었는데요. 하지만 이전 직장에서 엑셀과 데이터기반 업무를 하면서 데이터 관련 업무 환경이 낯설지는 않았습니다. 공무원이 된 후, 그 경험을 행정에 녹이게 되면서 현재와 같은 업무를 할 수 있게 된 것 같습니다.

과제를 추진하면서 어려웠던 점에 관해 듣고 싶습니다.

안전취약계층을 분석할 때 기존 사례가 없어서 어떤 데이터를 가지고 어떤 방식으로 분석해야 하는지 도무지 감을 잡을 수가 없었습니다. 특히 재난 종류, 분석 방법에 따라 취약 정도를 집계하는 식이 있는데, 어느 정도로 가중치를 둬야 할지 기준점이 전혀 없었죠. 여러 차례 분석하고 그 값의 평균을 도출하는 방식으로 가중치를 보정하는 것이 너무 힘들었습니다.



안전취약계층을 재난대응모델 구축의 중심에 뒀던 특별한 이유가 있었나요?

정부 지침에 따라 재난 유형에 따른 안전취약계층 대책을 수립해야 했지만, 안전취약계층이 정확히 누구를 지칭하며, 우리 지역에 얼마나, 어떤 방식으로 거주하고 있는지 알 수 없었습니다. 또, 인력이나 장비를 어느 정도로 준비해야 이분들을 충분히 대피시킬 수 있고, 만약 자원이 부족하다면 안전취약계층 내에서 우선순위를 어떻게 정해야 할지 등 데이터가 너무 부족했죠. 이 문제를 명확히 하는 게 재난대응모델 구축의 시작이라고 여겼고 울주군 안전취약지역 분석을 기획하게 된 것입니다.



공간다듬이사업의 도움을 받아 공간 데이터를 확보하셨다는 내용이 있던데요.

울주군은 법정리와 행정리가 같이 운용되는 도농복합 도시인데, 행정리 공간정보는 행정적으로 필요한 것이고, 매년 조례에 따라 변경됩니다. 그런데 민간 포털사이트에는 해당 데이터가 제공되지 않아서 정확한 공간 데이터를 얻을 수 없었습니다. 그래서 한국국토정보공사와 우리 군 토지정보과에서 진행하고 있는 공간다듬이 사업의 협조를 받았고 덕분에 데이터 확보가 가능했죠. 이처럼 데이터 관련 업무가 협업이 참 중요합니다.



개발모델을 구체적으로 공유하고 활용하기 위해 어떤 노력을 하셨는지요.

분석 결과를 반영해서 재난대응 매뉴얼을 만들었고, 우리 군의 데이터 분석·활용 사례 등을 중앙부처에 제출해서 울주군의 재난대응모델을 알리려고 노력했어요. 제가 외부에서 적극행정에 관해 강의를 진행할 때도 적극적으로 사례를 공유했죠. 민감 자료가 일부 포함돼 있어서 자료 전체를 공개하지 못하지만, 다른 기관에서 쉽게 접근할 수 있게 데이터 분석 모델링을 제공하고 있습니다. 이번 데이터 분석·활용 우수사례집에 나오게 되면 다른 지역에서 연락이 올 수도 있지 않을까요?



데이터 활용을 통한 업무개선에 부담을 느끼는 분들께 도움 되는 말씀 부탁드립니다.

처음에 말씀드렸듯 제게 '데이터'란 통신사 요금제에 나오는 단어였고, 빅데이터 분석을 진행하면서 많은 어려움이 있었습니다. '하지 말까'라는 생각도 수십, 수백 번을 했습니다. 하지만 완성 이후 결과물을 보니 정말 뿌듯해지더군요.

어떤 정책을 수립할 때 감(感)이나 누군가의 생각 등은 객관적이지도, 정확하지도 않습니다. 하지만 데이터 기반 업무는 '불확실성'이라는 위험을 최소화할 수 있고, 또 공공정책을 위한 설득, 설명 시 근거로 제공할 수도 있습니다.

데이터 분석이 매번 성공하는 건 아닙니다. 방향을 잘못 설정했거나 방법을 틀리게 썼다면 실패하겠죠. 하지만, 그 실패는 실패가 아닙니다. 다음 분석을 발전, 성공시킬 수 있는 발판이 된다고 생각합니다.. '데린이(데이터를 잘 모르는 사람)'라고 두려워하지 말고 도전한다면, 업무를 개선할 수 있는 좋은 사례가 반드시 나올 것이라고 믿습니다.

노인을 위한 나라는 있다. 교통 빅데이터로 지정한 '노인보호구역'

교통 빅데이터 분석을 통한
노인보호구역 적정지역 선정

한국교통안전공단 디지털혁신처



이렇게 바뀌었습니다!

AS-IS	TO-BE
• 노인복지시설/체육시설의 장이 신청할 시, 주변 조사 후 노인보호구역 선정 • 우선순위 미설정 및 사업추진 당시 여건을 고려한 정책 수립	• 각종 데이터 기반으로 노인교통사고의 위험도가 높은 지점을 선정 • 우선순위를 비교하여 한정된 예산으로 과학적인 정책 수립 가능

추진 배경

우리나라는 고령화가 급속도로 진행되는 나라 중 하나로, 고령화에 대한 대비가 사회 전반적으로 요구되고 있다. 2020년 노인교통사고 사망률은 전체 교통사고 사망률 중 무려 43.6%를 차지했다. 특히 2019년 기준 65세 이상 노인 인구 10만 명당 교통사고 사망자 수는 19.8명으로, OECD 회원국 중 1위를 차지하는 불명예를 안았다. 이에 한국교통안전공단(이하 공단)은 노인교통사고 예방 방법을 찾아 연구를 진행하게 되었다.

'노인보호구역'은 어린이보호구역처럼 자주 눈에 띄지 않는다. 또한, 법령상 노인보호구역에 대한 명확한 데이터 분석 기반의 기준값도 없다. 따라서 공단은 과학적 정책 의사결정을 위해 데이터를 기반으로 노인보호구역 지정 필요성이 높은 곳을 선정해 조례를 반영하는 것을 목표로 과제를 추진했다.

활용데이터

연번	데이터명	기준연도	형태	내 용	보유기관
1	교통사고 정보	2018~2021	csv	교통사고 날짜, 위치, 사고 항목 등	한국교통안전공단
2	교통카드 정보	2019~2021	csv	교통카드 정보	
3	위험 운전 통계	2019~2021	csv	DTG기반 급가속, 급감속 등 11대 위험 운전 통계	
4	노인 관련 시설 정보	2022	csv	경로당, 요양시설, 생활체육시설	대구광역시
5	가로망 주정차	2021	csv	가로망 주정차 정보	
6	유동 인구	2020~2021	csv	통신사 유동 인구	
7	용도지역	2022	shp	용도지역 정보	국토교통부
8	지역구분	2023	shp	대구광역시 행정 시군구 경계	
9	격자 정보	2022	shp	(250m x 250m)	

도로 데이터에 노인교통사고 관련 각종 데이터 매핑

공단은 노인보호구역 적정지역 선정 연구를 위해 대구광역시 등 약 10개 기관으로부터 데이터를 수집했다. 2018년부터 수집된 교통사고 데이터(약 211만 건), 통신사 유동 인구 데이터, 병원·약국 등 시설 데이터, 도로 데이터, 격자 단위의 공간 데이터 등을 이용해 모델을 개발했다.

데이터 분석 및 모델링 과정은 복잡하지만, 쉽게 요약하자면 아래 그림과 같다.

〈그림 6. 모델 개발 과정〉

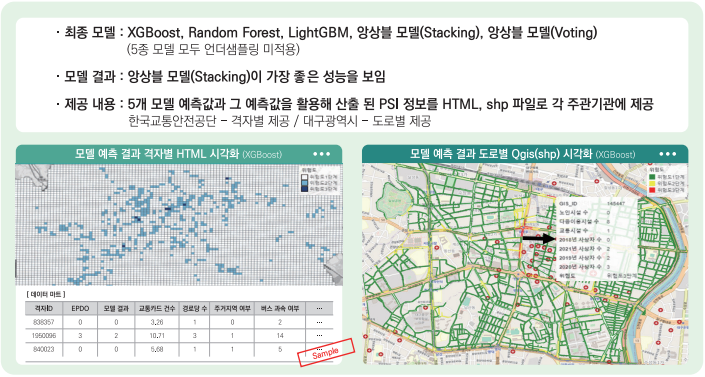


먼저, 공단은 전처리를 마친 교통정보, 시설 정보, 공간정보를 도로 데이터에 매핑한 후, 변수를 선정하고 이를 보완하기 위해 통계적 유의미 수준(T-TEST)을 검정했다. 그 결과, 1차로 78개의 변수를 선정하였고, 이후에 설명력을 높일 수 있는 중요변수 선택에 사용되는 알고리즘(Boruta)을 사용하여 15개의 변수를 선정하였다. 그리고 최종 데이터 마트(소규모 데이터 저장 및 분석시스템)를 생성한 후 예측 모델을 설계하기 위해 총 3개의 AI 기계학습 알고리즘을 적용해 보았다.

많은 데이터 수를 줄여나가는 작업(언더샘플링)을 통해 최종 모델을 선정하고, 생성된 다수의 예측 모델 결과를 결합(앙상블)해 학습시킨 후, 보다 정확도가 향상된 예측값을 도출하였다. 개발된 모델을 검증해 오차율을 개선했으며, 결과를 활용해 잠재안전지수(PSI)[5]를 생성하여 최종 결과를 산출하였다.

이 분석 결과를 QGIS로 시각화하였고, 운수안전컨설팅 지원시스템(COSAS) 내 보고서로 탑재하였다.

〈그림 7. 분석 모델 시각화 결과〉



5) 잠재안전지수는 Potential for Safety Improvement의 약자로, 예측된 사고에 비해 얼마나 많은 사고가 개선 가능한지 나타내는 지표이다. 값이 클수록 안전 개수 효과가 높은 지역이다.

노인교통사고 위험도로 한눈에 확인

노인보호구역 최적 입지 선정 모델은 머신러닝 모델 3종을 앙상블 기법을 활용하여 산출하였으며, 도로를 3단계의 위험도(위험/주의/안전)로 분류하였다.

〈표 2. 노인교통사고 발생 위험 도로 구분 결과〉

순위	위험도 구분	해당 도로 개수	해당 도로 비율
1	노인사고 위험	210개	0.1%
2	노인사고 주의	1,091개	0.6%
3	노인사고 안전	176,574개	99.3%

그 결과 210개의 도로가 노인교통사고 발생 가능성이 높은 위험 도로로 나타났다. 예시로, 대구광역시 북구를 중심으로 가장 위험한 도로 상위 5개를 산출해 보았는데 칠성시장역 인근으로 확인되었다. 이 결과를 지역 안전 보고서에 수록했으며, 운수안전컨설팅지원시스템에 탑재해 결과를 볼 수 있도록 했다.

〈표 3. 위험한 도로 순위〉

순위	지역	노인사고 주의 / 위험 도로 길이 합
1	칠성시장역 인근	632.48m
2	칠성시장역 인근	473.54m
3	태전근린공원 인근	429.96m
4	칠성시장역 인근	322.63m
5	침산공원 인근	227.11m

〈그림 8. 운수안전컨설팅지원시스템의 지역안전보고서〉



노인보호구역 지정 시 정책 결정 근거자료로 활용

노인보호구역 최적 입지 선정 모델은 도로별 교통사고 위험도를 제공해 노인교통사고 사상자 수를 감소시킬 수 있고 도로 위험도 결괏값을 참고해 교통사고 줄이기 캠페인 등을 진행할 수 있다.

모델에서 도출한 결과를 기반으로 행정안전부 안전개선과에서는 노인보호구역 지정 참고조례안을 작성해 전국 지자체에 배포하였고, 대구광역시에서도 조례 개정에 활용하였다.

노인 보행자에게 취약한 도로가 보호구역으로 선정되어 사고를 예방할 수 있다면 대한민국은 초고령화 사회를 대비하여 노인교통사고를 줄이고 교통안전 선진국으로 도약할 수 있을 것으로 기대한다.

기관소개(한국교통안전공단)

교통사고 예방을 위한 사업 및 교통 체계 운영, 관리 지원을 통한 안전한 교통환경을 도모하고자 1981년 설립되었다. 현재, 국토교통부 산하 준정부기관이며, 국민의 생명과 재산을 보호하는 역할과 함께 지속 가능한 미래 모빌리티 시대를 열어가기 위해 노력하고 있다.



Coffee Time (인터뷰)



현재 하시는 일을 포함해 자기소개 부탁드립니다.

저는 한국교통안전공단 미래모빌리티본부 디지털혁신처에서 공공데이터 개방 및 이용 활성화, 공모전, 데이터기반행정, 데이터 품질관리 업무를 맡고 있는 이유라 선임연구원입니다. 저는 교통공학을 전공했고, IT 전공은 아니었기 때문에 데이터를 잘 다루지 못했어요. 다만 2년째 공공데이터 업무를 하다 보니, 자연스럽게 관련 법이나 그 중요성에 대해서 알게 된 것 같습니다.

이번 과제는 어떻게 참여하게 되셨나요? 기존 일과 어느 정도 관련이 있었나요?

제가 근무하고 있는 한국교통안전공단의 디지털혁신처에서는 미래 모빌리티 시대에 발맞추어, 데이터 기반의 교통안전 정책연구, 교통사고 감소를 위한 각종 연구를 진행하고 있습니다. 저는 그중에서도 공단의 교통 관련 데이터를 총괄하면서 개방 및 이용 활성화를 위해 일하고 있습니다. 공단의 수많은 데이터를 활용해 교통사고를 줄일 방법이 없을까 고민하다가, 대구광역시에서 데이터 기반으로 노인보호구역을 선정하면 좋겠다는 의견을 주셔서 이번 분석 모델 개발에 참여하게 되었어요.



데이터는 어떤 방식으로 수집하셨나요?

먼저 공단이 보유하고 있는 데이터를 최대한으로 활용하기 위해 교통카드 빅데이터 통합시스템으로부터 데이터를 수집했고, 교통안전정보관리시스템(TMACS)에서 교통사고 데이터를 확보했습니다. 화물자동차의 초 단위 위험 운전 통계 데이터를 얻을 수 있는 디지털운행기록계 자료도 확보했죠. 대구광역시에서는 가로망 주정차 현황 자료, 각종 노인요양시설, 버스정류장 위치정보, 생활체육시설, 약국, 전통시장 정보 등 노인 유동에 영향을 미칠 수 있는 시설 데이터 10종을 수집하였고, 해당 데이터들을 지도 베이스로 매칭시켰습니다.

사실 데이터를 수집하는 것보다 가공하고 전처리하는 과정이 어려워요. 이때 전문 분석업체와 함께 데이터 세트(데이터마트)를 구성하여 총 29종의 변수를 도출했고, 데이터의 신뢰성과 적합성을 높이기 위해 교통 전문 교수님들을 찾아가 조언을 구했습니다.



과제를 진행하면서 어려움은 없었나요?

저도 처음에는 데이터에 관한 전문지식이 없다 보니, 용어들이 너무 생소했어요. 하지만 최종적으로 가고자 하는 방향을 제시하고 그 결괏값이 정확하게 데이터에 기반한 정책 수립에 활용될 수 있도록 정책입안자의 입장에서 과제에 참여하게 됐어요.

또, 모델의 정확도를 높이기 위해서는 실제로 '노인보호구역'을 지정해 보고 시행착오를 통해 개발하는 것이 정확하지만, 한정된 시간과 예산 내에서 이를 수행하는 것이 힘들었습니다. 학술적인 부분 외에 실무자들의 얘기를 좀 더 들어보지 못한 것도 아쉬움으로 남아 있어요.



개발 모델을 공유, 확산하기 위해 어떤 노력을 하셨는지 듣고 싶습니다.

분석 결과는 두 가지 버전으로 만들어졌습니다. 하나는 QGIS(지리정보 체계 응용 프로그램)를 통해 모든 도로의 위험/보통/안전 3가지 등급으로 분류된 모습과 격자 형태로 시각화된 모습이고요. 두 번째는 도출된 결과를 해석하여 위험예측 보고서 형태로 제작이 되었어요. 이 보고서는 저희 공단에서 운영 중인 운수안전컨설팅지원시스템(이하 COSAS) 내에 시범적으로 탑재되었습니다.

COSAS는 공단의 지역본부 직원, 지자체 교통행정과 담당자들, 운수회사에서 맞춤형 교통안전 컨설팅을 위하여 사용되는 시스템입니다. 분석 모델의 결과가 더 많이 활용되기 위해서는 COSAS의 홍보가 필요하고, 대구광역시에서 더 많은 이용과 피드백이 있어야 할 것 같습니다.



데이터 활용을 통한 업무개선에 부담을 느끼는 분들께 도움 되는 말씀 부탁드립니다.

최근 디지털플랫폼 정부가 출범하고, 데이터 초연결 시대가 되면서 '데이터'의 수요가 점점 늘어나고 있죠. 이제는 정책도 데이터 기반으로 과학적으로 수립해야 하므로 데이터가 없으면 업무가 불가능할 정도입니다. 그렇지만 아직은 과도기 단계에 있으므로 어떻게 활용해야 할지 담당자는 막막할 수 있어요. 이때, 데이터 분석을 도와주는 사업을 통해 도움을 받으셨으면 좋을 것 같고, 참여하는 과정에서 직접 분석하지 않더라도 데이터 역량이 증가하는 경험을 하실 수 있습니다.

한국교통안전공단에서는 이번 과제와 관련하여 노인교통사고를 줄이기 위한 다양한 노력을 진행하고 있으니, 정책을 수립하는 지자체 공무원분들이시라면 운수안전컨설팅지원시스템(kotsa.or.kr/cosas)에 한번 가입하여 둘러보시는 것도 좋으리라 생각합니다.

고속도로 낙하물 사고, AI 기술로 예방하겠습니다

AI 기반 적재 불량 자동 판단시스템

한국도로공사 ITS처



이렇게 바뀌었습니다!

AS-IS	TO-BE
• 대면 적발에 따른 민원 발생 및 업무 가중 (근무자 1인당 최대 7,500대/일 심사) • 적재 불량 차량 위반 기준 판단 시 개인 주관에 따른 정확도 저하	• 비대면 적발에 따른 민원 발생 감소 및 행정 소모 최소화 (심사시간 1/10로 단축) • AI 판별시스템을 통한 적재 불량 차량의 위반 기준 판단으로 정확도 향상

추진 배경

2019년 12월 기준 화물차량은 약 370만 대로 국내 등록차량수의 16% 정도를 차지하며, 그중 적재 불량 행위가 가능한 카고형·덤프형 차량이 72.5%이다. 적재 불량 화물차에 대한 단속기준은 「도로교통법」 제39조 제4항에 명시돼 있지만, 적재 방법에 대해서는 구체적으로 제시하고 있지 않다. 이런 기준의 모호함과 경찰 인력의 부족 등으로 효과적인 적재 불량 화물차 단속은 사실상 어려운 형편이다.

현재, 한국도로공사(이하 공사)에서는 대면 및 비대면 방법으로 적재 불량이 의심되는 차량[6]을 단속·신고(스마트 국민 제보)하고 있다. 대면 단속은 영업소 입구 근무자, 안전순찰원, 이동단속반이 직접 단속하는 방식으로, 사안이 경미한 차량은 시정조치 후 운행하도록 하고, 심한 차량이나 시정할 수 없는 차량은 관할 경찰서에 고발하거나 고속도로 밖으로 나오도록 조치한다. 비대면 단속은 영업소 입구와 고속도로 본선에서 적재 불량 판단용 CCTV를 활용하는 방식으로, 차량이 진입했을 때 CCTV 사진을 촬영하고 사후에 육안으로 검사한 후 적발 통보서 등 관계 서류를 작성해 고발한다.

6) 화물의 편중적재, 예비 타이어 고정 불량, 덮개를 씌우지 않았거나 묶지 않아 결속 상태가 불량한 차량, 액체 적재물 방류차량 등

그러나 하루에 고속도로를 이용하는 차량만 약 4백만 대 이상, 그중 10% 이상이 대형차량(3~5종)인 상황에서 적재 불량 차량을 효과적으로 걸러내기란 쉽지 않다. 감시자의 눈을 피한 차량은 고속도로에 매년 25만 건 이상 크고 작은 낙하물 사고를 일으키고 이는 더 큰 2차 사고를 유발한다. 이에 공사는 예산 및 인력 부족이라는 현실적 한계를 극복하고자 2020년부터 AI를 이용한 영상 분석기술을 도입해 사람을 대신하여 적재 불량 위반차량을 식별할 수 있는 시스템 개발에 착수하게 되었다.

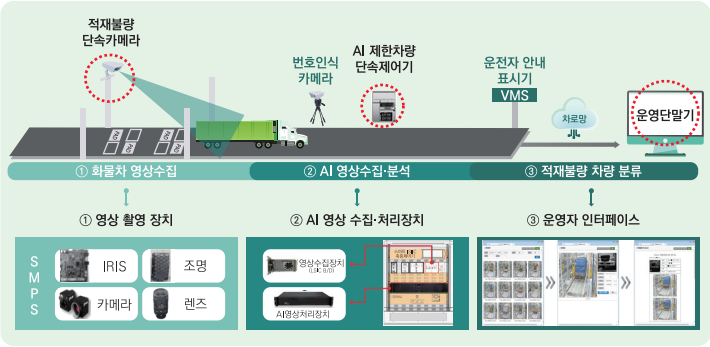
활용데이터

연번	데이터명	기준연도	형태	내 용	보유기관
1	적재 불량 판단용 CCTV 영상 이미지	2020~2023	JPG	화물차 적재함 영상, 차량번호판 영상, 차량 중량 등	한국도로공사

약 300만 장 차량 영상의 AI 학습을 통한 적재 불량 차량 자동 판단

'AI 기반 적재 불량 자동 판단시스템'은 공사에서 기운영 중인 제한 차량 단속시스템, 제한 차량 센터 시스템과 연동해 적재 불량을 학습화된 AI가 자동으로 판단하는 것이다. 전체 업무처리 과정은 아래 그림과 같다.

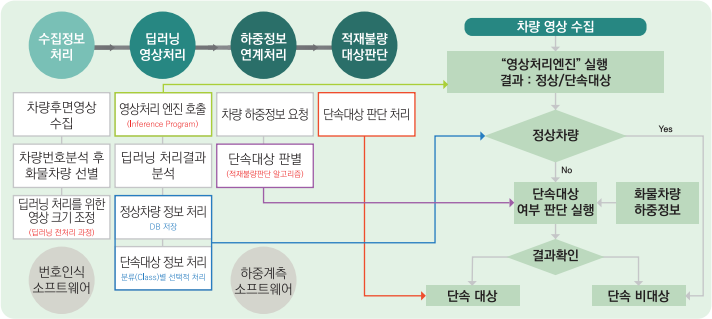
〈그림 9. AI 기반 적재 불량 자동 판단시스템 설치 구성도〉



화물차량이 입구 영업소 축중 차로[7]를 지나가면 적재 불량 단속카메라가 적재함 후면영상을 촬영하고, 번호 인식 카메라(AVI)는 전면 번호판 영상을 찍는다. 이를 AI 제한 차량 단속제어기가 적재 불량 의심 차량을 선별, 전면 번호판 영상과 매칭하여 해당 차량 정보를 제한 차량 센터 시스템으로 전송한다. AI 영상처리장치는 영상 수집 장치가 전달한 정보를 AI 딥러닝 알고리즘을 통해 분석하고, 적재 불량 차량을 자동 선별해 운영자가 효율적으로 단속 업무를 수행할 수 있도록 처리하는 지능형 장치이다.

적재 불량 자동 판단을 위한 AI 기반 화물차 적재함 영상분석을 위한 과정은 다음과 같다.

〈그림 10. 화물차 적재함 영상분석 과정〉



화물차 적재함 영상분석을 위해 CNN[8] 기반의 SSD[9] 딥러닝 기법을 적용했다. SSD 구조에서는 VGG16[10]을 베이스 네트워크로 사용하고,

7) 고속도로에 진입하는 차량의 축중량, 총중량 및 높이 등을 자동으로 계측하여 안전기준을 초과하는 차량을 단속하기 위한 차로(출처 : 한국도로공사, 통행료 수납차로 설계지침(안))
8) 영상 내 객체 분류, 객체 탐지 등 다양한 응용 분야에 폭넓게 활용되는 인공 신경망의 한 종류(출처 : TTA 정보통신용어사전)
9) 객체의 위치와 종류를 빠르게 알아내는 알고리즘
10) 16개의 레이어로 이루어진 컨벌루션(convolution) 신경망

특징 추출 층을 추가해 객체 검출을 위한 검출기를 생성했다. 그 결과, 자동화 도입 이후 분류 정확도는 93% 수준으로 나타났다.

영업소 차로에 시범운영용 시스템을 설치한 후 약 300만 장의 영상을 수집해 라벨링 작업을 거쳐 반복 학습을 실시했다. 또한, 적재 차량 유형을 화물차 적재함 형태 및 적재 상태에 따라 규격/비규격/미분류 차량[11]의 3개 그룹 23개 세부 유형으로 분류하여 학습효과를 강화했다. 적재 차량 유형을 반영하기 위해 딥러닝 기반 추론프로그램을 개발했고, 이를 통해 차량 형태를 정확히 분류(판별)할 수 있었다.

시스템을 기반으로 한 적재 불량 판별 프로세스는 ① 차량 진입 시 승용(합)차 제외 ② 그 외 차량 규격/비규격으로 분류 ③ 규격 차량은 단속 제외, 비규격 차량은 적재 불량 분류 알고리즘 실행 및 분석 ④ 비규격 차량 중 적재 불량이 아닌 차량 제외 ⑤ 적재 불량 의심 차량만 분류하는 형태로 진행된다.

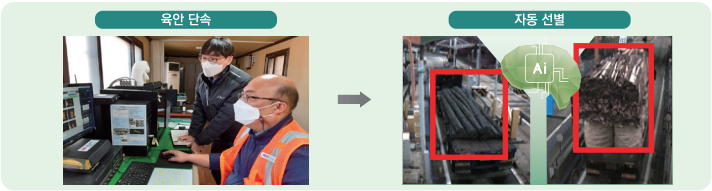
심사 시간 1/10로 단축, 적재 불량 신고 건수 320% 증가

개발한 시스템의 정확도를 확인하기 위해 공사는 동서울과 인천 두 곳을 시범 영업소로 지정했다. 동서울영업소에서 확인한 총 11,677대 화물차 중 규격 9,383대(미인식 44대·오인식 24대 포함), 비규격 1,917대, 미분류가 377대였고, 정인식률은 규격 99.28%, 비규격 97.86%였다. 그리고 인천영업소에서 확인한 총 12,829대 화물차 중 규격 10,489대(미인식 75대·오인식 68대 포함), 비규격 1,720대, 미분류가 620대였다. 정인식률은 규격 98.64%, 비규격 98.31%였다.

11) - 규격 차량 : 일반적으로 차량의 적재함에서 적재물 낙하가 발생할 수 없는 구조의 차량 (예 : 탑차, 탱크로리 등)
- 비규격 차량 : 적재함이 개방되어 적재물을 실은 화물차량(예 : 카고 등)
- 미분류 : 개조 및 특수차량

미인식 원인은 개방형 규격 차량의 덮개 인식 실패 및 야간 영상 중 빛에 의한 영향이었고, 오인식은 개방형 규격을 폐쇄형 규격으로 인식한 경우였다. 두 곳의 시범 영업소 간 차이를 살펴보면, 인천영업소는 동서울영업소에 비해 상대적으로 다양한 개방형 차량이 통행했고 거기에서 발생하는 미인식률과 오인식률이 높았다. 그러나 규격 개방형 화물차량이 많아지면서 비규격 차량에 대한 정인식률은 상승했다. 두 영업소의 규격 차량 인식률 평균값은 정인식 98.94%, 미인식 0.6%, 오인식 0.46%였고 비규격 차량 인식률 평균값은 정인식 98.08%, 미인식 0%, 오인식 1.92%였다.

적재 불량 자동 판단시스템의 운영 효과는 단속원이 화물차량의 적재 불량 여부를 검사하는 데 걸리는 시간이 얼마나 감소했는지를 통해 확인할 수 있다. 영업소별 한 달간(2020.12.) 통계자료를 기준으로 근무자가 통행하는 모든 차량을 맨눈으로 단속한다고 했을 때 인천영업소는 288,438대, 동서울영업소는 198,601대의 차량을 확인해야 했다. 그러나, 적재 불량 자동 판단시스템을 이용하면 인천영업소는 30,194대, 동서울영업소는 19,453대만 확인하면 되는 것으로 나타났다. 이는 기존 심사 시간을 1/10로 감소시킨 것이며, 한 달 신고 건수 역시 기존 약 130건에서 547건(320%↑)으로 대폭 늘어났다.



추가 검증 후 전국영업소 확대 예정

'AI 기반 적재 불량 자동 판단시스템'은 2020년에 동서울영업소 등 3개 영업소 14차로에 시범 구축 및 운영을 시작했고, 2023년 현재 서울영업소 등 19개 영업소 48차로(축중차로 기준)로 확대하여 시범운영 중에 있다. AI 판별 기능 향상 및 도입 효과 추가 검증 후에 전국의 영업소로 확대하여 적재 불량 차량의 도로 진입 방지 및 2차 사고 예방에 활용할 예정이다.

기관소개(한국도로공사)

한국도로공사는 도로의 설치 및 관리와 이에 관련된 사업을 통해 도로의 정비를 촉진하고 도로 교통의 발달에 이바지한다는 목적으로 1969년 설립되었다. 경부고속도로 건설을 시작으로 1970년 450km에 불과했던 고속도로는 현재 '5천km 시대'를 목전에 두고 있으며, 한국도로공사는 도로 교통 전문 공기업으로서 지난 반세기 동안 대한민국의 산업과 경제발전의 견인차 역할을 하고 있다.



Coffee Time (인터뷰)



현재 하시는 일을 포함해 자기소개 부탁드립니다.

저는 한국도로공사 ITS처에서 제한 차량 업무를 맡고 있는 안지성 차장입니다. 정보통신공학을 전공했고 컴퓨터, IT, 데이터 등에 관해 비전공자보단 나름대로 지식이 있는 편이었습니다. 하지만 학부 때 배운 지식을 현업에 바로 적용하기에는 무리가 있으므로 해박하다고 말할 만한 수준은 아니라고 할 수 있습니다.

이번 과제는 어떻게 참여하게 되셨나요? 기존 일과 어느 정도 관련이 있었나요?

저는 고속도로의 구조를 보전하고 통행 위험을 방지하며, 차량 흐름이 원활해지도록 운행 제한 차량을 판별하는 시스템을 유지관리·구축하는 업무를 맡고 있습니다. 운행 제한 차량으로는 과적 차량과 적재 불량 차량이 있는데, 하루에도 수만 대의 차량이 적재 불량 상태로 고속도로를 통행합니다. 그 많은 건수를 고속도로 영업소 근무자가 하나하나 맨눈으로 보고 판별하고 있었죠.

하지만, 영업소 근무자는 요금징수부터 운행 제한 차량 고발 업무까지 많은 일을 하고 있습니다. 그분들의 업무를 줄일 방법이 없을까 고민하게 되었고, 4차 산업혁명 시대에 아직도 사람을 통해 주먹구구식으로 일한다는 생각도 들었어요. '판별'한다는 건 결국 사람의 판단이 들어가는

부분으로, 개인의 주관에 따라 얼마든지 달라질 수 있습니다. 그래서, 보다 객관적이고 과학적인 방법으로 자동화하여 처리할 수 있다면 운행 제한 차량 업무가 대폭 개선될 것이라는 판단을 하게 되었죠.



데이터 수집과 정제는 어떻게 하셨나요? 그 과정에서 느낀 어려움을 어떻게 해결하셨는지 말씀해주세요.

필요한 데이터는 영업소에 설치돼 있던 화물차 후면 적재함 촬영 CCTV 영상을 활용해 수집했습니다. 사실 처음에 영상데이터를 가지고 적재 불량을 어떻게 판단할 건지 그 기준을 세우는 게 쉽지 않았어요. 도로교통법상 적재 불량이 추상적으로 정의돼 있었기 때문이었는데요. 이 부분을 해결하기 위해 경찰청 협조를 받아 고발한 차량 중 실제 단속된 차량 리스트를 활용하여 단속원들이 생각하는 위반차량의 물리적 특징이 무엇인지 파악할 필요가 있었죠. 이를 딥러닝 기법으로 학습시킨 후 AI 적재 불량 판별시스템에 반영해 기준을 세우고 정확도를 높일 수 있었습니다.



데이터 분석을 위해 어떤 알고리즘이나 방법을 사용하셨나요?

처음에는 화물차의 적재 형태에 따른 분류 트리별 영상 분류 및 특징점(바운딩 박스) 그리기, 라벨링 작업을 수동으로 진행했습니다. 하지만 AI 학습 데이터세트를 생성하기 위해선 더 효과적인 방법이 필요했습니다. 그래서 카메라 영상 이미지를 활용해 차량 분류, 검출 등에 가장 최적화된 AI 분석 모델을 활용해 영상 분류 및 특징점 추출을 자동으로 학습했습니다. 그 후 특징점을 추가해 2차 라벨링을 하였고 화물차량의 판단 정확도를 높이려고 노력했습니다. 특히 적재함에 화물이 조금 실린 차량의 오인식을 개선해 성능을 개선할 수 있었습니다.



이런 데이터 관련 프로젝트가 일회성으로 끝나지 않으려면 어떤 노력이 필요한지 의견을 듣고 싶습니다.

4차 산업 시대에 살고 있다고 모두가 말하지만, 막상 코앞에 닥친 문제에 그 기술들을 적용하는 것에는 다들 부담을 느낍니다. 저 역시 처음에는 그랬습니다. 왜 이런 걸 해야 하는지는 알지만, 데이터를 분석해 만든 모델을 과연 신뢰할 수 있을 것인지에 대해선 저 자신도, 주변에서도 반신반의했어요. 하지만 이러한 사례들이 계속해서 쌓여야 주변 사람들이나 조직의 인식이 개선될 수 있다고 생각합니다.



담당자로서 뿌듯했던 점에 관해서도 얘기해주세요.

시범운영 영업소 근무자들이 "개발한 시스템 덕분에 심사 업무가 한결 편해졌다"고 말씀하셨을 때가 가장 보람찬 순간이었습니다. 향후 이 시스템이 전국 영업소로 확산해 운영될 것을 생각하면 벌써 흐뭇해집니다.



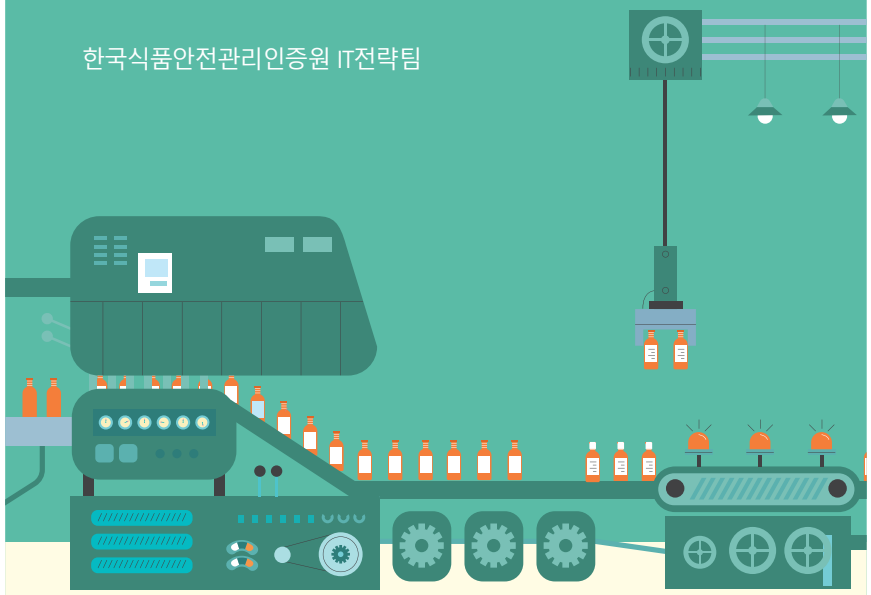
데이터 활용을 통한 업무개선을 망설이는 분들께 도움 되는 말씀 부탁드릴게요.

데이터를 분석하고 활용하는 것은 정말 쉽지 않아요. 당연히 어렵게 다가올 테지만, 그 어려움에 부딪히고 도전해야 합니다. 도전하는 사람만이 어떠한 결과든 만들 수 있다고 생각해요. 실패하더라도 다음 도전의 밑거름이 될 것이기에, 낙담하지 말고 도전해 보시길 바랍니다.

안심하고 먹을 수 있게, 빅데이터로 식품 제조공정을 관리하다

식품 제조공정 이상 징후
사전 탐지모형 구축

한국식품안전관리인증원 IT전략팀



이렇게 바뀌었습니다!

AS-IS	TO-BE
• 사고 발생 시 중요관리점 데이터 확인 후 대응 • 제품 폐기에 따른 업체 생산성 저하 및 폐기 누락 시 식품 안전사고 발생 가능성 존재	• 실시간 데이터 분석을 통한 중요관리점 공정 사고 예방 • 식품 제조업체 경쟁력 강화 및 식품 안전 신뢰도 제고

추진 배경

한국식품안전관리인증원(이하 식품인증원)은 중요관리점(Critical Control Point, CCP)[12] 관리일지의 위·변조 방지를 위해 센서로 중요관리점 데이터를 자동 기록 관리하는 스마트 HACCP 제도를 지난 2020년부터 도입하여 운영 중이다. 그러나 실시간으로 안전 데이터 수집·모니터링을 한다고 해도 사고가 발생한 직후에만 이상을 탐지할 수 있어 사고 자체를 원천적으로 차단하기는 어렵다.

또한 매출 5억 원 미만 영세업체가 전체 식품업체의 81.2%를 차지하고 있어 업체마다 중요관리점의 이상 탐지를 위한 전담 인력 확보 및 시스템 구축이 어려운 상황이다.

이에 식품인증원은 식품 안전 및 식품업체 경영성 제고를 위해, 가장 많은 업체가 지정하고 있는 중요관리점인 금속검출과 냉장/냉동 온도 관련 데이터를 활용한 식품 제조공정 이상 징후 사전 탐지모델 구축에 착수했다.

12) 식품 위해요소를 예방 제어 또는 허용 수준 이하로 감소시켜 식품의 안정성 확보를 위한 중요한 단계 또는 공정

활용데이터

연번	데이터명	기준연도	형태	내 용	보유기관
1	기상정보	2019~2022	API	식품업체 인근지역 기상정보	기상청
2	금속검출 및 온도 CCP 데이터	2020~2022	CSV	중요관리점에서 실시간 집계되는 금속검출 및 온도 정보	식품업체
3	금속 검출기 제품정보	-	DOC	금속 검출기 제품 종류, 검출 특성 등	식품업체
4	냉장/냉동창고 제원 정보	-	HWP	냉장/냉동창고 면적 등	한국식품안전관리인증원

중요관리점 데이터와 기상 데이터 상관관계 분석 후 모델 개발

'식품 제조공정 이상 징후 사전탐지 모델' 개발을 위한 과정은 ① 분석 대상 데이터 정의 및 수집 ② 데이터 현황 및 특성 분석 ③ 데이터 정제 ④ 분석 대상 중요관리점 특성 분석 ⑤ 중요관리점 이상 징후 예측 모델 개발 ⑥ 성능 검증 및 결과 도출 순서로 진행되었다.

〈그림 11. 분석 과정〉



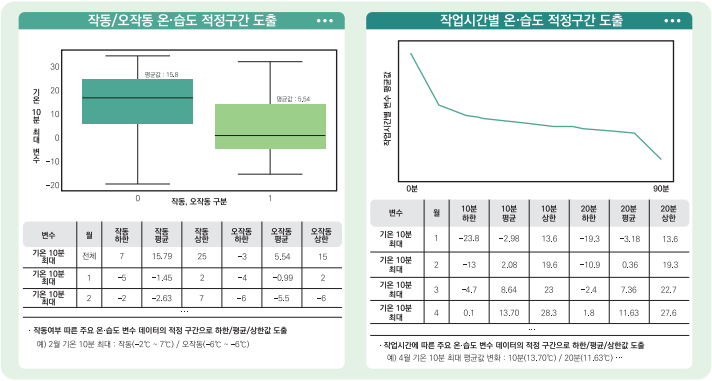
식품인증원은 먼저 계절 요인을 고려해 1년 이상 스마트 HACCP 실시간 데이터를 보유한 7개 식품업체를 대상으로 중요관리점 데이터(금속검출 데이터 14만 건, 온도 데이터 480만 건)와 해당 식품업체 소재 지역의 기상관측 데이터(680만 건)를 수집했다. 이후 각 중요관리점 공정 및 설비에 따른 데이터의 특성을 고려해 결측치 및 이상치 처리 등 데이터 전처리를 하였으며, 분석 효과를 높이고자 데이터 수집 시간 정보, 업체 구분 등 파생 변수를 추가로 생성했다.

수집한 데이터 중 기상 데이터는 중요관리점 공정 및 설비에 영향을 줄 수 있을 것으로 예측되어, 이를 확인하기 위해 금속검출 중요관리점(금속검출기에 의한 금속 이물 검출 공정)과 온도 중요관리점(미생물의 성장을 최소화할 수 있는 냉장·냉동 공정)을 기상정보와 연계한 상관관계를 분석하였다. 구체적으로 금속검출 중요관리점은 금속검출 로그 정보(로그파일 저장 정보)를, 온도 중요관리점은 온도 중요관리점 작업정보를 분석 요인으로 선정하여, 금속 검출기 오작동 발생 및 냉장/냉동창고 작업 시 한계이탈을 분석했다.

그 결과, 금속검출 설비는 기온 하강에 따른 오작동 증가 경향이 있었고, 냉장/냉동 설비는 기온이 하강하고 습도가 증가할수록 적정 작업시간이 증가함을 발견했다.

이처럼 데이터 전처리 및 데이터 간 상관관계 분석으로 최종 데이터 마트(소규모 데이터 저장 및 분석시스템)를 구성하였고, 중요관리점 공정 이상 징후 예측 모델 개발을 위한 DB를 구축하였다.

〈그림 12. 금속검출, 온도 CCP의 적정 온습도 구간 도출〉



중요관리점 이상 징후 사전탐지 모델 정확도 높은 수준으로 나타나

위의 분석 과정을 거쳐 '금속검출 오작동 유무 예측 모델'과 '냉장/냉동 온도 중요관리점 적정 작업시간 예측 모델'을 개발했다.

첫째, '금속검출 오작동 유무 예측 모델'은 Boruta 알고리즘[13]을 통해 변수를 검증했으며, 45개의 변수 활용으로 학습 알고리즘 모델을 개선해 1시간 단위로 오작동 유무 예측값을 산출했다. 그 결과, 성능 검증 시 평균 90% 이상의 정확도를 확인했다.

식품제조업체는 이 분석 모델을 통해 금속검출 설비 오작동이 일어날 수 있는 상황을 사전에 예측함으로써 금속검출 설비 가동 시간을 변경하거나 오작동이 일어날 수 있는 기상 상황에서 주변 온습도를 조절해 금속을 검출하는 등 업무 효율을 높일 수 있게 되었다.

둘째, '냉장/냉동 온도 중요관리점 적정 작업시간 예측 모델'은 적정 작업시간 범위를 90분까지로 제한해 다양한 학습과 성능 검증 과정을 거쳐 냉장/냉동창고 적정 작업시간의 예측값을 분(分) 단위로 산출했다.

그 결과, 평균 오차가 6분인 예측 모델이 도출되었으며, 이는 작업 진행 시 창고 온도가 기준치 이상으로 상승하기 시작하는 시간을 ±6분 이내로 예측한다는 것을 의미한다.

이 분석 모델을 통해 식품업체는 냉장/냉동창고 내 제품 반출·입 작업을 할 때 주변 기상 상황에 따라 창고 온도가 기준을 이탈하지 않는 적정 시간 내에 업무를 마무리하여 제품의 변질을 예방할 수 있게 되었다.

13) 변수 선택 알고리즘의 한 종류

〈그림 13. 식품 제조공정 이상 징후 사전탐지모델 결과〉



HACCP 데이터 활용 플랫폼에 모델 탑재 추진, 식품 안전 사고 예방 기대

식품인증원은 현재 개발 중인 스마트 HACCP 데이터 활용 플랫폼에 식품 제조공정 이상 징후 사전탐지 모델 탑재를 추진하고 있다. 이를 통해 식품업체들이 중요관리점 관리 시 발생할 수 있는 식품 안전사고를 예방할 수 있으리라 기대한다.

이번 분석을 통해 분석 모델 개발에 공장 내 온습도 데이터, 냉장/냉동창고 출입문 면적, 공기 조절 장치 등 온도 관리 설비 유무 등의 필요 변수가 추가 반영되어야 함을 느꼈으며, 향후 위 변수들을 포함하여 분석 모델을 고도화할 예정이다. 또한 신규 중요관리점(가열, 세척, 소독 등)에 대한 사전 탐지모델도 새롭게 개발해 스마트 HACCP을 활용한 식품 안전 상시 관리체계의 기반을 마련하고자 한다.

각종 식품 안전사고를 줄이고 국민의 부담을 덜기 위해 식품인증원은 더 많은 식품업체가 '식품 제조공정 이상 징후 사전탐지 모델'을 활용할 수 있도록 끊임없이 개선해 나갈 것이다.

기관소개(한국식품안전관리인증원)

한국식품안전관리인증원은 HACCP(식품안전관리인증) 제도 운영과 관리를 통해 국민에게 안전한 먹거리를 제공하기 위하여 지난 2017년 통합 설립된 식품의약품안전처 산하 공공기관이다. 한국식품안전관리인증원은 식품의약품안전처와 함께 스마트 HACCP 제도 도입 및 식품 안전 관련 데이터 공동 활용 등 디지털·데이터 중심의 과학적 식품안전관리 기반을 마련하여 국민에게 더욱 신뢰받는 안전한 식품과 축산물을 제공할 수 있도록 지속해서 노력하고 있다.



Coffee Time (인터뷰)



현재 하시는 일을 포함해 자기소개 부탁드립니다.

저는 한국식품안전관리인증원 기술혁신본부에서 데이터기반행정 업무를 담당하는 홍인배 심사관입니다. 대학교에서 경제학과 경영학을 전공했고, 데이터나 IT 관련 지식은 거의 없는 수준이었어요. 다만 대학원에서 경영학 공부를 하면서 통계와 관련된 지식은 조금 있었고, 식품인증원 입사 전 직장에서 경영관리 시스템 구축 업무를 수행하면서 IT 쪽 업무를 약간 경험했습니다.

이번 과제는 어떻게 참여하게 되셨나요? 기존 일과 어느 정도 관련이 있었나요?

식품인증원은 국민의 먹거리 안전을 제고하고자 2019년부터 HACCP(식품안전관리인증) 제도를 더욱 발전시켜 IT와 데이터, 디지털 기반의 스마트 HACCP 제도를 새롭게 구축하여 운영하기 시작했습니다. 식품 안전 데이터가 쌓이면서 이를 어떻게 활용할 수 있을지 고민하던 중 행정안전부와 NIA에서 공공 빅데이터 분석 사업을 하고 있다는 것을 알게 되었죠. 그래서 저희는 식품업계 대부분을 차지하는 중소 식품기업을 도울 수 있는 주제를 선정해서 사업에 참여하게 되었고, 이와 동시에 빅데이터 분석·활용의 기반을 조성하고자 했습니다.



일반 식품업체가 보유한 데이터를 활용하셨는데요, 어떤 방식으로 수집하셨나요?

일반 식품업체가 보유한 데이터는 사실 영업상 비밀일 수도 있는 부분이에요. 저희는 스마트 HACCP 도입업체들과 스마트 HACCP 솔루션을 공급하고 있는 업체를 만나 협의하여 데이터를 확보했는데요, 각종 데이터 제공에 난감해하는 업체들에는 중요관리점 이상 징후 사전 예측 모델을 활용해 업체가 향후 얻을 수 있는 다양한 이점을 설명하며 설득했습니다. 업체를 추측할 수 있는 정보 등은 모두 제외하고 사용하겠다고 약속하기도 했고요. 결국 신뢰의 문제라고 생각합니다.



과제를 진행하면서 느꼈던 어려움이 꽤 있었을 것 같습니다.

우리 기관이 식품 안전 관련 빅데이터 분석을 시도한 것이 처음이라, 사실 프로젝트를 시작하면서 막막함을 느꼈어요. 어떤 데이터를 활용해야 예측 정확도가 높은 모델을 만들 수 있을지 최초 데이터 선정에서부터 고려해야 할 점이 많았습니다. 그래서 NIA 담당자와 협의해 자문위원회를 개최해서 전문가 조언을 얻기로 했어요. 데이터 활용 전문업체와 데이터 분석 관련 학계 교수님들을 전문가 패널로 선정한 뒤 중요관리점, 기상 데이터의 분석 방향, 추가 분석이 필요한 데이터 등에 대한 의견을 청취했죠. 이를 바탕으로 추가 파생 변수를 생성해 분석하고, 기상 데이터 결측치를 보완하는 방안도 마련했습니다. 또한 우리 기관의 심사관들을 인터뷰해서 추가로 고려할 데이터 및 식품생산 프로세스 특성에 대한 정보도 얻을 수 있었습니다.

저와 같은 데이터 분석 초보자는 사업 담당자와 지속적인 협력을 통해 기초적인 절차에 대한 조언을 얻고, 다음으로 전문가 등을 활용해 데이터 분석 방향과 추가 분석이 필요한 데이터를 파악하는 것이 중요하다고 생각해요. 데이터 분석 업무가 쉽지는 않지만 그렇다고 하지 못할 업무도 아니기에 하나하나 차근히 배운다는 자세로 임하면 충분히 해낼 수 있다고 봅니다.



개발 모델을 공유, 확산하기 위해 어떤 노력을 하셨는지 듣고 싶습니다.

우선 분석 모델 사용을 위한 교육 계획을 수립하여 담당자 교육을 진행했고, 예측 모델 사용자들이 손쉽게 이용할 수 있도록 사용자용 매뉴얼을 별도로 제작했습니다. 그리고 현재 기관에서 구축 중인 스마트 HACCP 정보 활용 플랫폼에 예측 모델 탑재를 추진하고 있는데, 스마트 HACCP 제도를 운용 중인 다양한 업체의 중요관리점 데이터를 추가로 확보하여 예측 모델의 정확성을 제고하고 활용을 확대하고자 합니다.

빅데이터 분석 과제를 처음 추진하면서 대부분은 관련 부서에서 적극적으로 협조해 주셨으나, 일부는 데이터 분석의 신뢰성이나 활용 가능성에 의문을 가지기도 하셨어요. 이런 분들에게는 데이터 분석 업무의 중요성과 향후 데이터 기반 업무수행 시 이러한 경험이 큰 도움이 될 거라고 설명하면서 협조를 얻을 수 있었는데요. 향후 데이터 분석 사업을 계속 수행하고, 그에 관한 홍보, 교육 등을 반복한다면 해당 업무가 업무의 품질을 한 단계 높이는 방법이라는 사실을 모두 인정할 것으로 생각됩니다.



데이터 활용을 통한 업무개선에 부담을 느끼는 분들께 도움 되는 말씀 부탁드립니다.

데이터 분석은 시도해 보지 않으면, 어떤 데이터가 부족한지, 왜 분석의 신뢰성이 낮은지 원인을 파악하기 어렵습니다. 그래서 저는 일단 한 번 시도해 보기를 권합니다. 또한 NIA나 행정안전부를 통해 데이터 관련 교육을 시행하고 기관 내부의 데이터기반행정 추진 역량을 갖춘다면 기관의 업무 수준이 한층 업그레이드될 것으로 생각합니다.

PART 1-6. 국민안전

지역사랑상품권 부정 사용 NO! 숨어봤자 빅데이터 손바닥 위

지역사랑상품권 부정 사용 방지를 위한
이상거래 패턴 분석

한국조폐공사 기술처 기술개발부



이렇게 바뀌었습니다!

AS-IS	TO-BE
• 지자체가 특정 조건 데이터 요청 시 지자체에 해당 데이터 제공 • 조폐공사 관리시스템을 미활용한 지자체 현장점검	• 이상거래 탐지모델 결과를 지자체 제공 • 이상거래 탐지모델 모니터링이 가능한 조폐공사 관리시스템을 활용하여 현장점검 추진 가능

추진 배경

지역사랑상품권 발행 확대 및 지원 정책으로 인한 부정 유통 사례가 최근 들어 크게 늘고 있다. 지난 3년간 지역사랑상품권의 발행량은 2018년 0.37조 원에서 2019년 3.2조 원(전년 대비 8.6배↑), 2020년 13.3조 원(전년 대비 3배↑), 2021년 23조 원(전년 대비 1.7배↑)로 비약적으로 증가하는 추세이다. 따라서, 많아진 유통량에 대응하고 거래 관리의 효율성 향상을 위해서는 자동화된 부정 사용 탐지시스템이 필요했으며, 한국조폐공사(이하 공사)는 다년간의 지역사랑상품권 사업 운영을 통해 축적한 데이터를 활용하여 이상 거래를 탐지하는 모델을 만들기로 했다.

활용데이터

연번	데이터명	기준연도	형태	내 용	보유기관
1	모바일/지류 상품권 데이터	2021	CSV	지자체별 사용자 및 결제 내역 정보	한국조폐공사
2	가맹점 정보 데이터	2021	CSV	가맹점 관련 정보	한국조폐공사
3	상권분석 데이터	2021	CSV	지자체별 상권분석 정보	해당 지자체 및 공공데이터
4	유동 인구 데이터	2021	CSV	지자체별 유동 인구 정보	해당 지자체 및 공공데이터

이상거래 점수화 방식 사용, 정밀도 높은 분석 모델 선정

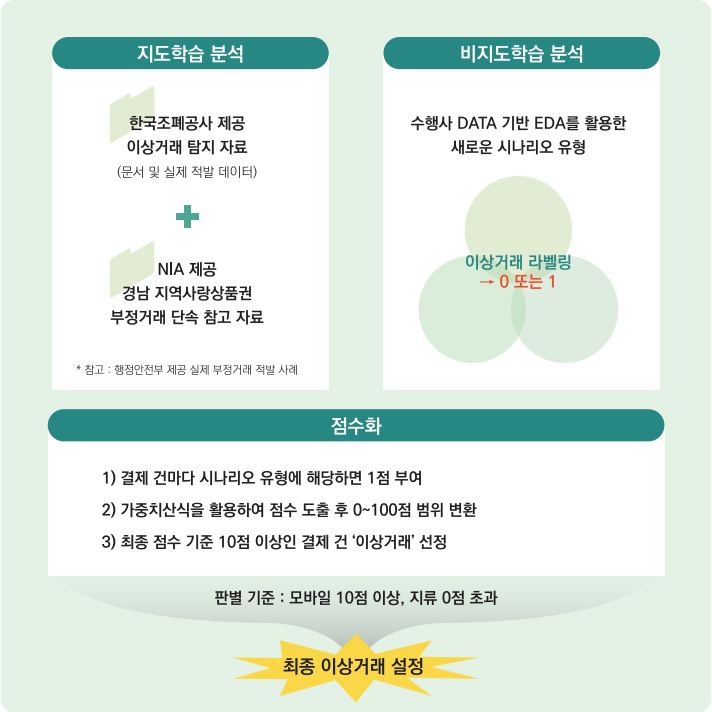
지역사랑상품권 데이터는 지방자치단체(이하 지자체)가 소유한 것으로, 개인정보가 포함된 민감 데이터를 사용하기 위해서는 지자체의 협조가 반드시 필요했다. 이를 위해, 공사는 지자체에 지역사랑상품권 이상거래 탐지 시스템 구축의 필요성을 설명하고 민감정보는 모두 가명처리한 후 데이터를 개발팀에 전달했다.

개발 환경도 그리 녹록지는 않았다. 공사는 외부망의 접속이 차단되어 있어, 데이터 분석을 위해 특정 권한을 가지고 접속할 수 있도록 사전 협의가 필요했다. 또한, 개발자들이 지역사랑상품권 DB와 결제 프로세스를 이해할 수 있도록 별도로 교육을 실시했고, 프로젝트 중간에 생기는 문제들에 대해선 수시로 Q&A를 통해 해결해 나갔다.

이러한 과정을 통해 처음 개발한 이상거래 탐지모델은 실제 부정거래 데이터 유형이 많지 않았으며, 공사에서는 이를 보완하기 위해 부정거래로 의심되는 이상거래 추정 모델 개발을 추가 진행하였다. 이를 위해, 먼저 데이터 정제와 가공 과정을 거친 후 몇 가지 모델을 만들고, 각 모델 성능을 비교하는 방식으로 프로젝트를 진행했다.

분석 대상 지자체는 경기 시흥시, 경기 성남시, 충남 서산시, 전북 군산시 네 군데로 정했고, 지역사랑상품권 모바일/지류 데이터를 활용했다. 지자체와 경남지역 부정거래 단속 데이터를 기반으로 한 인공지능 지도학습 분석과 공사 내부 데이터를 기반으로 거래 내역 현황분석의 EDA(탐색적 데이터 분석)를 통해 새로운 이상거래 시나리오 유형을 정의했다. 이후 점수화 방식을 이용해 이상거래 기준을 만들었다. 상품권 결제 건마다 시나리오 유형에 해당하면 1점을 부여했고, 가중치 계산식을 활용해 점수를 도출한 후 0~100점 범위로 변환하고 최종 점수가 10점 이상인 결제 건을 '이상거래'로 선정했다.

〈그림 14. 지역사랑상품권 이상거래 탐지 과정〉



다양한 분류 알고리즘 중 XGBoost[14]의 정밀도가 높아 최종 모델로 결정했다. 이후 가맹점별 이상거래 탐지 건과 회원 번호별 이상 거래 건을 집계하여 상위 의심 가맹점과 회원을 도출했다.

〈그림 15. 지역사랑상품권 이상거래 탐지분석모델 선정 과정〉



회원·가맹점 의심 순위 분석 통해 '이상거래 탐지모델' 판정 정확도 검증

지역사랑상품권 이상거래 탐지모델은 모바일 회원의 이상거래 의심 순위, 가맹점별 이상거래 의심 순위 분석을 통해서 14가지 케이스를 가지고 검증한다. 만일, 이상 거래 여부가 있으면, 이것을 가지고 예측 확률이 50% 이상일 경우 이상거래로 판정하게 된다.

14) 기계학습(머신러닝) 앙상블 기법 중 하나로, 의사결정 나무(Decision Tree) 모델을 여러 개 조합하여 분류(예측) 성능을 향상시키는 알고리즘

향후, 공사의 '지역사랑상품권 고도화 시스템'에 이상 거래 탐지모델이 탑재될 예정이며, 사건 발생 시 시스템 담당자는 사건 알림을 받고 즉시 지자체에 통보하는 방식으로 구현될 예정이다.

〈그림 16. 결제 내역에 따른 이상거래 예측 결과 예시〉

결제 내역에 따른 이상거래 예측 결과 상세

거래 ID	가맹점명	결제 일시	회원 번호	결제 금액	Case 1	…	Case 14	이상거래 여부	예측 확률	이상거래 판정
16661	한들식당	2021-09-04 10:52:41	6734	9,500	0		0	1	0.6	1
51271	이디커피	2021-10-03 14:52:52	4255	5,300	0		0	0	0.3	0
43436	요프레스	2021-09-14 12:32:57	1242	6,300	0		0	1	0.8	1
65745	하나마트	2021-11-08 10:42:14	8342	37,000	1		1	0	0.2	0

이상거래 사례집 발간해 가맹점 운영에 도움

지역사랑상품권 이상거래 탐지모델은 지난해 말부터 개발하여 2023년 10월부터 시범운영하고 있으며, 광역 단위 플랫폼에 실시간으로 반영해 추후 이상거래 탐지 시스템을 고도화할 예정이다.

아울러, 공사는 개발한 분석 모델을 바탕으로 구축한 '이상거래 탐지시스템'을 적극적으로 홍보하고, 시나리오 관련 사례집을 발간·배포함으로써, 사전 범죄 예방 및 가맹점 운영에 도움이 될 것으로 기대한다.

기관소개(한국조폐공사)

한국조폐공사는 화폐 및 유가증권 등 국가적 보안제품의 안정적 제조·공급 임무를 수행하기 위해 1951년에 설립되었다. 완벽한 제품 생산을 위하여 보안성 및 공신력을 최고의 가치 기준으로 설정하고, 공공 보안제품 생산을 위한 용지 제조, 인쇄 및 주화에 대한 일괄 생산 체제 구축, 특수기술 자체 개발, 보안성 유지를 위한 시스템, 국가 신분증 및 여권 등을 제조하는 기관으로 국민에게 안전하고 신뢰성 있는 보안제품을 제조하여 공급하기 위해 노력하고 있다.



Coffee Time (인터뷰)



현재 하시는 일을 포함해 자기소개 부탁드립니다.

한국조폐공사 기술처에서 빅데이터 기획 및 데이터 분석 업무를 맡고 있는 정대규 차장입니다. 전공은 컴퓨터공학이지만, 공사 기술연구원에서 펌웨어 개발 위주의 업무를 담당하다가 본사 기술처에서 빅데이터 관련 업무를 하면서 데이터를 조금 이해하는 정도가 되었습니다. 컴퓨터공학을 전공하면서 기본적인 통계 지식은 있었고, IT회사와 공공기관에서 업무를 하다 보니 데이터 관련 역량이 필요했습니다. 하지만, 데이터에 대해서는 사실 잘 모르고 있었습니다. 하지만 현업과 병행하면서 여러 온라인 교육과정과 행정안전부 데이터 역량 강화 교육과정을 수강하여 업무에 많은 도움을 받고 있습니다.

이번 과제는 어떻게 참여하게 되셨나요? 기존 일과 어느 정도 관련이 있었나요?

우리 공사는 지역사랑상품권 '착' 시스템을 2018년부터 운영하고 있습니다. 지역사랑상품권 이용 고객이 점차 늘어나면서 부정수급이나 이상거래도 함께 증가하는 추세였기에, 이상거래 패턴을 분석해 다양한 시나리오를 만들고 검증하는 방법으로 시스템을 구축하고자 하였습니다. 저는 당시 지역사랑상품권 프로젝트 관리자였으므로 인공지능의 기계학습(머신러닝) 기법을 도입하여 부정 사용 문제를 해결하고자 참여하게 되었습니다.



해당 주제를 선정한 이유는 무엇이고 어떤 이해관계자가 참여했나요?

지역사랑상품권은 지역 활성화를 위해 중요한 사업이므로 지자체 담당 공무원들과 협조가 매우 중요했습니다. 상품권 부정 사용이 늘고 있고 효과적으로 단속할 방안에 대한 고민도 있었기에 공감대가 쉽게 형성되었던 것 같습니다. 담당 공무원들도 지역사랑상품권을 사용하면서 발생할 수 있는 부정 사용에 대해 다양한 의견을 제시해 주어서 많은 아이디어를 얻을 수 있었습니다.



데이터 분석을 위해 어떤 방법을 쓰셨나요? 전문가나 관계자를 참여시켰나요?

공사 내부데이터로 가입자별, 연령별 구매 패턴과 시간대별 구매 이력을 분석하기 위해 간단한 시계열 분석을 진행하였고, 데이터 컨설팅 전문가들의 지원을 받아 경남지역 상품권 부정거래 단속 참고 자료도 추가해 더욱 풍부한 데이터 SET을 구성할 수 있었습니다.

전문가들과 같이 '지도학습' 분석 방식으로 확보한 지자체 데이터와 경남지역 부정거래 단속 참고 자료에, '비지도 학습' 분석으로 공사 내부데이터 기반의 EDA(탐색적 데이터 분석)를 만들어 새로운 시나리오 유형을 추가했습니다. 점수화 방식을 이용해 이상거래 기준을 만들었고, 다양한 분석 모델 중 XG Boost 방식이 가장 적합하다고 판단해 최종 모델로 결정했습니다. 이후 가맹점별 이상거래 탐지 건과 회원 번호별 이상거래 건을 집계하여 상위 의심 가맹점과 회원을 도출하였습니다.



개발 모델을 공유 및 확산하기 위해 어떤 노력을 하셨는지, 혹은 어떤 부분이 필요하다고 생각하는지 궁금합니다.

지역사랑상품권 고도화 구축에 선정된 시나리오 10가지를 적용해 실제 효용성을 평가하려고 합니다. 이는 올해 하반기 고도화 시스템 구축 완료 시 적용될 예정입니다. 아울러, 지역사랑상품권 고도화 프로젝트에서 운영자가 이상 패턴 시나리오를 수정할 수 있도록 현재 UI를 편리하게 설계하고 있는데, 이는 이상 패턴이 발견되면 지자체 담당자에게 공유·보고되도록 하는 것입니다.

이러한 데이터 분석 활용이 계속 유지되고 발전하기 위해서는 전담팀이 있어야 하고, 지속해서 투자할 수 있는 전사적인 안목이 필요하다고 생각합니다. 교육도 당연히 이루어져야 하고, 특히 임원진의 관심과 지원을 받기 위해서는 데이터 분석의 필요성을 어필하면서 관련 업무를 조직적으로 추진해야 합니다. 저 또한 타 부서에 도움이 될 수 있는 과제를 찾아 만들고 인사이트를 주는 방안을 계속 찾고 있습니다.



데이터 활용을 통한 업무개선을 망설이는 분들께 도움 되는 말씀 부탁드립니다.

데이터 자체에 먼저 관심을 가져야 합니다. 데이터에 문제점이 보인다면, 이유가 무엇인지 의문을 가짐으로써 해결의 실마리를 찾을 수 있을 겁니다. 그러나 데이터의 정의와 문제해결은 완전히 다른 층위의 문제입니다. 데이터만으로 "문제를 해결"하기는 쉽지 않습니다. 시스템과 전문가를 통한다면 해결할 가능성이 커집니다. 하지만 데이터에 관심이 없다면 데이터 활용을 통한 업무개선은 시작조차 할 수 없을 겁니다.

모두가 데이터 분석 역량을 키우기보다는 각자의 영역에서 데이터를 '이해'하는 '리터러시(literacy) 역량'이 필요하고, 이를 기반으로 문제에 접근하는 것이 저는 더 효과적이라고 생각합니다. 더 많은 공공영역에서 더 좋은 데이터기반행정 사례가 나올 수 있도록 많은 관심과 참여 부탁드립니다.

PART 2. 공공행정

PART 2

공공행정

1. CCTV 영상을 업로드하면 차종별 교통량이 보인다
행정안전부 통합데이터분석센터

2. 빅데이터 효자손으로 긁어드립니다. 지역이 균형있게 발전하도록
충청남도 예산군 전산정보팀

3. 차곡차곡 쌓인 소방 활동 데이터, 과학 소방정책 이끈다
경남소방본부 소방행정과

4. 중학교 배정으로 인한 고충민원, 민원 데이터 분석으로 개선책을 찾다
국민권익위원회 민원정보분석과

PART 2-1. 공공행정

CCTV 영상을 업로드하면 차종별 교통량이 보인다

CCTV 영상데이터를 활용한
도로 교통량 분석모델

행정안전부 통합데이터분석센터



이렇게 바뀌었습니다!

AS-IS	TO-BE
• 교통량 조사 장비 설치 비용 발생 • 사람의 오류, 과적, 가변축 등에 측정 오류 발생 • 필요시 간헐적으로 사람 또는 감압장치 설치 지점에서 조사	• 기존 CCTV 활용으로 별도 비용 발생 없음 • 사람 및 검지[15]센서 오류 없음 • 저조도 영상 활용을 통해 위험지역과 무관하게 상시 조사 가능

추진 배경

도로 교통량 정보는 도로 운영, 대중교통 설계, 교통안전 개선을 위해 꼭 필요하며, 부가적으로 미세먼지, 부동산, 상권, 온실가스 배출 등 다양한 문제 접근에 활용할 수 있다.

그동안 차종 분류 조사 장비, 차량 운행 기록장치, 내비게이션 등을 활용하여 교통량 정보를 수집하기도 하였으며, 그 외에 다양한 방면에서 교통량 정보 수집을 위한 노력이 이루어져 왔다. 그러나 이러한 방법으로 얻은 정보는 특정 도로 구간에 한정되거나, 일부 민간기관에서 비공개로 활용되는 경우가 대부분이었다.

행정안전부는 이러한 문제를 해결하기 위하여 별도의 장비 없이 지자체에서 사용 중인 CCTV 영상을 활용할 수 있는 방법을 연구했다.

고민 끝에 자동차 차축과 무게가 도로에 영향을 주는 방식에 따라 자동차를 분류할 수 있다는 점을 고려하여, 12종[16]의 차량 교통량을 측정할 수 있는 모델을 개발하기로 했다. 이 과제는 서울시 동작구와 경상북도 안동시의 협조를 받아 완료되었다.

15) 검사하여 알아냄
16) 승용차, 버스, 소형·중형·대형 트럭, 3~6 차축 유형별 세미·풀 트레일러

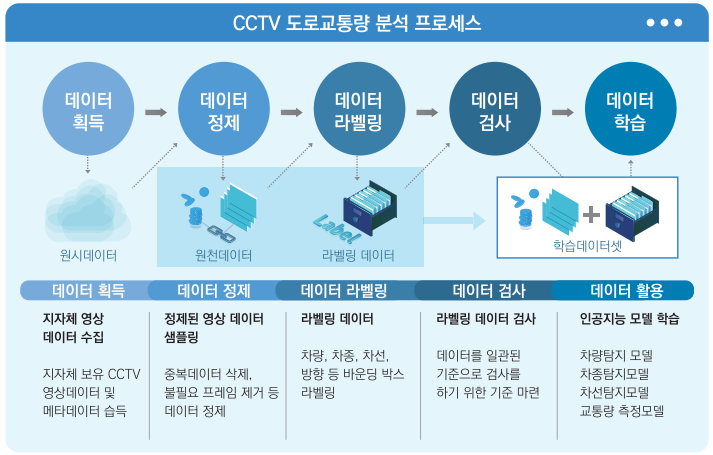
활용데이터

연번	데이터명	기준연도	형태	내 용	보유기관
1	CCTV 영상	2021, 2022	avi 등	안동시 CCTV 데이터, 동작구 스마트폴 CCTV 데이터	서울 동작구, 경북 안동시
2	차량 이미지	2020, 2021	jpg 등	양산시 12종 차량 학습 데이터, 트럭 10종 이미지 데이터	경남 양산시, 한국지능정보사회진흥원

차종의 정확한 분류를 위해 다양한 알고리즘 사용

분석 과정은 CCTV 영상 데이터를 확보하는 획득 과정, 불필요한 데이터를 제외하는 정제 과정, 차종을 사람이 입력해 주는 라벨링 및 검사 과정 그리고 AI가 분류 로직을 익히는 학습 과정으로 구분된다.

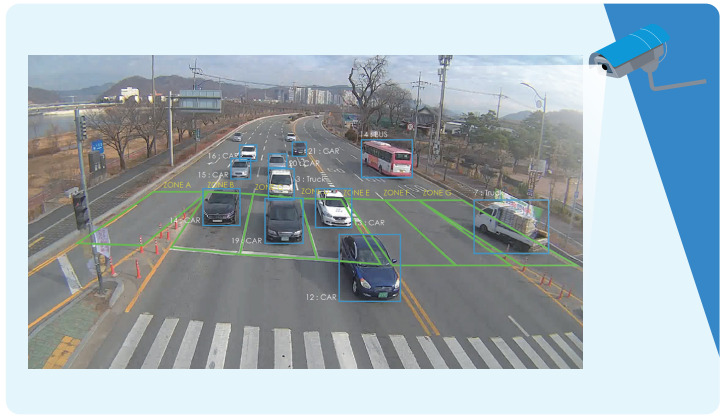
〈그림 17. CCTV 도로 교통량 분석 프로세스〉



먼저, CCTV 영상은 서울 동작구와 경북 안동시의 협조를 받아 확보했다. 개인정보보호 차원에서 차량번호판과 행인을 비식별 조치한 후 학습에 활용했다. 지자체가 관리하는 도로 특성상 컨테이너 차량과 카캐리어가 부족한 문제가 있었는데, 한국지능정보사회진흥원의 AI 허브가 보유한 차량 이미지 데이터로 보완했다.

데이터 정제 과정에서는 CCTV 설치 위치가 제각각이다 보니 영상별로 차량을 바라보는 각도가 다르기에 차량이 왜곡되어 보이는 문제가 있었으며, 이러한 문제를 해결하기 위해 영상 보정 작업이 필요했다. 차량이 통과하는 구간에 검지 구역을 사각형으로 설정하여 통과한 차량이 인식되었음을 알리는 바운딩박스(그림18의 하늘색 박스)가 표시되게 하였으며, 동시에 검지 구역을 CCTV가 바라보는 각도에 맞추어 사다리꼴 형태로 설정(아래 그림18의 연두색 박스)하여 왜곡을 보완했다.

〈그림 18. CCTV 영상 속 차량 인식 예시〉



그다음, 라벨링(labeling)[17]은 수작업으로 진행했으며, 영상을 보며 라벨러[18]들이 개개의 차량별로 차종을 기록했다. 이후 기록된 차종은 학습에 활용되었다.

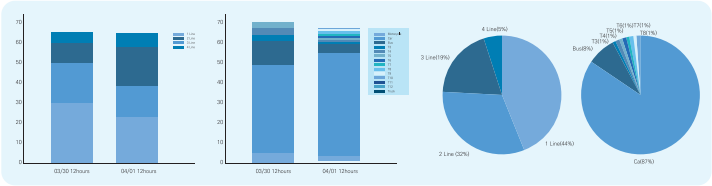
마지막으로, 12종 차종분류를 위한 인공지능 모델 학습을 진행했다. 실제 CCTV 영상을 판독하는 과정은 차량을 객체로 인식한 후에 라벨링된 영상을 기준으로 차종별 특징을 구분하여 12종 구분을 학습한다. 이 중 상당한 학습 시간이 요구되는 과정이 바로 은 차종분류이다. 같은 트럭이라도 크기와 차축에 따라 다른 차종으로 분류하기 위해 각각의 이미지를 학습하는 과정이 필요하다. 또한, 정밀한 분류작업을 위해 욜로(YOLO)[19], 레즈넷(ResNet)[20], 소트(SORT)[21]가 사용되었다.

정확도 98% 수준, 범정부 데이터 분석시스템 탑재

분석을 위해 20TB, 즉 DVD 영화 4,200편 분량에 해당하는 CCTV 영상에서 17만 대 분량의 차량 이미지를 추출·활용했다. 이 데이터들을 활용하여 서울 동작구와 경북 안동시의 51개소 CCTV를 학습한 결과, 그 정확도는 98% 수준으로 나타났다.

17) 해당 내용을 시스템에 학습시키기 위해 관련 데이터를 사람이 직접 입력·안내
18) 라벨링 하는 사람
19) 욜로(YOLO) : You Only Look Once 약자로, 객체 인식 모듈로서 다양한 사물을 빠르게 속도를 인식함
20) 레즈넷(ResNet) : Residual Network 약자로, 이미지 분류 및 감지 등에 사용되는 딥러닝 아키텍처
21) 소트(SORT) : Simple Online Realtime Tracking 약자로, 실시간으로 다수의 객체를 추적하는 알고리즘

〈그림 19. 차종별, 차선별 도로 교통량 분석 시각화〉



다양한 교통·환경 관련 정책에 활용 기대

개발된 분석모델은 현재 국토교통부가 주관하는 '전국 교통량 조사'에 시범적으로 활용되고 있으며, 지자체 및 공공기관에서도 자체적으로 사용 중이다.

또한, 행정안전부에서 운영 중인 '범정부 데이터 분석시스템'에도 탑재하여, 지자체를 비롯한 교통량 조사기관이 손쉽게 사용할 수 있도록 했다.

〈그림 20. 교통량 분석 결과 기반 신규 회전교차로 입지 선정〉



과거의 교통량 조사는 수작업으로 측정하거나 고가의 검지 센서를 설치하여 수행했으며, 지자체별 교통량 조사에 수천만 원의 예산이 소요됐다.

하지만, 교통량 분석모델 도입으로 인해 별도의 비용 발생 없이 교통량 측정이 가능해져 대폭적인 예산 절감 효과가 예상되는 한편, 행정 편의성 및 정확도 또한 대폭 향상할 것으로 기대된다.

기관소개(행정안전부 통합데이터분석센터)

행정안전부 통합데이터분석센터는 데이터기반행정법 제정(2020)에 따라 중앙부처, 지방자치단체, 공공기관을 아우르는 범정부적 관점에서 데이터 분석과 활용을 지원하기 위해 2021년에 신설되었다.

데이터 분석·활용 및 데이터 분석·활용 역량강화 지원, 범정부 데이터 분석시스템 구축·운영 등을 통해 각 기관의 객관적·과학적 의사결정과 이를 통해 좋은 정책이 만들어질 수 있도록 지원하고 있다.



Coffee Time (인터뷰)



자기소개 부탁드립니다. 데이터나 IT 분야 전공 여부도 함께 알려주세요.

저는 행정안전부 통합데이터분석센터에서 근무하는 김재만입니다. 우수 분석모델을 패키지화하여 공공기관에 공급하는 표준분석모델을 담당하고 있습니다. 전공은 산업공학이고 데이터 핸들링과 기계학습을 공부했습니다. 데이터에 관심이 많은 편이라 대학원 때 비전공 수업이라도 신경망이나 데이터웨어하우스 같은 과목을 찾아다녔습니다. 민간에서 주로 일했고 이십여 년 정도 경험을 쌓은 상태라 주요 업체들의 정보계 시스템과 다양한 도메인의 분석기법에 익숙한 편입니다.

이번 과제는 어떻게 참여하게 되셨나요? 기존 일과 어느 정도 관련이 있었나요?

본 과제는 표준분석모델 사업으로 진행되었습니다. 표준분석모델은 우수사례 선별이 필요한데, 전년도에 경상남도 양산시가 CCTV 차량인식 과제를 수행한 바 있습니다. 수작업을 줄이고 정확도를 개선해 그 결과를 교통량조사, 차로 조정, 도로파손, 차량 미세먼지 등 다양한 방면에 활용할 수 있겠다고 생각했습니다. 과제의 혁신성을 고려했을 때 이 분석모델을 표준화하여 보급할 필요가 있다고 판단하였습니다.

과제 추진 시 어려웠던 점이 있다면 말씀해주세요.

차종을 12종으로 분류하는 게 기술적으로 가장 어려웠습니다.. 한국도로공사의 자동 분류는 지면의 감압센서 4대가 차축과 무게를 인식하는 방식이었고, 고속도로 톨게이트 카메라는 번호판 인식 후 DB를 조회하여 차종을 파악하는 방식이었습니다. 일본은 2021년부터 교통량조사를 CCTV 조사로 바꿨지만, 이 역시도 소형차, 대형차, 오토바이로 한정하여 분류할 뿐이었습니다. 즉, 영상만으로 차종 12종을 분류해서 실용화한 사례가 없었기 때문에, 과제를 추진하면서 이 문제를 놓고 많은 토론이 있었습니다. 담당자로서, 위험도가 높은 목표에 불안감을 느낄 수밖에 없었으나 과제가 가진 의미와 영향력을 고려하여 과감히 인력 및 비용을 투입했고, 시도가 반복되면서 인식률이 개선되기 시작했습니다. 그 결과, 12종 분류를 완료하게 되었습니다.

또 하나를 더 얘기하자면, 이번 과제의 경우, 개인정보가 들어간 데이터를 사용해야 했기에 원천데이터를 가진 부서를 설득하고 안전장치를 마련하는 일련의 과정에 품이 많이 들었습니다. 데이터 담당자에게 갈 위험도를 제거하고 업무 내용을 구체적으로 안내했을 때 피드백을 조금 더 빨리 받을 수 있었는데, 앞으로 데이터 관련 업무를 하게 되실 분들은 참고하셔야 할 사항이라고 생각합니다.

데이터 분석 시 어떤 알고리즘을 사용했나요? 그 과정에서 담당자는 어떤 역할을 하셨나요?

CCTV 영상인식 알고리즘은 3가지 모듈로 구성되어 있습니다. 첫 번째는 자동차, 버스, 트럭을 인식해 차량 외곽에 사각형 바운딩 박스를 그려주는 알고리즘(Yolo)입니다. 두 번째는 인식된 트럭을 종과 차축 수에 따라 분류해 주는 알고리즘(ResNet)인데, 18개의 레이어를 가진 모델이 활용되었습니다. 세 번째는 이전 프레임과 다음 프레임의 박스 중 무엇이 같은 차량인지를 식별하는 알고리즘(SORT)입니다. 바운딩 박스의 각 꼭짓점과 다음 프레임에 있는 박스의 꼭짓점 중 거리가 가까운 것을 찾아 동일 차를 쫓고 원근감과 각도에 따른 형태 변화가 있는 경우 그로 인한 데이터 분산을 배제하는 역할을 합니다.

영상인식은 딥러닝에서도 별도의 전문성이 필요한 영역입니다. 분석을 직접 수행하지 않더라도, 담당자, 관리자로서 알고리즘에 대한 설명을 이해하고 목적에 따라 다른 알고리즘을 제안할 수도 있어야 하기에 영상인식과 관련된 최근의 알고리즘 대부분을 학습하고 다뤄보았습니다. 관련된 전문지식이 없다면 사업에 적합한 기술이 사용되고 있는지, 목표가 가까워지고 있는지, 사업 완료 후 어떤 문제가 생길 수 있는지를 예상하거나 대처하기 어렵습니다. 내외부의 관련 전문인력 섭외가 필요합니다.



개발 모델을 공유, 확산하기 위해 어떤 노력을 하셨는지 듣고 싶습니다. 또, 이런 프로젝트가 일회성으로 끝나지 않으려면 어떤 변화가 있어야 할까요?

교통량 분석모델은 범정부 데이터 분석시스템(insight.go.kr)에서 탑재하여 부처, 지자체, 공공기관에서 누구든 사용할 수 있습니다. 또한 국토교통부 주관 전국 교통량 조사(고속도로, 국도, 국가지원 지방도, 지방도 대상)에도 활용 중입니다.

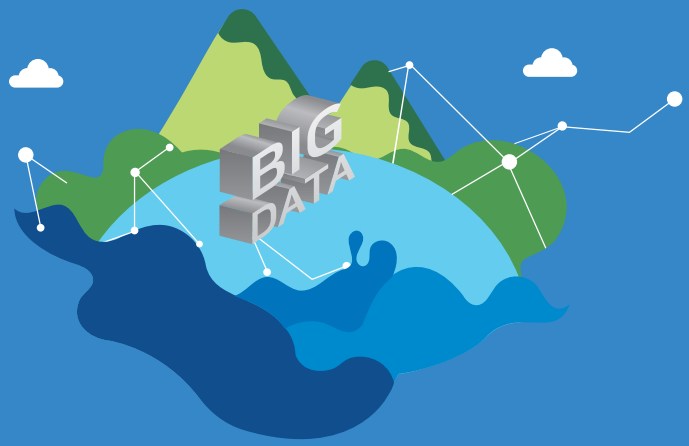
분석결과와 관련해서 한 가지 말하자면, 시스템으로 변환되어 업무 프로세스의 일부로서 정착되지 않으면 일회성으로 끝나는 경우가 있습니다. 데이터 분석 결과란 대부분의 사람에게는 이해하기 어려운 내용이며 가끔 이루어지는 영상교육이나 파견 교육만으로는 이러한 문제를 극복하기 어렵습니다. 오히려 분석에 대한 이해 없이도 사용할 수 있는 환경과 담당자의 피드백 없이도 변화를 받아들여 개선되는 시스템(데이터 유입부터 분석까지 자동화되는 데이터 옵스(ops) 환경)이 필요하다고 생각합니다.

PART 2-2. 공공행정

빅데이터 효자손으로 긁어드립니다. 지역이 균형있게 발전하도록

지방 신도시 현안 해결을 위한
지역 균형 발전모형 수립

충청남도 예산군 전산정보팀



이렇게 바뀌었습니다!

AS-IS	TO-BE
• 도시계획사업 구상 단계에서 필요 사업 중심으로 구도심 재생 뉴딜 마스터플랜 수립 • 인터뷰 및 현장답사 기반의 입지 선정 • 공간 범위를 예산읍 구도심 일원으로 한정	• 도시계획사업 계획 단계에 바로 적용할 수 있는 계량적 분석정보 제공 • 빅데이터 기반으로 객관적인 입지 선정 • 공간 범위를 예산군 전체 지역으로 확대하여 군 전체 균형 발전에 활용

추진 배경

전국의 지방자치단체를 중심으로 구도심 공동화 현상은 점점 심화하고 있다. 특히, 충청남도 예산군(이하 예산군)의 인구는 2000년 101,660명에서 20년 만에 80,114명으로 20,000명 넘게 줄어들었다. 또한, 인근 내포신도시가 분리 성장하면서 생활 SOC(Social Overhead Capital, 사회기반시설)[22]의 지역 불균형 현상이 발생해 지역 간 시설 보유 격차, 접근성 취약과 같은 문제점이 커지고 있다.

이를 해결하기 위해 예산군은 먼저, 신도시 성장에 따른 기존 도심의 중심 상권 쇠퇴, 청년층 인구감소 등 구도심 쇠퇴 원인과 현상을 먼저 분석했다. 그리고 그 결과를 예산군 혁신전략팀이 추진하는 '구도심 재생 뉴딜 마스터플랜 수립 사업'과 연계하여 입지 선정 분석모델의 활용성과 극대화하는 것을 계획했으며, 이를 실현하기 위해 지역 균형발전을 위한 생활 SOC 최적 입지 선정 모델 개발에 착수했다.

22) 일상생활에 필요한 필수 인프라

활용데이터

연번	데이터명	기준연도	형태	내 용	보유기관
1	유동인구	2022	xlsx	SKT 통신기반 유동인구 정보	충남도청
2	카드소비	2022	xlsx	신용카드 이용정보	
3	교통	2021	API	교통카드, 카드, 환승 정보	국토교통부
4	GIS 격자 통계	2022	DBF	격자별 통계정보	
5	행정구역 경계	2022	DBF	시군 행정동 경계 정보	통계청
6	시설	2022	xlsx	운영 중인 시설 정보	공공데이터포털
7	인구이동 현황	2022	xlsx	읍면별 전입 및 전출 현황	충남 예산군
8	부동산 거래 신고	2022	xlsx	부동산 거래 통계	
9	일반건축물대장	2021	xlsx	일반건축물 정보	
10	민원 원데이터	2022	xlsx	지자체 민원 접수 정보	
11	주민등록인구 현황	2022	xlsx	읍면별 인구 및 세대수 현황	
12	연령대별 인구현황	2022	xlsx	읍면별 연령별, 성별 인구 현황	

구도심 생활 SOC 취약 시설 중 '공영주차장' 입지 분석

충남 내포신도시 개발에 따라, 충남 예산군의 인구이동 현상이 발생했고, 인근 지역 간 시설 보유 격차가 발생하였다. 이러한 문제점을 해결하기 위해서는 생활 SOC 입지 분석모델 개발이 필요했으며, 이를 위해 데이터 수집 → 현황분석 → 분석모형 개발 → 결과 활용 단계 순으로 분석 사업을 추진했다.

단계별로 살펴보면, 첫 번째 분석데이터 수집 단계에서는 생활 SOC 분석에 필요한 시설 목록[23]과 위치 등의 예산군 내부 데이터와 SKT 유동인구 등의 외부 데이터를 확보하였다.

그리고 현황분석 단계에서는 탐색적 데이터 분석으로 GIS 기반의 분석 격자(국토정보플랫폼에서 제공하는 총인구 수 격자지도 적용)를 생성해 예산군 전체 지역의 인구수, 생활 SOC 개수 등 항목별로 통계

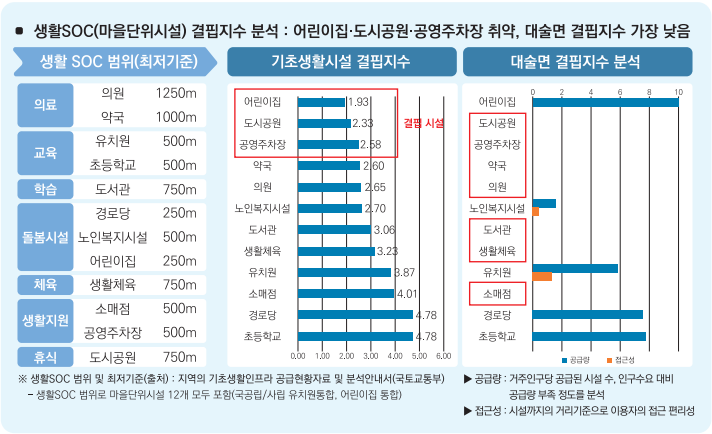
23) 국토교통부 생활 SOC 최저기준 : 의원, 약국, 유치원, 초등학교, 도서관, 경로당, 복지시설, 어린이집, 체육시설, 소매점, 공영주차장, 도시공원 (총 12개)

분석을 수행했다. 그리고 구도심 공동화 현황을 알아보기 위해 아파트 실거래 데이터, 인구 유동성 등을 분석했다.

그다음 분석모형 개발단계에서 예산군 내 전체 읍·면을 대상으로 도시쇠퇴 분석과 생활 SOC 결핍 지수를 평가했다. 생활 SOC 결핍 지수는 국토연구원의 분석체계를 적용한 지수로, 12개 생활 SOC 시설별로 공급량[24]과 이용접근[25]을 기준으로 각 읍면의 상대적 결핍도를 측정한 값[26]이다. 생활 SOC 결핍지수를 지역별로 분석 후 마을 단위로 적용하면, 해당 지역에 실제로 부족한 생활 SOC가 무엇인지 확인할 수 있다. 본 분석모델에서는 어린이집, 도시공원, 공영주차장이 결핍지수가 낮은 것으로 파악되었다.

이중 '공영주차장'을 분석 대상으로 선정했으며, ① 거주인구 ② 유동 인구 ③ Tmap 방문 건수 ④ 빅데이터 스코어 데이터를 바탕으로 분석모델을 개발했다.

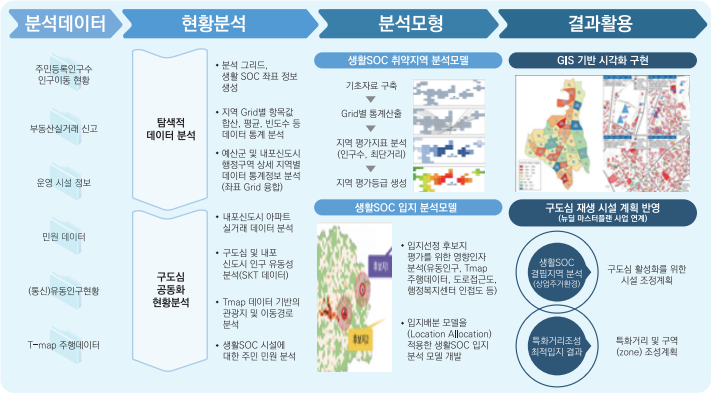
〈그림 21. 예산군 생활SOC 결핍지수 평가 결과〉



24) 공급량 기준을 말하며, 인구당 시설 수로 측정(인구수요 대비 공급량의 부족 정보를 비교)
25) 이용접근 기준을 말하며, 시설까지의 거리 기준으로 이용자의 접근성 측정(시설로부터 멀고 사각지대에 거주하는 인구가 많을수록 결핍도가 증가)
26) 결핍지수는 수치가 낮으면 낮을수록 실제 공급량과 이용접근이 높은 반비례 형식

마지막으로 '결과 활용 단계'에서는 생활 SOC 입지 분석 결과를 GIS 기반으로 시각화하여 구현하였으며, 생활 SOC 시설과 시설별 입지 영향 요인도 함께 확인할 수 있도록 구성했다.

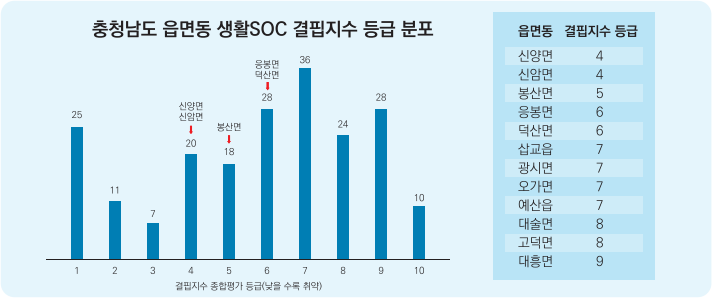
〈그림 22. 생활 SOC 입지 분석모델 분석과정〉



각 지역 최적 입지 도출, 향후 건축목적에 따라 활용

예산군 12개 읍·면의 생활 SOC 결핍 지수는 충청남도 전체 읍·면·동 중 중간 수준을 보이고 있으며, 신양면과 신암면이 4등급으로 가장 취약한 것으로 나타났다. 특히, 신양면, 봉산면, 대술면은 충청남도 전체와 비교해 보면, 12개의 생활 SOC 중 건강시설과 여가시설의 공급량과 접근성이 낮음을 알 수 있었다.

〈그림 23. 생활 SOC 결핍 지수 종합평가〉

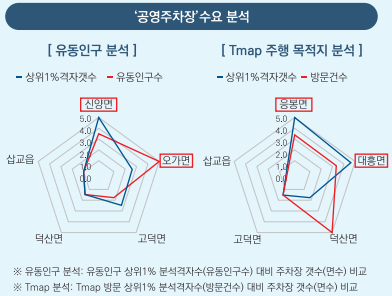


읍면동	결핍지수 등급
신양면	4
신암면	4
봉산면	5
응봉면	6
덕산면	6
삽교읍	7
광시면	7
오가면	7
예산읍	7
대술면	8
고덕면	8
대흥면	9

그리고 12개 읍·면의 생활 SOC 결핍 지수의 분석 내용, 도시쇠퇴 분석 결과 및 공영주차장 수요 분석을 종합해 본 결과, 신양면, 오가면, 대흥면, 응봉면을 공영주차장 입지 분석모델 대상 지역으로 선정하였다. 또한, 로지스틱 회귀모델[27]을 적용해 공영주차장 입지 영향 요인을 검증한 결과, 총 9개의 인자(유동인구, Tmap 주행데이터 등)가 선정되었다. 이 영향인자를 기준으로 총 6,181개의 격자가 공영주차장 입지 후보 지역으로 선정되었다.

〈그림 24. 공영주차장 입지 분석〉

	도시 쇠퇴지표	결핍지수 (충남전체)	결핍지수 (12개읍면)
봉산면	3	접근성취약	1.8
신양면	2	접근성취약	1.8
신암면	2		1.7
광시면	2		2.5
대흥면	2		3.0
응봉면	2		3.4
고덕면	2		3.6
삽교읍	2		4.8
예산읍	2		6.3
대술면	1	접근성취약	1.4
오가면	1		2.6
덕산면	1		4.7

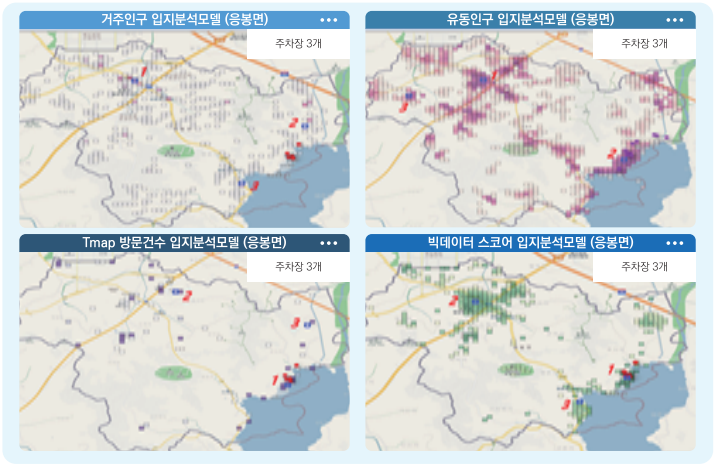


27) 두 개(0또는 1)의 값만 가지는 종속변수와 독립변수들 간의 인과관계를 로지스틱 함수를 이용하여 추정하는 통계기법이다.

공영주차장의 최종 입지는 입지 배분 모델을 적용하여 선정했는데, 입지 배분 모델을 통해 몇 개의 시설이 최적이며 각 시설에 어느 지역의 이용자를 배분 혹은 할당하는 것이 효율적인지 해답을 얻을 수 있다.

입지 배분 모델에 ①거주인구, ②유동인구, ③Tmap 방문인구, ④빅데이터 스코어 데이터를 각각 구분하여 반영한 4개의 입지 선정 모델을 개발하였다. 기존 주차장 대비 신규 주차장이 추가되는 경우의 접근성 개선 효과를 기준으로 분석모델을 검증하여 공영주차장 최적지를 선정했고, 각 지역에서 공영주차장 건축목적에 따라 최적의 결과를 활용할 수 있도록 하였다.

〈그림 25. 공영주차장 입지 분석모델 분석 결과(예시)〉



구도심 공동화 현상을 극복할 과학적 정책도구

예산군은 지방 신도시 현안 해결을 위한 지역 균형 발전 모형을 개발함에 따라, 향후 12개 생활 SOC 관련 사업을 추진할 때 이번 분석모델을 공유하여 입지 선정과 분석에 드는 비용을 절감할 예정이다.

또한, 혁신전략팀의 구도심 재생 시설 계획에 반영하여 구도심 활성화를 위한 시설 및 특화 거리·구역 조성을 위한 객관적인 근거자료로도 활용될 것이다.

이 밖에도, 예산군 2030 계획과 내포신도시 확장 기본구상 및 타당성 연구용역 결과에 객관적인 근거를 접목하여 예산 낭비를 최소화하고, 심화하는 구도심 공동화 현상을 극복하는 데도 큰 역할을 하리라 기대한다.

기관소개(충남 예산군청)

충청남도 예산군은 '새로운 내일 하나 된 예산'이라는 군정 구호 아래 군민 삶의 질 향상과 수준 높은 행정 서비스 제공을 위해 최선의 노력을 다하고 있다. 군민의 어려움을 적극 해결하고 군민의 의견이 빠짐없이 반영되는 행정서비스를 시행하기 위해 노력하고 있으며, 예당호, 예산시장 등 다채로운 볼거리, 즐길 거리, 먹거리가 가득하여 모든 이에게 사랑받는 지역으로 거듭나고 있다.



Coffee Time (인터뷰)



현재 하시는 일을 포함해 자기소개 부탁드립니다.

저는 충청남도 예산군청 총무과 전산정보팀에서 데이터기반행정과 빅데이터 분석 업무를 맡고 있는 신규호 주무관입니다. 정보통신공학을 전공했는데, 불과 몇 년 전만 해도 '데이터'는 지금만큼 익숙한 용어가 아니었어요. 게다가 데이터기반행정이란 말은 들어보지도 못했습니다. 데이터, DB, 빅데이터에 관한 약간의 지식이 있었지만, 소프트웨어보다는 하드웨어 관련 분야로 진로를 결정하였기에 데이터 업무에 적응하기까지는 다소 어려움이 있었습니다.

이번 과제를 진행하면서 어려웠던 점은 무엇인가요?

가장 큰 어려움은 데이터 수급과 기존에 없는 다중 로지스틱 회귀모형을 개발하는 것이었습니다. 기본적으로 필요한 데이터는 민간 데이터를 구매하거나 충청남도 데이터 포털(올담)을 사용했지만, 특정 데이터는 자체 보유 데이터를 활용하고 가공해야 했습니다. 그 과정에서 시간이 오래 걸려 어려움이 있었고, 활용할 수 있는 기존 분석모형이 없어 교수님들과 전문업체의 도움을 받아 새로운 모형을 만들 수 있었습니다.

이번 과제에 참여했던 이해관계자는 누구였나요? 중간에 가설이나 방향이 바뀌는 경우가 있었나요?

생활 SOC 관련 사업은 교육, 학습, 돌봄, 체육, 휴식, 생활 편의, 교통으로 대분류하는데, 군청 내 모든 부서와 관련이 있다고 해도 틀린 말이 아닙니다. 각자 다른 분야의 담당자를 모아 토의하고 협의하는 과정에서 서로 공통적인 문제점을 찾고 해결해 나가면서 힘든 점들을 극복할 수 있었죠. 그리고 담당자들과 회의를 통해 분석에 대한 방향이 보다 더 뚜렷해졌고 확장 방안을 많이 고민하게 됐던 것 같아요.

데이터 수집·정제 과정에서 느꼈던 어려움을 어떻게 극복하셨나요?

앞의 내용을 보면 아시겠지만, 저희가 수집한 데이터가 상당히 많았습니다. 양도 양인데, 내부 데이터는 담당자들이 직접 만든 데이터다 보니 점(.)이나 콤마(,) 등 잘못 표기된 경우가 많아서 하나하나 수작업으로 정제하느라 많은 시간이 소요됐어요. 원하는 데이터가 없을 때는 데이터를 가공해 만들거나 분석 결과를 활용해 데이터를 채워나가기도 했고요. 그리고 정제 과정에서는 관련 분야 교수님, 전문업체의 도움을 받았습니다.

개발 모델을 공유, 확산하기 위해 어떤 노력을 하셨는지 듣고 싶습니다.

분석 결과를 다른 지역, 기관에서도 참고할 수 있도록 '범정부 데이터 분석시스템(insight.go.kr)'과 충청남도 데이터 포털 '올담(alldam.chungnam.go.kr)'에 게시했습니다.

저는 아무리 좋은 아이디어로 모델을 만들어도, 사용자가 이해하지 못하거나 중요성을 인식하지 못하면 아무 소용이 없다고 생각해요. 그래서 과제가 선정될 때부터 데이터 분석의 목적과 중요성을 인식시켜 주는 노력이 중요하다고 생각합니다. 그런 점에서 홍보가 매우 큰 역할을 하는데요. 일반적인 홍보보다는 '어떠한 데이터 분석을 통해 어떠한 결과물이 나왔다'라는 걸 보여주고 그 결과물을 활용한 사업들에는 어떤 것이 있는지 눈으로 직접 보고 느낄 수 있도록 해주는 것이 가장 좋은 방법이라고 생각합니다.



데이터 활용을 통한 업무개선에 선뜻 나서지 못하는 분들을 위해 해주실 말씀이 있다면요?

데이터 분석을 통해 얻은 정보는 주관적·경험적 견해의 산물이 아닌, 객관적인 결과물입니다. 또, 다양한 업무에 데이터 분석을 적용할 수 있죠. 예를 들어, 입지 선정과 관련되면 '000 입지 분석', 지역축제를 분석하여 축제의 방향성을 잡는다면 '000 축제 분석', 작물 환경을 분석하는 데 써본다면 '농업환경 분석' 등 어떠한 업무에도 접목될 수 있으므로 하루라도 빨리 데이터 분석을 활용해 업무를 개선하고 의사결정에 참고하시는 게 좋으리라 생각합니다.

이외에도 저희는 충청남도와 '도시가스 우선 설치지역 선정' 분석을 진행하고 있는데요, 도시가스 배관데이터, 세대 주소 데이터, 취약계층 데이터를 가지고 도시가스 보급률이 저조한 지역과 취약계층 주민의 주소를 매칭해 에너지 복지와 경제성을 최대한 만족시킬 수 있는 지역을 추출하는 작업입니다. 이처럼 아이디어만 있다면, 빅데이터를 활용해 업무를 개선할 방법은 무궁무진하므로 주저하지 말고 일단 도전해 보라고 말씀드리고 싶어요.

PART 2-3. 공공행정

차곡차곡 쌓인 소방 활동 데이터, 과학 소방 정책 이끈다

소방수요 예측 모델 개발

경남소방본부 소방행정과



이렇게 바뀌었습니다!

AS-IS

- 내부 변수만을 고려하여 재난 발생의 원인을 추정할 수 있는 범위가 좁음
- 단순 통계의 평균을 활용하여 정책 근거로 활용

TO-BE

- 과거 일정 기간의 내·외부 변수를 활용하여 재난 발생원인 추정 범위 확대
- 다양한 사회적 변수(노령인구, 용도별 건축물, 산업시설 등)를 고려한 결과를 정책에 반영

추진 배경

오늘날 소방 수요는 국민의 생활양식 변화, 이상기후, 감염병 확산 등의 요인으로 인해 과거보다 훨씬 증가했으며, 더욱 복잡한 형태를 띠고 있다. 아울러, 각종 재난 사고에 효과적으로 대응할 수 있는 과학적인 접근방법 또한 요구받고 있다.

경상남도소방본부(이하 경남소방본부)는 이러한 환경변화에 적극 대응하고자, 그간 축적된 소방 활동 데이터를 바탕으로 다양한 예측 기법을 활용하여 소방수요 즉, 화재, 구조, 구급 등의 사고 발생을 예측하는 '소방수요 예측 모델'을 개발하게 되었다.

활용데이터

연번	데이터명	기준연도	형태	내 용	보유기관
1	소방 활동 데이터	2013~현재	API	화재·구조·구급 등 소방 활동내역	경남소방본부
2	사회통계 데이터	2013~현재	xls 등	인구, 산업구조 등 사회통계 변수	통계청 (국가통계포털)

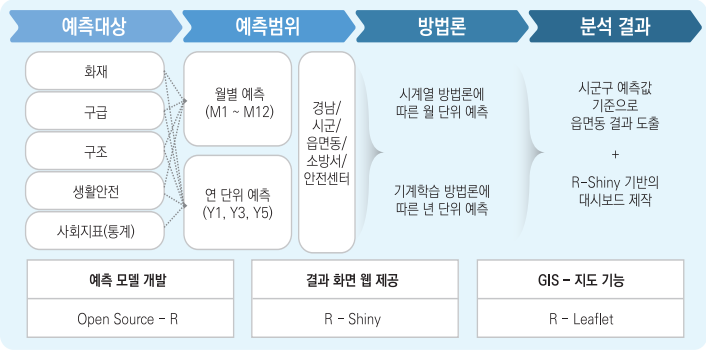
월별·연도별 예측 모델로 설계

소방 활동 관련 정보시스템(국가화재정보시스템, 구급 정보시스템 등)은 개인정보와 같은 민감정보가 포함되어 있어 별도의 폐쇄망을 운영 중이며, 위의 시스템들은 화재, 구조(생활 안전), 구급으로 분류되는 소방 활동에 따라 별도의 서버를 사용하고 있다. 다만, 소방 활동에 관한 내부데이터를 한곳으로 모으기 위해 경남소방본부는 폐쇄망 속에 별도의 서버를 구축했으며, 중계 서버를 통해 외부의 데이터를 확보할 수 있도록 정보보안 정책을 개선했다.

각종 사회변수를 나타내는 외부 데이터는 월별로 데이터가 정기적으로 업로드되는 통계청 국가통계포털을 통해 확보했다. 이처럼 내부 및 외부 데이터를 모은 다음, 예측 범위에 따라 월별 예측과 연도별 예측으로 구분하여 모델을 개발했다.

월별 예측 모델은 내부데이터인 화재, 구조, 구급, 생활 안전 출동자료만 활용해 시계열 방법론으로 분석 시점에서 향후 12개월을 예측하도록 구성했으며, 연도별 예측 모델은 내부데이터인 소방 활동과 각종 사회지표(통계)를 함께 사용하여 향후 1, 3, 5년을 예측하도록 설계했다. 해당 데이터를 바탕으로 월별 및 연도별 예측 범위를 설정한 분석 모델은 경남(창원제외) 전체와 각 시·군, 소방서, 119안전센터 및 읍·면·동 관할 지역단위로 분석 결과가 도출되도록 모델을 구성했다.

〈그림 26. 분석 과정〉

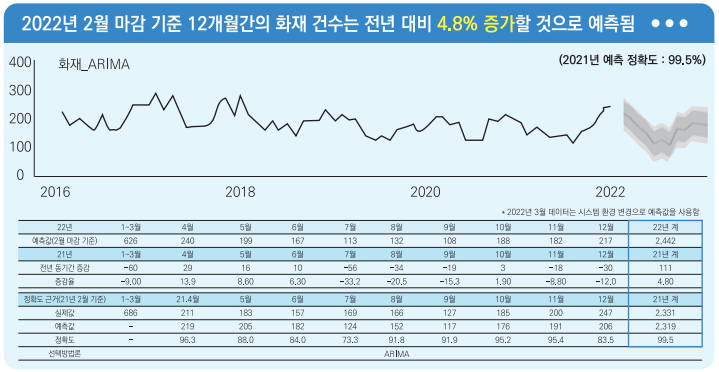


월별 예측 모델을 개발하기 위해 여러 개의 시계열 분석기법을 사용하여 최근 3개월에 대해 가장 안정적으로 예측한 모델을 추천하도록 구성했다. 연도별 예측 모델은 다양한 데이터를 활용하는 만큼 기계학습(랜덤 포레스트)을 채택했다. 최소 3년 이상의 관측치를 학습데이터로 사용했으며, 5년 후 예측은 학습데이터와 검증데이터로 사용할 수 있는 데이터가 부족해 분기별로 나눈 후 데이터를 4배로 증폭하여 분석에 사용했다.

연간 소방 활동 결과와 비교, 목표치 상회

소방수요 예측 모델의 목표 정확도는 80%로 설정했다. 월별 예측 모델은 검증 데이터인 2021년 월별 소방 활동 결과를 바탕으로 하였으며, 예측 정확도는 화재 99.5%, 구급 92.7%, 구조 96.3%, 생활 안전 88.1%로 목표치를 웃도는 결과를 달성했다. 연도별 예측 결과 역시 1년 후 예측 결과와 2021년 연간 소방 활동을 비교해보니, 화재 85.5%, 구급 96.2%, 구조 83.0%, 생활 안전 98.2%로 목표치를 달성했다.

〈그림 27. 월별 화재발생 예측 정확도〉



예측 모델 결과는 시각화 도구(R-Shiny)[28]와 Leaflet[29]를 사용해 웹 기반으로 구현하여 사용자가 결괏값과 정확도를 도표와 지도로 편리하게 확인할 수 있도록 했다.

위와 같은 분석 결과 외에도 연간 분석에 사용된 외부 데이터 변수의 중요도를 확인할 수 있도록 하고, 정책 활용이 가능한 후보군을 선별해 6개의 정책과제를 선정하여 정책안을 제시했다.

수요예측 모델 활용, 6개 중점 관리 정책안 도출

경남소방본부에서 최초로 시도한 '소방수요 예측 모델'의 정책적 목적은 중장기적인 시각에서 소방력 보강을 위한 방안을 마련하고, 나아가 다양한 데이터 활용으로 현장의 문제를 해결하여 도민의 안전을 지키는 조직으로 거듭나는 것이었다.

수요예측 모델을 정책에 활용하기 위해 현재의 재난환경과 소방 활동 프로세스에서 시사점을 발굴했다. 넓게는 소방서 관할지역, 좁게는 119안전센터 관할의 수요예측 결과와 분석 과정에서 확인한 변수 간 상관관계를 바탕으로 조직관리의 중점 관리 대상인 6개의 정책안을 도출했다.

28) R로 웹 어플리케이션을 제작할 수 있게 해주는 프레임워크
29) 매핑 어플리케이션을 만드는 데 사용되는 오픈소스 자바스크립트

〈표 4. 중점 관리 정책안〉

구분	세부내용(과제)	비고
1	센터/지역별 소방력 보강·조정안 제시	우선
2	다빈도 구조요청 건에 대한 예방·대응 홍보 강화	
3	구급 출동 시 중증질환에 대한 선별적 대응	우선
4	고령인구 증가지역에 대한 구급대원, 장비 우선 배치	우선
5	외국인 안전의식 개선을 위한 홍보활동 강화	
6	건축물 단위별 화재 위험도 평가 체계 개발	(중장기)

위 정책안을 기초로 업무별 시행부서에서는 '119안전센터/지역별 소방력 보강 조정', '제11차 소방력 보강 계획', '전담의용소방서 설치' 등의 계획수립에 활용했다. 또한, 월별 소방수요 예측 결과를 별도로 분석하여 코로나19 전후 변화한 소방 수요의 특성을 확인하고 예방 대응 정책에 반영할 수 있도록 했다. 아울러, 건축물 단위별 화재 위험도 평가를 위한 분석 및 애플리케이션 개발을 자체적으로 추진하고 있다.

기관소개(경상남도소방본부)

경상남도소방본부는 1992년 설치되어, 각종 화재, 구조, 구급뿐만 아니라 도내의 모든 소방 행정업무를 총괄·지휘·통제·관리 운영하고 있다.

Coffee Time (인터뷰)



현재 하시는 일을 포함해 자기소개 부탁드립니다.

경상남도소방본부 소방행정과에서 빅데이터 분석기획 업무를 맡고 있는 소방경 곽우섭입니다. 대학에서 법학을 전공했고, 컴퓨터활용능력 1급을 보유하고 있습니다. 2020년 7월 인사발령으로 현재 업무를 담당하게 되었고 온라인 강의, 유튜브 등을 통해 기초통계와 분석기법을 조금씩 배워나가고 있습니다.

이번 과제는 어떻게 참여하게 되셨나요? 기존 일과 어느 정도 관련이 있었나요?

데이터기반행정의 활성화로, 다양한 데이터 수집이 가능해지면서 소방안전분야에도 데이터기반행정을 적용해 봐야겠다는 생각이 들어서 시작하게 됐어요. 원래는 재난 사고가 데이터를 기반으로 어떤 인과관계로 발생하는지 밝히고 싶었지만, 너무 많은 변수가 각각의 상황에 다르게 영향을 미치고 있다는 재난의 복잡성 때문에 접근을 달리하게 되었죠.

우리가 수집할 수 있는 다양한 데이터를 바탕으로 화재, 구조, 구급, 생활안전 출동으로 구분되는 소방 활동의 수요를 예측하고, 거기에 영향을 미치는 변수를 하나하나 분석해보자는 결론을 내린 후에 해당 과제를 시작할 수 있게 되었습니다.

과제를 추진하면서 겪었던 어려움에 관해 듣고 싶습니다. 또, 이해관계자들은 어떻게 참여시키셨나요?

사실 가장 많이 고민했던 부분이 '활용'이었는데요. 소방수요, 즉 사고발생량을 예측해서 단순하게 도민에게 알리면 스스로 예방하고, 불가피할 경우 피해를 어떻게 줄일 수 있을까에 대한 의문이 해소되지 않았습니다. 그래서 예측 모델을 개발함과 동시에 재난사고 발생에 영향을 미치는 주요 변수를 살펴 시사점을 도출하고 정책적으로 활용할 수 있는 제안을 동시에 할 수 있도록 방향을 수정했어요.

저는 분석 모델 개발 사업의 실무감독 공무원으로 참여했고, 소방 활동 구분에 따라 각 실무담당자가 참여하도록 관련 부서에 요청해 분석 모델 개발을 함께 진행했습니다.

데이터 분석 때 사용한 알고리즘이나 방법에 관해 조금 더 자세한 설명 부탁드립니다.

수요예측 결과는 시간상으로 월(+1~12개월)·연(+1, 3, 5년) 단위로, 공간상으로 경남의 시·군, 읍·면·동, 소방서, 119안전센터 관할을 단위로 나타나도록 설계하였습니다.

분석 모델 개발 초기에 데이터 수집 및 초기 모델 설계 과정 분석 과정에서 문제가 생긴 적이 있었는데요. 월 단위 예측 모델을 구현할 때 전월 데이터까지 마감한 후 결과를 표출하도록 하여 매달 결괏값을 계속 갱신하도록 설계했는데, 통계청 데이터는 매년 2~4월 중에 전년도 통계를 공개하는 시스템이었습니다. 그래서 월별 수요예측은 소방 활동 데이터만 사용해 시계열 분석기법을 적용했죠.

연도별 예측 모델은 선정된 변수를 사용해 랜덤 포레스트를 적용하였고, 결괏값 외 모델에 사용된 변수의 중요도(분석 모델에 사용된 빈도) 또한 나타나도록 하였습니다. 해당 분석 결과는 앱으로 시각화하여 직관적으로 확인할 수 있도록 했습니다.



개발 모델을 구체적으로 활용하고 공유하기 위해 어떤 노력을 하셨나요?

처음 계획수립 단계부터 분석 모델을 지속해서 사용할 수 있도록 웹 애플리케이션으로 시각화하여 분석 결과를 서버에 적재하고, 정확도를 확인할 수 있도록 했습니다. 또, 매년 연간 분석에 사용되는 변수의 중요도를 확인해, 연관성을 찾을 수 있도록 리스트로 나타나도록 했습니다.

수요예측 분석 모델 개발과 동시에, 정책개선안 6개를 도출했고, 그중 부족한 소방력을 보완하기 위한 의용소방대 설치안은 수정·보완되어 전담의용소방대(소방관서가 미설치된 지역에 설치)가 설치되었습니다.



데이터 활용을 통한 업무개선에 부담을 느끼는 분들께 도움되는 말씀 부탁드려요.

정책을 입안하는 과정에서 데이터를 분석한다는 건 결국 문제를 파악하는 일에서부터 시작되는 것 같아요. 국민은 불편하게 느끼거나 위험하다고 생각하는 부분이 있으면 관련 공공기관에 민원을 제기하는데요, 민원 접수건 하나하나를 데이터로 변환하여 다양한 시각으로 바라보고, 다른 요인들과 결합하는 시도를 해본다면 보다 근본적인 문제 파악과 해결책이 도출될 것으로 생각합니다.

PART 2-4. 공공행정

중학교 배정으로 인한 고충민원, 민원 데이터 분석으로 개선책을 찾다

중학교 입학 배정 민원
분석을 통한 적극 행정

국민권익위원회 민원정보분석과



이렇게 바뀌었습니다!

AS-IS	TO-BE
• 배정 절차 및 방식에 대한 정보채널 산재	• 배정 절차 및 방식에 대한 정보채널 통합·일원화
• 주요 사항 변경 시 학부모 등 의견수렴 제한	• 배정방식에 대한 정보 접근성 향상 및 참여기획 확대
• 전년도 배정계획을 답습하여 유사 민원 반복	• 학생 인구 이동분포 조사를 통한 배정계획 수립
• 재배정 허용 사유 엄격히 제한	• 재배정 요청자에 대한 구제 기회 확대

추진 배경

중학교 배정은 교육청에서 '학교군[30]/중학구[31] 및 추첨 방법'을 고시하면 교육청 소속 각 교육지원청에서 그를 바탕으로 지역 실정에 맞는 배정계획을 수립한 후에 이루어진다. 그런데 이 배정 방법과 절차가 교육지원청별로 달라 학부모로선 이해하기가 쉽지 않기에, 민원이 계속 제기되었다. 교육청에서 학교군과 중학구를 지정한 후 교육지원청에서 실제 중학교 배정업무를 하기까지 3개월 이상 소요되고, 각 교육지원청에서 시행하는 우선 배정이나 재배정 방법의 차이가 크기 때문이다.

이에 따라, 교육부는 전국 중학교 입학 배정 관련 민원을 해소하고, 학부모의 양육 부담을 줄이기 위한 정책개선 방안을 모색했다. 가장 먼저 할 일은 학부모들이 제기한 민원 양상을 통계적으로 정확히 알아보는 것이었다. 그래서 교육부가 국민권익위원회(이하 권익위)에 중학교 배정과 관련된 민원 분석을 요청하면서 이번 데이터 분석을 시작했다.

30) 특정 지역 내 학생을 추첨방식으로 상급학교에 배정하기 위하여 복수의 학교들을 묶어서 구성한 학교의 군
31) 통학 상 거리·교통·지역 여건 등을 고려해 추첨 없이 특정 중학교에 입학하도록 설정된 구역

활용데이터

연번	데이터명	기준연도	형태	내 용	보유기관
1	중학교 입학 배정 민원데이터	2021	exl	2021.3.~2022.2. 기간 중 제기된 중학교 입학 배정 민원 5,036건	국민권익위원회
2	교육지원청별 배정업무 관련 자료	2015~2021	hwp, pdf	배정업무 시행계획 및 연구용역자료 (중학교 배정 방법 개선 연구, 학생 유발률 산출 연구용역 자료 등)	교육청

분석과제 추진을 위한 각종 데이터 수집

권익위는 2012년에 국민신문고와 자치단체 새올시스템의 민원 데이터를 수집·분석·모니터링할 수 있는 「민원 정보분석시스템」을 구축·운영하고 있으며, 매년 1,200 백만 여건의 민원 데이터가 축적되고 있다. 관련 분석 과제 수행을 위해 해당 시스템에서 2021학년도 중학교 입학 배정 관련 민원 약 14,000건을 추출했다. 아울러, 교육부로부터 전국 중학교 현황, 각 교육지원청 배정 업무 시행계획과 중학교 입학 배정 제도개선을 위해 실시한 각 교육청의 연구용역 자료 등을 제출받았다.

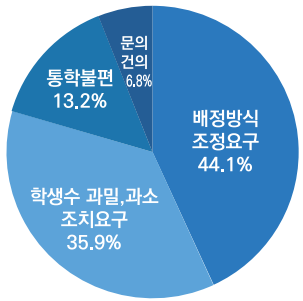
4가지 민원유형(배정방식·학생수·통학·일반문의)으로 분류

우선, 권익위는 분석 대상이 되는 데이터 세트를 확정하기 위하여 민원 관련 키워드, 기간, 조건 등을 입력하고, 해당 데이터를 엑셀 파일로 추출한 뒤 분석 주제와 상관없는 데이터를 삭제하는 정제작업부터 실시했다. 이를 통해 확정된 데이터 세트는 총 5,036건으로, 이 민원 데이터를 대상으로 민원인의 현황, 신청지역, 월별 민원 제출 추이 등을 분석했다. 또한 각각의 민원 내용을 일일이 검색해서 나온 언론·SNS 등 외부 데이터를 참고한 뒤 유사한 내용의 민원을 4개로 유형화했다.

4개 유형은 ① 배정방식 조정 요구 44.1%(2,219건) ② 학생 수 과밀·과소 조치 요구 35.9(1,810건) ③ 통학 불편 13.2%(667건) ④ 배정방식 및 현황 관련 문의 및 건의 6.8%(340건)였고, 권익위는 이 4개 유형을 다시 검토하여 세부 항목으로 재분류했다.

각 민원 유형에서 국민 불편을 유발하는 원인을 파악하기 위해 관련 법령 규정과 운영현황을 꼼꼼히 살펴봤다. 만약, 미비 사항이 도출되면 통계자료나 연구용역 자료, 언론자료 등을 참고해 정책 개선사항을 발굴했다. 권익위는 민원통계 분석 이후 발굴된 적극행정 방안의 적절성 및 수용가능성을 확인하기 위하여 교육청 중학 배정업무 담당자, 교육부 협업담당자와 업무협의를 거쳤다.

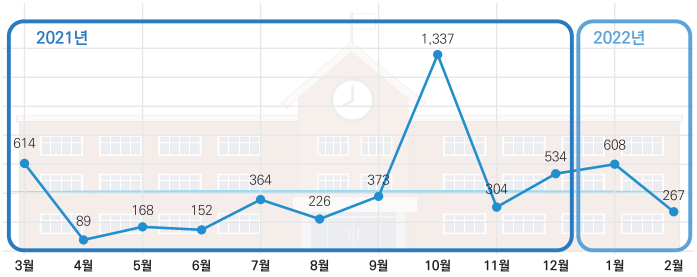
〈그림 28. 민원 유형〉



중학교 배정 정보에 대한 접근성 제고 등 유형별 민원해소 방안 제시

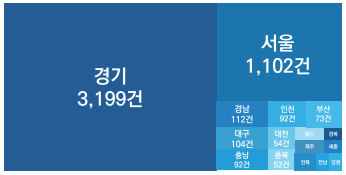
'중학교 입학 배정' 관련 민원의 1년간 월별 추이를 살펴봤을 때 학부모 대상 중학교 입학설명회가 개최되는 10월에 민원 신청 건수가 월평균 420건 대비 3.1배(1,337건)나 급증함을 알 수 있었다.

〈그림 29. 중학교 입학 배정 관련 민원 발생 현황〉

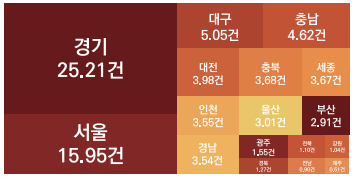


민원인의 연령별 유형을 보면, 30~40대가 전체의 85%를 차지했다. 지역별로는 경기도, 서울특별시, 경상남도 순으로 나타났는데 신도시 건설, 신규아파트 입주 등이 빈번한 수도권지역의 민원이 전체의 85.5%를 차지했다.

〈 그림 30. 지역별 민원 건수 〉



〈 그림31. 지역별 입학생수 1,000명당 민원 건수 〉



이를 토대로 권익위가 유형별 민원 해소를 위해 도출한 적극행정 방안은 다음과 같다. 먼저, 중학교 배정방식에 대한 정보접근성을 높이고 참여기회를 확대하도록 제안했다. 이를 위해 교육청과 교육지원청에서 각각 공개하는 배정 절차, 현황 및 결과 등에 대해 정보제공 채널을 일원화하는 방안을 제시했다.

두 번째로, 기존 학교군·중학구 및 추첨방식 등 주요 사항을 변경할 때는 교육청·교육지원청·학교·학부모가 협의절차(예 : 사전설명회, 공청회, 협의체 구성 등)를 마련해 정책 방향과 애로사항을 공유하며 합의점을 찾도록 했다. 또한, 학교 재배정은 한정된 사유 외에는 불허하고 있으나 부득이한 경우(예 : 가족 돌봄 필요, 통학 수단 부재로 자율적 통학 불가능 등)에는 허용할 수 있게끔 방안을 마련하도록 제안했다.

이외에 지역 내 인구 증감 요인(도시개발, 출산율 감소 등)과 지역 특성을 고려해 학생 인구 이동분포에 대한 상세한 조사를 거쳐 입학 배정제도를 마련하도록 했다. 그리고 통학환경 불편 호소 민원 관련해서 주기적으로 실태를 조사하고 통학 수단 확보 등 통학환경이 안정적으로 정착할 때까지 점검하도록 했다. 마지막으로 우선 배정대상자 여부를 학교에서 먼저 교육행정정보시스템(NEIS) 재학 정보 등으로 확인해 서류제출 부담을 줄이도록 했다.

지역 여건에 맞게 중학교 입학배정제도 개선

교육부는 지역에 따라 민원 내용의 차이가 있는 점을 고려해 유사 특성 지역을 모아 권역별 담당자 회의(총 3회)를 진행했다. 중학교 배정업무는 전국 176개 교육지원청별로 시행계획을 수립해 실시하므로 획일적인 기준을 적용하기는 어렵지만, 각 교육지원청은 지역 여건을 고려해 최대한 민원 해결 방안을 모색했다.

그 결과, 교육부는 민원 빅데이터 분석 결과를 바탕으로 지역에 맞게 자율적으로 중학교 입학배정 제도를 개선하도록 권고하였다. 2024학년도 중학교 배정 시에는 저출산 대책의 하나로 다자녀 혜택 확대, 배정 후 학교 맞교환 절차 도입, 제출 서류 간소화 등 지역별 요구에 맞는 제도개선을 할 수 있었다. 또한, 중학교 배정제도 개선 우수사례를 발굴한 업무 담당자에게 표창을 수여하고, 전국 교육지원청에 우수사례집을 배포해 제도개선 관련 아이디어를 얻을 수 있도록 지원했다.

기관소개(국민권익위원회)

국민권익위원회는 불합리한 행정으로 인한 국민의 권익침해를 신속히 해결하고 부패행위를 효과적으로 예방하기 위해 설립되었다. 또한, 범정부 국민 소통 플랫폼인 국민신문고 시스템을 운영하고 있으며, 축적되는 민원 빅데이터를 민원정보 분석시스템으로 분석하여 민원 예보를 발령하고 각급 기관에 정책개선 사항을 제공하고 있다.



Coffee Time (인터뷰)



현재 하시는 일을 포함해 자기소개 부탁드립니다.

국민권익위원회 민원정보분석과에서 민원 데이터 분석업무를 맡고 있는 김영수 사무관입니다. 언어학을 전공했기에 IT나 데이터 분석은 저에게 생소한 분야였는데요. 권익위에서 주로 민원 정보 분석을 통한 정책·제도개선 업무를 하다 보니 자연스레 데이터 업무를 많이 하게 되었고 분석 역량 역시 쌓이게 되었습니다.

해당 주제가 '해결해야 할 문제'라고 공감했던 이유는 무엇인가요?

교육부가 권익위에 제출한 민원 분석 수요신청서를 검토하고 관련 민원 데이터를 1차 스크린 해본 결과, 중학교 입학 관련 민원이 매년 전국적으로 동일하거나 유사한 형태로 발생하고 있음을 알 수 있었습니다. 그래서, 민원 빅데이터를 분석하고 유의미한 시사점을 찾아내어 중학교 배정 관련 정책을 개선하게 된다면, 학부모들의 자녀 양육 부담이 다소 완화되지 않을까 생각했어요. 또, 학교군 조정으로 인한 집단민원도 해소되고 학생들의 학교생활 만족도도 올라갈 수 있겠다고 판단했습니다.

데이터 수집·정제과정에서 겪은 어려움은 무엇이었나요? 또, 수집한 데이터의 신뢰성 혹은 적합성을 검증하기 위해 어떻게 하셨나요?

'민원정보분석시스템'에 있는 민원은 편지글 형식이라 하나하나 읽어봐야 해요. 예를 들어, 통학환경이 불만이라면 적절한 대중교통 수단이 없어서 그런 건지, 통학로에 공사 현장 등이 있어 위험해서 그런 건지는 분석관이 읽어보지 않으면 판단할 수 없습니다. 민원인의 고충이 정확히 무엇인지 파악해서 그 내용을 바탕으로 비슷한 민원들을 한데 묶어 유형화하는 작업이 가장 어려웠고 시간도 상당히 걸렸던 것 같아요.

중학교 배정 민원 데이터는 실제로 이와 관련된 사람들이 제출한 민원이 쌓인 것이고, 각 교육지원청에서 처리한 결과가 축적된 빅데이터이므로 신뢰성과 적합성이 확보되었다고 판단했어요. 다만, 집단민원이 제출되었을 때는 민원인의 주장과 민원 신청 건수 등 통계만을 보기보다는 언론보도 자료, 감사자료 등을 같이 검색하고 처리기관의 담당자와 면담해서 관련 현황을 객관적으로 파악하기 위해 노력했습니다.

데이터 분석을 위해 어떤 방법을 쓰셨고, 어떤 이해관계자를 참여시켰나요?

권익위의 민원정보분석시스템은 키워드 중심으로 관련 데이터를 검색·추출하도록 설계되어 있습니다. 그래서 중학 배정제도와 관련된 민원들의 키워드 풀을 만들고 그 키워드를 조합해 검색식을 만들고 민원 데이터를 추출했는데요. 관련 데이터를 정제한 후에는 데이터 세트를 확정하고 빈발 민원을 유형화했습니다. 이를 연관 키워드로 일단 그룹화한 뒤, 하나하나 읽어가며 내용을 상세히 확인해가며 정확히 재분류하였습니다.

그다음, 어떤 절차나 규정 때문에 민원이 발생했는지 역으로 다시 분석했습니다.

적극행정 방안을 도출한 뒤에는 교육청 중학 배정 담당자, 교육부 협업담당자와 업무협의를 여러 번 진행하면서 적극행정 방안이 적절한지, 수용할 수 있는지 확인해 나갔습니다.



해당 분석 결과를 공유·확산하기 위해 어떤 노력을 하셨는지 듣고 싶습니다.

교육부는 권익위가 제시한 적극행정 방안을 바탕으로 제도개선 계획을 수립했습니다. 그리고 전국 교육지원청 담당자를 대상으로 권역별 협의회, 워크숍을 개최하고 제도개선 우수사례집 제작해서 배포하는 등 각 지역의 요구에 맞게 중학교 입학 배정제도가 개선될 수 있도록 지원했어요.

그 결과, 교육부는 교육지원청과 협의를 통해 중학교 배정 시 다자녀 혜택 확대, 배정 후 학교 맞교환 절차 도입, 서류제출 간소화, 배정제도 홍보 동영상 제작 등의 제도개선을 이끌어낼 수 있었다고 합니다. 또한, 중학교 입학 배정업무 담당자들은 다른 지역에서는 유사한 민원을 어떻게 처리하고 줄여나가는지 서로의 사례들을 공유하면서 자발적으로 소관 기관의 제도들을 개선하게 되었어요.



데이터 활용을 통한 업무개선을 망설이는 분들께 도움 되는 말씀 부탁드립니다.

데이터 분석은 IT나 통계 관련 전문가들만 하는 것이 아닙니다. 분석 기술도 중요하지만, 데이터를 어떻게 분석해서 활용할 것인가에 대한 '기획'이 더 중요하다고 생각해요. 민원 빅데이터를 활용하게 되면 객관적인 근거자료를 확보할 수 있는데, 빅데이터를 분석한 숫자를 보여주는 것이 이해관계자의 동의를 이끌어내는 핵심 열쇠라고 할 수 있습니다.

저는 그동안 '민원'을 숙제처럼 '처리해야 할 대상'이라고 여기다가 데이터 분석업무를 맡은 후에는 전체 공무원이 공동으로 활용해야 할 '행정자산'이라고 생각이 바뀌게 되었어요. 제가 현장에서 직접 개별 민원을 해결하지는 않지만, 데이터 분석을 통해 민원 발생을 줄이고 정책의 변화를 끌어낸 경험이 있습니다. 꽤 많은 시간과 노력이 필요하기는 하지만, 데이터 분석은 매우 보람 있는 일이라고 말하고 싶네요.

PART 3. 산업경제

PART 3

산업경제

1. 흔들림 없는 천연가스 공급, AI 수요예측 모형으로 가능해졌습니다
   한국가스공사 스마트기술부
2. 정전을 막아라! 기상 데이터와 전력 데이터의 만남
   기상청 기상융합서비스과
3. 우리 경제, 지금 어떤 상태지? 가장 빠른 답은 통계속보에 있다
   통계청 경제동향통계심의관 서비스업동향과

PART 3-1. 산업경제

흔들림 없는 천연가스 공급, AI 수요예측 모형으로 가능해졌습니다

일별 천연가스 수요량 전망모형 구축

한국가스공사 스마트기술부



이렇게 바뀌었습니다!

AS-IS	TO-BE
• 계량경제 기반 단일 수요예측 모형 운용 • 다양한 환경 변화에 대처하기에 역부족	• 신규 AI 기반 모형 도입으로 수요예측 모형 다양화 • 외부 환경 급변 시 다각적 분석을 통한 수급 안정에 기여

추진 배경

한국가스공사(이하 공사)는 매년 천연가스 수요를 예측하여 LNG 도입량을 결정하고 있으며, 이는 안정적인 천연가스 공급에 있어 매우 중요한 사항 중 하나이다. 공사는 적정량의 천연가스를 들여오기 위해 계량경제 기반의 분석모형을 활용하고 있었다.

하지만, 최근 들어 급변하는 에너지 산업구조, 기상 조건들을 반영한 빅데이터 기반 수요예측의 필요성이 대두되었다. 아울러, 체계적인 가스 공급관리를 위한 단기 예측과 안정적인 수입을 위한 장기 예측 또한 필요했다.

따라서, 도시가스 및 발전용 천연가스 공급 물류의 최적화 달성과 수급 안정 도모를 위해 AI 기반의 일별 천연가스 수요량 전망모형을 시범적으로 구축하기로 했다.

활용데이터

연번	데이터명	기준연도	형태	내 용	보유기관
1	공급실적데이터	2013~2021	CSV	천연가스 공급량 실적	한국가스공사
2	기상데이터	2013~2021	CSV	지역별 기온데이터	기상청
3	세대수데이터	2013~2021	CSV	지역별 세대수	통계청
4	에너지원 가격데이터	2013~2021	CSV	경쟁에너지원 상대가격 자료	한국은행
5	전력수요데이터	2016~2021	CSV	에너지원별 전력수요	전력거래소
6	주유 가격데이터	2016~2021	CSV	전국 주유소 판매가격	한국석유공사

기존모형의 문제점 진단 및 개선책 마련

AI에 기반한 신규 수요량 전망모형을 구축하기 위해 우선 내부 요구사항을 수렴한 후, 기존모형인 도시가스 수요예측 모델과 LNG 발전량 예측 모델을 분석했다.

도시가스 수요예측 모델은 5대 권역을 대상으로 시간대별 도시가스 수요를 예측하고자 회귀모형을 적용했으며, 활용 변수는 기온을 사용했다. 해당 모델은 갑작스러운 외부 환경 변화를 반영할 수 있는 변수가 부족했기 때문에 이를 보완할 신규 모델로 학습형 모델을 선정하고, 기온 외 공휴일, 요일 등의 변수를 추가했다. 아울러, 지역 권역을 10개로 세분화하여 예측 정확도를 높이고자 했다.

LNG 발전량 예측 모델은 총발전량에서 석탄, 원자력, 신재생, 직수입 등의 에너지를 제외하여 산출하였으며, 에너지원별로 선형회귀모델을 적용하여 공휴일, 명절 등 일별 특성에 대한 고려 정도가 낮았다. 이를 보완하기 위해서는 에너지원별 외부 변수인 날씨, 전력 사용량, 유류값 등이 반영되어야 하며, 회귀모형을 포함한 학습형 모델 개발해야 했다. 아울러, 공휴일, 명절, 절기에 따른 특성을 반영하여 모델 오차율을 개선해야 했다.

데이터 세분화 및 현황분석

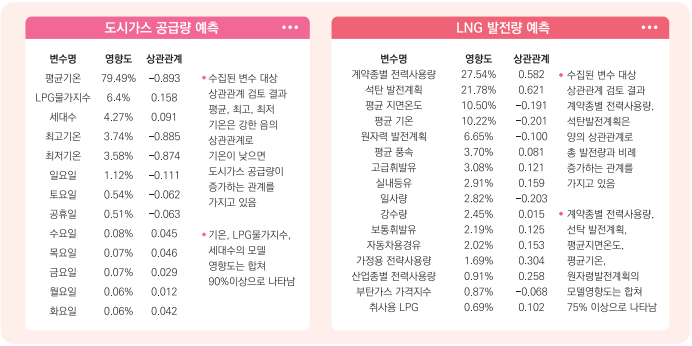
기존모형에 대한 문제점 진단 및 개선 모형 확정한 이후, 도시가스 공급사들의 데이터 결합을 통해 도시가스 관리지역을 10개로 세분화하여 데이터 그룹을 생성했다. 아울러, 한국가스공사에서 도시가스 공급량 데이터 수급(2013~2021) 관련, 도시가스 공급량 현황을 분석함과 동시에 LNG 발전량 현황도 분석했다.

예측 모델 활용 대상 변수 선정

전문가와 관계자 등의 의견을 취합하여 관련 변수를 수집했으며, 상관관계를 분석했다. 그 결과, 도시가스 공급량은 평균, 최고, 최저기온은 강한 음의 상관관계를, 기온이 낮으면 도시가스 공급량은 증가함을 확인했으며, LNG 발전량은 계약종별 전력사용량, 석탄 발전계획은 양의 상관관계로 총발전량과 비례 증가하는 관계임을 확인했다.

이에 따른 모델 영향도는 도시가스 공급량은 기온, LPG 물가지수, 세대수를 합쳐 90% 이상이 도출되었으며, LNG 발전량은 계약종별 전력사용량, 석탄 발전계획, 평균 지면 온도, 평균 기온, 원자력 발전계획을 더하여 75% 이상을 기록했다.

〈그림 32. 예측 모델 활용 대상 변수 선정 〉



최적의 개선 모델 도출

선정된 변수를 바탕으로 도시가스 수요는 분석 기간을 동·하절기 및 연간으로 설정하여 4개 사례로 분석하였으며, LNG 발전량 수요는 LNG 발전량, 총발전량, 에너지원을 바탕으로 5개 사례별로 분석·예측했다.

이러한 분석을 통해 도시가스용 수요예측은 학습형 모델인 랜덤 포레스트(Random Forest)를, LNG 발전량 예측 모델은 프로핏 (Prophet) 모델을 적용하기로 했다.

랜덤 포레스트(Random Forest)는 입력변수를 다양한 조합으로 구성 예측하여 각 조합에서 가장 많이 나온 값을 최종값으로 선정하는 방식인데 여러 모델 중 예측오차율이 가장 우수하며, Prophet(프로핏)은 페이스북(현 메타)에서 공개한 예측 모델로 계절성을 반영할 때 주로 사용하는 방식인데 계절별 발전 스케줄링을 반영한 모델 중 예측오차율이 가장 우수해 선택했다.

〈그림 33. 활용 예측 모델 선별〉

	LSTM (검토대상)	Prophet (검토대상)	SVR	AUTO.ARIMA	SARIMAX	Random Forest (검토대상)	CNN
학습 여부	학습	학습	학습	비학습	비학습	학습	학습
계절성 반영 여부	일부 반영	반영	미반영	반영	반영	미반영	미반영
기존 모델 중복 여부	중복X	중복X	중복X	중복	중복	중복X	중복X

기온 변화 심할 때 기존 모형 대비 정확도 높아

개선 모델을 도출한 직후, 기온 변동이 심했던 기간(2020.6.~2021.5.)에 대한 정확도는 도시가스 수요예측 부분과 LNG 발전량 예측 부분에서 모두 개선됨을 확인했다.

각종 외부환경 변화에 천연자원 수요예측 탄력적 대응 가능

급변하는 국제정세, 친환경 에너지 위주의 산업구조 개편, 기후 변동 등 다양한 외부 변수에 대비하고자 정확도를 올린 분석 모델을 앞으로 현장에서 적극 활용될 예정이며, 정확한 천연가스 수요예측을 통해 안정적인 수급 관리가 될 것으로 기대한다.

기관소개(한국가스공사)

한국가스공사는 청정에너지인 천연가스의 공급을 통한 국민 생활의 편익 증진 및 복리 향상을 위해 1983년에 설립되었다. 주요 업무는 LNG 인수기지와 천연가스 공급배관망을 건설하고, 해외에서 LNG를 수입하여 인수기지에서 재기화한 후 도시가스사와 발전소에 안정적으로 공급하는 것이다.



Coffee Time (인터뷰)



현재 하시는 일을 포함해 자기소개 부탁드립니다.

저는 한국가스공사 디지털혁신추진실 스마트기술부에서 데이터 관련 업무를 맡고 있는 김진영 대리입니다. 오랜 기간 가스설비 및 정보통신 인프라 업무를 담당하다가 데이터 업무를 맡은 지 얼마 되지 않아서 '데이터'를 잘 다루지는 못합니다만, 업무를 담당하며 교육, 검색 등을 통해 관련 지식을 익히고 있습니다. 전자공학 전공이지만 데이터 및 컴퓨터공학 관련 과목을 거의 수강하지 않아 비전공자에 가깝습니다.

이번 과제는 어떻게 참여하게 되셨나요? 기존 일과 어느 정도 관련이 있었나요?

우리 스마트기술부는 클라우드, 빅데이터 등 최신 4차산업 기술을 공사 업무에 적용하거나 추진하는 부서인데, 저는 그중 빅데이터 관련 사업 개념검증, 정보화전략계획 수립을 담당하고 있었기에 기존 업무와도 관련이 있었습니다. 우리 기관에서는 천연가스 수요예측 모델을 빅데이터 기반으로 구축해 보는 게 어떨지 검토하던 중 공공 빅데이터 분석사업을 알게 되었습니다. 천연가스 수요예측 모델 구축 전, 파일럿 모델을 구축하여 실제 어느 정도로 효과가 있을지 검증해 보기 위해 이번 과제에 참여하게 되었습니다.

이번 과제에 참여했던 이해관계자는 누구이며, 중간에 가설이나 방향이 바뀐 적이 있었나요? 그런 경우 어떻게 대처하셨나요?

주요 이해관계자는 천연가스 수요예측 업무를 담당하는 수급계획부였습니다. 새로운 수요예측 모델에 대한 개념검증 작업이 필요하다는 공감대가 형성되어 있었기에 해당 부서에서 적극적으로 참여했습니다. 또, 중앙통제실이나 도입처 등 공사 전반적인 업무에 영향을 미치는 분야다 보니 직접적인 이해관계를 가지고 있지는 않더라도 기관 내에서 관심을 가지고 지켜보는 분들이 많았습니다.

해당 과제를 진행하면서 교수 자문 등 각종 회의를 통해 변수가 미치는 영향을 분석하였고, 회의 결과에 따라 프로젝트의 방향이 처음과는 다소 달라진 부분이 있습니다. 한 번에 완벽한 답을 찾기보다는 사업 기간 내에 현실적으로 실현할 수 있는 부분만 반영하여 추진 속도를 조율했습니다.

과제를 진행하면서 어려움은 없었나요?

공사 보안 문제로 인터넷망과 분리된 업무망 PC에 시범 프로그램을 구축했을 때가 가장 어려웠습니다. 이에 따라 데이터를 수동으로 수집해야 하는 문제가 생겼고, 학습·테스트에 많은 시간이 소요되었습니다. 향후 프로그램의 효과성이 입증된다면 중계 서버를 설치해 오픈 API를 통해 개선할 부분이라 생각합니다.

또, 난해한 소스 코드에 비해 교육·운영자가 참고할 만한 상세한 매뉴얼이 부족하여 시스템 적응에 긴 시간이 걸렸다는 점도 힘들게 다가왔습니다. 앞으로 매뉴얼을 보강하고 계속해서 담당자 교육을 진행해야 한다고 봅니다. 과제 기간이 다소 짧은 것 역시 아쉬움으로 남아

있습니다. 탐색적 데이터 분석, 모델 선정, 프로그램 설치, 사용자 UI 개선, 시운전 및 오차율 테스트를 3~4개월 안에 모두 진행하기에는 시간이 부족했습니다. 하자보수, 오차율 테스트를 추가로 진행하기는 했지만, 사용자 UI 등은 점진적으로 개선해 나갈 계획입니다.



데이터에 전문지식이 없는 실무자라면 프로젝트에 어떤 역할로 참여할 수 있을까요?

저는 가스·전력 등 에너지 수요, 시계열 데이터 분석 분야 교수진 자문회의에 참석해 사업 추진 방향성에 관한 조언을 구하였습니다. 담당자가 데이터 활용에 대한 전문지식이 없다면 처음에 사업을 이해하는 데 어려움을 겪을 수는 있지만, 사업관리가 불가능한 것은 아닙니다. 담당자가 외부 데이터 전문가보다 기관 내부 프로세스를 훨씬 잘 이해하고 있으므로, 데이터에 대한 기초 지식을 습득한 후 현업부서 담당자와 데이터 전문가 사이 가교역할을 하는 게 중요하다고 생각합니다.



데이터를 활용한 이런 과제가 일회성 프로젝트에 그치지 않고 계속해서 활용되려면 어떤 노력을 해야 할까요?

빅데이터·AI 분야는 데이터 수집·분석·테스트·학습 등 일련의 순환과정이 필요합니다. 따라서 상당한 시간이 요구되는데 기관이 즉각적인 성과만 추구하는 분위기라면 절대 장기적인 목표를 이룰 수 없습니다. 사소한 실패와 관계없이 개념검증, 데이터 관리·축적 등 다소 지루한 과정이라도 계속 반복한다면 장기적으로 큰 성과를 거둘 수 있으리라 생각합니다.



끝으로 하고 싶은 말씀 자유롭게 해주세요.

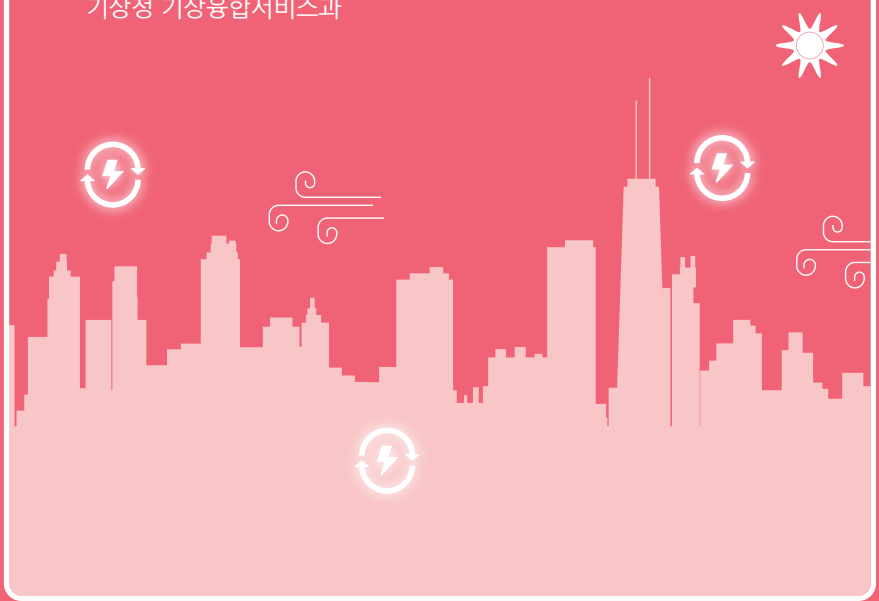
빅데이터·AI 사업은 대규모 자본이 투입되는 만큼 본 사업을 시행하기 전에 개념검증 단계가 필수라고 생각합니다. 행정안전부와 NIA에서 진행하는 각종 데이터 분석사업에 지원함으로써 빅데이터 기술을 사전에 검증해 볼 수 있고, 외부 산학연 전문가들의 질 높은 의견을 수렴할 수 있으므로 적극적으로 참여하시길 권합니다.

PART 3-2. 산업경제

정전을 막아라! 기상 데이터와 전력 데이터의 만남

아파트 정전사고 예방을 위한
전력기상지수 개발

기상청 기상융합서비스과



이렇게 바뀌었습니다!

AS-IS	TO-BE
• 날씨에 따른 전기 사용 증감량 예측 등 참고 지표 부재 • 기온, 습도와 같은 개별 기상자료를 활용한 전력 부하 예측	• 날씨에 따른 전기 사용 증감량 예측을 국민이 쉽게 이해할 수 있도록 지수화 및 시각화 • 전력기상지수를 활용한 48시간 전력 부하 예측으로 업무 효율성 및 정확도 향상

추진 배경

여름철 폭염 기간에 노후 아파트의 정전은 계속 증가하는 추세이다. 특히, 오래된 아파트일수록 세대별 설계 전력량이 낮아 변압기에 과부하가 걸리고 정전으로 곧바로 이어지는 사태가 벌어진다. 여기에 더해 냉난방 가전제품 보급률 증가는 전력 사용량을 매해 늘리고 있는 형편이다.

아파트 전력 사용량은 폭염, 열대야와 같은 기상 요인에 따라 급증하는 특성을 보인다. 그렇다고 해서 기온·습도가 올라갈수록 전력수요가 비례적으로 느는 건 아니다. 오히려 기온·습도는 특정 수치를 기점으로 전력 사용량이 급격히 증가하는 비선형적인 특징을 지니고 있으므로 단순히 기온·습도가 높아졌다고 해서 전력 사용량이 어느 정도 되리라고 예측하기는 어려운 실정이다.

이에 기상청은 기상변화에 따른 아파트 전력수요 예측 정확도를 높이는 한편, 전력 사용량 증감을 국민이 쉽게 인지할 수 있도록 수치화하여 제공하는 전력기상지수 개발을 한국전력공사(이하 한전)와 함께 추진했다.

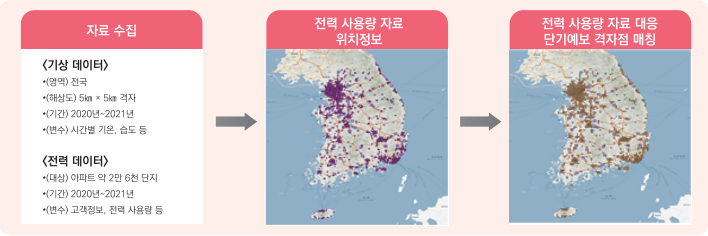
활용데이터

연번	데이터명	기준연도	형태	내 용	보유기관
1	기상 데이터	2020~2021	GRIB	시간별 기온, 습도 등	기상청
2	전력 데이터	2020~2021	CSV	고객정보, 전력 사용량 등	한국전력공사

예보 격자에 전력 데이터를 매칭한 뒤 체감온도를 활용한 모델 개발

기상청은 한전으로부터 최근 2년 치(2020~2021) 전력 데이터를 제공받아 기상청 단기예보[32] 격자에 맞춰 전국 약 2만 6천 개의 아파트 단지 지점들을 격자별로 매칭했다. 특히, 전력 데이터 중 고객정보는 개인정보로서 민감 자료로 분류되어 개인 식별이 되지 않도록 단기예보 1개 격자당(5㎞×5㎞) 아파트 개수가 3개 이상인 정보만 사용할 수 있도록 전처리가 필요했다.

〈그림 34. 단기예보 격자점 매칭 과정〉

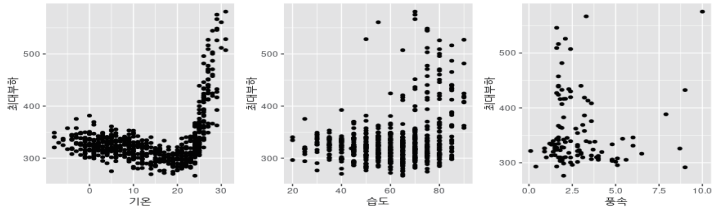


32) 단기예보: 오늘~글피까지 최대 4일 이내의 예보

단기예보 해상도를 기반으로 오늘과 내일 각 시간대의 아파트 전력 사용량 증감을 예측하는 지수를 개발하기 위해 기온, 습도, 풍속, 체감온도 데이터를 사용했다. 이 중 체감온도는 기온, 습도, 풍속의 기본 기상요소들을 이용해 산출되므로 기상요소를 각각 파악하지 않고도 단번에 인식하기 좋고 사람이 느끼는 더위나 추위를 잘 나타내는 요소이다.

기상요소와 아파트의 전력 부하 간에는 비선형적인 관계를 보이는데, 기상청에서는 이러한 시간대별 기상요소와 전력 부하량 간의 비선형적인 관계를 설명하기 위해 3가지의 통계분석모델인 ① 체감온도를 활용한 다항회귀모델[33] ② 기온, 습도, 풍속 요소를 활용한 다중회귀모델[34], ③ 기온, 습도, 풍속 요소를 활용한 랜덤포레스트 모델[35]을 구성했다.

〈그림 35. A 아파트의 기상 상태에 따른 시간대별 부하량 분포〉



아파트 전력 부하의 경우 계절에 따라 변동성과 그 분포가 상당히 다른 특성을 가진다고 볼 수 있으며, 특히 여름(7, 8월)에는 변동성이 상당히 커짐을 알 수 있다. 따라서, 3가지 통계 모델을 활용하여 예측

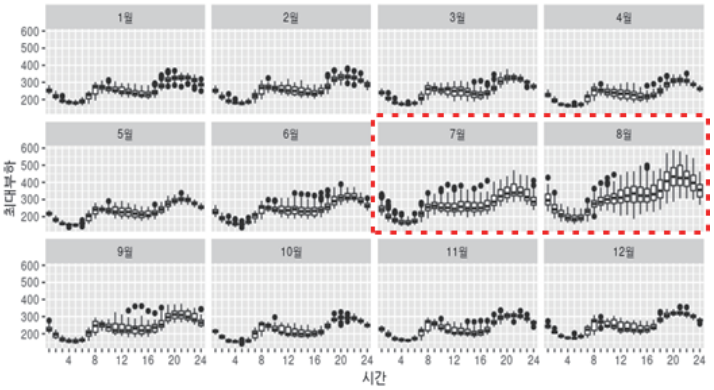
33) 주로 한 개 이상의 독립변수(입력변수)와 종속변수(출력변수) 사이의 관계를 나타내는 다항식을 찾는 데 사용

34) 여러 개의 독립변수가 종속변수에 영향을 미치는 경우에 사용하며, 결과를 예측하기 위해 하나 이상의 독립변수를 사용

35) 여러 개의 의사결정 나무(Decision Tree)를 결합한 앙상블(Ensemble) 모델이며, 각 의사결정 트리는 데이터의 특정 부분을 기반으로 예측을 수행하고, 이러한 예측 결과를 결합하여, 보다 정확한 예측을 제공

모델을 학습하는데 연중 전체 데이터를 사용하지 않고, 다항회귀모델은 동일 계절(하계/동계/춘·추계) 데이터, 다중회귀모델과 랜덤 포레스트 모델은 해당일 전후 ±15일 기간의 과거 데이터를 추출하여 학습에 활용했다.

〈그림 36. A 아파트의 월별 시간대에 따른 아파트 최대부하 분포〉



3가지 통계분석모델 중 성능 검증을 통해 가장 우수한 결과를 보인 것은 다항회귀모델이었는데, 전력기상지수 개발과 제공에 대한 전체 흐름도는 아래의 그림과 같다.

〈그림 37. 전력기상지수 개발 과정〉



전력기상지수 성능 검증, 대규모 행정단위일수록 상관도 증가

기상청은 다항회귀모델을 바탕으로 단기예보 해상도의 오늘과 내일(48시간) 시간대별 아파트 전력 사용량 증감을 예측하는 '전력기상지수'를 개발했다. '전력기상지수'란 해당 지역 아파트 등 공동주택의 연중 평균 전력 부하를 100으로 했을 때 기상예보를 기반으로 해당 시각의 예상되는 평균부하를 상대적인 수치로 표현한 지수이다.

예를 들어 A 지역의 전일 18시~19시 전력기상지수가 120이었고, 오늘 18시~19시 전력기상지수가 150이라고 한다면, 오늘 해당 시간의 최대 부하는 어제 대비 평균적으로 150/120=1.25배 증가한다고 예상할 수 있다. 이처럼 해당 지수의 전일(혹은 전주) 대비 시간대별 증감량을 활용해 전력 부하 증감을 예상할 수 있다.

한편, 전력기상지수의 성능을 검증하기 위해 2022년 7~9월 고압 아파트가 5개 이상 존재하는 지역에 대해 행정구역의 전력기상지수 예측자료와 2022년 실제 전력부하 자료 간 관계를 상세히 분석했다. 그 결과, 도·특별시·광역시처럼 행정구역 규모가 클수록 상관도가 높게 나타났다.

〈표 5. 규모에 따른 전력기상지수와 실제 전력 부하 자료 간 평균 상관계수[36](한전 제공)〉

구분	도/특별시/광역시	시/군/구	읍/면/동	5km 격자별
다항회귀모델	0.947	0.919	0.782	0.940

날씨와 연계한 전력 부하 정보 모바일 서비스 후 정전 사고 예방에 도움

기상청은 2022년 6월부터 전력기상지수를 API로 한전에 제공하고 있고, 한전은 전력기상지수를 포함한 전력 기상정보, 그리고

36) 두 변수 간 관계의 강도와 방향(양(+)의 상관, 음(-)의 상관)을 나타내며, 양의 상관계수는 한 변수가 증가하면 다른 변수도 증가함을 나타내며 1에 가까울수록 상관성이 높음.

전력기상지수와 개별 아파트 전력 계량데이터를 융합·분석해 48시간의 예상 최대 전력 수요정보와 부하 정보를 산출하여 2022년 7월부터 모바일 서비스(check.kepco.co.kr)로 제공하기 시작했다.

아파트 전기안전 관리자와 관리사무소는 이 정보를 활용해 해당 아파트의 최대 전력 사용 시간대와 사용량을 예측할 수 있게 되었다. 정전 사고 예방에 도움이 되었음은 물론 불가피하게 정전되었을 때는 신속한 대책 수립도 가능해졌다.

〈그림 38. 모바일 서비스 제공 화면〉



기관소개(기상청)

기상청은 국가 기상업무를 관장하고 있는 중앙행정기관으로서 1949년 설립됐다. 또한, 농수축산업, 환경, 국방, 스포츠, 에너지 등 사회 각 분야에 대한 기상지원과 융합을 통해 시너지를 극대화하고 있다.



Coffee Time (인터뷰)



현재 하시는 일을 포함해 자기소개 부탁드립니다.

저는 기상청 기상융합서비스과에서 기상기후 빅데이터를 활용한 기상융합 서비스 개발과 제공 업무를 담당하고 정선애 사무관입니다. 데이터 분석과 관련된 업무는 이번이 처음이지만, 예전에 기후자료 관리 업무를 하면서 기온, 강수량과 같은 기상통계자료를 다뤄본 적은 있었어요.

이번 과제는 어떻게 참여하게 되셨나요? 기존 일과 어느 정도 관련이 있었나요?

저는 기상과 타 분야 데이터를 융합해 수요자 맞춤형 기상정보를 제공하는 '기상기후 빅데이터 융합 서비스' 개발을 맡고 있으며 이번 과제도 그에 해당합니다. 우리 과에서는 매년 수요조사를 통해 신규 과제를 발굴하고 있는데 한전에서 날씨에 따른 아파트별 전력 사용량 증감을 예측하는 지수(전력기상지수) 개발 과제를 먼저 제안했어요. 활용도가 높고 구현 가능성도 있어 한전과 함께 과제를 시작하게 되었죠.



데이터는 어떤 방식으로 수집하셨나요?

이번 과제에 활용한 데이터는 기상청의 단기예보 자료(전국 5km×5km 해상도)와 한전의 전국 약 2만 6천 아파트 단지의 전력 사용량 자료입니다.

한전에서 보유한 아파트 단지의 위치정보, 계약 전력량 등은 개인정보에 해당합니다. 그래서 이 개인정보를 식별할 수 없도록 반드시 전처리가 필요했기에, 초기에는 한전에서 직접 처리 후 기상청으로 제공했어요. 하지만 전력기상지수의 원활한 개발과 검증을 위해서 개별 아파트 정보가 계속 필요했기에 기상청은 공동개발을 목적으로 한전과 업무협약을 체결하였으며, 이후 한전으로부터 아파트 정보를 공유받아 직접 처리하게 되었습니다.



데이터 분석을 위해 어떤 방법을 쓰셨나요? 전문가나 관계자를 참여시켰나요?

한전은 먼저 기상요소와 전력 사용량 간 관계를 분석해 본 경험이 있었어요. 그래서 한전은 기온·습도와 같은 개별 기상요소를 사용한 예측 모델과 체감온도를 사용한 예측 모델을 개발해 그 결과를 비교한 다음 정확도가 높은 쪽을 채택하자는 의견을 제시했죠. 기상요소와 전력 부하는 비선형적인 관계를 보이는데, 이러한 관계를 설명하기 위해 다항 회귀 모델, 다중회귀모델, 랜덤포레스트모델 3가지 통계분석모델을 활용하여 분석해 보고, 우수한 결과가 나오는 모델을 선택하도록 진행했습니다.

한전에서는 분석전문가가 참여해 함께 과제를 적극적으로 끌어갔으며, 기상청 담당 부서의 분석경험자, 기상청 융합 서비스 개발업체의 분석전문가도 함께 수시로 의견을 교환하며 알고리즘을 완성해 나갔어요.

예측 서비스 이후 모든 모델에서 공통으로 특정 시간대(오전 7~8시)에 상관성이 떨어지는 현상이 발생했는데, 이는 기상 상황과 관계없이 매일 아침 아파트 주민들이 외부 활동(출근, 등교 등) 준비를 위해 기본적으로 일정량의 전력을 사용하기 때문으로 판단되었으며, 이 부분은 추가 분석을 거쳐 2023년 9월 개선을 완료했습니다.



개발 모델을 공유, 확산하기 위해 어떤 노력을 하셨는지 듣고 싶습니다.

처음 기획 단계부터 전력기상지수가 개발되면 기상청은 한전에 API 형태로 실시간 제공하고, 한전에서는 '파워체크' 앱과 누리집을 통해 서비스하기로 정해놓은 상태였습니다.

한전의 '파워체크' 앱은 아파트 전기안전관리자와 관리사무소 등에서 회원으로 가입하여 사용하고 있는데요. 시범서비스 이후 한전에서 전기 관련 공공기관과 민간 전기회사에서 전력기상지수를 활용하고 싶다는 연락을 받은 적이 있다고 들었어요.

그래서 2023년에는 전력기상지수를 아파트 밀집도 등을 고려하여 도시 규모별로 모델을 상세화하여 정확도를 개선하고, 행정안전부 공공데이터포털(data.go.kr)과 기상청 API 허브(apihub.kma.go.kr)를 통해 데이터셋과 API를 9월 26일부터 국민에게 제공하고 있습니다.



끝으로 하고 싶은 말씀 있으시면 한마디 부탁드립니다.

단년도에 알고리즘과 서비스 개발까지 추진하는 것도 가능하겠지만, 개발된 융합 정보와 서비스가 현장에서 쓰이고 확산하기 위해서는 지속적인 개선과 검증 작업이 이어져야 한다고 생각합니다. 그리고 기상 데이터는 다양한 현안을 해결하는 데 최적화된 형태로 정리되어 있는 자료라고 생각해요. 기상 데이터를 사회 전반의 여러 분야와 융합해 활용한다면 더 많은 분께 도움이 될 것으로 기대합니다.

PART 3-3. 산업경제

우리 경제, 지금 어떤 상태지? 가장 빠른 답은 통계속보에 있다

신용카드 승인액 동향 분석

통계청 경제동향통계심의관 서비스업동향과



이렇게 바뀌었습니다!

AS-IS	TO-BE
• 서비스업, 소매판매 당월 동향 관련 통계자료를 익월말 공표	• 서비스업, 소매판매 당월 경제 동향 속보치를 익월 초·중반에 작성·공표
• 감염병 확산 등 긴급한 경제 위기에 신속한 대응 곤란	• 빠른 자료 분석 및 전파로 긴급한 경제 위기에 대응 가능

추진 배경

통계청은 경제의 단기적 양상을 파악하기 위해 '경제동향 통계'를 매월 작성하고 있다. 그중 각 월의 제조업, 서비스업, 소비 등을 포괄하는 산업활동동향은 익월 말 통계청 보도자료를 통해 공표한다. 하지만 실제 경제 상황과 해당 기간 통계자료 공표 사이에 약 1개월의 시차가 발생함에 따라 경제 현상에 신속히 대응하기 위한 속보 지표의 필요성이 꾸준히 제기되어 왔다.

특히 2020년 1월 이후 유행하기 시작한 코로나19의 부정적 충격은 국내 산업 전반에 큰 영향을 미쳤다. 감염병 확산 초기에 대면 접촉이 많은 서비스업 등 소비 분야에 경제적 피해가 단기간에 집중적으로 나타났는데, 이에 관한 통계지표 중 작성 주기가 월(月)보다 짧은 속보성 지표는 없는 상황이었다. 그래서 통계청은 현실을 반영한 시의성 높은 자료를 정책부처에 적시 제공하기 위하여 '카드 승인액 동향' 분석을 추진하기로 했다.

활용데이터

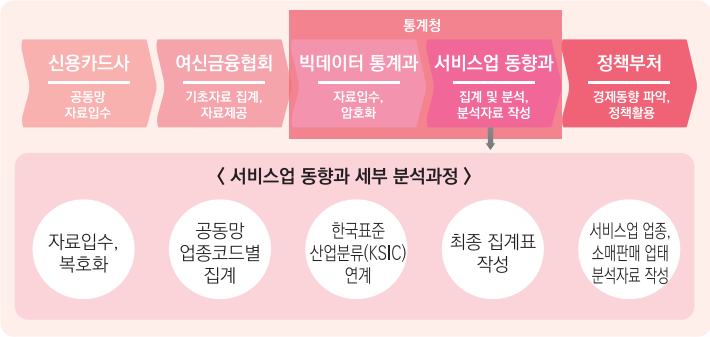
연번	데이터명	기준연도	형태	내 용	보유기관
1	신용카드가맹점 승인실적	2020~2022	xlsx	날짜, 지역, 업종코드, 매출금액 등	여신금융협회

신용카드 승인액 자료 업종·업태별 분류 후 분석

카드 승인액 동향 작성의 기초자료는 국내에서 영업 중인 8개 신용카드사의 주간 가맹점 승인 실적(이하 카드 승인액 자료)이다. 이는 해당 카드사 간에 공유되는 공동망을 통해 여신금융협회로 입수되는데, 통계청은 여신금융협회와 MOU를 체결하여 카드 승인액 자료를 주(週) 단위로 제공받았다.

카드 승인액 자료는 승인 일자, 공동망 업종코드, 지역, 매출액 등으로 구성되어 있다. 통계청은 카드 승인액 자료에 포함된 공동망 업종코드를 서비스업 8개 업종[37]및 소매 판매 9개 업태[38]로 분류하여 서로 연계한 후 카드 승인액의 추이, 변동률 등을 업종·업태별로 구분하여 분석했다.

〈그림 39. 분석과정〉



37) 음식·주점, 숙박업, 문화·여가, 개인서비스, 운수·창고, 보건·복지, 교육, 기타
38) 백화점, 대형마트, 슈퍼마켓, 편의점, 국산차, 연료소매, 전문소매, 인터넷쇼핑, 기타

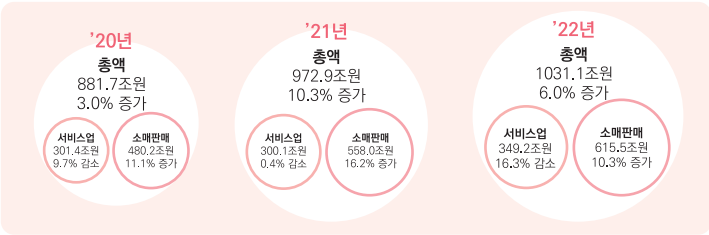
여신금융협회에서 받은 카드 승인액 자료를 더 살펴보면, 승인 일자, 시·도 구분, 카드 구분(신용/체크), 가맹점 구분(개인/법인), 공동망 업종코드, 매출 금액 정보가 포함되어 있으며, 자료 입수 부서(빅데이터통계과)는 6개 항목 값을 암호화한다.

담당 부서(서비스업동향과) 해당 자료를 공동망 업종코드 기준으로 집계한 다음, 한국표준산업분류를 토대로 서비스업·도소매업 업종과 매칭한 테이블을 활용해 업종·업태별 카드 승인액으로 재집계했다. 이때 확인된 값과 도소매업 통계작성 목적으로 입수한 타 행정자료, 소비자물가동향, 기작성한 당해 연도 계절조정인자 등을 참고해 서비스업·소매판매 하위 분류에 대한 분석자료를 작성했다.

분석자료와 최근 3년간(20~22년) 경제 동향 연계 확인

전체 카드 승인액 중 서비스업 및 소매판매에 해당하는 금액은 2020년 781.6조 원, 2021년 858.1조 원, 2022년 964.7조 원이었으며, 해당 금액의 전년 대비 증가율은 2020년 2.0%, 2021년 9.8%, 2022년 12.4%로 나타났다.

〈그림 40. 연도·부문별 카드 승인액 전년 대비 증감률('20–'22년)〉



코로나19 유행이 시작된 2020년의 전체 카드 승인액은 전년 대비 증가했지만, 하위 분류로 들어가면 소매판매는 전년 대비 약 11% 늘어난 데 비해 서비스업은 9.7% 감소한 것으로 나왔다. 이는 코로나19 유행에도 불구하고 전반적인 소비성향은 위축되지 않았으나 감염에 대한 우려, 사회적 거리 두기 등이 서비스업에 심리적·제도적 악재로 작용했음을 보여준다. 항공 여객 운송업, 음식점·주점업, 숙박업, 문화·여가 관련 서비스업, 개인서비스업 등 대면 서비스업종 소비 부진은 곧 서비스에서 재화로 소비가 이동했음을 드러내는 것이다. 이러한 소비의 이동은 2020년 상반기보다 하반기에 더 두드러지는 경향을 보였다.

2021년 또한 코로나19에 대한 심리적·제도적 제약이 완화되지 않아 서비스업 카드 승인액은 2020년의 악화한 수준과 비슷한 규모를 보였다. 반면 소매판매는 코로나19 유행 장기화에 대응해 가전제품 등 내구재를 교체하는 이른바 '보복소비'등의 효과로 전년 대비 약 16%가 증가했다. 이후 2022년에는 2분기 이후 단계적 일상 회복 등 방역 완화에 힘입어 서비스업 카드 승인액이 전년 대비 약 16%로 대폭 커졌고 소매판매 역시 약 10% 늘어나 전체 카드 승인액이 2021년보다 큰 값을 기록하였다.

급변하는 경제상황에 신속 대응, 정책 추진 기초자료 활용 기대

위와 같은 과정을 거쳐 작성된 카드 승인액 분석 결과는 매월 10일경 기획재정부, 중소벤처기업부 등 경제정책 유관 부처에 매주 또는 매월 제공하여 급변하는 경제 상황에 신속하게 대응할 수 있도록 지원했다. 아울러 해당 내용은 기획재정부에서 매월 중순 작성하는 「최근 경제동향」의 전망치에 일부 반영되어 정부·기업 등 경제주체의 의사결정 등을 돕고 있다. 또한, 담당 부서(서비스업동향과)에서 매월 물가 및 계절 요인 등을 고려해 작성하는 소매판매 및 음식숙박업 속보지표 작성의 참고 자료로 활용되어 정확한 경제 동향 예측에 직·간접적으로 기여하고 있다.

카드 승인액 자료 분석은 코로나19 발생 이후 속보성 통계지표를 신속하고 안정적으로 제공할 수 있는 기반을 마련하여, 향후 감염병 확산 등 긴급한 경제 위기 상황 극복을 위한 조기 대응 및 정책 추진의 기초자료로 유용하게 활용되리라 기대한다. 이 과정에서 통계청의 기초자료 분석 능력 향상 및 속보성 지표 작성의 노하우가 축적됨에 따라 타 부서 및 타 기관에 긍정적인 영향을 미칠 수 있을 것이다.

기관소개(통계청)

통계청은 1948년에 공보처 통계국으로 처음 신설되었으며, 1990년에 통계청으로 승격·개편되었다. 주요 업무는 통계의 기준설정과 인구조사 및 각종 통계 관련 사무이다. 아울러, 국민의 합리적인 의사결정을 위해, 명확한 통계와 데이터를 제공하고 있다.

통계청

Coffee Time (인터뷰)



현재 하시는 일을 포함해 자기소개 부탁드립니다.

통계청 서비스업동향과에서 서비스업 동향 조사를 담당하고 있는 최익서 주무관입니다. 통계학을 전공했지만 데이터 분석의 기회는 없었어요. 통계청 입사 후, 관련 업무를 수행하며 데이터의 가치와 처리기술의 중요성을 알게 되었고, 현재 배워나가는 단계입니다.

데이터 활용 시 가장 중요한 건 '정확한 문제정의'라고 합니다. 이번 과제를 주제로 선정했던 이유에 관해 조금 더 얘기를 들어볼 수 있을까요?

경제정책 수립과 경제적 지원이 필요한 대상을 선정할 때는 객관적이고 시의성 있는 '좋은 자료'가 필요합니다. 하지만 좋은 자료를 찾기가 쉬운 건 아니죠. 저희도 최대한 '좋은 자료'를 만들려고 방법을 찾다가 우리나라 국민의 높은 신용/체크카드 이용률(70% 육박)에서 착안하게 되었어요. 데이터는 앞서 밝혔듯 여신금융협회와 MOU를 체결하여 확보했습니다.

암호화한 자료를 받은 후 여신금융협회 기준으로 분류된 데이터와 표준산업분류의 연계, 카드 승인액과 소비 동향의 상관관계 등 자료 활용의 적절성을 평가하기 위해 학계, 연구계(서울, 부산, 인천 등 지역별로 설치되어 있는 지역연구원), 경제계 등으로 구성된 전문가 협의회를 구성하였고, 여러 차례 회의를 통해 카드 승인액 자료를 검토했습니다.

이번 과제를 추진하면서 겪었던 어려움에 관해 말씀해주세요.

여신금융협회 업종분류와 통계청 표준산업분류의 매칭 작업이 까다로웠습니다. 분류 기준이 달라 1:1 매칭이 힘들었는데 두 자료의 시계열 분석, 실무자들의 의견과 외부 자문을 종합해 업종연계표를 완성했습니다. 앞으로 공공기관의 행정자료를 제대로 활용하기 위해서는 업종분류의 표준화가 필요하다고 느낄 수 있었죠. 그리고 공식 통계보다 한발 앞서 정보가 제공된다는 점에서 참고 자료가 없는 부분이 조심스럽고 부담되기도 했습니다.

데이터 분석 시 특별히 중점을 뒀던 부분이 있나요?

신용카드 승인액 동향 분석에 사용하는 자료는 국내에서 카드로 결제되는 대부분(외국계 등 제외)의 사용 내역이 들어있습니다. 포괄 범위가 넓고 신뢰성 있는 자료라고 할 수 있죠. 그렇다고 해도 데이터 분석을 해보면 전혀 예상치 못한 결과가 나와서 검토 또는 분석 방법에서 보완이 필요한 경우가 종종 생깁니다.

데이터 자체에 자신이 있다고 해도 정확성 확인은 꼭 필요합니다. 저희는 확보한 데이터의 정확도를 위해 서비스업 동향 조사의 '소매 판매 통계'를 비교 대상으로 삼았어요. 과거 카드 승인액 데이터로 작성한 자료를 동일시기의 소매 판매 동향 자료와 비교하여 정확성을 검증했습니다.



해당 과제를 구체적으로 활용하고 공유하기 위해 어떤 노력을 하셨나요? 또, 데이터에 기반한 이런 프로젝트가 일회성으로 그치지 않으려면 앞으로 어떻게 해야 할까요?

이 과제는 국내 소비경향과 경기 상황을 이전보다 신속하게 파악하길 원하는 정책기관의 관심에서 시작되었기에 기획 전부터 활용성이 어느 정도 보장되어 있었어요.

이런 프로젝트가 일회성 이벤트로 종료되지 않으려면, 설득력 있는 데이터 분석 사례를 보여주는 게 중요하다고 생각합니다. 일단 비교적 접근하기 쉬운 과거 데이터를 분석한 다음 그와 관련된 이벤트를 매칭해 사전에 대비할 수 있었다는 것을 강조한다면, 데이터 분석이 중요하다는 인식이 자연스레 생겨나지 않을까요.

양질의 데이터 분석이 계속 이뤄지고 자료가 축적되면 자연스럽게 데이터 활용 문화가 확산할 것이라고 생각합니다. 이를 위해서는 데이터 분석 대회, 공모전 등이 꾸준하게 개최되어야겠죠. 또한, 의사결정 과정에서 데이터 활용을 명시해 관계자에게 데이터에 관한 인식을 꾸준히 높이는 게 중요할 것 같습니다.



데이터 활용을 통해 업무개선을 하고 싶으나 선뜻 나서지 못하고 주저하는 분들께 도움 되는 말씀 부탁드립니다.

데이터 활용이라 하면 데이터 수집, 통계 프로그래밍, 시각화 등등을 떠올리며 막연히 어렵다고 생각하기 쉽습니다. 하지만 막상 해보면 거창한 방법보다 심플한 분석이 의미 있는 결과를 보이는 사례도 많죠. 무겁게 발을 들이기에 앞서 누가 봐도 이해하기 쉽고 직관적인 분석을 먼저 해보는 게 어떨지 싶습니다.

PART 4. 관광활성화

PART 4

관광활성화

1. 지역관광을 부흥시킬 지름길, 여기 있습니다!
한국관광공사 관광컨설팅팀

2. 내 마음에 쏙 든 철도여행, 알고 보니 빅데이터 덕분이었다
한국철도공사 미래기술처

PART 4-1. 관광활성화

지역관광을 부흥시킬 지름길, 여기 있습니다!

빅데이터 기반 지역관광 활성화 지수 개발

한국관광공사 관광컨설팅팀



이렇게 바뀌었습니다!

AS-IS	TO-BE
• 한국관광데이터랩에서 제공하는 개별 데이터만 있고, 관광정책에 활용할 수 있는 대표 지수 부재 • 지역별 관광 현황 파악을 위한 데이터 선정과 지자체별 비교 및 해석이 어려움	• 한국관광데이터랩에 있는 데이터를 기반으로 관광정책에 활용할 수 있는 관광 활성화 지수 개발 • 관광 활성화 지수를 통해 지역별 강·약점 확인이 가능하며, 이를 통해 실질적인 관광정책 수립 가능

추진 배경

2021년 2월, 한국관광공사(이하 공사)는 관광 특화 빅데이터 플랫폼인 '한국관광 데이터랩(datalab.visitkorea.or.kr)'을 오픈해 관광 관련 빅데이터 및 융합분석 서비스를 제공하기 시작했다. 이에 따라 데이터에 기반해 관광정책을 수립하거나 관광 비즈니스를 진행하고자 하는 기업, 지방자치단체, 업계, 학계 등은 자유롭게 관광 빅데이터를 이용할 수 있게 되었다.

그런데 한국관광 데이터랩의 운영 과정에서 아쉬운 점이 생겼다. 이용자들이 이동통신, 신용카드, 내비게이션, 관광통계 등 다양한 데이터에 접근할 수는 있지만, 때때로 광범위한 데이터 중 어떤 것을 선정해 분석해야 할지 어려움을 겪게 되는 경우가 많았다. 성격도 중요도도 각기 다른 데이터를 어떠한 논리로 해석해야 할지에 대한 고민 또한 적지 않았다.

특히 지자체들은 지역관광 활성화와 관련한 데이터 사용에 어려움이 있었으며, 공사는 이러한 지자체의 관광정책 수립에 도움을 주고자 시기별, 지역별 관광 활성화 정도를 복합적으로 표현할 수 있는 단일 지표와 이에 영향을 주는 세부 지표로 구성된 관광 활성화 지수[39]를 개발하게 되었다.

39) 방문객의 숙박, 외식, 쇼핑, 관광서비스 이용 등 관광 관련 활동을 통해 관광산업이 원활하게 유지되는 정도 및 성과로 이어지는 수준으로 정의

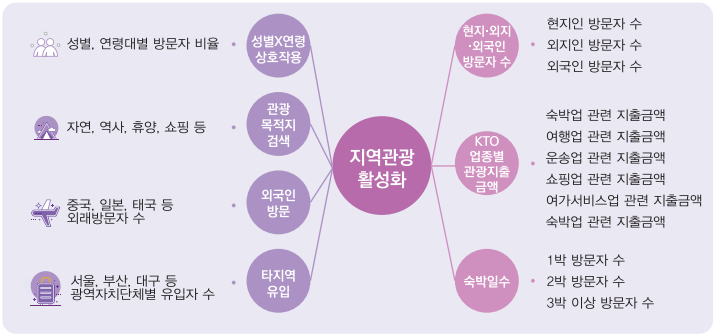
활용데이터

연번	데이터명	기준연도	형태	내 용	보유기관
1	성별/연령별	2019~2021	csv	성별/연령별 내국인 방문객 수	SKT
2	관광목적지	2019~2021	csv	Tmap 관광지 검색 건수	SKT
3	방한 외국인 수	2019~2021	csv	주요 방한 외국인 방문자 수	SKT
4	숙박일수	2019~2021	csv	내국인/외국인 숙박일 수 합산	SKT
5	방문자 수	2019~2021	csv	현지·외지·외국인 방문자 수	KT
6	유입 방문자 수	2019~2021	csv	시군구별 유입 방문자 수	KT
7	시군구간 거리	2020,2021	xlsx	국가교통DB 도로망 데이터와 카카오API 행정구역 위치정보에 기반한 거리 데이터	국가교통DB 카카오맵API
8	관광 지출	2019~2021	csv	KTO 업종별 관광 지출 금액 – 현지·외지·외국인 소비 금액 합산	신한카드

관광 활성화 지수 산출을 위한 중요 변수 탐색과 분석

관광 활성화 지수 산출은 이미 2021년에 시도했다. 당시에는 강도·다양성·거리의 3개 지수 및 세부 요소로 관광 활성화 정도를 확인했지만, 여러 한계점이 있어 좀 더 고도화된 분석 모델이 필요했다. 마침 2022년 신규 데이터가 쌓였으며, 해당 데이터를 기존모형과 더하여 새로운 분석 모델을 구축하기로 했다.

〈그림 41. 최종 연구모형 요인별 변수 구성〉



기존 모델에서는 KT 유동인구, BC카드, Tmap 내비게이션 데이터를 활용했는데, 새로운 모델에는 SKT 유동인구와 신한카드 데이터를 추가로 적용해 분석을 진행했다.

먼저, 선행연구 및 보고서 등을 조사해 관광 활성화 정도에 영향력이 큰 변수를 1차로 선정했다. 그리고 한국관광데이터랩에 적재된 데이터 중에서 해당 변수를 설명할 수 있는 데이터를 모아 데이터 세트를 최종 선정했다.

선정된 데이터는 분석에 맞게 전처리한 후 탐색적 요인분석과 신뢰성 검사, 자문회의를 거쳐 적합성을 검증했으며, 지역관광 활성화 단일지수를 산출할 때 필요한 하위 요인들의 가중치를 도출하기 위해 고차 확인적 요인분석[40]을 실시했다. 분석 결과, 각 요인이 관광 활성화에 미치는 영향력은 이론적 검토를 통해 설정한 가설과 일치하여 분석을 진행하기에 적합하다고 판단되었고, 이렇게 도출된 가중치를 기반으로 지역관광 활성화 지수를 산출했다.

지역관광 활성화 지수 및 관광 활성화 지수의 세부 지표 값은 누적분포함수[41]를 적용하여 산출되었으며, 250개 시군구 중에서 조회한 지역의 지수 및 지표 값의 상대적 위치를 나타낸다(범위 : 0이상 100이하).

〈그림 42. 관광활성화 지수 산출 방법〉

$$\text{관광활성화 지수} = P[Z < z_{FAC}] \times 100 = [\int_{-\infty}^{z_{FAC}} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{Z^2}{2}\right) dZ] \times 100$$

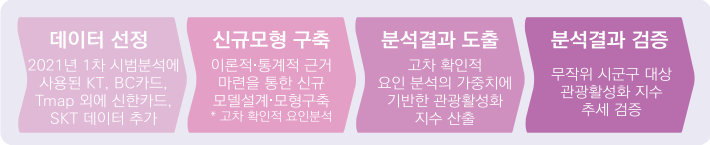
$$FAC = (\sum_{i=1}^{14}((\frac{X_{Ai}-\overline{X_{Ai}}}{S_{Ai}})\times W1_{Ai})\times W2_A) + (\sum_{i=1}^{9}((\frac{X_{Bi}-\overline{X_{Bi}}}{S_{Bi}})\times W1_{Bi})\times W2_B)$$
$$+(\sum_{i=1}^{12}((\frac{X_{Ci}-\overline{X_{Ci}}}{S_{Ci}})\times W1_{Ci})\times W2_C) + (\sum_{i=1}^{17}((\frac{X_{Di}-\overline{X_{Di}}}{S_{Di}})\times W1_{Di})\times W2_D)$$
$$+(\sum_{i=1}^{3}((\frac{X_{Hi}-\overline{X_{Hi}}}{S_{Hi}})\times W1_{Hi})\times W2_H) + (\sum_{i=1}^{6}((\frac{X_{Ki}-\overline{X_{Ki}}}{S_{Ki}})\times W1_{Ki})\times W2_K)$$
$$+(\sum_{i=1}^{3}((\frac{X_{Mi}-\overline{X_{Mi}}}{S_{Mi}})\times W1_{Mi})\times W2_M) + ((\frac{X_T-\overline{X_T}}{S_T})\times W2_T) + ((\frac{X_Q-\overline{X_Q}}{S_Q})\times W2_Q)$$

A = 성별/연령 B = 관광목적지 C = 방한외국인수 D = 타지역유입 H = 방문자 수 K = 지출금액 M = 숙박일수
T = 성수기여부 Q = 코로나19여부 $\overline{X}$ = 평균 S = 표준편차 W1 = 하위요인 가중치 W2 = 상위요인 가중치

40) 구성개념이 직접적으로 관찰되거나 측정이 되지 않는 변수인 잠재 변수를 다층적으로 살펴보는 것
41) 주어진 확률 변수가 특정한 값보다 작거나 같을 확률을 나타내는 함수

또한, 관광 활성화 지수에 시기별 주요 사건 등이 적절히 반영되었는지 확인하기 위해 무작위로 시군구를 추출했다. 그 시군구를 대상으로 추세를 검증한 결과, 시기별 실제 관광 현황과 일치하는 것을 확인할 수 있었다.

〈그림 43. 분석과정〉

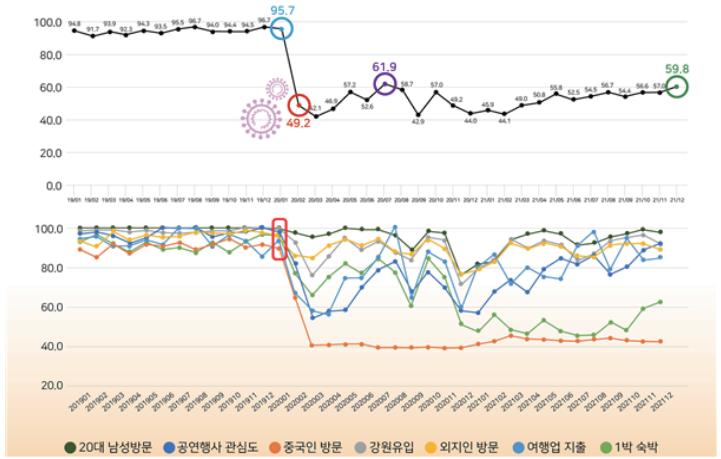


지역관광 실태 진단과 정책 방향 제시

이렇게 산출된 지역관광 활성화 지수는 주요 요인 간 비교를 통해 지자체별 관광 활성화 수준을 진단하고 데이터에 기반한 관광정책을 수립하는 데 활용된다.

예를 들어, 2020년 1월 관광 활성화 지수가 95.7인 A 시군구는 타 시군구와 비교해 상대적으로 관광 활성화 수준이 높은 것으로 나타났다. 좀 더 세부적으로 살펴보면 '20대 남성 방문' 항목 값이 99.8로 상대적 수준이 매우 높아, A 시군구의 관광 활성화 지수에 긍정적인 영향을 크게 준 것으로 해석할 수 있다. 그러나 '자연관광 관심도' 항목 값은 34.6으로 관광 활성화 지수에 비해 상대적으로 낮은 수준을 보임에 따라 자연관광 콘텐츠를 개선하고 홍보에 힘을 써야 함을 알 수 있다.

〈그림 44. A시의 3년간 관광활성화 지수 월별 변화 예시(2019.01~2021.12)〉



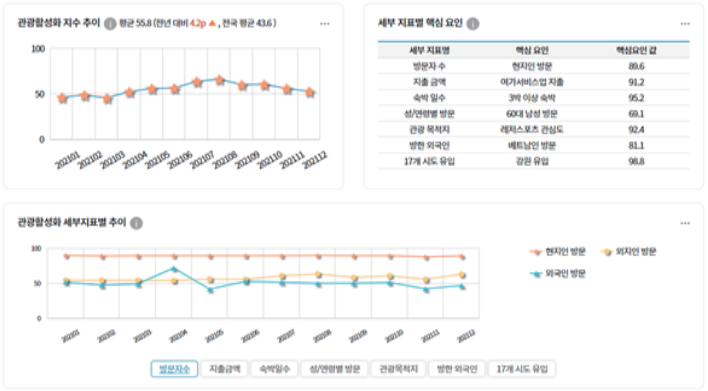
* 출처 : 한국관광공사, 빅데이터 기반 지역관광 활성화 수준진단 보고서(2022)

관광 활성화의 길잡이

지역관광 활성화 지수는 기초지자체가 지역의 관광 활성화 수준을 월별로 파악하고, 64개 주요 관광 요인을 타 지자체와 비교하여 상대적으로 우위 혹은 열위에 있는 요인을 찾아 지역관광 활성화 방아 도출에 활용할 수 있다.

'한국관광데이터랩'에서 시각화 서비스로 제공되고 있는 관광 활성화 지수는 향후 빅데이터 및 지역 인프라 데이터 기반 분석모형의 고도화를 통해 더욱 효율적으로 지역관광 활성화 지원을 할 수 있을 것으로 보인다.

〈그림 45. 관광활성화 지수 시각화 예시〉



* 출처 : 한국관광데이터랩

기관소개(한국관광공사)

한국관광공사는 1962년에 제정한 국제관광공사법(현 한국관광공사법)에 근거하여 설립되었다. 주요 업무로는 관광 진흥, 관광자원 개발, 관광산업 연구, 개발 및 관광 전문인력 양성 등이 있다. 또한, 관광 빅데이터 플랫폼 구축 및 강화, 관광 빅데이터 품질 제고 및 이용 활성화, 디지털 마케팅 지원, 데이터 기반 종합 관광 컨설팅 제공 등의 업무를 함께 수행함으로써 관광생태계의 디지털 전환을 선도하고 있다.



Coffee Time (인터뷰)



현재 하시는 일을 포함해 자기소개 부탁드립니다.

한국관광공사 관광컨설팅팀에서 관광 빅데이터 기반 내외부 컨설팅 및 자문을 맡고 있는 홍주연 전임입니다. 전공은 호텔관광경영학이라서 IT 분야와는 거리가 있었지만, 석사·박사과정을 밟으면서 통계를 공부하다가 데이터 분석을 접할 수 있었어요. 관광빅데이터실인 현재 부서로 발령받은 후에 본격적으로 관광데이터 분석 업무를 하기 시작했습니다.

이번 과제는 어떻게 참여하게 되셨나요? 기존 일과 어느 정도 관련이 있었나요?

한국관광공사에서 운영 중인 관광 빅데이터 플랫폼(한국관광데이터랩)에서는 이동통신, 신용카드, 내비게이션 등의 관광 빅데이터를 시기별·지역별로 조회할 수 있도록 서비스하고 있습니다. 이를 활용해 한눈에 관광 현황을 파악할 수 있는 새로운 시각화 서비스는 어떤 것이 있을까 고민하다가 관광 활성화 지수를 개발하게 되었어요. '지역관광 활성화 수준 진단(가칭 관광 활성화 지수) 분석 고도화 연구용역'으로 추진된 과제의 최종 결과물을 바탕으로 현재 한국관광데이터랩에서 자동 시각화 서비스를 제공하고 있습니다.



데이터는 어떻게 수집하셨나요? 원하는 형태로 정제하기까지 어려움이 있었을 것 같은데요.

데이터마다 적재 방식이나 시간·공간 범위 등 특성이 달라 분석에 모두 활용할 수 없다는 어려움이 있었어요. 하지만, 보유하고 있는 모든 데이터를 사용해야 한다는 욕심보다는, 지역별 관광 활성화를 진단할 수 있는 대표성 있는 변수와 이를 설명하기에 적합한 데이터 세트 후보군 중에서 공통된 조건을 만족시키는 데이터를 선정하는 데 중점을 두었습니다.

앞서 사례에서 언급된 대로, 우선 기존 연구 등 문헌조사를 통해 변수를 탐색했고 한국관광데이터랩에 적재된 데이터에서 데이터 세트를 최종 선정했어요. 탐색적 요인분석과 신뢰성 검사, 자문회의를 통해 변수들과 사용된 데이터의 적합성을 검증했고요.

내부에서 확보할 수 없는 데이터는 때로 본 과업과 관련 없는 다른 산업에서 찾는 경우가 많이 있어요. 이 과제에서는 기초지자체 간 이동 거리 데이터가 필요했지만, 저희가 보유하고 있는 건 직선거리 데이터뿐이었습니다. 그래서 저희는 공간분석 전문가의 조언을 통해 도로망 데이터에 기반한 거릿값을 산출해 분석에 사용했습니다.



데이터 분석을 위해 사용한 알고리즘이나 방법은 무엇인가요? 그 과정에서 어떤 전문가나 관계자가 참여했나요?

같은 시간·공간 범위를 갖추고 있는 데이터 세트로 변수를 설정하고, 관광 활성화 단일지수를 산출할 때 필요한 하위 요소들의 가중치를 도출하기 위해 고차 확인적 요인분석을 실시했어요. 관광 활성화 단일지수와 세부 지푯값은 250개 기초지자체 중에서 조회한 지역의 상대적 활성화 수준을 나타내기 위해 누적분포함수를 적용했고, 그 범위는 0 이상 100 이하로 정했어요.

이러한 분석을 진행하기 위해, 관광경제·통계 분야의 외부 전문가 및 연구진과 협업했고, 공간분석, 지표개발 분야 등 관련 전문가에게 자문을 얻기도 했습니다.



프로젝트 결과를 활용, 공유하기 위해 어떤 노력을 기울이셨는지 말씀해주세요.

앞서 말씀드린 대로, 한국관광데이터랩에 관광 활성화 지수 자동 시각화 서비스를 제공하고 있습니다. 일회성 과업이 아니라 개선된 서비스를 지속해서 제공하기 위해 올해도 분석 고도화를 계획하고 있어요. 2022년에 수행한 분석과의 차별점은 지수별 가중치를 산출할 때 사용된 데이터를 분석 이후의 가장 최신 데이터로 업데이트하는 것, 그리고 보유하고 있는 관광 빅데이터 외에 관광 활성화를 설명하기에 적합한 민간·공공데이터를 추가해 더 다각적인 인사이트를 제공하는 겁니다.



데이터 활용을 통한 업무개선에 부담을 느끼는 분들께 도움 되는 말씀 부탁드려요.

빅데이터에 기반한 이슈 발견과 문제 해결 방식은 분명 다른 방법으로 얻을 수 없는 인사이트를 제공하곤 합니다. 하지만 때로는 현장에 대한 지식 없이 데이터만을 통해 도출된 결과는 현실과 큰 괴리감이 있는 방향성 제시로 혼란을 빚기도 하죠.

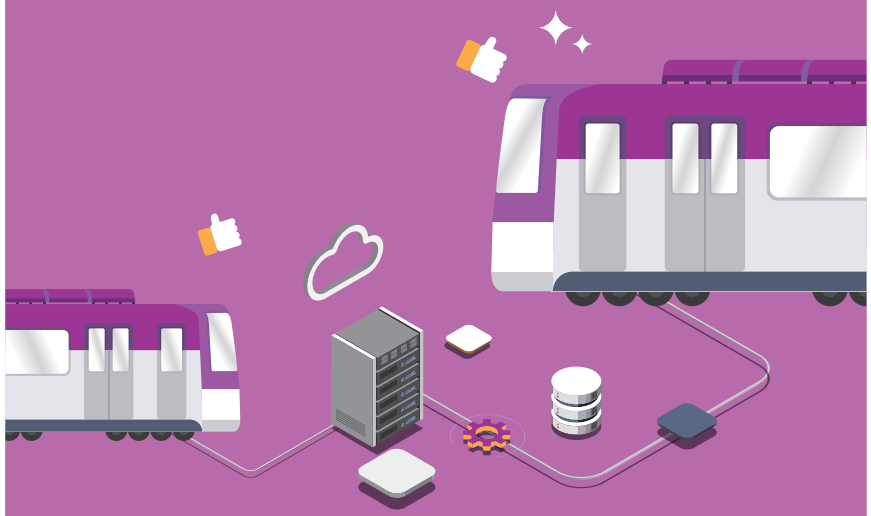
그래서, 데이터를 통한 문제 해결에 매몰되기보다는 다른 분야 전문가, 관련 업계 종사자, 학계, 민간, 공공기관 등 필요에 따라 다양한 견해를 수집하고, 그 모든 것들이 데이터를 뒷받침했을 때 좀 더 폭넓게 의사결정을 할 수 있다고 생각합니다.

PART 4-2. 관광활성화

내 마음에 쏙 든 철도여행, 알고 보니 빅데이터 덕분이었다

철도 & 민간 가명 정보 활용
5대 도시 관광패턴 분석

한국철도공사 미래기술처



이렇게 바뀌었습니다!

AS-IS	TO-BE
• 경험 기반의 여행상품 개발 및 마케팅 • 협력여행사에 단기적 지원 (수수료 조정 등)	• 데이터 기반의 상품개발 방식으로 혁신 유도 • 상품 운영 지원으로 기업 및 공사 수익 증대 기여

추진 배경

코로나19로 인한 관광산업 전반의 침체로 국내여행사는 매출이 급감하거나 폐업하는 등 경영난이 심화하였다. 한국철도 협력여행사가 판매한 철도여행상품 역시 이용객 수가 뚝 떨어져 코로나19가 본격적으로 확산하던 2020년에는 전년 대비 68% 수준밖에 되지 않았다.

또한, 철도 협력여행사 대부분은 종사자 수가 3~5명 수준이고 사업 규모도 영세하여, 자체적으로 여행상품을 개발하기 위한 핵심적인 정보와 역량을 갖추지 못하고 있었다. 이에 한국철도공사(이하 공사)는 협력여행사가 코로나19에 의한 극심한 경영난을 빠르게 극복하고, 데이터를 기반으로 여행상품을 설계할 기회를 제공함으로써 장기적인 경영 체질 개선을 지원하기로 했다.

활용데이터

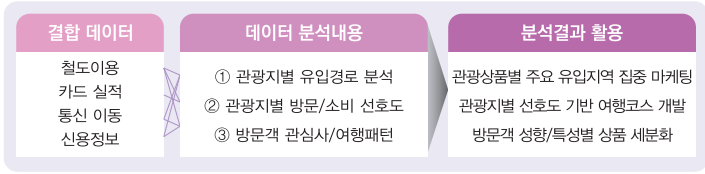
연번	데이터명	기준연도	형태	내 용	보유기관
1	철도 고객 열차 이용 내역에 대한 가명 정보	2019	csv	연간 열차표 발권량·정기권, 선호 열차 종·노선·객실등급·구간·계절·사용시간대, 회원 상세분류 정보 및 분석 대상 지역 방문 여부 등(1,500만 명)	한국 철도공사
2	동선 데이터 (가명, 통계)	2021	csv	지역별 관심 영역 내 30분 이상 체류 지점·시간 등(2,700만 명)	SKT
3	유입경로 데이터	2021	csv	분석 대상 도시 및 지점에 대한 유입경로 정보	SKT
4	개인 선호 분석 정보	2021	csv	앱 사용 내역을 통한 개인 선호(반려동물, 골프, 게임 등에 관한 관심도) 분석 정보	SKT
5	카드 소비데이터 (가명, 통계)	2021	csv	지역별 관심 영역 내 업종별 카드 사용 데이터(2,500만 명)	신한카드
6	개인 선호 분석 정보	2021	csv	카드 이용 내역을 통한 개인 선호(취미, 선호 브랜드, 스포츠, 문화, 여행지 등) 분석 정보	신한카드
7	개인 신용 관련 가명 정보	2021	csv	거주지, 성별, 연령대, 직업군, 소득수준 등(4,400만 명)	신용 정보사

국내 5대 관광도시 중 선호 방문지, 소비 행태 분석

공사는 협력여행사가 원하는 데이터 생산을 위해 4개 분야(철도 이용 정보, 카드 사용 실적, 통신 이동 정보, 신용정보) 총 625개 항목(속성)의 데이터를 확보하여, 가명 정보 형태로 결합했다. 그리고 이 결합 데이터를 활용하여 5대 관광도시(강릉·부산·경주·전주·여수)의 주요 관광지점 99곳 중 선호 방문지, 소비 행태 등에 관한 분석을 수행했다.

세부적으로 분석한 내용은 ① 5대 도시 방문객의 교통수단별, 거주지별 유입경로 ② 5대 도시 방문객의 성별, 연령별 비중 및 취미, 관심 등의 선호도 ③ 5대 도시 여행지 방문객의 성향별 선호도에 따른 주요 소비·이동패턴과 교통수단에 따른 지역별 이동패턴 간 차이이다.

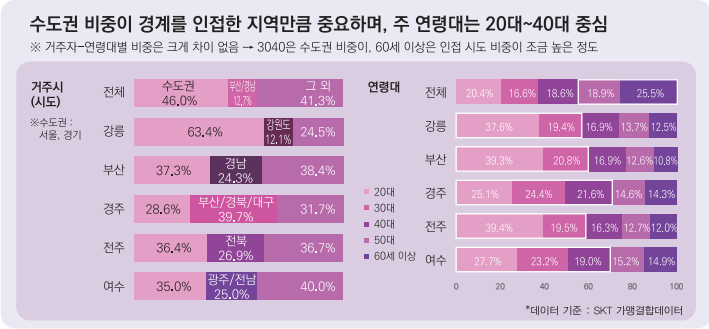
〈그림 46. 분석 과정〉



여행객 유형·선호 방문지·성향 등 고려 여행상품 세분화 필요

국내 5대 관광도시 중 선호 방문지, 소비 행태 등을 분석한 결과, 첫째, 유입경로 분석을 통해 알 수 있었던 것은 전체 5개 관광도시 방문객의 거주지 비중은 경주시를 제외하고 통상 수도권이 가장 높다는 사실이었다. 이는 여행상품을 구성할 때 수도권 거주자를 끌어들일 만한 요소가 포함되어야 함을 시사한다.

〈그림 47. 거주지, 연령대에 따른 분석 결과〉

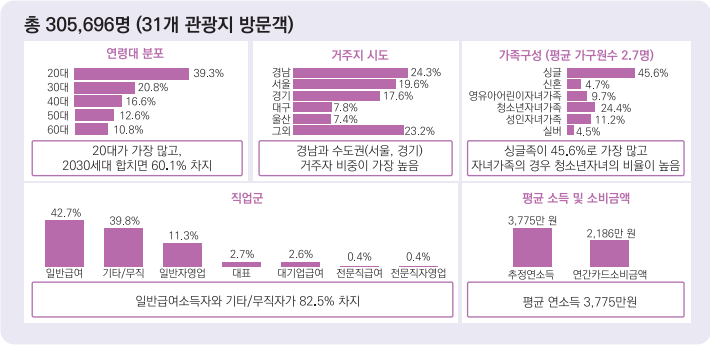


20~30대가 많이 방문하는 도시는 부산·전주·강릉이었고, 20~40대가 고루 방문하는 곳은 경주·여수였다. 30~40대는 초중고 자녀를 둔 연령대로 주로 가족 단위 여행객이 많았으며, 이들의 방문 비중이 높은 도시는 가족 단위의 여행상품을 기획해야 함을 알 수 있다.

이처럼, 여행사들은 유입경로 분석을 통해 집중적으로 마케팅할 대상이 누구인지 데이터를 통해 결정할 수 있게 되었다.

둘째, 부산광역시 분석 결과를 중심으로 관광지별 방문객 특성을 분석하면, 부산지역 31개 관광지 방문객은 20~30대 비율이 60.1%로 싱글족과 일반급여자, 기타/무직자(대학생 포함)의 비율이 높은 것으로 나타났다. 따라서, 대학생-사회초년생-중간관리자 등 20~30대는 사회적 지위 변동과 함께 경제 사정의 변화가 크다는 점을 고려할 때, 각 고객이 선호할 만한 여행코스를 개발하여 상품을 구성한다면 여행사의 수익은 대폭 증가할 것이다.

〈그림 48. 관광지별 방문객 특성 분석 결과(부산)〉



셋째, 방문객 특성에 따른 선호 방문지를 분석한 결과, 개인의 취미·성향에 따라 선호 방문지가 다르게 나타남을 확인할 수 있었다. 예시로 부산을 살펴보면, 쇼핑을 선호하는 방문객은 레저 선호 방문객과 비교해 상대적으로 서면역, 남포동을 더 선호하는 것으로 나타났다. 따라서, 여행사에서는 방문객들의 성향을 고려하여, 맞춤형 여행상품을 개발해야 함을 알 수 있다.

협력여행사 대상 '철도 여행상품 공모전' 개최, 9천만 원 수익 올려

공사는 협력여행사가 가명 정보 분석 결과를 실제 여행상품 개발 업무에 활용할 수 있도록 '철도 여행상품 공모전'을 개최했으며, 공모전에는 총 19개의 여행상품이 제출되었다. 그중 반려견 동반 상품, 부산 핫플레이스 체험 등 강릉·부산 각 3개 상품, 경주·전주·여수 각 1개 상품 등 총 9개가 선정되어 6개월간 해당 상품별 전용 좌석을 배정(1일 20석)받았으며, 해당 상품을 통해 총 9천만 원의 수익을 올리기도 했다.

또한, 일반 국민도 활용할 수 있도록 국토교통 데이터 오픈마켓에 분석보고서를 무료로 개방했으며, 철도 협력여행사 35개를 대상으로 '찾아가는 데이터 컨설팅'을 시행해 데이터 분석 결과에 관한 설명과 다양한 여행상품을 컨설팅했다.

기관소개(한국철도공사)

한국철도공사는 1894년 철도국이 창설되면서 대한민국 철도의 역사와 함께 출발했다. 100년이 넘는 역사를 이어오다 2005년 지금의 한국철도공사로 출범하였으며, 기관의 주요 업무는 여객과 화물의 안전한 철도 수송이다. 철도 운송을 통하여 국민의 보편적 이동권 보장과 함께 지역 균형발전을 직·간접적으로 지원함으로써 국민경제발전과 철도 산업 발전에 이바지하고자 한다.



Coffee Time (인터뷰)



현재 하시는 일을 포함해 자기소개 부탁드립니다.

한국철도공사 미래전략실에서 빅데이터 기획과 분석 업무를 맡고 있는 이형일 차장입니다. 저는 전기통신 직렬로 입사해 전기시설물 유지보수를 담당했기 때문에 데이터를 처음부터 잘 다루는 사람은 아니었어요. 제가 처음 데이터 분석에 관심을 가지게 된 계기는 학부에서 경영학 전공 당시 통계프로그램과 마케팅 리서치 쪽을 공부하면서부터인데요. 그러다 빅데이터 전공으로 대학원에 진학하면서부터 본격적으로 데이터 분석에 대한 전문적 지식을 습득하게 되었습니다.

다행히 우리 공사에서 IT 조직에 빅데이터 부서를 신설하면서 전공을 살릴 기회를 얻게 되었는데, 지금은 공사의 다양한 분야에서 필요한 데이터 분석 업무를 지원하는 데 제가 가진 역량을 발휘하고 있습니다.

이번 과제는 어떻게 참여하게 되셨나요? 기존 일과 어느 정도 관련이 있었나요?

2019년도부터 한국철도공사는 과학기술정보통신부에서 지원하는 '빅데이터 플랫폼 및 센터 구축' 사업에 교통 분야 컨소시엄으로 참여하고 있었습니다. 컨소시엄 내에서 저희의 역할은 공사에서 보유한 데이터와 민간·공공 데이터를 융합한 신규 콘텐츠를 생산하는 것이었어요.

3차년도(2021)에 코로나19로 인한 관광수요 감소로 경영난을 겪고 있는 민간여행사를 데이터로 지원하기 위한 사회문제 해결형 분석 과제의 목적으로 본 과제를 시작하게 되었습니다.

이번 과제를 추진하면서 겪었던 어려움에 관해 듣고 싶습니다.

이번 과제는 한국철도공사 협력여행사들의 데이터 기반 상품구성 역량 확충과 경영체질 개선에 주목적이 있었습니다. 이를 위해 우리는 여행사들이 원하는 관광지 방문객의 성향과 선호 방문지, 소비 행태 등에 관한 데이터를 제공하고자 민간업체와 데이터 협력이 꼭 필요했죠. 그런데, 마침 2020년 8월에 데이터 3법이 개정되면서 개인정보를 활용할 수 있는 법적 근거가 마련되었고, '가명 정보 결합'이라는 새로운 방식을 활용해 여행사들이 원하는 데이터를 확보할 수 있게 되었습니다.

여기까지만 보면 별 어려움 없이 잘 흘러간 듯 보이나 실상은 전혀 그렇지 못했어요. '가명 정보 결합'이라는 신생 제도와 관련 법에 대한 이해 부족으로 난관에 봉착했거든요. 한 번도 가보지 못한 길을 가는 건 절대 쉽지 않았습니다. 하지만, 다행히 개인정보보호위원회에서 제공하는 '가명 정보 활용 컨설팅' 프로그램에 참여하여 가명 정보 전 과정에 대한 지원을 확보할 수 있었습니다.



데이터 수집과 정제 과정에서 겪은 어려움과 극복 방법에 대해 알려주세요.

가명 정보 결합에 쓰이는 시간과 비용적 한계 때문에 분석에 필요한 변수들을 한 번에 잘 선정하는 게 매우 중요했습니다. 그러나 민간기업에서 보유한 정보에 관해 모두 알기 어려워 어떤 변수들을 요구해야 하는지, 개인 식별 가능성이 커지지 않게 정보를 결합하려면 어느 정도 수준으로 정보를 요청해야 하는지 등 어떤 경계를 정하는 게 상당히 어려웠어요. 그래서 아예 기획 단계에서부터 데이터 제공 기업들을 적극 참여시켜 도메인 지식의 부족과 관련 사업 경험의 한계를 극복해 나갔습니다.



데이터 분석을 위해 어떤 전문가나 관계자를 참여 시켰나요?

이 분석 과제는 가명 데이터의 결합부터 반출까지의 단계는 공사에서 추진하고, 분석은 전문 분석기업에 위탁해서 결과물의 완성도를 높일 수 있도록 설계하였습니다. 또한, 도시별 관광지에 대한 분석 영역 설명, 분석 결과에 대한 시사점 도출 등 중요한 의사결정 시점에 해당 도시의 전문가(지자체, 여행사, 공사의 여행센터)들을 참여시킴으로써 분석기업에만 의존하지 않고 실제 사용자들이 활용할 수 있는 결과물이 나오도록 추진했습니다.



데이터 활용을 통한 업무개선에 부담을 느끼는 분들께 도움 되는 말씀 부탁드려요.

저희는 기존에 보유하고 있던 데이터로 문제를 해결할 수 없어서 민간의 데이터 제공기업들과 6개월 이상 협업했고, '가명 정보'라는 새로운 제도를 이해하고 활용하기 위해 내부 관리계획을 수립하고 데이터를 결합하기까지 꽤 오랜 시간이 걸렸어요. 시간이 많이 흐를수록 이 과제를 완수할 수 있을까 하는 걱정과 고민도 함께 커졌죠. 하지만 공공기관으로서 사회적 책임과 새로운 시도가 우리 기관에 대한 외부 인식의 변화를 끌어 내리라 기대하는 마음이 컸기에 끝까지 달릴 수 있었습니다.

우리가 수행한 과제는 성과를 인정받아 사회문제 해결 과제로 국토교통부에서 예산 지원을 받게 되었고, 개인정보보호위원회가 주최한 '가명 정보 활용 우수사례 경진대회'에서 대상으로 선정되기도 했습니다. 내가 데이터로 만들어 낸 결과물이 국민에게 더 좋은 서비스로 다가가고 기업에 큰 도움이 될 수 있다는 기대감을 품는다면 조금 더 용기를 내어 적극적인 행정을 펼칠 수 있지 않을까요.

2023년 공공부문 데이터 분석·활용 우수사례집

발 행 처 | 행정안전부 통합데이터분석센터, 한국지능정보사회진흥원

엮 은 이 | 행정안전부 통합데이터분석센터
센터장 김철
담당 정현곤
담당 김현우
한국지능정보사회진흥원 데이터기반행정팀
팀장 강경훈
선임 정혜린

발 행 일 | 2023년 12월 14일

© 행정안전부, 한국지능정보사회진흥원, 2023

본 사례집 내용의 무단 전재를 금하며, 가공 및 인용 시 반드시 출처를 명기해주시기 바랍니다.