| 연구보고서 2021- 12 |

공공도서관의 수요추정 모형 개발

김제국 · 함윤주

한국지방행정연구원
Korea Research Institute for Local Administration

공공도서관의 수요추정 모형 개발
한국지방행정연구원

KOREA RESEARCH INSTITUTE FOR LOCAL ADMINISTRATION

공공도서관의 수요추정 모형 개발

연구진 김 제 국 (부연구위원)
함 윤 주 (부연구위원)

발행일 2021년 12월 31일

발행인 김 일 재

발행처 한국지방행정연구원

주 소 (26464) 강원도 원주시 세계로 21(반곡동)

전 화 033-769-9999

판매처 정부간행물판매센터 02-394-0337

인쇄처 문화공감 02-2266-1897~8

ISBN 978-89-7865-511-8

서 문

PREFACE

한국지방행정연구원 지방재정투자사업관리센터(LIMAC)에서는 2015년부터 지방자치단체에서 추진하는 다양한 공공시설 건립에 대한 타당성조사를 수행해왔으며, 해당 공공시설에 대한 지역 수요가 충분한지 검토하기 위해 다양한 기법을 활용하여 조사해왔습니다.

그러나 각 공공시설 유형별 특성을 고려한 방법론에 대한 가이드라인은 부재한 상황이며, 특히 문화기반시설로서 공급이 지속적으로 확대되고 있는 공공도서관에 대한 수요예측을 위한 기초연구가 부족한 상황입니다.

이러한 상황에서 본 연구는 타 분야에서 활발하게 적용되고 있는 기계학습방법을 활용하여 연구의 가능성과 방법론적 지평을 확대하는데 기여하고 있습니다. 특히 그동안 국가도서관 통계시스템에 축적된 방대한 데이터를 활용하여 도서관 운영 시 다양한 특성들이 이용수요에 미치는 영향을 고려하여 궁극적으로 이용수요를 예측할 수 있는 수요추정 모형을 개발하고, 공공도서관 정책 수립의 근거자료로 활용할 수 있도록 신규 공공도서관 건립이 지역 내 수요에 미치는 효과를 분석하였습니다.

이 연구가 향후 공공도서관 수요추정의 신뢰성과 정확성을 제고하는 데 기여할 수 있기를 기대하면서 실용적인 연구결과 도출을 위하여 노력한 연구진과 연구 진행 과정에서 많은 조언을 해주신 원내 연구심의위원회 위원 및 외부 자문위원들께도 감사의 말씀을 전합니다.

2021년 12월

한국지방행정연구원 원장 김일재

요 약

오늘날 공공도서관은 전통적인 기능인 도서관 자료를 중심으로 한 지식정보 제공 및 학습의 기능뿐만 아니라 지역사회 구성원을 위한 문화공간으로서 역할도 수행하고 있다. 이러한 수행 역할과 기능의 확장에 따라 국민의 삶의 질 향상을 높이기 위한 국민기초시설인 생활SOC의 한 축으로 자리매김하게 되었다. 그 결과 2020년 기준, 생활SOC 사업에 선정된 289개 사업 중 160개 사업이 공공도서관 및 작은 도서관 사업이 포함된 사업이었다. 특히 공공도서관이 포함된 사업은 73개이며, 국비지원 규모는 2,031억 원에 달하였다.

최근에는 생활SOC뿐 아니라 기존 도서관의 노후화 및 리모델링, 신도시 건설에 따른 도서관 신규 건립 등에 따라 지방자치단체에서 공공도서관 추진 사업 건수가 증가하였다. 이에 따라 지방자치단체가 추진하는 대규모 사업에 대해 투자심사 전에 이행해야 하는 「지방재정법」 제37조에 의한 타당성 조사 의뢰 건수도 증가하고 있다.

따라서 타당성 조사의 핵심이 되는 도서관 수요 추정에 대한 신뢰성과 정확성을 높이기 위한 적절한 방법론 개발이 필요하다. 공공도서관 수요추정 방법론과 관련해서는 『문화체육관광 타당성 조사를 위한 지침 연구』(한국지방행정연구원, 2016)에 제시된 다양한 방법을 참고할 수 있으나, 공공도서관의 특성을 고려한 방법론에 대한 가이드라인은 부재한 상황이다. 함윤주 외(2019)에서는 중력 모형으로 안양시에 위치한 공공도서관을 대상으로 실증분석을 실시하였으나, 공공도서관 관련 수요 추정에 대한 기초연구는 상당히 부족한 실정이다.

본 연구는 도서관의 이용수요의 특징과 기존에 연구된 수요추정방법을 검토하고, Microeconometrics(Program evaluation) 접근법과 Data science 접근법을 통하여 도서관 운영 시 다양한 특성들이 이용수요에 미치는 영향을 고려하여 궁극적으로 이용수요를 예측할 수 있는 수요추정 모형을 개발하였다. 국가공공도서관통계에서 방대한 양의 도서관 관련 데이터를 제공하고 있으므로, 이를 활용하여 XGboosting 등을 포함하는 기계학습 방법을 사용하여 앞으로 타당성 조사에 사용이 가능한 도서관 방문자 수 예측 모형을 구축하였다. 또한 예측치 생성에 미치는 영향을 분석하여 데이터 내 변수들의 중요성을 분석하였다.

그리고, 방문자 수를 공공도서관 제공의 충격을 나타내는 지표로 보아 공공도서관 건립의 정책 충격을 측정하였다. 이를 위해서 문화체육관광부에서 제공하는 도서관에 관한 패널 자료가 미처 고려하지 못한 변수 중 분석 단위에 따라 변하나 시간에는 변하지 않는 변수와 시간에 따라 변하나 지정된 시간에는 분석 단위에 따라서는 변하지 않는 변수들을 제어하여 이 변수들의 제외 편의를 다루는 fixed effects 모형을 사용하여 목표한 분석을 하였다. 분석결과 현재 서울에서는 도서관 방문자 수 증가를 목표로 할 때 신규 도서관 제공보다는 먼저 기존 도서관의 좌석 수와 장서 수를 늘리는 것을 고려하고 현재 해당 구 내 도서관의 수용 능력이 그 고려에 미치지 못할 때 신규 도서관 건립을 고려해야 한다는 결론이 도출되었다.

차 례

표목차

그림목차

CONTENTS

제1장

서 론

공공도서관의
수요추정 모형 개발

KRILA
KOREA RESEARCH INSTITUTE FOR
LOCAL ADMINISTRATION

제1장 서론

제1절 연구의 배경과 목적

1. 연구배경

공공도서관은 「도서관법」 제2조(정의)에 따라, 도서관자료를 수집·정리·분석·보존하여 공중에게 제공함으로써 정보이용·조사·연구·학습·교양·평생교육 등에 이바지하는 시설로 정의된다. 그러나 오늘날 공공도서관은 상기 제시된 바와 같이 도서관자료를 중심으로 한 지식정보 제공 및 학습의 기능뿐만 아니라 지역사회 구성원을 위한 문화공간으로서 그 기능과 사회적 역할에 많은 변화가 있었다.

지역사회에서 공공도서관의 기능 및 역할 확대는 문재인 정부가 도입한 "지역밀착형 생활SOC사업"에서도 드러난다. 생활SOC는 국민의 삶의 질 향상을 높이기 위한 국민기초시설로서 정의된다. 문재인 정부에서 발표한 『생활SOC 3개년 계획(안)(2020~2022)』(국무조정실, 2019.4.15.)에서는 매년 지방자치단체들이 주도하여 지역주민 수요에 기반한 생활SOC복합화 사업을 추진하도록 하였다. 특히 공공도서관은 생활SOC 포함된 10개의 다양한 공공시설 중 상당 부분을 차지하고 있다. 2020년 기준, 생활SOC 사업에 선정된 289개 사업 중 160개 사업이 공공도서관 및 작은도서관 사업이 포함된 사업이었다. 특히 공공도서관이 포함된 사업은 73개이며, 국비지원 규모는 2,031억 원에 달하였다.[1)]

최근에는 생활SOC뿐 아니라 기존 도서관의 노후화 및 리모델링, 신도시 건설에 따른 도서관 신규 건립 등에 따라 지방자치단체에서 공공도서관 추진 사업 건수가 증가하였다. 이에 따라 지방자치단체가 추진하는 대규모 사업에 대해 투자심사 전에 이행해야 하는 「지방재정법」 제37조에 의한 타당성조사(LIMAC 타당성조사[2)]) 의뢰 건수도 증가하고 있다.

1) (보도자료) "20년부터_생활SOC_복합화_사업_본격_추진한다"(191002, 국가균형발전위원회)
2) 「지방재정법」 제37조 제2항에 따라 지방자치단체의 장은 총 사업비 500억 원 이상의 신규사업에 대해서는 행정안전부장관이 정하여 고시하는 전문기관으로부터 타당성 조사를 받고 그 결과를 토대로 투자심사

LIMAC 타당성조사에서는 앞서 제시한 바와 같이 공공도서관 기능의 다변화에 따라 경제성 분석 수행시 지역주민에게 제공하는 다양한 형태의 서비스에 대한 이해를 토대로 수요 및 편익을 추정해야 하는 과제에 직면해 있다. 특히 도서관의 유형이 생활SOC 복합화 사업에 포함된 공공도서관 및 작은도서관을 비롯하여 지역의 대표도서관 및 거점도서관 등 다양하게 나타나고 있어, 유형별 특징을 고려하여 지역 내 수요 및 공급 여건을 반영한 수요 및 편익 추정방법 개발이 필요한 실정이다.

기존에 LIMAC 타당성조사 대상 도서관은 복합시설물에 포함된 공공 또는 작은도서관으로 규모가 작거나, 전체 사업에서 차지하는 비중이 적은 경우가 상당 부분을 차지하였다. 규모나 사업 비중 등을 고려할 때 한정된 시간 내 효율적인 조사방법 선정 차원에서 해당 도서관들은 유사시설 실적 또는 중력모형 등을 활용하여 수요를 추정하였다. 일부 단일 시설로서 거점도서관, 중앙도서관, 지역대표도서관 등 일정 규모 이상의 도서관의 경우에는 중력모형 외에 회귀분석으로 수요를 추정하기도 하였다.

따라서 타당성조사별로 일관성 측면 제고 및 지역대표도서관이나 거점도서관 관련 사업에 대한 조사도 수행하면서 전반적으로 도서관 수요 추정에 대한 신뢰성과 정확성을 높이기 위한 적절한 방법론 개발이 필요한 시점이다. 편익 또한 각 수요 추정 결과에 의거하여 다양한 형태로 구득된 서비스의 가치를 적용하여 추정하여 왔으나, 조사의 일관성을 제고하기 위해 수요에서 논의된 다양한 도서관 서비스에 대한 가치를 검토할 필요가 있다.

공공도서관 수요추정 방법론과 관련해서는 『문화체육관광 부문 타당성조사를 위한 지침 연구』(한국지방행정연구원, 2016)에 제시된 다양한 방법을 참고할 수 있으나, 공공도서관의 특성을 고려한 방법론에 대한 가이드라인은 부재한 상황이다. 함윤주 외(2019)에서는 중력모형으로 안양시에 위치한 공공도서관을 대상으로 실증분석을 실시하였으나, 공공도서관 관련 수요 추정에 대한 기초연구는 상당히 부족한 실정이다.

다양한 수요추정 방법 중 중력모형의 경우 함윤주 외(2019)에서 안양시 사례를 들어 실증분석 실시 및 수요추정 예측 오차를 줄이기 위한 적정 중력모형을 제시하였다. 그러나 지역내 공공도서관의 분포가 상당히 균일한 안양시 사례만으로 모든 유형의 공공도서관

를 하여야 한다. 전문기관은 [행정안전부고시 제2020-21호, 2020. 4. 29., 일부개정]에 따라 한국지방행정연구원으로 정하며, 이에 2014년 말 한국지방행정연구원은 지방투자사업관리센터(LIMAC)을 설치하여 타당성조사(이하 LIMAC 타당성조사)를 수행하고 있다.

사례에 적용가능한 방법론을 도출하기에는 무리가 있었다. 또한 근본적으로 거리조락현상을 전제한 중력모형은 영향권 설정, 준거시설 선정 등의 측면에서 객관적 기준 마련이 가장 핵심적인 사항이다. 이러한 측면에서 도서관의 경우 여러 공공시설 중에서도 기존에 지역 내 다양한 유형의 시설 비교적 촘촘하게 공급되고 있는 시설이기 때문에 지역별로, 공급여건별로 중력모형 적용이 모든 타당성조사에서 적용 가능한 방법론인지에 대한 의문점이 제기된다. 다만 함윤주 외(2019)를 비롯하여 여러 학술연구에서 도서관의 이용실적에 영향을 미치는 다양한 요인이 있다고 제시하고 있으므로, 이러한 영향요인들을 감안하여 수요를 예측할 수 있는 모형 개발이 가능한지 연구해 볼 필요가 있겠다.

2. 연구목적

본 연구는 도서관의 이용수요의 특징과 기존에 연구된 수요추정방법을 검토하고, 미시계량경제학(microeconometrics)의 프로그램 평가(Program evaluation) 접근법과 데이터 과학(Data science) 접근법을 통하여 도서관 운영시 다양한 특성들이 이용수요에 미치는 영향을 고려하여 궁극적으로 이용수요를 예측할 수 있는 수요추정 모형을 개발하고자 한다.

특히 국가공공도서관통계에서 방대한 양의 도서관 관련 데이터를 제공하고 있으므로, 이를 활용하여 데이터 과학(Data science) 접근법으로 수요에 영향을 미치는 주요 요인들에 대한 분석을 실시할 수 있다.

본 연구의 목적은 최적의 수요추정 모형을 개발하고, 이를 통해 LIMAC 공공도서관 사업타당성조사 신뢰성을 확보하고 투자심사의 의사결정을 지원하는데 목적이 있다. 이를 통해 사후적으로 공공도서관의 운영 활성화 및 평가 측면에서도 참고할 수 있는 수요예측방법이 제시될 수 있기를 희망한다.

제2절 연구의 범위와 방법

1. 연구범위

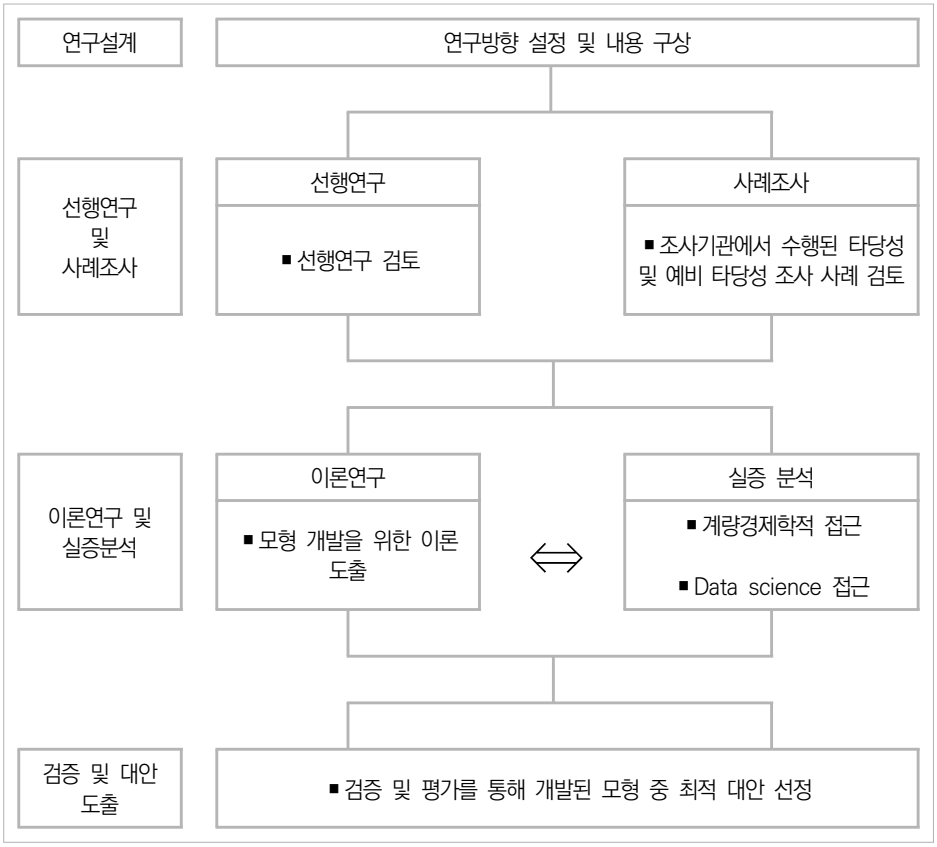
본 연구의 대상은 서울시에서 개관되어 운영 중인 도서관 중 문화체육관광부 국가도서관통계시스템에 등록된 공공도서관을 대상으로 한다. 이를 바탕으로 공공도서관 건립사업 타당성 조사의 핵심이 되는 도서관 수요에 관한 실증적인 연구를 수행한다.

2. 연구방법

본 연구에서는 도서관 관련 제도 및 법령에 대한 분석을 통해 공공도서관의 개념, 유형, 기능 등을 검토하고, 현재 도서관 공급과 관련한 기준 및 현황을 살펴보고자 한다. 이후 2015년부터 수행한 LIMAC 타당성조사 중 도서관 관련 사례를 조사하고자 한다. 또한 국내외 학술논문, 학위논문 및 연구보고서 등 선행 학술연구 및 PIMAC 사례 등을 통해 그동안 공공도서관 수요와 관련하여 논의 및 적용되어 온 수요 관련 쟁점들을 검토하고자 한다.

또한, 국가도서관통계를 활용하여, 기계학습을 통한 방문자 수 예측 모형 개발과 미시계량경제학(microeconometrics) 방법을 이용하여 도서관 제공의 정책 효과를 측정한다.

그림 1-1 연구의 흐름도


연구설계
연구방향 설정 및 내용 구상
선행연구 및 사례조사
선행연구
■ 선행연구 검토
사례조사
■ 조사기관에서 수행된 타당성 및 예비 타당성 조사 사례 검토
이론연구 및 실증분석
이론연구
■ 모형 개발을 위한 이론 도출
실증 분석
■ 계량경제학적 접근
■ Data science 접근
검증 및 대안 도출
■ 검증 및 평가를 통해 개발된 모형 중 최적 대안 선정

제2장

도서관 수요에 대한 이론적 논의

공공도서관의
수요추정 모형 개발

KRILA
KOREA RESEARCH INSTITUTE FOR
LOCAL ADMINISTRATION

제2장 도서관 수요에 대한 이론적 논의

제1절 도서관의 개념 및 현황

1. 도서관의 정의 및 기능

1) 법적 정의

도서관은「도서관법」제2조 1항에서 "도서관자료를 수집·정리·분석·보존하여 공중에게 제공함으로써 정보이용·조사·연구·학습·교양·평생교육 등에 이바지하는 시설"로 정의하고 있다. 또한 동법 제2조 4항에서는 공공도서관을 "공중의 정보이용·독서활동·문화활동 및 평생교육을 위하여 국가 또는 지방자치단체 및..(중략) 교육감이 설립·운영하는 도서관"으로 정의하고 있다. 한편,「도서관법」제4조(국가 및 지방자치단체의 책무)에서는 국민이 자유롭고 평등하게 지식정보에 접근하고 이를 이용할 수 있도록 국가와 지방자치단체가 도서관의 발전을 지원하여야 하며 이에 필요한 시책을 강구하여야 한다고 제시하고 있다. 따라서 도서관은 도서를 매개로 여가 및 문화향유권을 보장하는 한편 정보접근권이라는 기본적인 권리를 보장하기 위해 제공되는 시설로 보아야 한다.

2) 도서관의 기능 변화

「도서관법」에서도 도서관은 도서를 중심으로 한 서비스 외에 교육, 여가, 문화 관련 활동을 지원하는 시설로 정의된다. Philip Gill(2002)은 도서관을 정보, 교육 등 기본권 관련 서비스뿐만 아니라 오락 및 여가활동을 포함한 인간적 발전 욕구를 포괄적으로 충족시켜주기 위한 다양한 유형의 자원과 서비스를 제공하는 기관이며, 인종, 국적, 연령, 성별, 종교, 언어, 장애, 경제력, 교육수준에 관계없이 누구나 이용할 수 있는 시설로 정의하였으며 문화체육관광부의『2019년 공공도서관 건립·운영 매뉴얼』에서도 공공도

서관의 기초기능으로 지식정보센터, 확장기능으로 사회문화센터, 여가생활센터, 평생교육센터, 생활편의센터 등을 제시하였다. 또한 최근 공공도서관이 지역사회의 지식향상 및 문화발전뿐만 아니라 지역사회 커뮤니케이션 공간을 제공함으로써 그 기능과 사회적 역할에 많은 변화가 있었다. 따라서 도서관의 수요는 도서에 한정된 서비스 수요를 넘어서 지역 사회 다양한 지식 및 문화발전을 위한 활동에 대한 수요와 연계하여 살펴볼 필요가 있다.

표 2-1 도서관의 정의 및 기능

구분	기능
도서관법	도서관자료를 수집·정리·분석·보존하여 공중에게 제공함으로써 정보이용·조사·연구·학습·교양·평생교육 등에 이바지하는 시설
Philip Gill (2002)	정보, 교육 등 기본권 관련 서비스뿐만 아니라 오락 및 여가활동을 포함한 인간적 발전 욕구를 포괄적으로 충족시켜주기 위한 다양한 유형의 자원과 서비스를 제공하는 기관
공공도서관 건립·운영 매뉴얼	기초기능으로 지식정보센터, 확장기능으로 사회문화센터, 여가생활센터, 평생교육센터, 생활편의센터 등을 제시
한국도서관협회 (2003)	정보이용, 문화활동, 평생교육의 증진을 통한 기본권의 신장과 지역사회 발전에 기여하는 시설

2. 도서관 유형

도서관의 유형은『공공도서관 건립·운영 매뉴얼』과「도서관법 시행령」에 따라 분류할 수 있다. 문화체육관광부의『2019년 공공도서관 건립·운영 매뉴얼』에서는 봉사대상인구 규모에 따라 도서관 규모 및 지위가 결정되며, 지역대표도서관〉지역중앙관〉거점도서관〉분관〉작은도서관 순으로 분류한다. 도서관 지위별 역할에 따른 규모계획 기준은 다음과 같다.

표 2-2 공공도서관의 유형별 규모체계

구분	2,500㎡ 미만	2,500~5,500㎡ 미만	5,500㎡ 이상
도서관	소규모 분관 중규모 분관	대규모 분관 거점도서관 지역중앙관	지역중앙관 거점도서관
평균 연면적	2,000㎡ 정도	4,000㎡ 정도	6,000㎡ 정도 9,000㎡ 이상 최대규모

출처 : 문화체육관광부, 『2019년 공공도서관 건립·운영 메뉴얼』

「도서관법 시행령」[별표1](공공도서관 시설 및 도서관 자료 기준)에서는 봉사인구 수에 따라 건물 면적, 열람 좌석 수, 장서 수 등을 제시하고 있으며 이는 시군구의 인구수를 기반으로 책정된다.

표 2-3 공공도서관 시설 기준

유형	봉사대상 인구 (명)	시설		도서관자료	
		건물면적 (제곱미터)	열람석 (좌석 수)	기본장서 (권)	연간증서 (권)
공립 공공 도서관	2만 미만	264 이상	60 이상	3,000 이상	300 이상
	2만 이상~5만 미만	660 이상	150 이상	6,000 이상	600 이상
	5만 이상~10만 이하	990 이상	200 이상	15,000 이상	1,500 이상
	10만 이상~30만 이하	1,650 이상	350 이상	30,000 이상	3,000 이상
	30만 이상~50만 이하	3,300 이상	800 이상	90,000 이상	9,000 이상
	50만 이상	4,950 이상	1,200 이상	150,000 이상	15,000 이상

주 : 1) 봉사대상 인구란 도서관이 설치되는 해당 시[구가 설치된 시는 제외하며, 도농복합형태의 시는 동(洞)지역에만 해당한다]·구(도농복합형태의 시는 동지역에만 해당한다)·읍·면지역의 인구를 말한다.

2) 봉사대상 인구가 2만 명 이상인 공립 공공도서관에는 열람실 외에 참고열람실·연속간행물실·시청각실·회의실·사무실 및 자료비치시설 등의 시설을 갖추어야 한다.

3) 전체 열람석의 20퍼센트 이상은 어린이를 위한 열람석으로 하여야 하고, 전체 열람석의 10퍼센트 범위의 열람석에는 노인과 장애인의 열람을 위한 편의시설을 갖추어야 한다.

4) 공립 공공도서관에는 기본장서 외에 다음 각 목에서 정하는 자료를 갖추어야 한다.

가. 봉사대상 인구 1천 명 당 1종 이상의 연속간행물

나. 봉사대상 인구 1천 명 당 10종 이상의 시청각자료를 갖추되, 해마다 봉사대상 인구 1천 명 당 1종 이상의 시청각자료를 증대할 것

다. 그 밖의 향토자료·전자자료 및 행정자료

출처 : 『도서관법 시행령』 [별표 1]

작은도서관의 경우, 동 시행령에 별도로 면적, 열람석 수, 장서 수 등을 제시하고 있으며 장애인도서관의 경우, 동 시행령에 별도로 면적, 기계·기구, 장서 수 및 녹음테이프를 제시하고 있다.

표 2-4 작은도서관 및 장애인도서관 시설 기준

유형	시설		도서관자료	
	건물면적 (제곱미터)	열람석 (좌석 수)	기본장서 (권)	연간증서 (권)
작은도서관	33 이상	6석 이상	1,000권 이상	
장애인도서관	66 이상	점자제판기 점자인쇄기 점자타자기 1대 이상 녹음기 4대 이상	1,500권 이상(장서) 500점 이상(녹음테이프)	

주 : 1) 장애인도서관은 시각장애인의 이용을 주된 목적으로 하는 경우에만 해당함
2) 작은도서관과 장애인도서관의 건물면적에 현관·휴게실·복도·화장실 및 식당 등의 면적은 포함되지 아니함
출처 : 『도서관법 시행령』 [별표 1]

3. 도서관 공급 기준

유네스코가 1994년에 발표한 '공공도서관의 목적에 관한 선언(Public Library Manifesto)'에서 도서관은 기본적으로 누구에게나 평등하게 제공되어야 한다는 점에서 보편적 접근성이 보장되어야 한다고 강조된다. 또한 한국도서관협회는 '한국도서관 기준(2003)'에서 공공도서관의 사명으로 정보이용, 문화활동, 평생교육의 증진을 통한 기본권의 신장과 지역사회 발전에 기여하는 것을 제시하고 있다(조권중, 2004). 즉 도서관은 정보 및 지식에 대한 접근권을 시민의 기본권으로 설정 및 보장하기 위한 필수시설로, 다른 문화시설과 비교하여 공공성이 높은 시설로 인식되고 있는 것이다. 「도서관법 시행령」에서도 이용자의 접근성을 고려하여 배치하도록 큰 틀에서 도서관 설치 기준을 제시하고 있다. 문화체육관광부(2013)의 『공공도서관 건립·운영 매뉴얼』에서는 도서지역의 경우, 도서관을 1차 지역(1km 반경 내) 봉사대상 인구가 10분 이내, 2차 지역(2km 반경 내)

인구는 도보 20분 이내로 접근 가능한 곳에 배치하도록 제시하고 있다. 이원태(2004)는 전국 기초자치단체를 대상으로 인구 1인당, 면적 1㎢ 당 도서관 개수를 산출한 후 도시 및 농촌지역의 특수성과 문화체육관광부의 공급 목표량을 감안하여 기초자치단체 유형에 따른 최소 공급기준을 제시하였다. 구체적으로 공공도서관의 경우, 인구 6만 명당 1개소 비율을 적용하여 공급목표를 800개소로 설정하였다.

표 2-5 도서관 최소기준

구분	기준	구분	기준
특별광역시	인구 6만 명당 1개+지역거점 1개	기초시	인구 6만 명당 1개
대도시	인구 6만 명당 1개	기초군	자치단체당 1개
도청소재지	인구 6만 명당 1개+지역거점 1개	낙후지역	자치단체당 1개(복합문화공간)

출처 : 이원태(2004),『전국문화기반시설 최소기준수립 연구』

국토교통부는 2019년 1월 기초생활인프라의 범위 및 국가적 최저기준을 개정 및 시행하기 위해 국가도시재생기본방침을 개정하였다. 기초생활인프라 국가적 최저기준에서는 도서관을 지역거점형과 마을형으로 구분하며, 지역거점형은 차량 10분 이내, 마을형은 도보 10~15분 이내로 최저기준을 제시하였다.

표 2-6 기초생활인프라 범위 및 최저기준

단위	분류	시설	세부시설	최저기준
마을 (도보)	학습	도서관	공공, 사립, 작은도서관	10~15분
지역거점 (차량)	학습	공공도서관	국공립도서관 (국립, 도립, 시립, 교육청 설립)	10분

출처 : 국토교통부(2019).『국가도시재생기본방침 일부개정 재공고』

4. 도서관 공급 현황

주요 국가와 비교하면 국내 도서관 발전종합계획 수립과 지속적인 도서관 양적 성장 노력에도 불구하고, '19년 기준 대한민국의 공공도서관 1관당 서비스 인구는 일본·미국의 1.2~1.3배, 호주의 3배, 독일의 4배 정도 높은 수준인 것으로 나타나 주요 선진국 대비 공급이 부족한 상황이다.

표 2-7 대한민국 공공도서관 1관당 서비스 인구

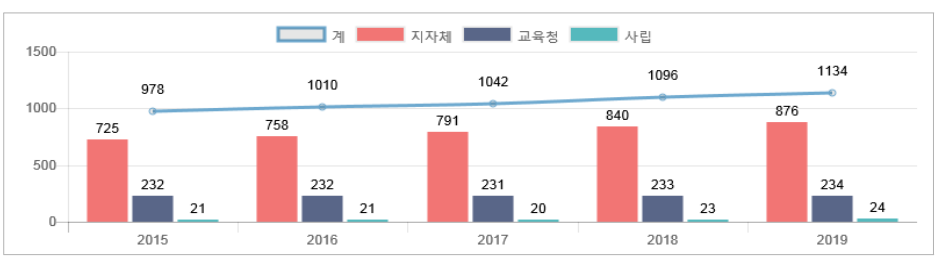
(단위 : 명)

구분	한국 (2019)	미국 (2018)	호주 (2018-2019)	독일 (2019)	일본 (2019)
서비스인구(명)	45,723	35,331	14,963	11,614	38,664

출처 : 국가도서관통계시스템(www.libsta.go.kr), 문화체육관광부

그러나 전국의 도서관 수는 꾸준히 증가하고 있으며, 2019년 기준 1,134관으로 지자체가 운영하는 도서관이 약 77%를 차지하고 20%가 교육청 소속 도서관, 나머지가 사립도서관인 것으로 나타났다.

그림 2-1 최근 5년간 공공도서관 현황



출처 : 국가도서관 통계시스템(www.libsta.go.kr), 문화체육관광부

2015년 978개 대비 156개관 증가했으며, 2015년 대비 약 16%가 증가하였다. 특히 지자체가 운영하는 도서관은 2015년 대비 약 21%가 증가하였다.

표 2-8 연도별 전국 공공도서관 현황

(단위 : 개관)

구분	2015	2016	2017	2018	2019
전체	978	1,010	1,042	1,096	1,134
서울	146	147	160	173	180
부산	36	40	40	43	44
대구	33	35	36	41	43
인천	46	47	48	50	53
광주	21	22	23	23	23
대전	24	24	24	24	26
울산	17	17	18	19	19
세종	4	5	5	10	11
경기	228	244	250	264	277
강원	53	54	56	57	58
충북	42	44	45	45	48
충남	58	59	59	62	63
전북	56	58	58	58	59
전남	64	64	67	69	70
경북	64	64	65	65	66
경남	65	65	67	71	72
제주	21	21	21	22	22

출처 : 국가도서관 통계시스템(www.libsta.go.kr), 문화체육관광부

2019년 기준 전국 광역지방자치단체별 공공도서관 현황을 살펴보면, 부산이 1관당 평균 인구수가 77,587명으로 17개 광역지방자치단체 중 가장 높은 수준인 것으로 나타났으며, 가장 낮은 곳은 강원도로 약 26,578명이었다.

표 2-9 국내 광역지자체별 공공도서관 현황 비교

(단위 : 개관, 권, 명)

구분	공공 도서관수	1관당 장서 수	1관당 인구수	1관당 방문자수	1관당 대출도서 수	1인당 장서 수
전체	1,134	101,477	45,723	250,804	117,962	2.22
서울	180	83,206	54,051	370,706	134,143	1.54
부산	44	134,895	77,587	352,960	157,728	1.74
대구	43	106,071	56,698	295,338	159,542	1.87
인천	53	90,286	55,793	305,682	106,803	1.62
광주	23	115,565	63,325	308,571	108,366	1.82
대전	26	116,524	56,726	279,145	129,653	2.05
울산	19	110,396	60,422	312,436	137,845	1.83
세종	11	44,421	30,961	122,820	105,899	1.43
경기	277	118,354	47,797	285,473	161,833	2.48
강원	58	89,060	26,578	128,387	52,297	3.35
충북	48	92,783	33,333	154,139	84,830	2.78
충남	63	92,611	33,710	158,932	89,999	2.82
전북	59	83,901	30,829	139,552	59,732	2.72
전남	70	92,570	26,696	126,894	58,540	3.47
경북	66	95,963	40,391	174,294	76,725	2.38
경남	72	107,950	46,702	218,275	113,639	2.31
제주	22	118,909	30,500	151,907	92,935	3.9

출처 : 국가도서관 통계시스템(www.libsta.go.kr), 문화체육관광부

제2절 도서관 수요추정법

공공도서관 같은 문화시설의 이용객에 대한 수요추정의 기법은 크게 정량적 분석과 정성적 분석방법이 있으며, 두 가지를 혼합한 결합기법과 수요조사 및 간편법 등과 같은 여타의 방법이 있다.

1. 정량적 분석법

관련 정보를 이용하여 도서관 시설의 수요를 분석하는 방법으로 이동평균법 및 지수평활법, 추세분석법 및 ARIMA 기법 등과 같은 시계열 분석 방법과 회귀분석과 중력모형을 이용한 방법 등이 있다. 이러한 정량적 분석은 미래 수요를 예측하는 데 필요한 변수들을 충분히 확보하거나 상대적으로 충분한 시계열 자료가 요구된다.

이동평균법과 지수평활법은 과거 수요자료를 기반으로 단순 평균하여 미래 수요를 예측하거나, 과거 관측치에 대한 서로 다른 가중치를 두어 추정하는 방법으로 과거 수요자료에 대한 시계열 자료를 통하여 산출한다. 시계열 자료를 활용한 추세분석법은 총수요에 대한 과거 추세만을 이용하여 분석하거나, 미래 수요에 영향을 줄 것이라 예상되는 변수들을 투입하여 분석하는 방법으로 장기 시계열 자료를 필요로 한다. 특히, ARIMA 분석방법은 자기회귀모형(AR : Autoregressive Model) 및 이동평균모형(MA : Moving Average)을 결합하여 수요를 추정하는 방법을 말한다. 회귀 분석모형은 수요에 영향을 주는 다양한 변수들을 고려하여 최소자승법을 이용하여 계수를 추정한 후 미래 수요를 예측하는 모형으로 시계열 분석 방법과 혼합하여 사용할 수 있다. 또한 중력모형(Gravity Model)은 수요에 영향을 미치는 거리와 규모를 고려하여 분석하는 모형으로, 시설의 수요 및 규모 파악과 이용권역의 분할이 가능한 사업에 적용할 수 있다.

2. 정성적 분석법

정량적 분석을 수행하기 위해서는 과거 시계열 자료나 다양한 변수들에 대한 자료가 확보되어야 수요예측의 정확성을 높일 수 있다. 그러나 과거 데이터나 유사한 시설물이

없는 경우에는 연구자나 전문가들의 주관적 의견에 의존하는 시나리오 설정법이나 해당 전문가들의 의견을 반복적으로 수집, 발전시켜 수요를 예측하는 델파이 기법 등이 있다. 시나리오 설정법이나 델파이 기법은 객관성이나 보편적 측면에서 단점에 노출될 수 있으므로, 객관성을 담보하기 위하여 잠재 이용자들을 대상으로 직접 이용의향을 조사하여 수요를 확률적으로 추정하는 진술 수요추정법이 있다.

제3절 선행연구 검토

1. 기존 LIMAC 타당성 조사

LIMAC에 도서관 시설 관련 타당성 조사는 총 13건으로, 복합사업으로 추진된 도서관이 9건, 단일시설로 추진된 도서관이 4건이다. 복합건축물 내 도서관 사업은 공공청사, 공연장 시설, 복지시설 등과 복합시설로 추진되었다.

표 2-10 LIMAC 수행 도서관시설 타당성조사 현황

No	지자체	조사의뢰	사업명	유형	사업비 (억 원)	B/C
1	서울 본청	2016-2차	○△ 건립사업	복합	737	0.69
2	경기 본청	2017-4차	○○도서관 건립	단일	999	0.52
3	경기 시흥	2018-1차	×※ 조성사업	복합	1,821	0.28
4	경남 창원	2018-2차	○※ 건립사업	복합	618	0.29
5	서울 본청	2018-3차	∨× 건립사업	복합	1,663	0.37
6	경기 용인	2019-2차	◇☆ 건립사업	복합	592	0.34
7	대구 본청	2019-3차	※△도서관 건립사업	복합	1,003	0.46
8	경기 김포	2019-3차	◇※센터 건립사업	복합	591	0.46
9	서울 본청	2020-1차	△△도서관 건립사업	단일	2,341	0.23
10	서울 본청	2020-1차	××도서관 건립사업	단일	805	0.31
11	서울 본청	2020-1차	※※도서관 건립사업	단일	822	0.67
12	서울 중랑	2020-2차	☆× 건립사업	복합	876	0.51
13	대구 달성	2020-2차	※◇ 사업	복합	850	0.11

주 : 1) 사업비는 의뢰서 기준이며, 검토안 기준으로 본다면 다음과 같음 : ※△도서관(716억 원), ∨×도서관(2,210억 원), ○○도서관(1,223억 원), ××도서관(602억 원), ※※도서관(813억 원)

2) ※△ 건립사업은 도서관 건립, 공원조성, 공영주차장 건립으로 구성된 사업으로, 전체 사업비는 1,003억 원이며, 도서관 사업비는 694억 원임

3) ×※ 조성사업은 시의회, 보건소, 중앙도서관, 시민문화복지관, 공원, 지하주차장 건립으로 구성된 사업으로, 전체사업비는 1,821억 원이며, 이 중 도서관 사업비는 350억 원임

도서관 단일 시설로 의뢰된 사업의 평균 B/C는 0.43이며, 방법론 상 2건은 조건부가치측정법(CVM)으로 추정하였고, 나머지 2건은 객단가 방식으로 추정하되 사용가치를 시장단가와 설문에서 구한 이용에 대한 지불의향금액을 사용하였다. 이에 따라 객단가 방식의 조사에서는 이용수요를 서비스별로 추정하였으며, 2건은 중력모형으로, 2건은 회귀분석을 활용하여 추정하였다. 또한 복합사업 내 포함된 도서관의 경우 중력모형으로 4건, 나머지는 유사시설이용법(원단위법 등)을 활용하여 수요를 추정하였다. 아래의 표에 정리된 바와 같이 그간 LIMAC 타당성조사에서 각 사업별 특성을 고려하여 다양한 방법을 사용한 것으로 판단되나, 현 시점에서 조사의 일관성과 객관성 측면에서 수요 및 편익 추정방법에 대한 검토가 필요한 것으로 판단된다. 예를 들어, 중력모형을 활용하여 수요 추정시 쟁점은 준거시설 및 각 준거시설별 영향권과 분석대상시설의 영향권과 관련된다. 회귀분석의 경우, 영향권 문제에서 자유로울 수 있는 장점이 있으나, 분석시 모형, 변수의 선정 등에 있어서 추가적인 연구가 필요한 상황이다.

표 2-11 LIMAC 수행 도서관시설 타당성조사 현황(단일시설)

사업명	연면적(㎡)	세부항목	수요추정		편익추정				
			총수요(준거시설)	신규수요	방법	세부내용			
						이용수요	객단가		비고(유사시설 등)
OO도서관	16,500	도서관 이용	중력모형 (인천 미추홀도서관, 대전 한밭도서관, 대구시립중앙도서관	설문조사	CVM	-	-	-	경기도 전체 세대 수, 세금
△△도서관	35,000	도서관 이용	중력모형 (국립중앙도서관, 국회도서관, 서울도서관)	설문조사	CVM	-	-	-	서울시 전체 세대수, 세금
××도서관	10,520	열람실(자료실)	회귀분석 (지역내 11개 공공도서관)	설문조사	객단가	추정수요-기존시설 이용자수	WTP	시장단가	내 사설 스터디센터 또는 독서실
		도서대출서비스				추정수요	WTP	시장단가	국가상호대차서비스 이용료
		행사/교육(강좌)					WTP	시장단가	백화점 교육·문화강좌
		도서관 내 특화기능					WTP	설문조사	-
※※도서관	11,020	자료이용(도서)	설문조사, 회귀분석 (서울시 공공도서관)	설문조사	객단가	추정수요	WTP	설문조사	-
		공간이용(열람실)							
		교육/전시 프로그램이용							

표 2-12 LIMAC 수행 도서관시설 타당성조사 현황(복합시설)

사업명	연면적(㎡)	세부항목	수요추정		편익추정				
			총수요(준거시설)	신규수요	방법	세부내용			
						이용수요	객단가		비고(유사시설 등)
※△ 대표도서관	14,953	도서대출	중력모형 (지역 내 5개 도서관)	설문조사	객단가	방문인원 대비 세부시설 이용비율 (준거시설)	WTP	시장단가	국가상호대차서비스 이용료
		도서열람실					WTP	시장단가	프리미엄 독서실
		문화프로그램					WTP	시장단가	백화점 인문 프로그램
		기타 부대시설					WTP	설문조사	-
		보존서고 확보	-		임대료 절감편익				상업용부동산 임대동향, 오피스 임대료(한국감정원, 2019)
○△ 건립사업	550	장애인도서관 도서열람실 이용	장애인 인구 및 장애인도서활동실태조사 등 통계 활용	-	객단가	추정수요	WTP	선행연구	한국문화관광정책연구원(2006) 문화체육관광부(2009) 표순희(2012)
○※ 건립사업	4,094	도서관 시설이용	유사시설[원단위법] (지역 내 15개 도서관)	설문조사	객단가	신규수요	WTP	설문조사	-
		교육/행사 참여							
×※ 조성사업	10,112	시설 이용	유사시설[원단위법) (유사 자치단체 중앙도서관 3개)	설문조사	객단가	신규수요	WTP	시장단가	민간 스터디센터 및 스터디카페
		교육 참여	유사시설 (기존 중앙도서관)				WTP	설문조사	-

사업명	연면적(㎡)	세부항목	수요추정		편익추정				
			총수요(준거시설)	신규수요	방법	세부내용			
						이용수요	객단가		비고(유사시설 등)
∨× 건립사업	325	도서관 이용	유사시설[원단위법] (●● 소재 공립작은도서관)	-	객단가	추정수요	WTP	선행연구	정혜경·정은주(2007). 공공도서관의 가치평가를 위한 가상가치평가법
◇☆ 건립사업	3,000	열람석 이용	중력모형 (지역 내 4개 도서관)	설문조사	객단가	신규수요	WTP	시장단가	독서실 일일 이용료
		자료실 이용					WTP	설문조사	-
		독서/문화프로그램					WTP	시장단가	백화점 인문학 강좌 수강료
		도서대출 이용					WTP	시장단가	국가상호대차서비스 이용료
◇※ 건립사업	2,770	시설 이용	유사시설[원단위법] (지역 내 2개 도서관)	설문조사	객단가	신규수요	WTP	시장단가	민간 스터디센터 및 스터디카페
		프로그램 참여					WTP	설문조사	-
☆× 건립사업	1,856	도서대출 이용	중력모형 (지역 내 3개 도서관)	설문조사	객단가	신규수요	WTP	시장단가	국가상호대차서비스 이용료
		자료실 이용					WTP	설문조사	-
		단순시설 이용					WTP	시장단가	독서실 이용료
		프로그램 이용					WTP	시장단가	민간문화센터 수강료
※◇ 사업	2,872	자료 및 열람실 이용	중력모형 (지역 내 1개 도서관)	설문조사	객단가	신규수요	WTP	설문조사	-
		문화교육프로그램							-

2. PIMAC 타당성 조사

1999년부터 최근까지 수행한 PIMAC 타당성조사에서 도서관 사업은 총 4건으로 파악된다. 2005년에 수행한 헌법재판소 도서관을 제외한 나머지 도서관은 국립중앙도서관 분관 사업으로 모두 중력모형으로 수요를 추정하였으며, 준거시설로 모두 서울 국립중앙도서관 사례를 활용하였다. PIMAC에서는 다른 유형의 문화사업과 마찬가지로 영향권 설정시 전국을 기준으로 하였다.

표 2-13 PIMAC 수행 도서관시설 타당성조사 현황

No	시설	조사시기	지역	총사업비	수요추정방법	편익추정방법	B/C
1	헌법재판소도서관	2005	서울	353	추세분석법	CVM	0.32
2	국립중앙도서관 부산분관	2007	부산	1,048	중력모형	CVM	0.81
3	국립중앙도서관 광주분관	2010	광주	1,370	중력모형	CVM	0.38
4	국립중앙도서관 부산분관	2010	부산	829.6	중력모형	CVM	0.58

3. 도서관 수요 관련 선행연구

선행연구 중 도서관의 이용수요를 직접적으로 분석한 학술연구는 없는 것으로 파악되며 예비타당성조사 및 지방재정투자사업 타당성조사에서만 직접적으로 시설에 대한 수요추정을 하고 있는 실정이다.

다만 함윤주 외(2019)는 이용자의 거리조락현상을 반영한 중력모형을 안양시 공공도서관 사례에 적용하여 실증분석을 시도하였다. 분석 결과, 지역 내 많은 수의 시설이 기운영되고 있는 도서관의 경우 거리조락현상보다는 각 도서관의 운영특성을 반영하는 것이 수요추정의 예측력을 높이는 것으로 나타났다. 즉 중력모형에서 전제하는 도서관의 이용자 수요가 거리에 반비례한다는 가정은 유효하지 않은 것으로 나타났다. 이는 해당 연구에서 도서관의 지리적 분포를 고려하여 영향권을 도보 15분 내로 설정하였기 때문에 영향권

내에서는 거리에 무차별하게 수요가 발생하는 것으로 추정되었다. 이는 안양시와 같이 도서관 공급이 균일하게 이루어진 경우 거리 외에 다른 요인이 더 중요할 수 있음을 시사한다. 특히 기존 중력모형에서 매력도로 규모만을 반영하였다면 해당 연구에서는 규모 측면에서뿐만 아니라 장서 수, 인력, 예산, 운영 프로그램 등 다양한 특성을 반영하였으며, 그 결과 예측 수요량의 오차율이 낮아지는 것으로 나타났다.

〈표 2-14〉에 요약된 바와 같이 학술연구에서는 도서관 수요에 영향을 미치는 요인을 분석하기 위해 주로 한국도서관협회의 『한국도서관연감』, 문화체육관광부의 『공공도서관 통계』 또는 자체적으로 수행한 설문조사 및 수집자료를 분석하였다.

특히 이학준·이용관(2019)은 개관 직후 도서관 이용자 수와 대출권 수에 미치는 영향을 분석하였다. 분석 결과, 개관 직후에는 이전수요가 나타날 수 있으나 점차 이전효과가 상쇄되고 신규 이용자가 창출하는 것으로 나타났다.

도서관의 경우 상기 수요 관련 요인분석 외에 운영과 관련한 효율성을 분석한 자료포락분석((Data Envelopment Analysis, 이하 DEA) 연구가 다수인 것으로 나타났다. 포락분석은 유사조직 간의 상대적 능률성을 측정하는 대표적인 기법으로 도서관의 경우 도서수, 예산 등 투입요소가 이용자수 등 산출요소에 얼마나 영향을 미치는지 분석이 가능한 방법이다. 도서관 DEA 분석 연구에서도 앞서 살펴본 선행연구에서 주로 선정한 설명변수인 건물면적, 도서자료수, 직원수, 예산 등 도서관의 운영상 특징과 지역주민수 등을 설정하여 분석하고 있는 것으로 나타났다(함윤주 외, 2019, p. 186).

따라서 상기 선행연구를 바탕으로 도서관의 다양한 특성을 반영하여 수요를 추정할 수 있는 방법을 모색할 필요가 있겠다. 특히 그간 국가도서관 통계시스템에 축적된 도서관 관련 데이터를 탐색하고, 이를 바탕으로 수요추정 방법을 모색해 볼 필요가 있다.

표 2-14 도서관 이용자 수 및 수요추정 관련 선행연구

선행연구	활용자료	종속변수	영향을 미치는 변수
최희곤 (2009)	한국도서관연감 (2008)	도서관 이용자 수	사서 수(+), 연간 증가 책수(+), 자료구입비 예산(+), 도서관 전체 예산(+), 도서관 및 독서관련 프로그램 실시횟수(+), 좌석 수(+), 건물연면적(+), 개관일수(+), 개관경과년(+)

선행연구	활용자료	종속변수	영향을 미치는 변수
원종준·안건혁 (2010)	한국도서관연감 (2006) 서울시중심지 체계 현황도(2006)	도서관 이용자 수	도서관의 건물면적(+), 일회성 문화행사(+), 봉사대상 행정동의 인구밀도(+), 500m 내 지하철역의 수(+), 대형점포시설과의 연계성(+), 지역 중심 입지(+)
장훈 (2018)	온라인 조사 (1,000부)	도서관 이용여부	연령(+), 영유가 가구원(+), 아동 가구원(+)
전계형·권선영 (2018)	공공도서관 통계(08~17), 인구총조사 자료, 지역총생산, 실업률 자료	도서관 방문자 수	도서관 수(−), 인쇄자료 수(+), 개관시간(+), 성인 회원 수(+), 15세 미만 인구비율(+), 65세 이상 인구비율(+)
		자료실 이용자 수	도서관 수(−), 인쇄자료 수(+), 개관시간(+), 성인 회원 수(+), 15세 미만 인구비율(+), 65세 이상 인구비율(+), 고학력자 비율(−)
		대출자 수 (어린이 / 청소년 / 성인)	도서관 수(+), 개관시간(+), 어린이 회원 수(+), 성인 회원(−), 실업률(+), RGDP(−), 15세 미만 인구비율(+), 고학력자 비율(+)
이학준·이용관 (2019)	도서관 통계조사 (07~18)	이용자 수	공공도서관 개관(+), 1관당 도서관 자료 수(+), 미성년자 비중(+), 인구수(+)
		대출권수	공공도서관 개관(+), 1관당 도서관 자료 수(+), 1관당 도서관 총예산(+), 인구수(+), 미성년자 비중(+), 고령자 비중(−),

주 : 통계적으로 유의미하게 추정된 변수에 한하여 정리함
출처 : 함윤주·홍근석·주운현(2021) 〈표 3〉 일부 수정

제3장

통계적 학습을 통한 도서관 방문자수 예측

공공도서관의
수요추정 모형 개발

KRILA
KOREA RESEARCH INSTITUTE FOR
LOCAL ADMINISTRATION

제3장 통계적 학습을 통한 도서관 방문자수 예측

제1절 개요

본 보고서의 주요 목적은 도서관의 수요 특성을 고려한 적절한 수요추정 방법 모색에 있다. 앞서 본 보고서 제2장에서 살펴본 바와 같이 도서관 이용 수요 추정을 위한 학술적 연구는 부재하며, 다만 각 도서관의 운영상 다양한 특성을 고려하여 수요에 영향을 미치는 요인 분석이 이루어져 왔다. 또한, 타당성조사에서는 중력모형, 회귀분석 등의 방법으로 수요를 추정하였으나, 여러 모형과 방법론 사이에 최소 오차와 최대 정확도를 갖춘 수요 예측을 위해 기초연구가 더 필요한 실정이다.

특히 2008년도부터 매년 국가도서관 통계시스템[3]에 전국 도서관(「도서관법」 제2조에 정의한 도서관)의 각종 현황(시설규모, 설립년도 등 기본정보, 소장자료, 예산현황, 이용 및 이용자 등) 통계가 구축되어 이를 활용한 다양한 수요 예측 모형 탐색이 가능해졌다. 국가도서관 통계시스템상 도서관은 국립도서관, 공공도서관, 작은도서관, 학교도서관, 대학도서관, 기타도서관(장애인도서관, 병영도서관, 교도소도서관, 전문도서관)으로 구분되어 통계가 취합 및 관리되고 있다. 상기 도서관 유형 중 지방재정투자사업으로 추진되는 도서관은 단일 건축물로서 주로 공공도서관이 해당되며, 작은도서관이나 기타도서관 중 장애인도서관 등도 복합청사나 복합공공복지시설 내 일부 시설물로 추진되기도 한다.

본 연구 제3장에서는 도서관 유형 중 지방재정투자사업에서 대규모 사업으로 수요 예측이 타당성 평가에서 중요한 부분을 차지하고 있는 공공도서관를 대상으로 국가도서관 통계시스템 상의 데이터를 활용하여 수요 예측을 위한 효율적이고 효과적인 방법을 모색하

3) 국가도서관 통계시스템은 국가승인통계인 "전국도서관통계"의 작성과 결과를 서비스하기 위해 구축된 시스템이다.

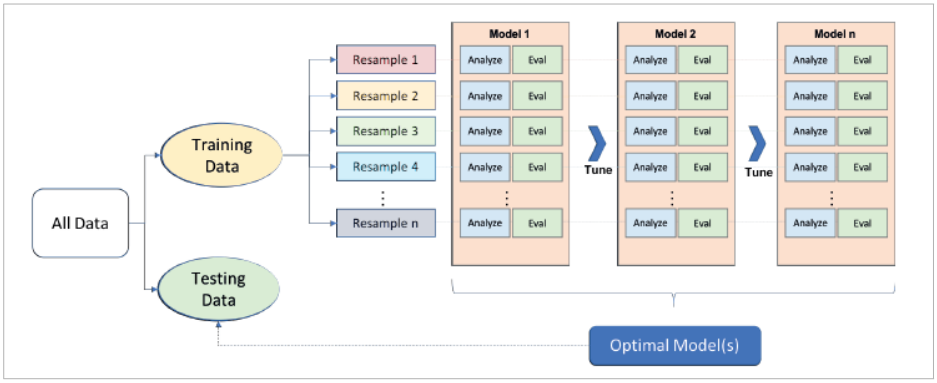
고자 한다. 특히 새로운 컴퓨팅 기술 발전과 함께 빅데이터 분석에 널리 활용되고 있으며, 특히 데이터 예측 분석(predictive data analysis)를 위해 주로 활용되고 있는 머신러닝(machine learning) 또는 통계적 학습(statistical learning)을 이용하여 공공도서관 수요예측을 시도하고자 한다.

수요 예측방법으로는 타당성조사에서 활용되는 중력모형 외에 전통적으로 시계열모델이나 회귀모델이 있으며, 이를 활용하여 도서관 데이터의 수요 패턴 분석을 시도할 수 있다. 그러나 이러한 전통적인 방법은 다양한 분야에서도 수요예측기법으로 사용되고 있으나, 적절한 변수와 모델을 선택하는 것이 어려운 일이다. 머신러닝 등 인공지능기반 수요예측법은 예측의 정확도를 최대한 높이면서, 독립변수와 종속변수 간 관계를 나타내는 모델을 설정할 필요가 없는 장점이 있다. 따라서 최근 다양한 분야에서 이러한 기존 방법론의 문제를 극복하고 동시에 수요예측의 오차를 줄이고자 인공지능 기반 수요예측기법을 시도해왔다(정혜린·임창원, 2019).

머신러닝(machine learning) 또는 통계적 학습(statistical learning)을 통한 데이터 예측 분석(predictive data analysis)은 알고리즘을 이용하여 과거 데이터의 패턴을 자동으로 학습(learn)하고, 이를 바탕으로 미래의 결과를 예측하는 기술이다. 수많은 머신러닝 알고리즘 중 어떤 알고리즘이 수요예측을 위한 최적의 모델(optimal model)인지는 직접 적용해 보기 전까지는 알 수 없으며, 여러 모형을 반복적으로 적용 및 평가하는 과정을 통해 결정된다(Boehmke and Greenwell, 2020). 머신러닝 알고리즘이 데이터를 학습하는 과정에서 모델이 얼마나 잘 학습되는지 평가하기 위해서는 전체 데이터를 학습 데이터(training dataset)와 테스트 데이터(testing dataset)으로 나누어 다음 [그림 3-1]과 같은 과정을 거친다.

본 장에서는 머신러닝의 일환인 통계적 학습(Statistical Learning)의 주요 개념에 대한 소개 후, 다양한 머신러닝 알고리즘을 통해 도서관 수요예측을 시도하여 최적의 모델을 찾아가고자 한다.

그림 3-1 머신러닝을 활용한 데이터 분석 예측 과정



출처 : Boehmke and Greenwell(2020), Figure 2.1(p. 14)

제2절 통계적 학습의 주요 개념

통계적 학습(statistical learning)은 알고리즘을 통하여 자동으로 목표 데이터로부터 의사결정에 필요한 통찰력을 제공하는 일련의 과정을 말한다. 이 과정은 현대 컴퓨터의 계산 능력 확장과 광범위한 보급에 따라 과거와 달리 국가에 버금가는 기관이 아니더라도 수행이 가능한 일이 되었다. 따라서 통계학습 수행을 위한 지식이 어느 정도 있다면 개인부터 조직까지 여러 차원에서 데이터 기반으로 최적의 의사결정을 할 수 있게 되었다.

본 연구의 목표인 도서관 방문자 수 예측도 데이터 기반 최적 의사결정의 영역에 들어간다. 따라서, 본 연구를 위한 통계학습 수행을 위한 기본 지식을 본 절에서 제공하고자 한다.

예를 들어, 어떤 Y(quantitative response)와 p개의 설명변수(predictors) $X=(X_1, X_2, ..., X_p)$를 가지고 관련된 예측(prediction)이나 추론(inference)을 하기 위해서는 둘 사이의 관계를 규명하는 것이 필요하다. X와 Y 사이의 있을 수 있는 관계는 가장 일반적인 형태로 다음과 같이 나타낼 수 있다.

$$Y=f(X)+\epsilon$$

여기서 ϵ는 무작위 오차(random error)를, f는 X가 제공하는 Y에 관한 체계적인 정보를 나타낸다. 본 연구에서는 ϵ의 평균이 0이고 X와 서로 독립적이라고 가정한다.

통계적 학습(statistical learning)은 체계적인 정보를 제공하는 부분인 f를 추정하는 일련의 접근법을 통칭하는 말이다. 통계적 학습이 수행하는 바를 X와 Y를 가지고 우리가 하고자 하는 예측(prediction)이나 추론(inference) 맥락에서 구체적으로 살펴보면 다음과 같다.

1) 예측(prediction)

통계적 학습 과정에서 대부분의 경우 X는 쉽게 얻을 수 있으나 대응하는 Y는 얻기가

쉽지가 않다. 이런 경우 X를 통해 Y에 대한 예측을 하고자 하는 것은 자연스러운 과정이나, X를 통해 Y를 추정하는 관계식인 f를 사전에 정확하게 알기 어렵다. 다만 어떤 경로를 통해 추정된 $\hat{f}$가 있다면 무작위 오차(random error)의 평균이 0이기 때문에, 다음과 같이 Y의 예측치 $\hat{Y}$를 구해볼 수 있다.

$$\hat{Y} = \hat{f}(X)$$

여기서 주목할 점은 예측(prediction) 맥락에서는 f의 정확한 형태를 알아내는 것 보다는 추정된 $\hat{f}$가 가능한 한 정확한 추정을 하는 것이 중요하다는 것이다. 즉, $\hat{f}$는 블랙박스(black box)로 볼 수 있다. 이를 좀 더 정확히 표현해 보기 위해서 $\hat{f}$가 어떤 방법을 통해 추정되었다고 가정하자. 그러면 우리는 예측치 $\hat{Y}= \hat{f}(X)$를 얻고 실측치와 예측치의 차이에 관한 다음과 같은 관계식을 얻는다.

$$\begin{aligned} E[(Y-\hat{Y})^2] &= E[(f(X)+\epsilon-\hat{f}(X))^2] \\ &= [f(X)-\hat{f}(X)]^2 + Var(\epsilon) \end{aligned}$$

예측의 정확도를 제고하기 위해서 결국 우리의 목적은 실측치와 예측치의 차이를 최소화하는 $\hat{f}$를 구하는 방법을 알아내는 것이다. 다만, $Var(\epsilon)$는 무작위 오차(random error)가 전체 오차에 공헌하는 부분이기 때문에 어떠한 $\hat{f}$를 사용하더라도 줄일 수 없다. 즉 $Var(\epsilon)$는 줄일 수 없는 오차(irreducible error)로 예측치의 정확도에 관한 하한치(lower bound)가 된다.

따라서 예측의 맥락에서 우리의 최종 목표는 줄일 수 있는 오차(reducible error)를 최소화 하는 것이다. 즉, $E[(f(X)-\hat{f}(X))^2]$를 최소화하는 $\hat{f}$을 제시하는 방법을 찾아내는 것이다. 이 방법을 찾아가는 과학적인 접근이 통계적 학습(statistical learning)이다.

2) 추론(inference)

관측치 X, Y를 가지고 실제값 Y에 대한 예측을 하는 것이 아닌 X가 변할 때 Y가 어떻게 영향을 받는지를 아는 것이 중요할 때가 종종 있다. 예를 들어, 어떤 제품을 생산하는 회사가 제품광고를 TV, 라디오, 그리고, 스폰서쉽을 통해서 한다고 가정해보자. 이때, 광고 선전비 포트폴리오를 어떻게 구성하는지는 중요한 의사결정 문제다. 이 문제에 대한 최적의 의사결정을 위해서는 구성비(X) 변화에 대해 제품 판매량(Y)이 어떻게 영향을 받는지 알아야 한다. 즉, X, Y를 통해 추론(inference)을 해야 한다. 여기서 주목할 점은 예측(prediction)과 달리 추론(inference)을 할 때는 Y의 변화가 X의 함수로 변화하기 때문에 $\hat{f}$를 블랙박스(black box)로 보지 않는 것이다. 이러한 맥락에서 통계적 학습(statistical learning)은 효과적인 추론이 가능하면서 설명력이 강한 $\hat{f}$을 찾아가는 과학적인 접근을 말한다.

3) 추정(estimation)

실제 $\hat{f}$을 추정할 때는 목적이 예측(prediction)인지 아니면 추론(inference)인지에 따라서 이에 맞는 통계 학습 방법론(statistical learning method : SLM)을 사용해야 한다. 이런 관점에서 SLM을 크게 모수적(parametric) 방법과 비모수적(non-parametric) 방법으로 구분할 수 있다.

(1) 모수적 방법(parametric method)

모수적 방법(parametric method)은 2단계 모형 기반의 접근법으로, 첫 번째 단계에서는 함수형태를 결정한다. 예를 들어, 함수형태를 다음과 같이 가정해 보자.

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

모수적 방법의 두 번째 단계는 모수 $\beta_0, \beta_1, .., \beta_p$를 추정하는 것이다.

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

모수적 접근법은 결정된 함수형태에 따라 다르겠으나 많은 경우 추론(inference)에 적합한 강한 설명력을 갖는다. 다만, 예측력은 떨어지는 경우가 많다. 특히 위에서 예를 든 선형모형의 경우 사용되는 데이터에 따라 차이가 있겠으나 근본적으로 선형모형은 예측력보다는 설명력이 강한 모형이다.

(2) 비모수적 방법(non-parametric method)

비모수적 방법(non-parametric method)은 함수형태에 대한 구체적인 가정을 하지 않는다. 그렇기 때문에 모수적(parametric) 접근법 대비 추정시 더 많은 대안을 고려할 수 있어 더 강한 예측력을 보여줄 잠재력이 있다. 다만 이를 위해서는 일반적으로 모수적 접근법을 통한 추정보다 훨씬 많은 관측치가 필요하다. 또한, 함수형태를 특정하지 않기 때문에 설명력도 모수적 방법 보다 낮을 가능성이 크다. 즉, 그것이 모수적이든 비모수적이든지 간에 목적과 무관하게 모든 데이터에 대해 다른 방법 보다 좋은 성과를 보이는 방법은 존재하지 않는다(no free lunch theorem[4]). 그러므로, 목적에 맞게 설명력과 예측력 사이의 균형을 설정할 필요가 있다.

4) 예측 정확성과 모형 해석력 사이의 균형상황(trade off) 설정

비모수적(non-parametric) 방법과 모수적(parametric) 접근법의 차이를 보면 결국 두 접근법이 가지는 대안 집합의 크기가 결정적임을 알 수 있다. 예를 들어, 선형모형에서는 결국 유한차원의 벡터공간이 대안 집합이고 그 선택 기준이 최소자승(least square)을 포함하여 무엇이든지 간에 결국 그 유한차원의 벡터공간에서 하나의 점을 선택하는 것이다.

이에 비해, 비모수적(non-parametric) 접근법은 무한차원의 벡터공간이 대안 집합이고 이 공간은 결국 유한차원의 대안 집합을 포함한다. 따라서 예측 오차를 줄이는 정량적인

4) David Wolpert와 William Macready가 1997년에 발표한 "No Free Lunch Tehrems for Optimization"에 실린 것으로 머신러닝에서 특정 문제에 최적화된 알고리즘이 다른 문제에서는 좋은 결과를 보여줄 수 없다는 것을 수학적으로 정리하였다.

기준에 의해서 선택된 해는 비모수적 접근법을 통하는 것이 모수적 접근법을 사용하는 것보다 더 예측 오차가 작을 수밖에 없다. 즉, 덜 제한적이거나 더 유연한 접근법을 사용할수록 예측 오차가 작을 가능성이 있는 것이다. 그러나 모형의 설명력으로 보면 제한적인 접근법, 예를 들어, 선형모형 등이 더 나을 수 있다. 그러므로 접근법에 따른 상충되는 측면(trade off)을 이해하고 이를 고려하여 목적에 맞는 통계 학습 방법(statistical learning method)을 선택해야 한다.

5) 편향-분산 분해(bias-variance decomposition)

예측모형에 있어서, 예측치와 실측치와의 관계인 편향(bias)과 예측치 간의 관계인 분산(variance)은 모형의 과대적합(overfitting) 또는 과소적합(underfitting)과 연결되어 서로 상충되는 역할을 한다. 따라서, 이 둘의 영향을 받아 수요예측을 위한 알고리즘의 성능이 떨어지는지 확인하는 과정은 매우 중요하다. 다음에서는 편향-분산 분해(bias-variance decomposition)를 통해 이러한 수요예측 과정에서 상충관계(trade-off)를 살펴보고자 한다.

이를 위해 $P(X, Y)$로부터 독립동일분포($i.i.d.$: independently and identically distributed)로 추출된 데이터 집합 $D = \{(X_1, y_1), (X_2, y_2), ..., (X_n, y_n) : y \in R\}$이 주어졌다고 가정하자. 이때 X는 데이터를, y는 목표값(target value) 또는 라벨(label, 정답값)을 의미하며, 다음은 기댓값과 기대함수를 정의한다.

① 기댓값(Expected Label)

주어진 $X \in R^d$에 대하여 우리가 얻을 수 있는 라벨(Label)의 기댓값 $\bar{y}$를

$$\bar{y}(X) = E_{y|X}[Y] = \int_y yPr(y|X)\partial y$$

로 정의한다.

② 테스트 오차 기댓값(Expected Test Error)

주어진 데이터 집합 D에 선택된 알고리즘 A를 학습(training)시키면, $h_D = A(D)$이다. 이렇게 해서 주어진 h_D를 가지고 일반화 오차(generalization error)를 다음과 같이 계산할 수 있다.

$$E_{(X,y)\sim P}[(h_D - y)^2] = \iint_{(x,y)} (h_D(X) - y)^2 \Pr(X,y)\partial y \partial x.$$

③ 분류기 기댓값(Expected Classifier)

데이터 집합 D 역시 결합확률분포(joint distribution) P^n을 확률분포(distribution)로로 갖는 확률변수(random variable)이고 h_D는 D의 함수이므로 h_D 역시 확률 변수로 볼 수 있다. 따라서 h_D의 기댓값을 다음과 같이 구할 수 있다.

$$\overline{h} = E_{D\sim P^n}[h_D] = \int_D h_D \Pr(D)\partial D.$$

마지막으로 선택된 알고리즘의 성능(quality)을 측정하기 위해 다음과 같이 테스트 오차(Test Error)를 정의한다.

④ Expected Test Error

$$E_{(X,y)\sim P, D\sim P^n}[(h_D(X) - y)^2] = \iiint_{(D,X,y)} (h_D(X) - y)^2 \Pr(X,y)\Pr(D)\partial X \partial y \partial D.$$

앞서 정의된 값들을 가지고 다음과 같이 편향(bias)과 분산(vaiance)을 분해(decomposition)하는 과정을 전개할 수 있다.

⑤ Bias-Variance Decomposition

$$E_{X,y,D}[(h_D(X)-y)^2]$$
$$= E_{X,D}[(h_D(X)-\overline{h}(X))^2] + E_{X,y}[(\overline{y}(X)-y)^2] + E_X[(\overline{h}(X)-\overline{y}(X))^2]$$

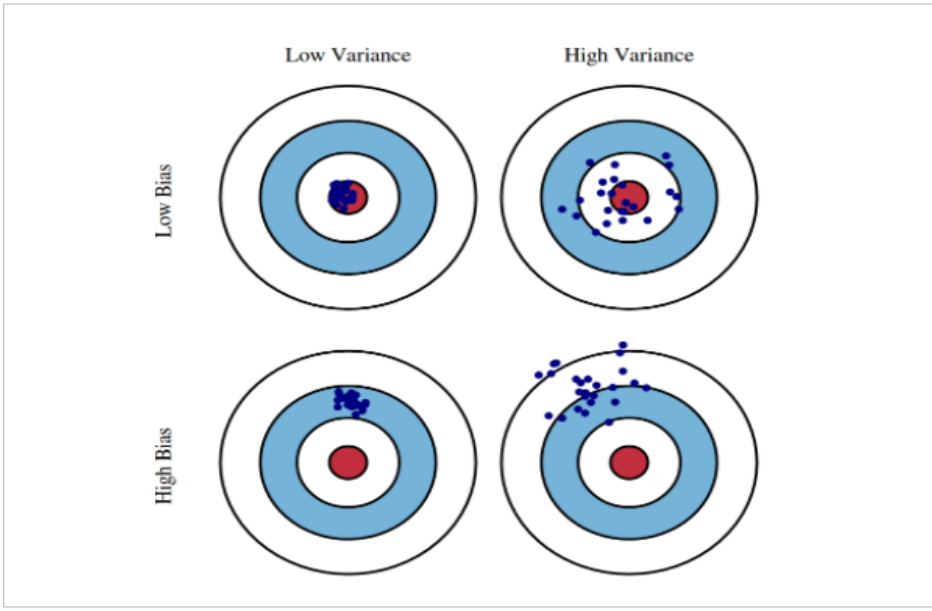
각 항이 가지는 의미를 다음과 같이 표현할 수 있다.

$$\text{Expected Test Error} = \textit{Variance} + \textit{Noise} + \textit{Bias}$$

분산(variance)은 "만약에 선택된 알고리즘을 주어진 데이터 집합이 아닌 다른 집합에 훈련(traning)시킨다면 함수 h_D가 어떻게 얼마나 변하는가?"에 대한 의미있는 답을 주며, 편향(bias)은 "선택된 알고리즘에 의한 h_D에 내재된 오차(error)는 무엇인가?"에 대한 답을 준다. 잡음(noise)은 우리가 다룰 수 없는 데이터 고유의 모호성의 크기를 나타낸다. 결국, 통계적 학습(statistical learning)을 사용하여 예측을 한다는 것은 분산(variance)과 편향(bias)을 최대한 줄여가며 테스트 오차의 기댓값(expected test error)을 최소화하는 알고리즘을 선택하는 문제로 귀결된다.

다음 그림에서는 분산(variance)과 편향(bias)으로 인해 발생하는 과대적합(overfitting) 또는 과소적합(underfitting)의 관계를 볼 수 있다. 우선 분산(variance)이 크다는 것은 예측 모델이 학습 데이터(training data)에 비해 지나치게 복합하다는 것을 의미하며, 학습 데이터에 지나치게 적합(fitting)시켜 일반화가 되지 않았을 시 발생한다. 편향(bias)은 반대로 지나치게 단순한 모델을 사용하여 예측값이 정답으로부터 상당히 떨어져 있음을 의미한다. 즉 모델이 복잡해질수록 편향은 작아지나 분산이 커지고, 반대로 모델이 단순해지면 분산은 작아지나 편향이 커지는 현상이 발생하므로 오류를 최소화하기 위해서는 편향과 분산의 합, 즉 테스트 오차의 기댓값(expected test error)을 최소화하기 위한 노력이 필요하다.

그림 3-2 분산-편향 상충관계(Bias-Variance tradeoff)



출처 : http://scott.fortmann-roe.com

제3절 분석 결과

1. 분석 과정

머신러닝(machine learning)의 알고리즘은 크게 지도학습(supervised learning)과 비지도학습(non-supervised learning), 강화학습(reinforcement learning) 세 가지 유형으로 나뉜다.

지도학습은 정답(label)이 있는 데이터를 활용하여 학습(training)시킨 후, 테스트 데이터(test data)를 분류하거나 예측하는 학습이다. 지도학습의 대표적인 알고리즘으로는 의사결정트리(decision tree), 회귀분석(regression analysis), KNN(K-nearest neighbor), 서포트벡터머신(support vector machine : SVM) 등이 있다. 비지도학습은 지도학습과 달리 정답(label)이 없는 데이터를 군집화하여 특징을 분류하는 등 결과를 도출하는 방법이다. 대표적인 방법으로 클러스터링(clustering)이 있다. 강화학습은 보상(reward)에 기반하여 컴퓨터가 스스로 문제를 해결하는 방법으로 알파고 등이 대표적인 예이다.

본 연구에서는 도서관의 과거 데이터를 활용하여 수요의 정확성을 높일 수 있는 예측모델 개발에 있으므로, 지도학습의 주요 알고리즘인 의사결정트리, 회귀분석 등을 활용하여 주어진 과거 데이터를 수요 예측의 정답(label)으로 설정하고, 예측 모델의 수정과정을 통해 최적의 수요예측 모델을 찾고자 한다.

2. 분석 자료

본 연구에서는 문화체육부의 국가도서관 통계시스템에서 공개되고 있는 전국 공공도서관 자료 중 데이터의 크기 및 품질 측면에서 머신러닝을 활용한 수요예측 결과가 유의미하게 해석될 수 있는 서울시 내 총 25개 자치구에 위치한 공공도서관의 2008~2018년 데이터를 활용하고자 한다. 2008~2018년 내 새로 설립되거나 폐관된 곳을 포함하여, 서울시 내 공공도서관은 총 210개소이며, 총 관측치 수는 1,378개이다. 해당 공공도서관 데이터에 포함된 도서관은 설립주체별로 지자체도서관, 사립도서관, 교육청도서관으로 나뉜다.

조사된 도서관의 연간 평균 방문자수는 486,000여명이었다. 도서관 연간 방문자 수의 분포는 정규분포를 따르지 않고, 아래 그래프와 같이 비대칭이 심한 분포(right-skewed)를 보였다. 도서관의 연면적은 2,196㎡ 이며, 좌석 수는 328석, 도서관 당 평균 8만여권의 장서를 보유하고 있는 것으로 집계되었다. 자료구입비, 운영비, 그리고 인건비는 통계 연도 집행 결산액이며, 천 원 단위이다. 도서관당 연간 개관일 수는 300일에 조금 못 미치며, 일주일 평균 개관시간은 71시간 정도였다. 공공도서관이지만 설립 주체가 사립인 경우는 3.7%, 어린이 도서관은 총 도서관 수의 12.1%를 차지했다.

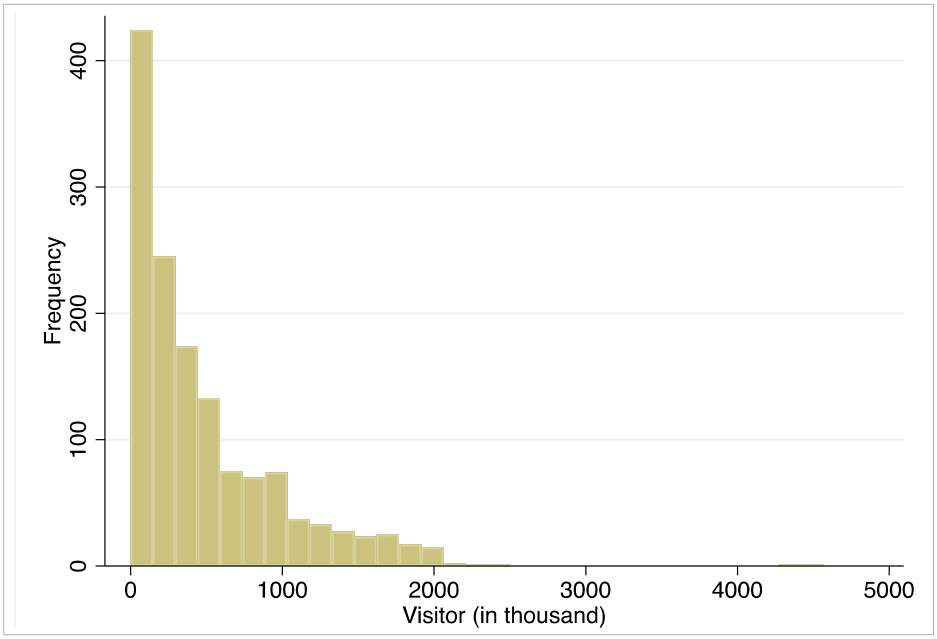
지역구 별 인구 수와 도서관 수는 지역구가 관측 단위이지만, 해당 지역구 내 도서관에 같은 값을 부여하므로, 도서관이 관측 단위인 다른 변수들과 관측 횟수가 동일하다. 인구 변수는 구 별 총 인구 수와, 15~64세 사이의 인구 수를 모두 사용하였다. 구 내 도서관 수는 최소 1관에서 최대 13관이었고, 평균 도서관 수는 6.276개인 것으로 나타났다.

표 3-1 기술통계량

변수	내용	샘플 수	평균	표준편차	최소값	최대값
visitor	연간 방문자 수	1,378	486,170.3	500,984.18	0	4,570,261
area	연면적(㎡)	1,378	2,196.823	2,450.085	0	20,229
seats	좌석 수	1,378	328.914	327.55	0	2,146
books	장서 수	1,378	83,818.035	87,643.505	0	532,536
computers	이용자 컴퓨터(대)	1,378	27.52	26.467	0	141
cost material	자료구입비(천 원)	1,378	84,887.536	95,939.953	0	1,683,260
cost op	운영비(천 원)	1,378	290,705.08	379,049.47	0	6,651,878
wage	인건비(천 원)	1,378	431,811.71	698,362.82	0	17,607,610
days	연간 개관일 수	1,378	298.768	51.282	6	365
hour	평균 주간 개관 시간	1,378	71.226	17.001	0	142
private	사립	1,378	0.037	0.189	0	1
child	어린이 도서관	1,378	0.121	0.326	0	1
center	평생학습관	1,378	0.04	0.196	0	1
eng	영어도서관	1,378	0.01	0.1	0	1

변수	내용	샘플 수	평균	표준편차	최소값	최대값
pop	구 당 인구 (15-64세)	1,378	324,190.64	103,781.27	93,096	529,637
totalpop	구 당 총 인구	1,378	428,042.23	134,080.76	125,249	685,279
nolib	구 당 도서관 수	1,378	6.276	2.88	1	13

그림 3-3 방문자수 히스토그램



3. 도서관 방문자 수 예측

1) 의사결정나무 (decision tree)

의사결정나무 모형(Tree-based model)은 일련의 분할 규칙 또는 특정 기준에 의해 데이터를 구분한다. 이때, 해당 규칙이나 기준에 의해 종속변수 값이 유사하며 겹치지 않게

최대한 작은 영역으로 구분하는 알고리즘을 사용한다. 각 영역에서 예측값은 그 영역에서 종속변수 값의 평균 등 매우 간단한 모형을 통해 얻게 된다.

의사결정나무를 만들어내는 많은 방법론 중에서 가장 잘 알려진 것은 Breiman(1984)이 제안한 CART(classification and regression tree) 알고리즘이다. 이 알고리즘을 통해 만들어진 의사결정나무는 학습 데이터(traning data)를 유사한 종속변수 값을 갖는 겹치지 않는 하위그룹으로 나누고 각 하위그룹에 그 그룹에 속한 종속변수 값의 평균 등의 추정값을 부과한다. 각 하위그룹은 알고리즘 중단 조건이 만족할 때까지 미리 정해진 분할 규칙을 가지고 반복이진분류(binary recursive partitioning)를 통해 생성된다. 분류가 완료된 후 새로운 설명변수(predictor)를 가정하고, 이에 대응하는 종속변수에 대한 예측치는 새 설명변수(predictor) 값을 가지고 분할규칙에 따라 CART 나무를 타고 도착한 하위그룹에 부과된 값이 된다. 다음에서는 의사결정나무 모형을 구성하는 요소와 개념들을 살펴보고, 도서관 데이터를 사용하여 도서관 방문자 수 예측을 위한 의사결정나무를 도출하였다.

(1) 분할

CART는 각 하위그룹으로의 분할 방법으로 반복이진분류(binary recursive partitioning)을 사용하여, 각 분기점(node)에서 남아있는 데이터를 양분하면서 나무를 만든다. 연속형 변수에 대한 구체적인 분류기준은 다음과 같다. 먼저, 각 설명변수(predictor)에 대해서 총 오차제곱합(sum of squared error : SSE)[5]을 가장 작게 만드는 분할 영역을 찾고 그렇게 설명변수(predictor)별 계산된 SSE를 비교하여 가장 작은 SSE를 주는 설명변수를 선택한다. 다만, 이 과정에서 한 설명변수가 분할을 위해 여러 분기점에서 사용될 수 있다는 점을 주의해야 한다. 극단적으로 나무를 만들어 갈 때 첫 분기점에서 마지막 분기점까지의 모든 분할에 하나의 설명변수만이 사용되는 것도 가능하다.

(2) 나무의 크기(complexity)

앞서 설명한 분할규칙을 가지고 다른 조건 없이 만들어진 나무는 일반적으로 그 나무 생성에 사용된 데이터에 속하지 않는 새로운 설명변수(predictor)값에 대응하는 종속변수

5) $SSE = \sum_{i \in R_1} (y_i - c_1)^2 \sum_{i \in R_2} (y_i - c_2)^2$

값에 대한 예측력이 떨어지게 된다. 이를 과대적합(overfitting)이라고 하며, 이러한 과대적합을 막기 위한 가장 일반적인 접근법으로는 사전적으로 정해진 규칙에 따른 이른 종료(early stopping)와 사후적인 가지치기(pruning)가 있다.

이른 종료(early stopping)는 명시적으로 나무의 크기에 제한을 두는 접근법이다. 가장 널리 쓰이는 접근법으로는 나무의 깊이(maxdepth)에 구체적인 제한을 두는 것과 각 잎(leaf) 또는 최종 노드(termianl node)에 있을 수 있는 관측치 개수(minsplit)에 제한을 두는 것이다. 가지치기(pruning)는 먼저 어떤 제한을 두지 않고 큰 나무(T_0)를 만든 이후에 사후적으로 적절하게 가지를 잘라내는 방법이다. 가지를 적절하게 잘라내는 데 여러 가지 방법이 있으나 가장 널리 쓰이는 방법은 비용복잡도 가지치기(cost complexity pruning)이다. 이 방법은 다음과 같은 관찰에서 출발한다. 주어진 임의의 α(0보다 크거나 같은 실수)에 대해서 다음의 값(E)을 최소화하는 subtree $T(T \subseteq T_0)$가 유일하게 존재한다고 하자(Breiman, 1984).

$$E = SSE + \alpha|T| = \sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha|T|,$$

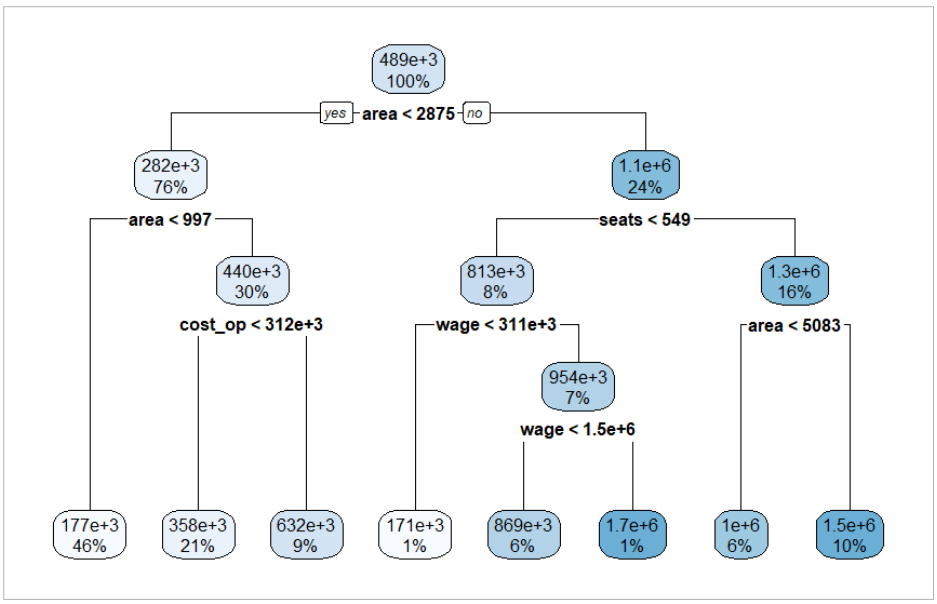
여기서 $|T|$, R_m, $\hat{y}_{R_m}$은 각각 나무 T의 노드(leaf 또는 termial node) 수, m번째 잎(leaf)의 영역, 그리고, 그 영역에서의 추정 종속변수를 나타낸다. 여기서 비용복잡도 매개변수(cost complexity parameter) α는 잎(leaf)의 개수가 많은 나무를 가지는 데 따른 대가로 해석할 수 있다. 즉, α가 커짐에 따라 E를 최소화하는 나무는 그 크기가 작아지려는 경향이 있다. 즉, 나무의 크기가 커지면 오차제곱합(SSE)의 감소하는 양이 비용복잡도에 따른 패널티(cost complexity penalty) 보다 커야 한다.

(3) 실행 결과

위에서 언급된 방법 및 절차를 본 연구 대상인 서울시 공공도서관 데이터 집합에 적용하자. 이를 위해 도서관 데이터 집합을 학습(training)과 테스트(test) 집합으로 나눈다. 학습(training) 집합은 나무 생성에 사용하고, 나무가 완성되면 테스트(test) 집합의 설명변수

(predictor)를 사용하여 예측치를 구하고 다시 테스트(test) 집합 내 설명변수(predictor)에 대응하는 실제 방문자 수와 비교를 한다. 비교 수단으로는 평균 제곱근 편차(root mean squared error : RMSE)를 사용한다. 다만, 실행에 앞서 CART의 초모수(hyperparameter)[6]인 maxdepth, minsplit, 그리고, α를 RMSE를 줄이는 방향으로 미세조정을 해야 한다. R을 통한 자동화된 격자 탐색(grid search)을 통해 maxdepth=12, minsplit=8, 그리고 α=0.01이 실험 중 가장 낮은 오차를 제시하는 초모수값으로(<표 3-2>) 이를 가지고 CART 알고리즘을 학습(training)시킨 후 테스트한 결과 RMSE가 254,473.9이 도출되었다. [그림 3-4]는 CART 방문자 예측 나무 도출 결과이다.

그림 3-4 CART 방문자 예측 나무



6) 주어진 자료로부터 추정되지 않고 알고리즘 실행을 위해서 고정되어야 하는 모수를 말한다.

표 3-2 CART 초모수 격자 탐색

Top 10	minsplit	maxdepth	cp	error
1	16	12	0.01	0.3169820
2	12	6	0.01	0.3216894
3	15	11	0.01047189	0.3231612
4	8	4	0.01	0.3236970
5	16	11	0.0104789	0.3237405
6	13	15	0.01166318	0.3239150
7	7	8	0.01166318	0.3239332
8	14	14	0.01047189	0.3239868
9	10	10	0.01047189	0.3248563
10	8	5	0.01	0.3255435

[그림 3-5]와 [그림 3-6]은 앞서 언급한 초모수(hyperparameter)를 찾기 위하여 자동화된 격자탐색(grid search)을 알고리즘이 수행하는 동안 축적된 초모수(hyperparameter) 사이의 관계에 대한 축적된 정보를 보여준다.

그림 3-5 가지치기 복잡도 매개변수(Pruning complexity parameter : cp)

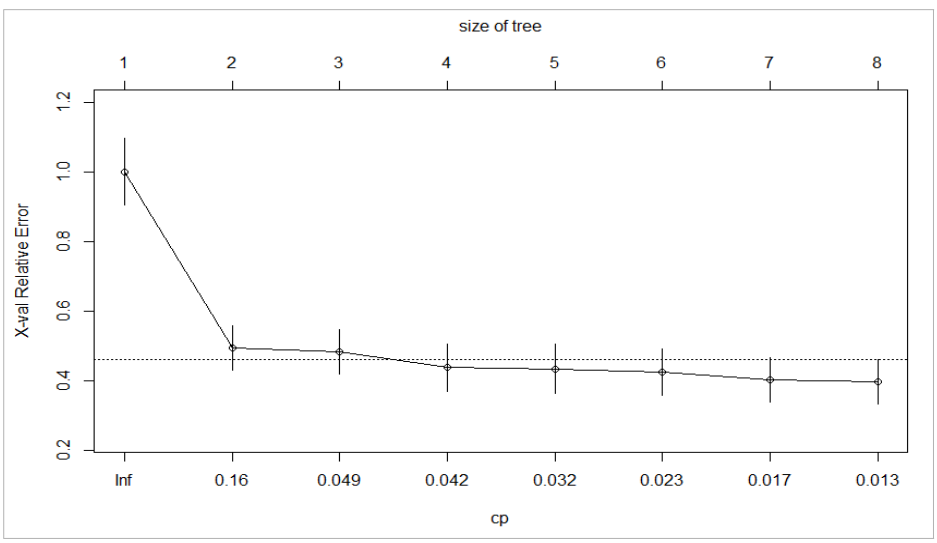
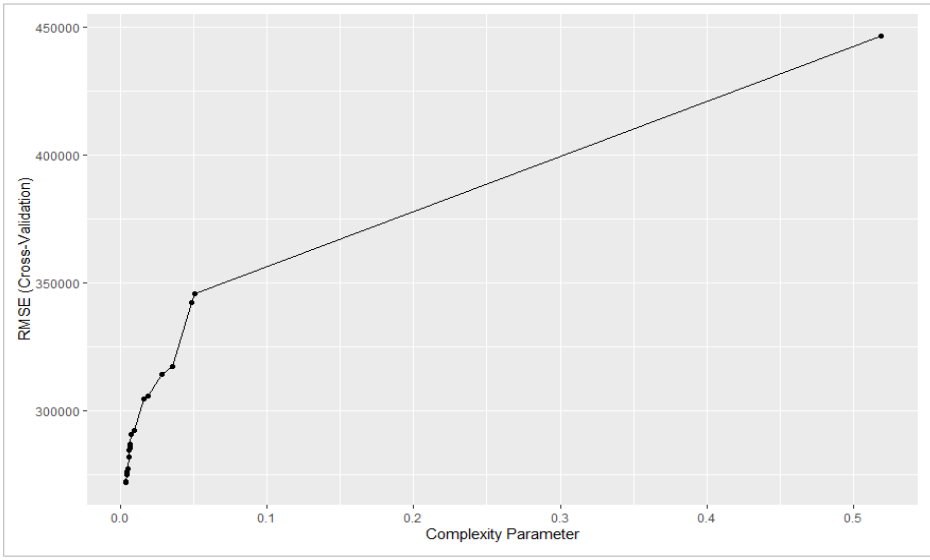
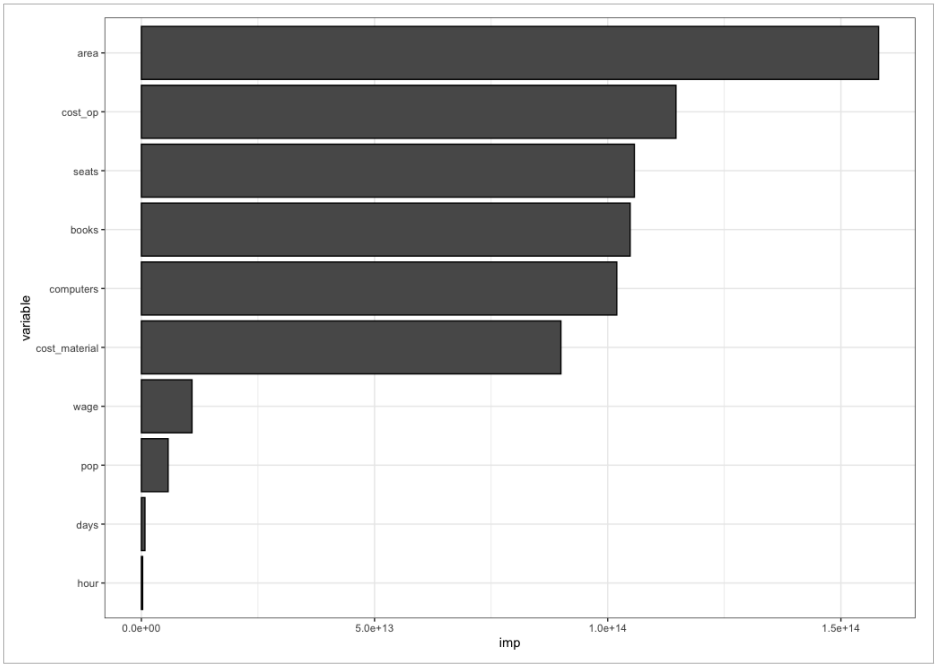


그림 3-6 교차 검증 정확도율(Cross-validated accuracy rate)



마지막으로 CART를 사용하여 도서관 방문자 수 예측을 위한 나무를 생성할 때 자연스러운 질문은 각 설명변수(predictor)가 알고리즘이 제시한 예측치에 미치는 중요성(공헌정도)이 어떻게 되는가이다. 즉, 예측 결과에 대한 해석을 추정된 나무 모형이 줄 수 있는가는 중요한 질문이다. 이에 대한 답을 위해 각 분기점에서 선택된 설명변수와 계산된 SSE를 표준화한 결과를 [그림3-7]에서 제시하였다. 설명변수 중 area(면적)이 공공도서관 수요예측을 위한 나무 구축에 가장 큰 기여를 한 것으로 나타났다. 그 외에 운영비(cost-op), 좌석 수(seats), 장서 수(books) 등의 중요도가 그 다음으로 높은 것으로 나타났다. 인건비(wage), 인구(pop), 연관 개관일수(days), 개관시간(hours) 등은 나무 구축에 거의 기여하지 않는 것으로 나타났다. 특히 개관시관과 개관일수는 도서관별 운영시간과 운영일수가 크게 다르지 않다는 점을 고려하면 직관에 부합하는 결과로 해석된다.

그림 3-7 CART 설명변수 중요도



이러한 의사결정나무 알고리즘은 실행이 간단하고 해석이 쉽다는 강점이 있으나 예측력은 다른 여러 알고리즘에 비교해 낮은 편이다. 의사결정나무 기반(tree-based) 모형의 이러한 부족한 예측력을 보완하기 위하여 여러 가지 방법이 제안되었고 가장 널리 받아들여지는 아이디어는 많은 나무들을 생성하여 이들의 앙상블(ensemble)을 통하여 주어진 설명변수(predictor) 값이 대응하는 종속변수의 예측치를 만드는 것이다. 여러 나무를 만들어내는 방법과 만들어진 여러 나무를 합치는 방법에 따라 많은 앙상블(ensemble)이 있다. 본 연구에서는 배깅(bagging)(Breiman 1996a), 랜덤포레스트(random forest)(Breiman 2001), 그리고 부스팅(boosting)(Friedman 2001, 2002), 이 세 가지 방법을 가지고 도서관 수요를 예측하였다.

2) 배깅(Bagging)

(1) 소개

부트스트랩 집계(bootstrap aggregating) 또는 배깅(bagging)의 아이디어는 간단하다. 먼저 1절에서 본 편향-분산 분해(bias-variance decomposition)를 다시 생각해보자.

$$E_{X,y,D}[(h_D(X)-y)^2]$$
$$= E_{X,D}[(h_D(X)-\overline{h}(X))^2] + E_{X,y}[(\overline{y}(X)-y)^2] + E_X[(\overline{h}(X)-\overline{y}(X))^2]$$

해석하면, 기대 테스트 오차는 분산과 잡음, 그리고 편향의 합으로 나타난다.

$$\text{기대테스트오차} = \text{분산} + Noise + \text{편향}$$

그러면 의사결정나무를 통한 예측에 대해 이 방법의 분산(variance) 또는 편향(bias)을 줄여서 결국 테스트 오차(test error)를 줄일 수 있는 의사결정나무 알고리즘의 변종을 탐색하는 것은 자연스러운 일이다. 여러 변종 중 편향(bias)은 변하지 않으나 분산(variance)은 줄이는 방법이 바로 배깅(bagging)이다.

구체적으로, 약한 대수의 법칙(Weak Law of Large Numbers)에 의해 평균이 $\overline{x}$인 확률변수($i.i.d.$ random variables) x_i가 m이 커짐에 따라 $\frac{1}{m}\sum_{i=1}^{m} x_i$은 $\overline{x}$로 수렴한다. 이것을 우리의 경우에 적용해 보자. 확률분포 P^n으로부터 추출된 m개의 학습(training) 집합 $D_1, D_2, ..., D_m$이 주어졌다고 가정하자. 그러면 우리가 기대하는 것은 m이 커짐에 따라 $\hat{h} = \frac{1}{m}\sum_{i=1}^{m} h_{D_i}$가 $\overline{h}$로 수렴하는 것이며, 분산(variance)은 m이 커짐에 따라 사라질 것이다. 문제는 우리에게는 오직 하나의 D만 주어졌다는 것이다. 이를 해결하기 위해서 부트스트랩(bootstrap)을 통하여 대체(replacement)를 허락하며 균등분포에 따라 $D_1, D_2, ..., D_m$를 생

성한다. 그리고 이를 기반으로 배깅 추정치(bagging estimator) 또는 앙상블(ensemble)을 앞의 $\hat{h}$과 같이 정의한다. 다만 여기서 주의할 점은 이 앙상블(ensemble)이 직접 약한 대수의 법칙에 기댈 수는 없다는 것이다. 이는 생성된 부트스트랩(bootstrap) $D_1, D_2, ..., D_m$이 서로 독립이 아닐 수 있기 때문이다. 다만 D_i는 D와 같은 확률분포을 가진다는 것을 증명할 수 있다. 그러나, $D_1, D_2, ..., D_m$는 분산(variance)을 줄이는 게 가능할 정도로 서로 '독립'되어 있다.

배깅 알고리즘의 실행은 간단하다. 주어진 학습 데이터를 가지고 m개의 부트스트랩 복사본(bootstrap cop)를 만들고 각 복사본에 대해 CART 알고리즘을 실행하여 가지치기를 하지 않고 완전히 성장된 기본 학습자 나무(base learner tree)를 생성한다. 그러면 기본 학습자 나무별 예측치는 덜 편향되고 더 분산되는 특징을 갖는다. 그러면 앞의 논의에 따라 최종적인 예측치를 각 기본 학습자(base learner)로부터의 예측치의 평균으로 정의하면 덜 편향되고 동시에 덜 분산된(low bias and variance) 예측을 하게 된다. 그러므로 배깅(bagging) 알고리즘은 데이터의 변화에 강건하지 않은 의사결정나무 기반(tree-based) 방법에 효과적이다.

그러나 배깅(bagging)도 문제가 있다. 모형을 만들어가는 과정은 독립적이지만 개별 부트스트랩(bootstrap)으로부터 생성된 나무들은 서로 완전히 독립적이지 않고 비슷한 구조를 갖게 된다(tree correlation). 이는 각 나무 생성에서 모든 설명변수(predictor)가 분할에 사용되기 때문이다. 이를 극복하기 위한 변종은 다음 절에서 소개하고자 한다.

(2) 실행 결과

이제 배깅(bagging)을 서울시 공공도서관 데이터 집합에 적용하자. 이를 위해 우선 도서관 데이터 집합을 학습(training)과 테스트(test) 집합으로 나눈다. 학습(training) 집합은 모형을 만드는 데 사용되고, 테스트(test) 집합은 모형의 정확성을 판단하는 데 사용한다. 설명변수(predictor)를 이용하여 예측치를 구하고 다시 테스트(test) 집합 내 설명변수(predictor) 값에 대응하는 실제 방문자 수와 비교하였다. 비교 수단으로는 평균 제곱근 편차(root mean squared error : RMSE)를 사용한다. 실제 실행에서는 몇 개의 나무를 만들 것인가를 결정하는 것이 핵심이다. 이론적으로는 나무를 많이 만들수록 예측치의 분산

이 줄어들어 더 정확한 예측을 하게 된다. 그러나 이론을 실행하는 컴퓨터는 본래 계산에 근사치를 만들어내는 한계가 있다. 이런 한계는 이론과는 별개로 줄일 수 없는 오차가 있음을 뜻한다. 그래서 나무를 늘려가는 과정의 초반에는 급격한 오차의 감소를 보이다가 나무 수가 일정 수준을 넘어가면 오차의 감소가 일어나지 않는 현상이 일반적이다. 그러므로 목표 데이터집합에 대해서 컴퓨터 실험을 통해서 최적의 생성 나무 개수를 정해야 한다. [그림 3-8]는 나무 개수를 50개에서 500개까지 증가시켜 갈 때 RMSE 감소를 보여준다. 이를 바탕으로 100개에서 200개 사이에 어딘가에서 배깅 알고리즘 실행을 위한 최적 나무 개수가 있다는 것을 보여준다. [그림 3-9]는 100에서 200 구간에서 RMSE의 감소를 보여준다.

그림 3-8 배깅 나무수 증가와 RMSE 감소(50~500개)

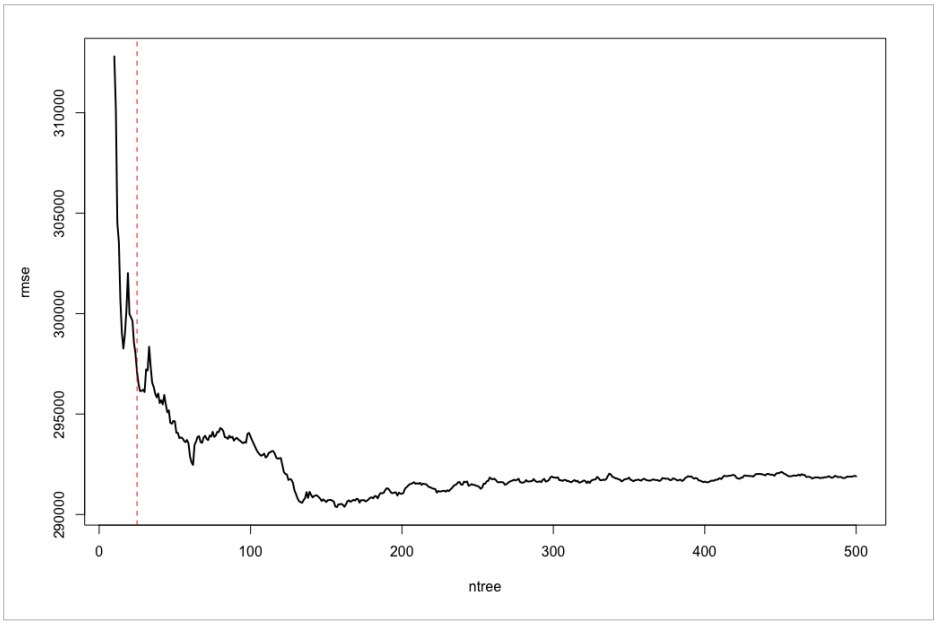
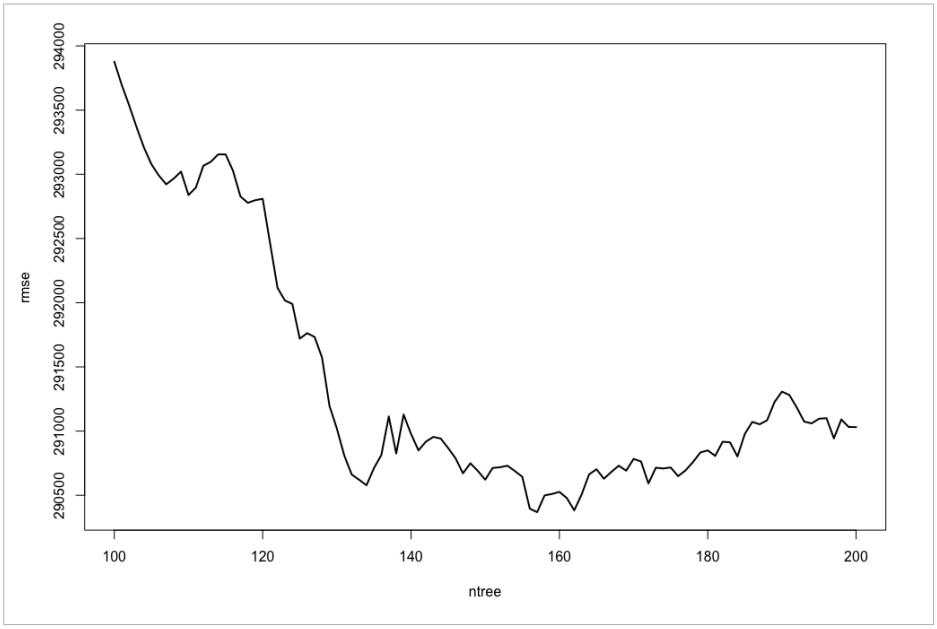
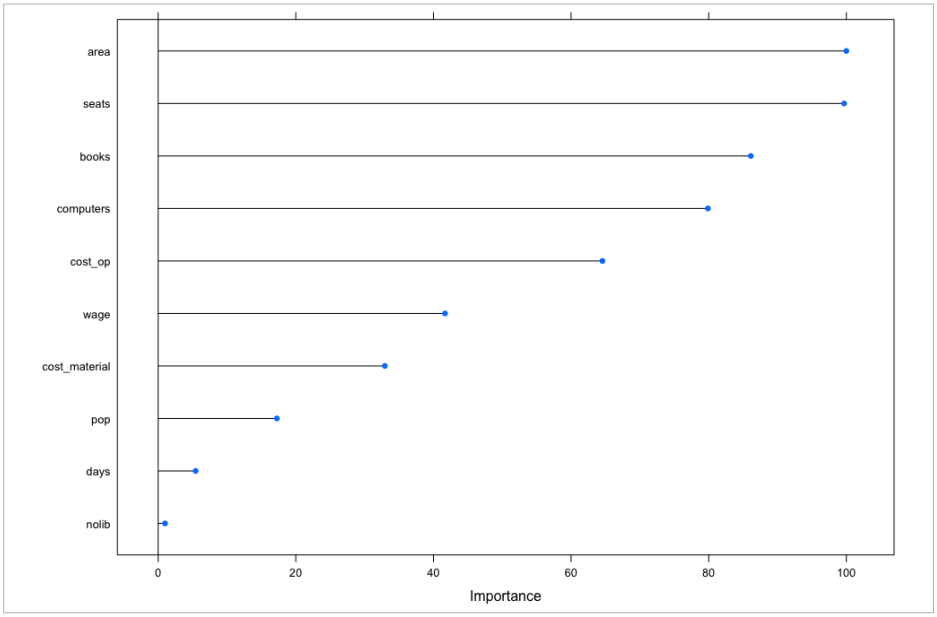


그림 3-9 배깅 나무수 증가와 RMSE 감소(100~200개)



본 연구에서는 그림의 결과를 바탕으로 도서관 방문자 수 예측을 위한 배깅 알고리즘에서 필요한 나무 개수를 160으로 정하였다. [그림 3-10]은 배깅 알고리즘을 통한 도서관 방문자 수 예측치 추정에 설명변수들의 중요도를 보여준다. 배깅은 이와 같은 절차를 따라서 단일 나무 생성을 통한 방문자 예측보다 더 분산이 줄어든 예측치를 제공한다. 그러나 앞서 언급했듯이 각 나무 분할에 사용되는 설명변수 집합이 동일하고 각 나무들에 사용되는 분할기준이 똑같아서 나무들이 서로간에 독립이기 힘든 현상이 나타난다. 이러한 현상을 나무상관성(tree-correlation)이라고 한다. 이론적으로 독립적일수록 분산의 감소가 더 커지므로 배깅 맥락에서 나무들이 서로 간 더 독립적이게 만들 수 있는 방법으로 더 큰 분산의 감소를 추구하는 것은 자연스러운 다음 단계가 될 것이다. 그러한 방향으로 나무상관성을 다루기 위한 방법 중 하나가 랜덤포레스트(random forest)이다. 본 연구의 다음 절에서 이 랜덤포레스트를 방문자 수 예측에 적용해 볼 것이다.

그림 3-10 배깅 설명변수 중요도



3) 랜덤 포레스트(random forest)

(1) 소개

랜덤 포레스트(random forest)는 앞서 언급된 나무 상관관계(tree correlation)를 줄이고자 배깅(bagging)을 위한 나무 생성 과정에 소개된 무작위성(randomness)에 더해 추가적인 무작위성(randomness)인 분할 변수 무작위화(split-variable randomization)을 도입한다. 분할변수 무작위화(split-variable randommization)란 배깅(bagging)을 거쳐 나무를 생성하는 동안 분할을 할 때마다 사용되는 분할 변수를 전체 설명변수(predictor) 집합의 무작위(random) 부분집합에서 고르도록 제한하는 것이다. 그러면, 배깅을 이루는 각 나무가 추가적인 무작위성을 도입하지 않았을 때보다 더 '독립적'(de-correlated)이게 되어 예측치 추정에서 분산 감소의 효과가 있게 된다. 다만, 분할 변수 무작위화를 위해 사용되어야 하는 설명변수의 부분집합을 결정하는 데 일반적인 방법은 없다. 그래서, 무작위로

추출하게 되는 설명변수의 개수를 초모수로 두고 전산실험을 통해 실험적으로 분할 무작위화를 위한 최적 설명변수 개수를 결정한다.

(2) 실행 결과

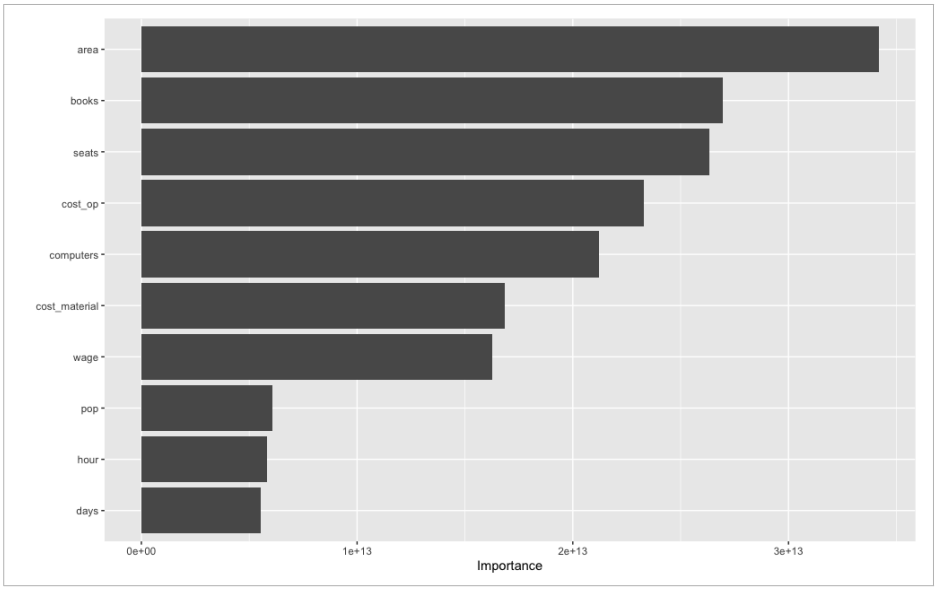
랜덤포레스트를 서울시 공공도서관 데이터 집합에 적용하기 위해 우선 도서관 데이터 집합을 학습(training)과 테스트(test) 집합으로 나눈다. 학습(training) 집합은 모형을 만드는 데 사용되고, 테스트(test) 집합은 모형의 정확성을 판단하는 데 사용한다. 설명변수(predictor)를 이용하여 예측치를 구하고 다시 테스트(test) 집합 내 설명변수(predictor) 값에 대응하는 실제 방문자 수와 비교를 한다. 비교 수단으로는 평균 제곱근 편차(root mean squared error : RMSE)를 사용하기로 한다. 다만. 실행에 앞서 랜덤포레스트의 초모수(hyperparameter)인 mtry(각 분할에서 사용되는 설명변수의 개수), min.node.size(각 나무의 복잡도), numtree(나무의 개수), sample.fraction 등을 RMSE를 줄이는 방향으로 미세조정을 해야 한다. sample.fraction은 학습집합에 있는 모든 관측치를 사용하지 않고 일부를 사용할 때 그 양을 정하는 모수이다. R을 통한 자동화된 격자 탐색(grid search)을 통해 〈표3-3〉의 OOB(Out-Of-Bag) RMSE 기준 상위 10개의 초모수 조합을 얻었다. OOB는 부트스트랩(bootstrap) 샘플링 과정에서 추출되지 않는 관측치로, 테스트(test) 집합에서의 오분류율을 예측하는 용도 및 변수 중요도를 추정하는 용도로 사용된다. 표의 첫 번째 행의 초모수 값을 가지고 랜덤포레스트 알고리즘을 학습(training)시킨 후 테스트한 결과 RMSE는 171,242로 보고 되었다. [그림 3-11]은 랜덤포레스트 알고리즘을 통한 예측에서 그 예측치에 계산에 기여한 중요성을 보여준다. 배깅과 마찬가지로 예측치 생성에 area(연면적) 변수가 가장 크게 기여하여 중요성이 제일 높음을 알 수 있다.

표 3-3 랜덤포레스트 초모수 격자탐색

Top 10	mtry	min.node.size	sample.fraction	numtree	OOB_RMSE
1	2	3	0.8	400	246,229.4
2	2	3	0.8	600	246,576.9
3	2	3	0.8	500	246,690.5

Top 10	mtry	min.node.size	sample.fraction	numtree	OOB_RMSE
4	4	3	0.8	1000	246,916.5
5	2	3	0.8	300	246,997.6
6	2	3	0.8	700	247,033.2
7	0	3	0.8	400	247,174.6
8	3	3	0.8	400	247,174.6
9	4	3	0.8	900	247,191.3
10	2	3	0.8	800	247,313.4

그림 3-11 **랜덤포레스트 설명변수 중요도**



4) 그래디언트 부스팅(gradient boosting)

(1) 소개

예측을 위한 머신러닝(ML) 알고리즘은 선형회귀와 같이 예측을 한번 수행하는 모형과 배깅(bagging)이나 랜덤포레스트(random forests)와 같이 여러 번 예측을 수행하여 앙상블

(ensemble)하는 방법이 있다. 부스팅(boosting)은 둘 중 앙상블(ensemble) 방법의 한 종류이다. 그러나 부스팅(boosting)방법은 다른 앙상블(ensemble)방법인 배깅(bagging)이나 랜덤포레스트(random forests)와 본질적인 차이가 있다. 부스팅(boosting)은 점진적이고 순차적으로(sequentially) 새로운 모형들을 생성하고 더해간다. 즉, 부스팅은 강력한 단일 모형을 생성하는 대신 여러 단순 모형을 추가하면서 강력한 복합 모형을 향해 점진적으로 나아가는 것이다. 핵심은 약한 모델을 더해 갈 때 '어느 방향으로 얼마만큼'을 더해 갈 것인지 잘 정하는 것이다.

그래디언트 부스팅(gradient boosting)은 부스팅(boosting) 아이디어를 다음과 같이 구현한다. 먼저 약한 학습자(weak learner)로 초기 예측치들을 생성하고 이에 따르는 잔차(현재 예측치와 실측치 사이의 차이)를 가지고 약한 학습자(weak learner)를 학습시켜 잔차에 대한 예측치를 생성한다. 이렇게 약한 학습자(weak learner)에 의해 예측된 잔차를 기존 모형의 예측치에 추가하면 그만큼 모형은 올바른 목표치를 향해 움직인다. 이러한 단계를 여러 차례 거치면서 전체 모형의 예측 정확도를 개선한다. 이를 정리하면 다음과 같다.

1. 데이터를 의사결정나무에 적합(fit)시킨다 : $F_1(x) = y$,
2. 다음 의사결정나무에 이전에 생성된 잔차를 적합(fit)시킨다 :
   $h_1(x) = y - F_1(x)$,
3. 이렇게 생성된 새로운 나무를 우리의 알고리즘에 더한다 :
   $F_2(x) = F_1(x) + h_1(x)$,
4. F_2의 잔차에 다음 의사결정나무를 적합(fit)시킨다 : $h_2(x) = y - F_2(x)$,
5. 새로운 나무에 우리의 알고리즘을 더한다 : $F_3 = F_2(x) + h_1(x)$,
6. 메커니즘이 끝날 때까지 이러한 과정을 반복한다.

결과적으로 다음과 같은 모형을 얻게 된다.

$$F_m(x) = F_{m-1}(x) + h_m(x)$$

여기서 h가 벡터임을 상기하면 h의 방향은 목표한 강력한 복합모형을 향한 방향으로 볼 수 있다. 그러면 그 방향으로 얼마만큼 움직이는 것이 목표한 복합모형을 찾아가는데 효과적인가는 자연스러운 질문이다. h방향으로 움직임의 크기를 반영하는 모형은 다음과 같이 표현할 수 있다.

$$F_m = F_{m-1} + \eta h_m$$

η는 움직임의 크기를 나타내는 모수이나 목표 데이터 집합으로 최적값을 구할 수 없는 격자탐색 등의 컴퓨터 실험을 통해서 적절한 값을 찾아야 하는 초모수(hyperparameter)이다.

마지막으로 위에 소개된 절차 중 잔차 적합 과정에 사용되는 최소화 기준을 소개한다. 가장 흔히 사용되는 기준(loss function)으로는 평균제곱오차(Mean Squared Error : MSE)와 평균절대오차(Mean Absolute Error : MAE)가 있다. 잔차($y - \hat{y}$)에 대해서 MSE와 MAE는 다음과 같이 각각 정의된다.

$$MSE := \sum_i (y_i - \hat{y_i})^2$$

$$MAE := \sum_i |y_i - \hat{y_i}|$$

이 연구에서 MSE와 MAE에 대해서 각각 그래디언트 부스팅 방법을 실행해 본다.

(2) 실행 결과

그래디언트 부스팅을 서울시 공공도서관 데이터 집합에 적용하기 위해 우선 도서관 데이터 집합을 학습(training)과 테스트(test) 집합으로 나누었다. 학습(training) 집합은 모형을 만드는 데 사용되고, 테스트(test) 집합은 모형의 정확성을 판단하는 데 사용한다. 설명변수(predictor)를 이용하여 예측치를 구하고 다시 테스트(test) 집합 내 설명변수(predictor) 값에 대응하는 실제 방문자 수와 비교하였다. 비교 수단으로는 평균 제곱근 편차(root mean squared error : RMSE)를 사용한다. 다만. 실행에 앞서 그래디언트 부스팅의 초모수(hyperparameter)[7]인 Number of trees(NOT), Depth of trees(DOT), Learning

rate(LR), Number of minimum observation in node(NOMO), Subsampling(Sub)을 RMSE를 줄이는 방향으로 미세조정을 해야 한다. Number of trees는 목표 복합 모형을 구성하는 단순 모형의 수, Depth of trees는 각 단순모형에 적용될 분할 횟수, Learning rate은 앞서 소개한 η, Number of minimum observation in node는 나뭇잎에 있어야 할 최소 관측치 수, Subsampling은 학습집합 중 실제 사용되는 비율을 나타낸다[8]. R을 통한 자동화된 격자 탐색(grid search)을 통해 MAE에 대해서는 〈표 3-4〉, MSE에 대해서는 〈표 3-5〉의 실험을 통한 최적 초모수를 얻었다.

표 3-4 그레디언트 부스팅 초모수 격자탐색(MAE)

LR	DOT	NOMO	Sub	NOT	RMSE
0.1	9	5	0.65	3829	316.4846
0.05	9	1	0.65	1847	318.3953
0.1	9	1	0.5	1458	318.7736
0.01	8	1	0.65	4942	319.2039
0.05	7	5	0.8	4989	319.6270
0.05	8	1	0.8	1770	319.8193
0.05	7	1	0.5	2914	319.8566
0.05	8	5	0.65	4926	320.4108
0.05	8	10	0.5	4925	320.6626
0.01	9	1	0.5	4755	320.8267

표 3-5 그레디언트 부스팅 초모수 격자탐색(MSE)

LR	DOT	NOMO	Sub	NOT	RMSE
0.01	9	10	0.5	1298	199499.2
0.01	8	10	0.5	1298	200545.8
0.05	9	5	0.5	408	200949.0

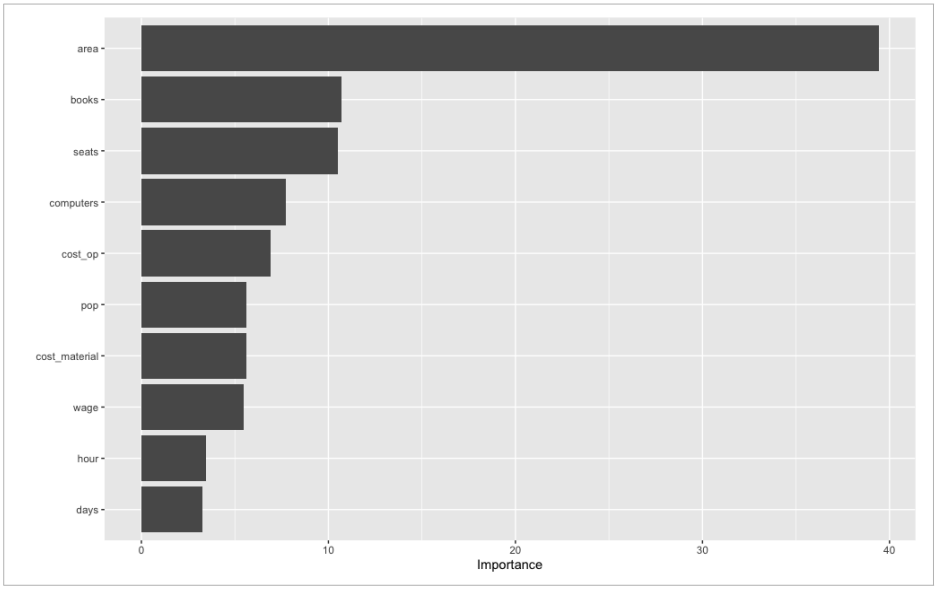
7) 주어진 자료로부터 추정되지 않고 알고리즘 실행을 위해서 고정되어야 하는 모수를 말한다.

8) 학습집합의 관측치를 모두 사용하지 않고 정해진 비율에 따라 무작위로 사용할 관측치를 추출하여 MAE나 MSE 최소화 시 발생 가능한 문제들(local minima 등)을 극복한다.

LR	DOT	NOMO	Sub	NOT	RMSE
0.1	8	15	0.65	163	201943.4
0.1	8	15	0.8	160	202009.9
0.05	8	15	0.8	167	202083.2
0.01	7	10	0.5	1298	202087.0
0.05	8	5	0.5	411	202104.3
0.05	9	15	0.65	277	202384.3
0.01	9	10	0.65	994	202691.0

표의 첫 번째 행의 초모수 값을 가지고 그래디언트 부스팅을 학습(training)시킨 후 테스트한 결과 MAE 사용 시 RMSE가 155,544.5, MSE 사용 시 165,379.4가 보고 되었다. [그림 3-12]로부터 그레디언트 부스팅을 통한 알고리즘 학습에서도 예측치 생성에 area(연면적) 변수가 가장 중요성이 높음을 알 수 있다.

그림 3-12 그레디언트 부스팅 설명변수 중요도



5) 익스트림 그래디언트 부스트(extreme gradiance boost : XGboost)

(1) 소개

본 절에서는 그레디언트 부스팅을 효율적으로 실행하는 방법인 익스트림 그래디언트 부스트를 사용하여 도서관 방문자 수 예측 모형을 만들어 보았다. 익스트림 그래디언트 부스트(extreme gradiance boost : XGboost)는 그레디언트 부스팅 방법을 구성하고 실행하는 이론(예를 들면, 사용된 최적화 모형 등)과 컴퓨터를 통한 실행, 두 가지 측면에서 계산상의 효율성을 높일 수 있는 거의 모든 방법을 생각하여 기존 알고리즘에 반영한 결과물이다. 따라서, 구체적인 개선 사항을 다루는 것은 순전히 수학적이거나 전산과학적인 논의가 되어 본 연구의 방향과는 맞지 않아 참고문헌(Chen과 Guestrin(2016))을 소개하는 것으로 대신한다. 다만, 그러한 개선의 결과는 그레디언트 부스팅과 비교 시 유용하므로 정리가 필요하다. 구체적인 내용은 다음과 같다.

1. k-fold cross validation을 기본 제공한다.
2. 모든 나무의 모든 분할에서 stochastic 그래디언트 부스팅을 실행한다.
3. 효율적인 선형 모형 solver와 나무 학습 알고리즘을 포함한다.
4. 한 대의 컴퓨터에서 parallel computation이 가능하다.
5. regression, classification, ranking 등 다양한 목적 함수를 지원한다.
6. XGboost는 확장가능(scalable) 시스템으로 사용자가 필요한 목적 함수를 스스로 쉽게 정의할 수 있다.

(2) 실행 결과

XGboost를 서울시 공공도서관 데이터 집합에 적용하는 절차는 다음과 같다. 우선 도서관 데이터 집합을 학습(training)과 테스트(test) 집합으로 나누었다. 학습(training) 집합은 모형을 만드는 데 사용되고, 테스트(test) 집합은 모형의 정확성을 판단하는 데 사용한다. 설명변수(predictor)를 이용하여 예측치를 구하고 다시 테스트(test) 집합 내 설명변수(predictor) 값에 대응하는 실제 방문자 수와 비교하였다. 비교 수단으로는 평균 제곱근 편차(root mean squared error : RMSE)를 사용한다. 다만. 실행에 앞서 XGboost의 초모수(hyperparameter)

인 eta, max_depth, min_child_weight, subsample, colsample_bytree, optimal_trees 등을 RMSE를 줄이는 방향으로 미세조정을 수행한다. optimal_trees는 목표 복합 모형을 구성하는 단순 모형의 수, max_depth는 각 단순모형에 적용될 분할 횟수, eta는 학습률(learning rate), min_child_weight는 나뭇잎에 있어야 할 최소 관측치 수, Subsampling은 학습집합 중 실제 사용되는 비율을 나타내고, colsample_bytrees는 각 나무의 생성에서 수행된 모든 분할에서 사용되는 독립변수의 비율[9]을 나타낸다. R을 통한 자동화된 격자 탐색(grid search)을 분석한 결과 〈표 3-6〉의 최적 초모수를 얻었다.

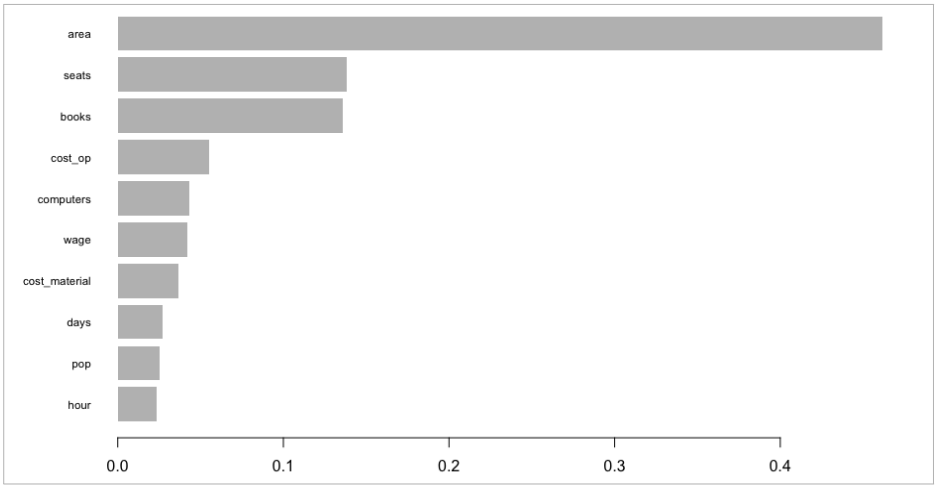
표 3-6 XGboosting 초모수 격자탐색

eta	maxdepth	minchildweight	subsample	colsample	optimaltrees	RMSE
0.05	9	3	0.8	0.8	107	231,073.8
0.05	9	3	0.9	0.65	106	232,329.6
0.03	9	3	0.8	0.65	170	232,475.4
0.05	9	3	0.9	0.9	95	232,587.3
0.05	9	3	1.0	0.65	121	232,807.8
0.01	9	3	0.9	0.8	438	232,829.1
0.01	9	3	0.9	0.65	458	232,830.8
0.07	9	3	0.9	0.65	77	232,845.9
0.01	9	3	0.8	0.65	429	232,862.4
0.03	8	3	0.9	0.65	162	232,910.7

표의 첫 번째 행의 초모수 값을 가지고 XGboost를 학습(training)시킨 후 테스트한 결과 RMSE가 148791.1이 보고되었다. [그림 3-13]은 XGboosting 알고리즘을 통한 예측에서 각 설명변수(feature)가 예측치의 계산에 기여한 공헌도(중요성)을 보여준다. 표로부터 앞서 사용된 알고리즘과 마찬가지로 예측치 생성에 area(연면적) 변수가 가장 큰 공헌을 하여 중요성이 제일 높음을 알 수 있다.

9) 모든 독립변수를 사용하지 않고 정해진 비율에 따라 무작위로 사용할 독립변수를 추출한다.

그림 3-13 XGboosting 설명변수 중요도



6) 회귀분석

(1) 선형 회귀분석

선형 회귀분석은 데이터를 통한 모형 중 고전적인 모형이지만, 가장 널리 사용되는 방법이다. 머신러닝 관점에서 선형 회귀분석은 학습을 수행하는 가장 간단한 알고리즘 중 하나이다. 그러나 현재는 실제 예측에 사용되는 알고리즘으로서의 역할보다는 선형 회귀분석이 예측에 한계를 보이는 복잡한 데이터를 위한 정교한 알고리즘 개발의 좋은 출발점 역할을 하고 있다.

(2) 정규화(Regularization)

선형모형은 간단하고 다루기 쉬운 예측 모형을 제공하고, 모형에 필요한 특정 가정(예 : 상수 분산)이 충족되면 추정된 계수는 편향되지 않으며 모든 선형 편향 추정치 중에서 분산이 가장 낮다는 바람직한 성질을 가지고 있다. 그러나 오늘날 분석이 요구되는 데이터 집합은 대부분의 경우 특정 가정을 만족하지 않는다. 예를 들어, 대부분의 경우 설명변수(feature)가 많이 포함되어 있다. 문제는 설명변수(feature)의 수가 증가함

에 따라, 바람직한 선형모형을 위한 가정은 무너지고 과대적합(overfitting) 등의 문제를 일으키며 결국 모형의 다른 데이터 집합에의 적용, 즉 일반화에 문제가 생긴다는 점이다. 정규화(regularization)는 추정된 계수에 제한을 두어 위와 같은 문제를 해결하는 수단을 제공한다.

구체적으로, 선형 회귀분석의 목적은 관측된 값과 예측된 값 사이의 오차제곱합을 최소화하는 초평면을 찾는 것이다. 그리고 최소자승법(Ordinary Least Squares : OLS)은 특정 가정을 만족하면 매우 바람직한 모형을 제시한다. 그러나 현대의 데이터 과학은 그러한 가정을 만족하지 못하는 데이터 집합을 분석해야 하는 경우를 자주 만난다. 예를 들어, 설명변수(feature)의 개수 p와 관측치 개수 n의 다음과 같은 관계 $p > n$를 가정하자. 그러면, 우리는 OLS 문제에 대해 무한개의 해를 얻게 된다. 그러한 경우 설명변수(feature) 공간에 제약을 두고 OLS 문제를 조건부 최적화 문제로 바꾸면 문제를 해결할 수 있다. 다만. 그런 방식을 따라 얻게 된 모형이 원래의 목적인 예측에도 부합할 것인지는 주어진 데이터 집합의 특성에 달려있다. 더 간단하게 제약이 목적에 부합하는지 미리 알 수가 없다. 다만 $p > n$인 경우와 다중공선성이 강력하게 존재하는 경우나 설명변수가 종속변수와 관련이 뚜렷하지 않은 경우에 조건부 최적화 접근이 모형의 일반화에 성과가 좋은 방법이라는 점은 이해가 가능하다. 그리고 수학적으로 그러한 접근법이 제약이 없는 다음과 같은 최적화 문제와 동일한 것으로 알려져 있다.

$$\min(SSE + P)$$

여기서 SSE는 오차제곱합(Sum of Squared Error)을 말하고, P는 페널티 항을 나타낸다. 그리고 P의 정의에 따라 Ridge, Lasso, 그리고 Elastic net, 이상의 세 가지 정규화된(regularized) 회귀분석이 있다.

(3) Ridge

Ridge 회귀분석은 계수는 다음의 최적화 문제의 해로 주어진다.

$$\min(SSE+\lambda\sum_{j=1}^{p}\beta_j^2)$$

여기서 λ는 주어진 데이터 집합에 따라 정해지지 않는 초모수(hyperparameter)로 페널티의 크기를 결정하는데 결국 테스트 데이터 집합을 이용한 교차검증(cross validation) 등을 통해 실험적으로 정해진다. $\lambda=0$이면 평범한 OLS 문제임을 알 수 있고, λ가 커질수록 페널티는 커지고 따라서 많은 경우 β은 영에 가까워 짐을 알 수 있다. 즉, ridge 회귀분석은 상관관계가 있는 설명변수(feature)끼리 서로 가까워지도록 강제하게 된다. 그러나, Ridge 회귀분석은 설명변수(feature)를 선택하지는 않아서 최종 모형에 모든 설명변수가 남게 된다. 그러므로, 만약에 데이터 집합의 크기가 작고 심각한 다중공선성이 있다고 판단되거나 설명변수의 부분집합이 종속변수와 관계가 없거나 뚜렷하지 않을 때 Ridge 회귀분석은 좋은 선택일 수 있다. 다만, 그렇지 않은 경우, 많은 설명변수(feature)가 의미가 없다고 판단되면 Lasso가 올바른 선택일 수 있다.

(4) Lasso

Lasso 회귀분석은 다음의 최적화 문제로 정의된다.

$$\min(SSE+\lambda\sum_{j=1}^{p}|\beta_j|)$$

Ridge와의 차이점은 수학적으로 Ridge는 L^2norm을, Lasso는 L^1norm을 사용한다는 것이다. 그러면 대응하는 동치인 조건부 최적화 문제에서 실현가능한 영역(feasible region)이 각각 원과 마름모가 되고 Lasso의 경우 최적해, 최종 모형에서 계수가 영이 되어 배제되는 설명변수(feature)들이 생기는 것이다. 즉, Lasso는 모형의 개선과 더불어 설명변수 선택을 수행하는 알고리즘이다.

(4) 실행 결과

선형회귀모형과 이를 바탕으로 개선된 모형인 Ridge와 Lasso 모형을 서울시 공공도서관 데이터 집합에 적용하기 위해 앞서 다른 방법의 실행에서와 같이 우선 도서관 데이터 집합을 학습(training)과 테스트(test) 집합으로 나누었다. 학습(training) 집합은 모형을 만

드는 데 사용되고, 테스트(test) 집합은 모형의 정확성을 판단하는 데 사용한다. 설명변수(predictor)를 이용하여 예측치를 구하고 다시 테스트(test) 집합 내 설명변수(predictor) 값에 대응하는 실제 방문자 수와 비교하였다. 비교 수단으로는 평균 제곱근 편차(root mean squared error : RMSE)를 사용한다. 계산 환경으로 R의 관련 패키지(plmnet 등)을 사용하여 실험한 결과 선형모형은 평균 제곱근 편차가 225,382.2가 보고 되었다. Ridge와 Lasso를 위해서는 컴퓨터 실험을 통한 최적 λ를 먼저 탐색하였다. 그 결과 63095.73와 15848.93이 본 연구의 목표 데이터집합에 대하여 Ridge와 Lasso를 위한 최적 λ로 각각 보고되었다. 보고된 최적 λ를 가지고 알고리즘을 실행한 결과 Ridge는 226,180.5, Lasso는 229,514.9가 테스트 집합의 RMSE로 각각 보고되었다.

선형모형과 비교하여 실행 성과가 낮은 이유는 앞서 소개한 Ridge와 Lasso의 개발 배경에 있다. Ridge와 Lasso는 $n < p$이며 설명변수간 다중공선성이 높거나 설명변수의 부분집합이 종속변수와 관계가 뚜렷하지 않아 잡음(noise) 수준일 때 효과가 있다. 그런데 현재 우리의 목표 데이터집합은 $p << n$인 구조이다. 그러므로 선형모형과 Ridge, Lasso가 그 성과에 큰 차이가 없는 것은 소개된 이론과 부합하는 결과로 볼 수 있다.

제4절 종합

본 장에서는 기계학습을 위한 알고리즘을 통한 도서관 방문자 수 예측 모형을 구축하였다. 사용된 모든 알고리즘은 메타 알고리즘으로서 초모수를 가지고 있다. 따라서 목표 데이터집합인 도서관 통계로부터 컴퓨터를 이용한 격자탐색을 수행하여 최적 초모수 집합을 찾아 고정시키면 그것이 바로 도서관 방문자 수 예측모형이 된다. 구축된 모형이 아래 〈표 3-7〉에 기록되어 있다. 즉, 도서관 방문자 수를 예측하고자 할 때 표에 정리된 초모수를 가지고 학습된 알고리즘에 예측하고자 하는 도서관의 속성을 대입하여 알고리즘을 실행하면 된다.

표 3-7 도서관 방문자 수 예측 모형

구분		tree	bagging	random forests	boosting		xgboosting
					Gaussian	Laplace	
초모수	minsplit	16		3	10	5	3
	maxdepth	12			9	9	9
	nbag		160				
	ntree			400	1,298	3,829	107
	mtry			2			
	shrinkage				0.01	0.1	
	samsize			0.8			
	eta						0.05
	lambda						1
	bag.fraction				0.5	0.65	
	cp	0.01					
	Subsampling						0.8
	colsample_by trees						0.8
구분							1
초모수	lambda	regression	ridge	lasso			
			63,095	15,848			

〈표 3-8〉에서는 test set을 이용하여 수요 예측 결과를 요약하였다. 종합하면, 익스트림 그래디언트 부스팅(extreme gradient boosting)이 가장 좋은 예측력을 보이는 것으로 보고되었다.

표 3-8 평균제곱편차(RMSE)

구분	tree	bagging	random forests	boosting	
				Gaussian	Laplace
Root Mean Squared Error	254,473.9	218,695.8	171,242	165,379.4	155,544.5
구분	**xgboosing**	**regression**	**ridge**	**lasso**	
Root Mean Squared Error	148,791.1	225,382	226,180	229,514	

제4장

도서관 건립의 효과 측정

공공도서관의
수요추정 모형 개발

KRILA
KOREA RESEARCH INSTITUTE FOR
LOCAL ADMINISTRATION

제4장 도서관 건립의 효과 측정

제1절 개요

제3장에서는 도서관 추가 건립시 수요 추정의 정확도를 높이기 위한 분석을 수행하였다. 특히 기계학습 방법을 도서관 수요추정에 적용하여 타당성 조사에 사용할 수 있는 도서관 방문자 수 예측 모형을 구축을 시도하였다. 다만 기계학습 방법의 경우 수요추정에 영향을 미치는 각종 변수들의 인과관계를 설명하기 어려운 단점이 있다.

또한 도서관 신규 건립 사업을 추진할 시 앞서 기계학습을 통해 수요를 추정할 수 있으나, 정책 결정자는 개별 도서관 수요 자체에 한정하기 보다 지역 사회 전체 도서관의 수급 현황을 고려해야 한다는 점에서 추가적인 분석이 요구된다.

이에 따라 본 장에서는 첫째, 도서관 수요와 관련한 각종 요인에 대한 인과관계를 분석하고, 둘째, 본 장에서는 도서관 추가 건립으로 인하여 지역사회 전체에 미치는 영향 및 효과를 분석하고자 한다. 특히 개별 도서관의 수요 추정을 넘어서 지역사회 전체로 논의를 넓히는 이유는 도서관과 같은 공공시설 건립시 이전수요 여부 및 그 규모에 대한 논의가 지속되고 있기 때문이다. 지역 내 도서관이 다수 운영 중인 경우 기존 도서관 이용자가 신규로 건립하는 도서관 수요로 전환될 가능성이 있다. 따라서 도서관 신규 건립 시에 예상되는 수요를 정확하게 추정하는 것도 중요하나, 지역 사회 도서관 공급 및 운영 정책의 판단근거를 제공하기 위해서는 추가적인 분석이 요구된다.

따라서 본 장에서는 방문자 수를 공공도서관 제공의 충격을 나타내는 지표로 보아 공공도서관 건립의 정책 충격을 측정하였다. 본 연구 제2장의 선행연구 검토에서 제시한 바와 같이 기존 도서관 수요 관련 연구에서는 도서관 방문자 수 또는 대출권 수, 도서관 이용여부 등에 영향을 미치는 다양한 요인에 대해 분석하였다. 그러나 도서관이 도서 대출뿐만 아니라 지역사회에서 다양한 문화 및 여가 관련 기능을 담당하면서 대출권 수 혹은 자료실 등 일부 공간에 대한 이용자 수로 도서관의 수요를 한정하는 것보다는 시설 전체 방문자

수를 도서관의 포괄적인 수요를 나타내는 지표로 보는 것이 더 타당하다고 볼 수 있다. 수요에 영향을 미치는 요인, 즉 설명변수으로는 여러 연구에서 공통적으로 장서 수, 좌석 수 등 도서관의 규모 및 특성을 파악할 수 있는 지표를 활용하였으며, 그 외 배후 인구 특성을 파악하기 위해 인구 수, 고령자 또는 미성년자 비중 등 지표를 활용하였다.

본 연구에서도 기존 연구에서 활용한 도서관 규모 관련 변수 및 인구 변수를 활용하고자 한다. 또한 문화체육관광부의 국가도서관 통계시스템의 10년치 자료를 활용하여 패널분석을 수행하고자 한다. 다만 본 장에서 목표한 지역 내 신규도서관 공급 효과를 분석하기 위해서는 지역 내 도서관 수, 분포 등이 충분하게 확보되어야 하므로, 본 연구에서는 서울시 도서관 자료에 한하여 분석을 수행하였다.

분석 모형은 제공된 자료가 미처 고려하지 못한 변수 중 분석 단위에 따라 변하나 시간에는 변하지 않는 변수와 시간에 따라 변하나 지정된 시간에는 분석 단위에 따라서는 변하지 않는 변수들을 제어하여 이 변수들의 제외 편의를 다루는 고정효과(fixed effects)모형을 사용하여 목표한 분석을 수행하였다.

제2절 자료와 모형

1. 분석 자료

본 연구에서는 국가도서관 통계시스템에서 제공하는 2009년부터 2018년도까지 10개년 서울시 내 개별 도서관 데이터로 총 관측치는 1,296개이다.

〈표 4-1〉에서 'distid'는 서울시 내 25개 구를 표시한다. 'visitor'는 도서관별 연간 방문자 수를 나타내며, 도서관별 연평균 방문자 수는 48만 명 수준이다. 'nolib'는 구별 도서관 수를 나타내며 평균 구별 도서관 수는 6.4개였다. 'pop'는 구별 인구수를 나타내며 평균 구별 인구수는 약 32만 명 수준이었으며, 'books'는 도서관별 장서 수를 나타내고 평균 장서 수는 83,800여권, 마지막으로 'seats'는 도서관별 좌석 수를 나타내며 평균 319개 정도였다. 다만 분석 단위가 도서관이기 때문에, 구별 도서관의 수에 대한 가중치가 부여된 평균이다.

표 4-1 개별 도서관 데이터 기초통계량

변수	내용	평균	표준편차	최소값	최대값
distid	자치구 ID	11.6	7.5	1	25
visitor	연간 방문자수	482,973.2	489,931.5	1,422	4,349,438
seats	좌석 수	319.1	314.1	10	2,146
books	장서 수	82,836.2	87,352.5	723	532,536
pop	자치구별 인구	324,076.3	104,046.2	93,096	529,637
nolib	구별 도서관 수	6.4	2.9	1	13

주 : N=1,296

〈표 4-2〉는 자치구별 자료에 대한 기술통계량을 보여준다. '$visitor_{gu}$', '$books_{gu}$', '$seats_{gu}$'는 별 도서관 자료 중 'visitor', 'books', 'seats'를 구별로 합한 값들을 기술하고 있다. '$visitor_{gu}$'는 구별 연간 도서관 방문자 수를 나타내며, 연평균 방문자 수는 2,500,550만 명 수준이다. '$books_{gu}$'는 구 내 도서관별 장서 수를 모두 더한 값이며 구별 평균 장서 수는 434,156여권, 마지막으로 '$seats_{gu}$'는 구내 도서관별 좌석 수를 모두 더한 값으로 구별

평균 1,654개 정도였다.

표 4-2 자치구별 도서관 데이터 기초통계량

변수	내용	평균	표준편차	최소값	최대값
visitorgu	자치구내 방문자수	2,500,550.0	1,195,454.0	65,300	5,947,984
seatsgu	자치구내 좌석 수	1,654.4	800.1	34	4,242
booksgu	자치구내 장서 수	434,156.9	225,399.6	9,943	1,160,477
pop	자치구별 인구	305,039.9	99,982.2	93,096	529,637
nolib	구별 도서관 수	5.2	2.5	1	13

주 : N=250

2. 분석 모형

본 장에서는 앞서 제시한 서울 공공도서관 데이터를 활용하여 도서관 공급 정책의 충격을 측정하기 위하여 개별 도서관 단위와 자치구별 도서관 자료 단위로 구분하여 모형을 분석하였다. 개별 도서관 단위의 분석결과는 기계학습 방법에서 제시하지 못하는 인과관계에 대한 분석을 가능케 한다. 자치구별 단위 분석에서는 자치구내 전체 도서관 이용자수를 종속변수로 분석하여 도서관 추가 건립이 지역 내 도서관 수요에 미치는 영향을 살펴보고, 다른 설명변수(좌석 수, 장서 수, 인구)의 영향 및 효과와도 비교 검토하고자 한다.

1) 개별 도서관 분석 모형

서울시 내 개별 도서관 자료를 활용하여 분석한 모형은 다음과 같다. 분석에 사용한 변수 중 nolib을 제외한 변수의 분포를 확인한 결과, 모두 자연로그로 변환된 후에 정규분포에 근사한 분포를 나타내었다. 또한 분포의 왜도 및 첨도도 자연로그 변환 이후 정규분포에 가까운 모습을 나타내므로 해당 변수에 대해서는 자연로그로 변환 후 모형에 적용하였다.

(1) $\log(visitor) = \beta_0 + \beta_1 nolib + u$

(2) $\log(visitor) = \beta_1 nolib + \alpha_i + \lambda_t + u$

(3) $\log(visitor) = \beta_1 nolib + \beta_2 \log(pop) + \alpha_i + \lambda_t + u$

(4) $\log(visitor) = \beta_1 nolib + \beta_2 \log(pop) + \beta_3 \log(seats) + \alpha_i + \lambda_t + u$

(5) $\log(visitor) = \beta_1 nolib + \beta_2 \log(pop) + \beta_3 \log(books) + \alpha_i + \lambda_t + u$

(6) $\log(visitor) = \beta_1 nolib + \beta_2 \log(pop) + \beta_3 \log(seats) + \beta_4 \log(books) + \alpha_i + \lambda_t + u$

(1)은 선형모형으로 OLS 회귀분석을 위한 모형이다. (2)는 시간('year')과 지역('distid')를 기준으로 한 고정효과(fixed effects) 모형이다. (3)부터 (6)까지는 기본 모형인 (2)에 다른 잠재적인 설명변수들인 'pop', 'seats', 'books' 등을 추가한 모형이다.

2) 자치구별 도서관 분석 모형

서울시 내 자치구별 단위로 분석한 모형은 다음과 같다.

(1) $\log(visitor_{gu}) = \beta_0 + \beta_1 nolib + u$

(2) $\log(visitor_{gu}) = \beta_1 nolib + \alpha_i + \lambda_t + u$

(3) $\log(visitor_{gu}) = \beta_1 nolib + \beta_2 \log(pop) + \alpha_i + \lambda_t + u$

(4) $\log(visitor_{gu}) = \beta_1 nolib + \beta_2 \log(pop) + \beta_3 \log(seats_{gu}) + \alpha_i + \lambda_t + u$

(5) $\log(visitor_{gu}) = \beta_1 nolib + \beta_2 \log(pop) + \beta_3 \log(books_{gu}) + \alpha_i + \lambda_t + u$

(6) $\log(visitor_{gu}) = \beta_1 nolib + \beta_2 \log(pop) + \beta_3 \log(seats_{gu}) + \beta_4 \log(books_{gu}) + \alpha_i + \lambda_t + u$

개별 도서관 모형과 마찬가지로 모형(1)은 선형모형으로 OLS 회귀분석을 위한 모형이다. 모형(2)는 시간('year')과 지역('distid')를 기준으로 한 고정효과(fixed effects) 모형이다. 모형(3)부터 모형(6)까지는 기본 모형인 (2)에 다른 잠재적인 설명변수들인 'pop', 'seats', 'books' 등을 추가한 모형이다.

제3절 분석 결과

본 절에서는 앞 절에서 소개한 데이터와 모형으로 분석한 결과를 제시하고자 한다. 주요 분석결과는 모형들의 계수 추정, 모형의 유의성(significance), 계열상관(serial correlation), 그리고 다중공선성(multicollinearity) 등을 포함한다. 분석에는 R의 plm 패키지를 주로 사용하였다.

1) 개별 도서관 모형 분석 결과

개별 도서관을 대상으로 패널회귀분석한 결과는 다음과 같다.

표 4-3 개별 도서관 패널분석 결과

	Dependent varilable : log(visitor)					
	(1)	(2)	(3)	(4)	(5)	(6)
nolib	-0.053*	-0.007	-0.009	-0.006	-0.013	-0.010
	(0.027)	(0.024)	(0.023)	(0.022)	(0.022)	(0.021)
lob(pop)			3.139***	3.109***	2.750**	2.738***
			(1.069)	(1.058)	(1.071)	(1.062)
log(seat)				0.901***		0.581***
				(0.079)		(0.085)
log(books)					0.422***	0.411***
					(0.145)	(0.147)
constant	12.841***					
	(0.218)					
Observations	1,296	1,296	1,296	1,296	1,296	1,296
R2	0.015	0.0001	0.014	0.019	0.056	0.059
Adjusted R2	0.014	-0.202	-0.187	-0.182	-0.137	-0.135
F Statistic	19.199***	0.138***	7.564***	7.008***	21.316***	16.828***
	(df=1;1,294)	(df=1;1,077)	(df=2;1076)	(df=3;1,075)	(df=3;1075)	(df=4;1074)

주 : ***p<0.01, **p<0.05, * p<0.1

모형(1)은 pooled ols 회귀분석을 보여준다. 계수는 음의 값(-0.053)을 가지고 유의 수준 10%에서 통계적으로 유의미하다. 이는 도서관이 하나 추가되면 각 도서관 방문자수가 평균적으로 5.16% 감소하는 것을 의미한다.

모형(2)는 시간과 지역 고정효과(fixed effects) 모형으로 계수는 음의 값(-0.007)을 가지나 통계적으로 유의미하다는 증거는 없는 것으로 나온다.

모형(3)은 설명변수로 인구수(log(pop))를 추가하였다. 분석결과 nolib의 계수는 여전히 음의 값(-0.009)을 보였고 통계적으로 유의미함을 보이는 데 실패하였다. log(pop)의 계수는 양의 값(3.139)을 가졌고 1% 유의수준에서 통계적으로 유의미하였다. log(pop)의 계수 값은 도서관이 위치한 구의 인구가 1% 증가하면 해당 도서관의 방문자 수가 약 3.14% 증가함을 의미한다.

모형(4)는 좌석 수(log(seats))가 추가된 모형이다. 분석결과 nolib(-0.006), log(pop)(3.109), log(seats)(0.142)가 보고되었다. nolib의 계수는 통계적으로 유의미함을 확보하는 데 실패하였으나 log(pop)sms 유의수준 1%에서, log(seats)는 유의수준 10%에서 각각 통계적으로 유의미하였다. 이 모형은 도서관의 방문자 수가 도서관이 위치한 구의 인구가 1% 증가하면 3.109%, 좌석 수를 1% 증가하면 0.142% 증가함을 의미한다.

모형(5)는 장서 수(log(books))가 추가된 모형이다. 분석결과 nolib(-0.013), log(pop)(2.75), log(books)(0.422)가 보고되었다. nolib의 계수는 통계적으로 유의미함을 확보하는 데 실패하였으나 log(pop)는 유의수준 5%, log(books)는 유의수준 1%에서 통계적으로 유의미하다. log(pop)의 계수 값은 도서관이 위치한 구의 인구가 1% 증가하면 해당 도서관 방문자 수가 2.75% 증가하고 log(books)의 계수 값은 도서관의 장서 수를 1% 늘리면 그 도서관의 방문자 수가 0.422% 증가함을 의미한다.

모형(6)은 좌석 수(log(seats))와 장서 수(log(books))가 추가된 모형이다. 분석결과 nolib(-0.01)의 계수는 음의 값을 보고 하였으나 통계적 유의미함을 확보하지 못하였고, log(seats)(0.104)의 계수는 양의 값을 보고 하였으나 역시 통계적 유의미함을 확보하지 못하였다. log(pop)(2.738)와 log(books)(0.411)는 모두 양의 계수 값을 보고 하였고 둘 다 유의수준 1%에서 통계적으로 유의미하였다. 통계적으로 유의미한 두 계수 값은 도서관이 위치한 구의 인구수가 1% 증가하면 해당 도서관의 방문자 수가 2.74% 증가하고 도서관의

장서 수를 1% 늘리면 그 도서관의 방문자 수가 0.41% 증가함을 의미한다.

혼다 라그랑주 승수 테스트(Honda Lagrange Multiplier Test)는 시간과 지역 고정효과(fixed effects)가 통계적으로 유의미함을 보여주고, Breusch-Godfrey / Wooldridge test를 통하여 계열 상관(serial correlation)이 통계적으로 유의미하게 존재함을 보여준다. 이에 따라 계열 상관(serial correlation)을 다루기 위해서 강건표준오차(clustered standard error)를 사용하여 모형(1)~(6)까지 분석을 진행하였다.

강건표준오차(clustered standard error)의 사용은 계열 상관(serial correlation)을 다루는 데 효과적일 뿐만 아니라 이분산성(heteroscedasticity)도 동시에 다룰 수 있다. 모형(3), (4), (5), (6)은 다중공선성(multicollinearity)이 없는 것으로 테스트 결과가 나왔다.

2) 자치구별 도서관 모형 분석 결과

자치구별 도서관을 대상으로 패널회귀분석한 결과는 다음과 같다.

표 4-4 자치구별 도서관 패널분석 결과

	Dependent varilable : log(visitorgu)					
	(1)	(2)	(3)	(4)	(5)	(6)
nolib	0.133***	0.098**	0.095*	0.014	0.006	-0.010
	(0.017)	(0.040)	(0.040)	(0.022)	(0.023)	(0.017)
lob(pop)			4.520	0.316	2.395**	0.553
			(3.135)	(1.257)	(1.643)	(1.020)
log(seatgu)				0.901***		0.581***
				(0.156)		(0.213)
log(booksgu)					1.219***	0.719***
					(0.084)	(0.237)
constant	13.890***					
	(0.107)					
Observations	250	250	250	250	250	250
R2	0.248	0.046	0.093	0.607	0.580	0.712
Adjusted R2	0.245	-0.105	-0.055	0.541	0.509	0.662
F Statistic	81.838***	10.412***	11.025***	109.704***	97.874***	130.913***
	(df=1;248)	(df=1;215)	(df=2;214)	(df=3;213)	(df=3;213)	(df=4;212)

주 : ***p<0.01, **p<0.05, * p<0.1

모형(1)은 pooled OLS 회귀분석을 보여준다. 계수는 양의 값(0.133)을 가지고 유의 수준 1%에서 통계적으로 유의미하다. 이는 도서관이 하나 추가되면 해당 구에서 도서관 방문자 수가 평균적으로 14.2% 증가함을 보여준다.

모형(2)는 시간과 지역 고정효과(fixed effects) 모형으로 계수는 양의 값(0.098)이며 유의수준 5%에서 통계적으로 유의미하다. 이 계수 값은 도서관이 하나 추가되면 해당 구에서 도서관 방문자 수가 평균적으로 0.098% 증가함을 보여준다.

모형(3)은 설명변수로 인구수(log(pop))를 추가하였다. 분석결과 nolib의 계수는 여전히 양의 값(0.095)을 보였고 유의수준 10%에서 통계적으로 유의미하였다. log(pop)의 계수도 양의 값(4.52)을 가졌으나 통계적으로 유의미함을 보여주는 증거는 없는 것으로 나왔다.

모형(4)는 좌석 수(log(seatsgu))가 추가된 모형이다. 분석결과 nolib(0.014), log(pop)(0.316), log(seatsgu)(0.901)가 보고 되었다. nolib와 log(pop)의 계수는 통계적으로 유의미함을 확보하는 데 실패하였으나 log(seats)는 유의수준 1%에서 통계적으로 유의미하다. log(seats)의 계수 값은 각 구가 구 내 도서관의 좌석 수를 1% 늘리면 구 내 도서관 방문자 수가 약 0.9% 증가함을 의미한다.

모형(5)는 장서 수(log(booksgu))가 추가된 모형이다. 분석결과 nolib(0.006), log(pop)(2.395), log(booksgu)(1.219), 모두 양의 계수 값을 보고 하였다. nolib와 log(pop)의 계수는 통계적으로 유의미함을 확보하는 데 실패하였으나 log(booksgu)는 유의수준 1%에서 통계적으로 유의미하다. log(books)의 계수 값은 각 구가 구 내 도서관 장서 수를 1% 늘리면 구 내 도서관 방문자수가 약 1.22% 증가함을 의미한다.

모형(6)은 좌석 수(log(seatsgu))와 장서 수(log(booksgu))가 추가된 모형이다. 분석결과 nolib(-0.01)의 계수는 음의 값을 보고 하였으나 통계적 유의미함을 확보하지 못하였고, log(pop)(0.553)의 계수는 양의 값을 보고 하였으나 역시 통계적 유의미함을 확보하지 못하였다. log(seats)(0.581)와 log(books)(0.719)는 모두 양의 계수 값을 보고 하였고 둘 다 유의수준 1%에서 통계적으로 유의미하다. log(seats)의 계수 값은 각 구가 구 내 도서관 좌석 수를 1% 늘리면 구 내 도서관 방문자수가 약 0.58% 증가함을 의미하고 장서 수를 1% 늘리면 0.72% 증가함을 의미한다.

좌석 수와 장서 수를 모형에 추가하는 것은 그렇지 않은 경우에 비해 R^2가 0.215에서

0.712로 극적으로 증가하였다.

고정효과(fixed effects)의 통계적 유의미성(significance)를 테스트하기 위해 혼다 라그랑주 승수 테스트(Honda Lagrange Multiplier Test)를 사용하였다. 그 결과 시간과 지역에 대한 고정효과(fixed effects)가 통계적으로 유의미한 것으로 나타났다. 또한 Breusch-Godfrey/ Wooldridge test를 통하여 계열상관(serial correlation)을 테스트한 결과 통계적으로 유의미한 것으로 분석되었다. 이에 따라 계열 상관(serial correlation)을 다루기 위해서 강건표준오차(clustered standard error)를 사용하여 모형(1)~(6)까지 분석을 진행하였다. 강건표준오차(clustered standard error)의 사용은 계열상관(serial correlation)을 다루는데 효과적일 뿐만 아니라 이분산성(heteroscedasticity)도 동시에 다룰 수 있다. 모형 (3), (4), (5), (6)은 다중공산성(multicollinearity)이 없는 것으로 테스트 결과가 나왔다.

제4절 종합

종합하면 개별 도서관 분석에서는 통계적으로 유의하지 않으나 모든 모형에서 한 구의 도서관 수의 증가는 해당 구 내 개별 도서관의 방문자 수에 부(-)의 효과가 있는 것으로 분석이 되었다.

개별 도서관 분석에서는 도서관이 위치한 구의 인구수 증가는 방문자 수에 통계적으로 유의미한 정(+)의 효과가 있는 것으로 분석이 되었다. 또한, 개별 도서관의 좌석 수를 증가시키면 통계적으로 유의미한 방문자 수 증가가 발생하는 것으로 분석되나 장서 수와 동시에 고려하면 통계적으로 유의미하지 않은 것으로 분석되었다.

반면에 도서관 장서 수는 어떠한 모형에서도 통계적으로 유의미한 방문자 수 증가를 불러왔다. 구별 분석을 보면 해당 구의 도서관 수를 늘리면 다른 설명변수를 고려하지 않을 경우 통계적으로 유의미한 해당 구 내 도서관 방문자 수가 증가하는 것으로 분석되었다.

그러나, 인구, 좌석 수, 장서 수 등을 고려하면 해당 구 내 도서관 수는 방문자 수 증감에 통계적으로 유의미한 영향이 없거나, 있어도 미약한 것으로 분석되었다. 인구 증가도 통계적으로 유의미한 영향이 없는 것으로 분석이 되었다.

구내 도서관의 좌석 수와 장서 수를 증가시키는 것은 개별 또는 함께 모형에 고려하는 모든 경우에 통계적으로 유의미한 구내 도서관 방문자 수의 증가를 가져왔다. 그러므로, 도서관 제공의 정책 효과를 방문자 수의 증가로 정의한다면 신규 도서관 제공보다는 먼저 기존 도서관의 좌석 수와 장서 수를 늘리는 것을 고려하고 현재 구 내 도서관의 수용능력이 그 고려에 미치지 못할 때 신규 도서관 건립을 고려하는 것이 올바른 정책 방향으로 판단된다.

다만 도서관마다 수용능력 확충 가능성 및 입지여건이 다르기 때문에 이러한 결과를 모든 도서관 신규 건립사업에 일반화하는 것은 어려울 수 있다. 예를 들어 신도시 내 추진하는 도서관 건립 사업의 경우, 기성 시가지 등에 운영 중인 도서관의 수용 능력을 늘리더라도 인구가 감소하는 기성 시가지보다는 신도시 내 인구가 증가하는 지역의 주민 편의를 고려하여 신규 건립 사업을 추진해야 할 필요성이 있을 수도 있다.

따라서 타당성 조사 과정에서는 기존 도서관의 운영현황 등을 살펴보고, 신규 도서관 사업대상지와 경쟁관계에 있는 기존 도서관 존재 여부 및 수용능력 확충 가능성 등을 일차적으로 살펴본 이후 해당 신규 도서관의 수요 추정을 진행할 필요가 있다.

제5장

결 론

제1절 연구의 요약

제2절 연구의 한계 및 향후 연구

공공도서관의
수요추정 모형 개발

KRILA
KOREA RESEARCH INSTITUTE FOR
LOCAL ADMINISTRATION

제5장 결론

제1절 연구의 요약

오늘날 공공도서관은 전통적인 기능인 도서관 자료를 중심으로 한 지식정보 제공 및 학습의 기능뿐만 아니라 지역사회 구성원을 위한 문화공간으로서 역할도 수행하고 있다. 이러한 수행 역할과 기능의 확장에 따라 국민의 삶의 질 향상을 높이기 위한 국민기초시설인 생활SOC의 한 축으로 자리매김하게 되었다. 그 결과 2020년 기준, 생활SOC 사업에 선정된 289개 사업 중 160개 사업이 공공도서관 및 작은 도서관 사업이 포함된 사업이었다. 특히 공공도서관이 포함된 사업은 73개이며, 국비지원 규모는 2,031억 원에 달하였다.

최근에는 생활SOC뿐만 아니라 기존 도서관의 노후화 및 리모델링, 신도시 건설에 따른 도서관 신규 건립 등에 따라 지방자치단체에서 공공도서관 추진 사업 건수가 증가하였다. 이에 따라 지방자치단체가 추진하는 대규모 사업에 대해 투자심사 전에 이행해야 하는 「지방재정법」 제37조에 의한 타당성 조사 의뢰 건수도 증가하고 있다.

따라서 타당성 조사의 핵심이 되는 도서관 수요 추정에 대한 신뢰성과 정확성을 높이기 위한 적절한 방법론 개발이 필요하다. 공공도서관 수요추정 방법론과 관련해서는 『문화체육관광 부문 타당성 조사를 위한 지침 연구』(한국지방행정연구원, 2016)에 제시된 다양한 방법을 참고할 수 있으나, 공공도서관의 특성을 고려한 방법론에 대한 가이드라인은 부재한 상황이다. 함윤주 외(2019)에서는 중력 모형으로 안양시에 위치한 공공도서관을 대상으로 실증분석을 실시하였으나, 공공도서관 관련 수요 추정에 대한 기초연구는 상당히 부족한 실정이다.

본 연구는 도서관의 이용수요의 특징과 기존에 연구된 수요추정방법을 검토하고, 미시계량경제학(microeconometrics)의 프로그램 평가(Program evaluation) 접근법과 데이터 과학(Data science) 접근법을 활용하여 도서관 운영 시 다양한 특성들이 이용수요에 미치는 영향을 고려하여 궁극적으로 이용수요를 예측할 수 있는 수요추정 모형을 개발을 시도하였다. 국가공공도서관통계에서 방대한 양의 도서관 관련 데이터를 제공하고 있으므로,

이를 활용하여 기계학습 방법을 사용하여 서울시 내 공공도서관에 대하여 앞으로 타당성 조사에 사용이 가능한 도서관 방문자 수 예측 모형을 구축하였다. 또한 예측치 생성에 미치는 영향을 분석하여 데이터 내 변수들의 중요성을 분석하였다.

기계학습 알고리즘 중에서도 의사결정나무(decision tree), 배깅(bagging), 랜덤 포레스트(random forest), 그래디언트 부스팅(gradient boosting), 익스트림 그래디언트 부스트(extreme gradiance boost : XGboost), 회귀분석 등 다양한 기계학습 알고리즘을 활용하여 도서관 방문자 수 예측 모형을 구축하였다. 즉, 각 모형분석 과정에서 격자탐색을 수행하여 최적 초모수 집합을 식별하였으며, 이를 고정하여 모형에 적용하면 향후 도서관 방문자 수 예측 모형으로 적용가능하다. 또한 기계학습 알고리즘은 전체 데이터 표본을 학습(training)과 테스트(test) 집합으로 나누어 자체적으로 검증을 수행하며, 검증 결과 상기 모형 중에서 익스트림 그래디언트 부스팅(extreme gradient boosting)이 가장 좋은 예측력을 보이는 것으로 보고 되었다.

이상 본 연구에서는 도서관 추가 건립시 해당 신규 도서관의 수요 추정 정확도를 높이기 위한 기계학습 방법을 적용하여 분석을 수행하였다. 다만 기계학습 방법의 경우 수요추정에 영향을 미치는 각종 변수들의 인과관계를 설명하기 어려운 단점이 있다.또한 도서관 신규 건립 사업을 추진할 시 앞서 기계학습을 통해 수요를 추정할 수 있으나, 정책 결정자는 개별 도서관 수요 자체에 한정하기 보다 지역 사회 전체 도서관의 수급현황을 고려해야 한다는 점에서 추가적인 분석이 요구된다.

이에 따라 본 연구에서는 추가적으로 도서관 수요와 관련한 각종 요인에 대한 인과관계를 분석하고, 도서관 추가 건립으로 인하여 지역사회 전체에 미치는 영향 및 효과를 분석하였다. 이를 위해서 방문자 수를 공공도서관 제공의 충격을 나타내는 지표로 보고, 문화체육관광부에서 제공하는 도서관에 관한 패널 자료를 활용하여 고정효과(fixed effects) 모형을 사용하여 분석을 수행하였다. 분석결과 현재 서울에서는 도서관 방문자 수 증가를 목표로 할 때 신규 도서관 제공보다는 먼저 기존 도서관의 좌석 수와 장서 수를 늘리는 것을 고려하고 현재 해당 구 내 도서관의 수용 능력이 그 고려에 미치지 못할 때 신규 도서관 건립을 고려해야 한다는 결론이 도출되었다. 이를 모든 도서관 신규건립 사업에 일반화하기는 어려우나, 도서관의 입지나 경쟁관계에 있는 기 운영 중인 도서관 존재 여부 및 수용능력 확충 가능성을 우선적으로 살펴볼 필요가 있음을 시사한다.

제2절 연구의 한계 및 향후 연구

공공도서관과 같이 문화기반시설을 신규로 건립하는 사업에서 가장 중요하게 고려해야 할 사항은 중복 투자 여부와 적절한 이용수요의 존재이다. 본 연구에서는 국가도서관 통계시스템에서 제공하는 공공도서관 관련 방대한 데이터를 활용하여 최근 다른 분야에서 활발하게 사용되고 있는 기계학습 알고리즘을 적용하여 수요 추정 분석을 시도하였다. 기계학습 알고리즘은 학습과 검증을 동시에 진행하여 가장 최적의 모형을 구축한다는 장점이 있으나, 기존에 타당성 조사에서 일반적으로 사용되었던 중력모형, 회귀분석과 다르게 인과 관계를 밝혀내는 데 취약한 단점이 있다. 즉, 모형의 예측력은 강하나 설명력은 약한 단점이 있다(Athey(2017), Varian(2014)).

따라서 향후 연구로는 수요예측 시 분석된 주요 설명변수의 중요도(importnace)를 바탕으로 설명력을 보완하고 상호 호환가능한 수요 모형 구축을 위해 모수적인 방법으로 추가적인 연구가 필요하다.

또한 국가도서관 통계시스템의 방대한 자료가 있기 때문에 본 연구가 가능하였으므로 향후 다른 유형의 공공시설에서도 유사한 데이터가 생성된다면 기계학습을 통해 충분히 수요예측을 시도해볼 수 있을 것으로 보인다. 다만 현재 국가도서관 통계시스템의 경우에도 일부 데이터 검증 과정에서 오류 등이 확인되어 전체 데이터를 사용하는데 어려움이 있었다. 따라서 데이터 집계 및 검증과정에서 보다 자료의 정확성을 높이기 위한 노력이 필요한 것으로 보인다.

참고문헌

국무조정실. (2019).「생활SOC 3개년계획(안) (2020~2022)」.
국토교통부. (2019).「국가도시재생기본방침 일부개정 재공고」.
문화체육관광부. (2019).「공공도서관 건립·운영 매뉴얼」.
문화체육관광부.「국가도서관 통계시스템」.
원종준·안건혁. (2010). 공공도서관 입지 및 시설특성이 이용활성화에 미치는 영향 연구.「대한건축학회 논문집 - 계획계」, 26(2) : 79-86.
이원태. (2004).「전국문화기반시설 최소기준수립 연구」. 한국문화관광정책연구원.
이학준·이용관. (2019). 공공도서관 공급의 효과성 분석 : 이용자 수와 도서 대출권수 변화를 중심으로.「예산정책연구」, 8(2) : 226-256.
전계형·권선영. (2018). 경제요인이 도서관 이용에 미치는 영향에 관한 연구.「한국융합학회논문지」, 9(11): 299-306.
정혜린·임창원. (2019).「인공지능 기반 수요예측 기법의 리뷰. 응용통계연구」, 32(6): 795-835.
장훈. (2018).「국민여가활성화를 위한 문화서비스 개선 연구」. 한국문화관광연구원.
조권중. (2004).「서울시 공공도서관 발전 방향에 관한 연구」. 서울연구원.
최희곤. (2009). 공공도서관의 이용자 수에 영향을 미치는 요인 분석.「정보관리학회지」, 26(4) : 129-146.
한국개발연구원. (2007).「국립중앙도서관 부산분관 건립사업 예비타당성조사 보고서」.
한국개발연구원. (2010a).「국립중앙도서관 광주분관 건립사업 타당성 재조사 보고서」.
한국개발연구원. (2010b).「국립중앙도서관 부산분관 건립사업 타당성 재조사 보고서」.
한국개발연구원. (2005).「헌법재판소 도서관 신축사업 타당성 재검증 보고서」.
한국도서관협회. (2003).「한국도서관기준」.
한국지방행정연구원. (2016).「문화체육관광부문 타당성 조사를 위한 지침 연구」.
함윤주·조현민·김지선. (2019).「문화시설 수요추정 개선방안 연구」. 한국지방행정연구원.
함윤주·홍근석·주우현. (2021). 지방자치단체 공공도서관 수요추정 개선방안: 중력모형을

중심으로. 「정책분석평가학회보」, 31(2): 35-58.

Boehmke and Greenwell. (2020). Hands-On Machine Learning with R. CRC Press, Boca Raton.

Gill, Philip. (2002). 「공공도서관 서비스 개발을 위한 IFLA/UNESCO 가이드라인」. 서울 : 한국도서관협회.

T. Chen and C. Guestrin. (2016). *XGBoost : A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. pages 785–794, San Francisco, CA.

Leo Breiman. (1984). *Classification and regression trees*. Chapman & Hall/CRC,

Leo Breiman. (1996a). Bagging Predictors. *Machine Learning*, 24(2) : 123–140.

Leo Breiman. (1996b). Heuristics of Instability and Stabilization in Model Selection. *The Annals of Statistics*, 24(6) : 2350–2383.

Leo Breiman. (1996c). Technical Note : Some Properties of Splitting Criteria. *Machine Learning*, 24(1) : 41–47.

Leo Breiman. (2000). Randomizing Outputs to Increase Prediction Accuracy. *Machine Learning*, 40 : 229–242.

Leo Breiman. (2001). Random Forests. *Machine Learning*, 45 : 5–32.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. (2013). *An introduction to statistical learning. Vol. 112*. New York : Springer.

Jerome H Friedman. (2001). Greedy function approximation : a gradient boosting machine, *The Annals of Statistics*, 29(5) : 1189-1232.

Jerome H Friedman. (2002). *Stochastic gradient boosting*, 38(4) : 367-378.

Susan Athey. (2017). Beyond prediction : Using big data for policy problems. *Science*, 355 : 483-485.

Guido W. Imbens. (2015). *Causal Inference for Statistics, Social and Biomedical Sciences*. Cambridge University Press.

Max Kuhn and Kjell Johnson. (2013). *Applied Predictive Modeling*. Springer.

Matt Taddy. (2019). *Business Data Science*, McGraw-Hill Education.

Hal R. Varian. (2014). Big Data : New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2) : 3-28.

Abstract

A Study on Public Library Demand Estimation

Public libraries nowadays are playing a role of cultural space for local communities while still performing their traditional function of providing local people with knowledge information through library materials. This expansion in library functions makes a public library an axis of life. As a result, 160 of the 289 projects selected for the SOC project in 2020 included public library and small library projects. In particular, 73 projects among them included public libraries, and the government aid amounted to 233.1 billion won.

Moreover, the need to remodel old and aging libraries and the demand for new libraries in new town have increased the number of public library projects with local governments. Therefore, the number of requests for feasibility studies under Article 37 of the Local Finance Act, which must be implemented before the investment examination, has increased.

Therefore, it is necessary to develop an appropriate methodology to enhance the reliability and accuracy of library demand estimation which is the core of feasibility study. Although one can find a list of methods for demand estimation in 『Guidance for feasibility study for cultural / sports / tour investment projects』(Local Investment Management Center, 2016), it would be better if the list included methods specific to public library feasibility study than otherwise. Ham Yun-ju et al.(2019) conducted an empirical analysis of demand estimation for public libraries located in Anyang city through a gravity model, but there is a lack of research on the estimation of demand related to public libraries.

This study reviewed the features imposing impacts on the library demand and the existing method of estimating demand, and developed models that predict the demand for public libraries through the microeconometrics (program evaluation) and the data science. Since the National Public Library Statistics provides a vast amount of library-related data, we have constructed a library-visitor-prediction model that can be used for feasibility studies in the future by using machine learning methods including XGboosting. Also, the importance of variables in data was considered through measuring their contribution on making the prediction value.

Furthermore, this study analyzed the impact of a library provision using the number of visitors as a measure of the impact. To this end, we used a fixed effect model that is a method for controlling for omitted variables in panel data when the omitted variables vary across entities but do not change over time. As a result, we concluded that when investing in the library system of interest we should first consider increasing the number of seats and collections of the existing libraries. Further, the construction of a new library should be legitimate only if the capacity of the library system is short.