군집분석 기법을 이용한 공공도서관 그룹화에 대한 연구*

A Study of Library Grouping using Cluster Analysis Methods

곽 철 완 (Chul Wan Kwak)**

초 록

이 연구의 목적은 공공도서관 그룹화를 위해 적합한 군집분석 모델을 파악하고 그 특징을 분석하는데 있다. 국가도서관통계시스템의 공공도서관 통계 데이터를 사용하였으며, 군집분석 기법의 3가지 모델을 적용하였다. 공공도서관 규모를 기준으로 군집분석을 실시한 결과 크게 2가지 군집으로 구분되었으며, 군집의 크기는 크게 한쪽으로 치우쳤다. 그룹화 모델로 도서관 규모를 기준으로 삼으면, 계층적 군집분석의 와드측정법과 k-평균군집분석 모델이 적합하였다. 공공도서관 그룹화 연구 결과에 대한 시사점은 다음과 같다. 첫째, 통계 데이터 외에 도서관 서비스 관련 다양한 데이터 수집이 진행되어야 한다. 둘째, 분석 대상이 되는 데이터 세트에 적합한 분석 모델이 적용되어야 한다. 셋째, 도서관 서비스 향상을 위해 군집분석 기법의 다양한 분야 적용 가능성에 대한 적극적인 연구가 필요가 있다.

ABSTRACT

The purpose of this study is to investigate the model of cluster analysis techniques for grouping public libraries and analyze their characteristics. Statistical data of public libraries of the National Library Statistics System were used, and three models of cluster analysis were applied. As a result of the study, cluster analysis was conducted based on the size of public libraries, and it was largely divided into two clusters. The size of the cluster was largely skewed to one side. For grouping based on size, the ward method of hierarchical cluster analysis and the k-means cluster analysis model were suitable. Three suggestions were presented as implications of the grouping method of public libraries. First, it is necessary to collect library service-related data in addition to statistical data. Second, an analysis model suitable for the data set to be analyzed must be applied. Third, it is necessary to study the possibility of using cluster analysis techniques in various fields other than library grouping.

키워드: 공공도서관, 도서관 통계, 군집분석, 계층적 군집분석 와드측정법 모델, 도서관 그룹
Public Library, Library Statistics, Cluster Analysis, Hierarchical Clustering Ward Linkage Method, Library Group

---

* 본 연구는 (2018 학년도) 강남대학교 교내연구비 지원에 의해 수행되었음.
** 강남대학교 산업데이터사이언스학부 교수(ckwak@kangnam.ac.kr)
논문접수일자 : 2020년 8월 19일 논문심사일자 : 2020년 9월 3일 게재확정일자 : 2020년 9월 3일
한국비블리아학회지, 31(3): 79-99, 2020. http://dx.doi.org/10.14699/kbiblia.2020.31.3.079

1. 서 론

1.1 연구의 목적

우리나라 공공도서관은 지역대표도서관을 비롯하여 각 지역마다 중앙관과 분관 등으로 구분하여 서로 다른 역할을 수행하고 있다. 일반적으로 지역의 중앙관은 봉사대상 지역에서 가장 큰 규모의 도서관이며 분관들은 상대적으로 소규모로 건립되어 운영된다. 도서관을 다양한 규모로 건립하여 운영하는 것은 지역주민에게 효과적이며 경제적으로 폭 넓은 서비스를 제공하기 위함일 것이다.

매년 문화체육관광부에서 실시하는 전국 도서관 운영평가는 공공도서관을 규모에 따라 그룹별로 구분하여 도서관 서비스의 우수 사례를 발굴하여 보급하고 있다. 이는 서로 유사한 도서관 사이에서 우수한 도서관 서비스를 발굴하고 서로 공유하여, 전국적으로 도서관 서비스를 향상시키는데 그 목적이 있다고 볼 수 있다. 예를 들어, 자관과 유사한 도서관의 우수 서비스 사례를 참조하여 지역 환경에 적합한 서비스를 개발한다면 지역 주민들에 대한 서비스 향상에 커다란 도움이 될 수 있을 것이다. 이런 측면에서 유사한 도서관을 파악하는 일은 매우 중요하다고 본다. 이를 위해 유사한 공공도서관들을 그룹으로 모아주는 작업은 도서관 서비스 향상에 커다란 의미가 있다.

문화체육관광부는 2020년 전국 도서관 운영평가를 위해 공공도서관을 단위도서관의 봉사대상인구 1인당 시설, 장서, 인력, 예산항목을 기준으로 삼아 5개 그룹으로 구분하였다(국가도서관통계시스템 2020a). 그룹 구분을 위해 4개 항목을 30점에서 10점으로 배정하고, 항목별 점수를 산정한 후, 총점을 기준으로 전국 공공도서관을 5개 그룹으로 배정하였다.

최근 빅데이터 확산에 따라, 그룹을 나누는 방법으로 머신러닝 기반의 비지도학습 기법인 군집분석이 활용되고 있다. 군집분석은 데이터 간의 유사성을 측정하여 상호 유사성이 높은 대상을 동일 집단으로 분류하는 기법이다(조민호 2019, 330). 군집분석 기법은 복잡한 과정을 거치지 않고 간단하게 처리할 수 있는 방법으로 공공도서관을 그룹으로 구분하는데 사용될 수 있을 것이다. 이에 이 연구는 공공도서관 그룹화를 위한 적합한 군집분석 모델을 파악하고 그 특징을 분석하는데 목적이 있다.

1.2 선행연구

공공도서관 그룹화에 대해 차미경, 표순희(2015)는 공공도서관 평가에 관한 연구에서 도서관 그룹화에 대해 언급하고 있다. 이 연구에서는 공공도서관을 대규모 및 중규모 도서관이 포함되는 A그룹과 소규모 및 어린도서관이 포함되는 B그룹으로 구분하였다. 공공도서관을 군집분석을 통하여 그룹으로 구분한 연구도 있다. 장철호(2009)는 k-평균 군집분석 기법을 이용하여 공공도서관의 면적, 직원 수, 장서 수 데이터로 565개 도서관을 3개 그룹으로 구분하였다. 각 그룹에는 도서관 10개관, 148개관, 407개관이 각각 포함되었다.

통계학 분야에서는 다양한 군집분석 모델에 대한 비교 연구와 반복적인 군집분석 연구가 있다. 다양한 군집분석 모델에 대한 연구로 김재희, 고윤실(2009)은 계층적 군집분석의 최장연

결법과 와드법, 비계층적 군집분석의 k-평균기법 등을 이용하여 모의실험을 통해서 각 모델들을 비교하였다. 또한 반복적인 군집분석을 활용한 연구로 조용준(2009)은 수산업 분야의 통계자료인 조업정보를 이용하여 2단계 군집분석을 통하여 데이터를 8가지 군집으로 구분하였다.

2. 연구방법

2.1 연구의 기본 틀

공공도서관을 그룹으로 묶기 위해서는 크게 3가지 측면을 고려할 수 있다. 첫째, 그룹 구성의 기초가 되는 데이터에 대한 내용으로, 어떤 데이터를 이용하여 그룹을 만들 것인가 결정해야 한다. 공공도서관과 관련된 대표적인 데이터로 국가도서관 통계시스템에서 제공하는 공공도서관 통계가 있다. 그밖에 도서관 사서 혹은 지역주민들을 대상으로 한 설문조사 결과 데이터 등을 생각할 수 있다. 이 연구에서는 규모를 기준으로 공공도서관을 구분하는데 목적이 있기 때문에 공공도서관 통계 데이터를 사용하였다.

둘째, 공공도서관 관련 변수에 대한 부분으로 어떤 변수를 사용하여야 목적에 적합한 그룹을 만들 수 있는지 선택해야 한다. 공공도서관 통계 데이터에는 도서관 시설, 장서, 인력, 예산, 서비스 등에 관한 세분화된 변수들을 포함하고 있다. 이 연구에서 사용한 변수는 전국도서관 운영평가에서 공공도서관을 구분하기 위해 사용한 변수를 기준으로 공공도서관 통계 데이터에서 선택하였다. 또한 도서관 서비스와 관련된 변수로 도서관 회원 수, 대출자 수, 대출자료 수, 이용자 교육, 문화프로그램 운영, 직원의 전문 교육 등을 선택하여 사용하였다.

셋째, 그룹화 방법에 대한 부분으로 여러 가지가 있을 수 있으나, 이 연구에서는 머신러닝의 군집분석 기법을 사용하였다. 군집분석이 데이터 간의 유사성을 기준으로 군집을 작성하기 때문에 연구의 목적과 일치한다고 판단된다.

2.2 연구의 질문

이 연구의 목적을 달성하기 위해 연구의 질문은 두 가지로 구분된다. 첫째, 공공도서관 그룹화를 위해 어떤 군집분석 모델이 적합한가? 둘째, 공공도서관 통계를 사용하여 군집분석을 실행했을 때, 특징은 무엇인가?

2.3 분석 대상 및 절차

분석에 사용할 데이터는 문화체육관광부 국가도서관통계시스템(2020b)에서 제공하는 2018년도 공공도서관 통계 데이터로 엑셀 형식으로 저장되어 있다. 엑셀 파일에서 분석에 필요한 필드만을 추출하여 csv 형식의 파일로 변환시켜 분석하였다. 파일에 포함된 공공도서관 수는 1,096개관이며, 분석을 위해 추출한 필드는 순번, 지역, 개관년도, 도서관명 등의 기본 정보를 담고 있는 필드와 도서관 규모와 도서관 서비스에 관련된 필드 등이 포함되었다(〈표 1〉 참조).

데이터 분석은 RStudio를 통하여 R 프로그램을 이용하였다. 군집분석을 위해 군집분석용 패키지인 NbClust, cluster 등의 라이브러리를 이용하였다. 데이터 처리는 데이터 정규화 작

〈표 1〉 그룹화를 위해 사용한 변수

구분	사용한 필드(변수)
규모 구분용 필드 (6개 필드)	국내서, 국외서, 비도서, 도서관 면적, 인력자원, 예산(자료구입비)
서비스 분석용 필드 (20개 필드)	직원 전문교육 건수, 직원 전문교육 참가자 수, 직원 전문교육 시간, 어린이 회원 수, 청소년 회원 수, 성인 회원 수, 대출자 수, 어린이 대출 책 수, 청소년 대출 책 수, 성인 대출 책 수, 이용자교육 횟수, 이용자교육 참가자 수, 이용자교육 시간, 문화프로그램 강좌 횟수, 문화프로그램 강좌 참가자 수, 독서프로그램 강좌 횟수, 독서프로그램 참가자 수, 독서 동아리 수, 학습동아리 수, 지식취약대상 봉사 수

※ 여러 필드로 구분되어 있는 변수는 적합한 명칭으로 합하여 사용

업을 통해 필드에 따라 범위가 다양한 데이터 값의 분포를 표준화하였다. 적용한 군집분석 모델은 여러 가지 군집분석 모델 중 가장 일반적으로 사용되는 계층적 군집분석 중 완전측정법과 와드측정법, k-평균 군집분석 3가지 모델을 사용하였다. 각 모델들의 주요 내용은 〈표 2〉와 같다.

3. 분석 결과

3.1 공공도서관 그룹화

3.1.1 군집분석 기법 적용

공공도서관을 그룹으로 나누기 위해 공공도서관 통계 데이터를 이용하여 군집분석을 실행한 결과 전국의 1,096개 공공도서관은 크게 2개 클러스터로 구분되었다. 분석을 위해 사용한 군집분석의 3가지 모델 모두 클러스터를 2~4개로 구분하였는데 각 클러스터에 포함된 도서관의 수는 매우 편향적인 모습을 보여주었다. 계층적 군집분석의 완전측정법 모델은 한 클러스터에는 1,095개 도서관이 다른 클러스터에는 오직 1개 도서관만 포함되었다. 반면에 다른 두 가지 군집분석 모델은 한 클러스터에 전국 공공도서관의 90% 정도가 포함되어 있었다.

군집분석을 사용하여 전국 공공도서관을 그룹화 할 때, 100개의 대규모 도서관과 그 외의 도서관이 포함되어 있는 클러스터로 구분된 이유는, 공공도서관 통계 데이터의 특징이 그 원인이라 생각된다. 공공도서관 통계 데이터는 양적 데이터로 대부분 필드(변수)의 값이 수량

〈표 2〉 군집분석 기법 구분

구분	모델	주요 내용
계층적 군집분석	완전측정법	두 군집 사이의 거리가 최대인 값을 측정하여 군집을 합하여 그룹화 하는 방법
	와드측정법	군집에 속한 점들의 중심으로부터 오차 제곱합을 기초로 군집을 만들어 그룹화 하는 방법
비계층적 군집분석	k-평균 군집분석	주어진 군집 수 k에 대해서 군집 내 거리 제곱 합의 합을 최소화하는 형태로 데이터 내의 개체들을 서로 다른 군집으로 그룹화 하는 방법

※ 참고: 한국직업능력개발원, 명지대학교 산학협력단. [2017]. 07 머신러닝 기반 데이터 분석. p. 86.

으로 표시되어 규모에 밀접한 영향을 받는다. 우리나라 공공도서관은 100여개 대형 도서관과 그 밖에 유사한 규모의 도서관으로 구성되어 있기 때문에 데이터 분석 결과 한쪽으로 치우친 결과를 얻을 수밖에 없었다. 다음은 각 군집분석 모델을 사용하여 분석한 결과이다.

3.1.2 계층적 군집분석-완전측정법

계층적 군집분석 중 완전측정법(complete) 모델을 이용하여 데이터를 분석한 결과, 9가지 지표가 2개 클러스터 사용을 제안하였고, 5개 지표가 3개 클러스터 사용을 제안하였다(〈그림 1〉 참조). 일반적으로 군집분석에서는 가장 많은 지표가 제안한 클러스터 수를 사용한다.

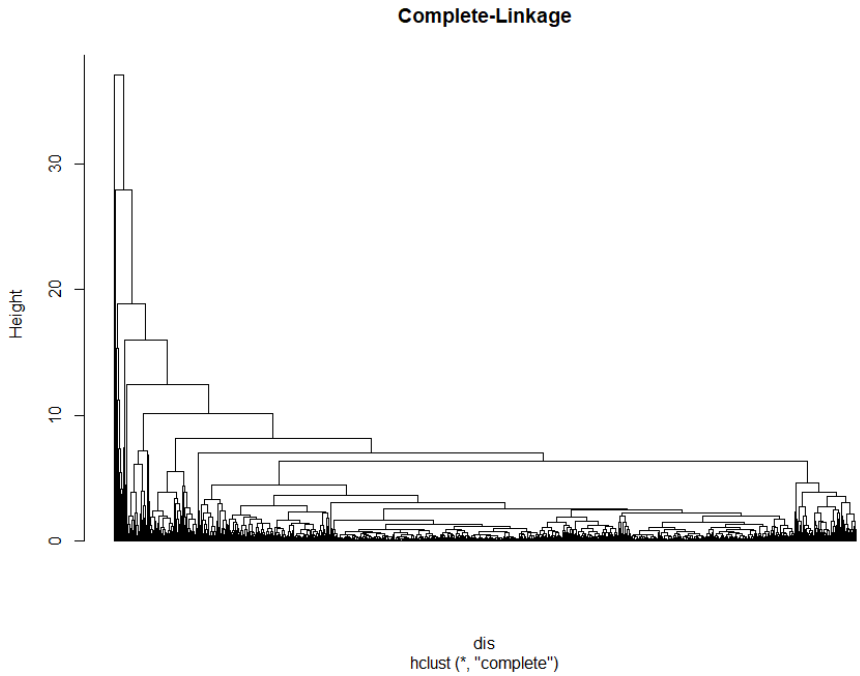
유클리디안(euclidean) 거리행렬 모델을 적용하여 계층적 군집화와 덴드로그램(dendrogram)을 그렸으며, cutree 함수를 사용하여 클러스터 수 확인을 위한 표를 작성하였다. 덴드로그램 그림과 표로 보면 2개와 3개로 구분된 클러스터는 매우 편향된 모습을 보여주고 있다(〈그림 2〉, 〈그림 3〉 참조). 군집분석 결과 제안한 2개의 클러스터를 기준으로 각 클러스터에 포함된 도서관 수를 살펴보면, 클러스터 1에 1,095개 도서관이 포함되고, 클러스터 2에는 오직 1개의

```
> numComplete <- NbClust(librarydata, distance = "euclidean", min.nc = 2,
                         max.nc = 20, method = "complete", index = "all")
------ 중략 ------
* 9 proposed 2 as the best number of clusters
* 5 proposed 3 as the best number of clusters
```

〈그림 1〉 계층적 군집분석 완전측정법 R 코드 및 결과

```
> dis <- dist(librarydata_matrix, method="euclidean")
> hc <- hclust(dis, method="complete")
> plot(hc, hang=-1, labels = FALSE, main="Complete-Linkage")
> comp2 <- cutree(hc, 2)
> table(comp2)
comp2
   1    2
1095    1
> comp3 <- cutree(hc, 3)
> table(comp3)
comp3
   1    2    3
1093    2    1
```

〈그림 2〉 계층적 군집분석 완전측정법 모델의 군집 테이블 작성 R 코드 및 결과



〈그림 3〉 계층적 군집분석 완전측정법 모델의 덴드로그램

도서관만 포함되었다. 3개의 클러스터를 제안한 경우도 이와 유사하게 클러스터 1에 1,093개 도서관, 클러스터 2에 2개 도서관, 클러스터 3에 1개 도서관이 포함되었다. 이처럼 극단적으로 군집이 나뉘는 이유는 1~2개 도서관이 다른 도서관에 비해 규모 면에서 크게 차이가 나는 값을 가지고 있기 때문이다.

3.1.3 계층적 군집분석-와드측정법

계층적 군집분석의 와드측정법(ward) 모델을 이용하면 완전측정법 모델과 다른 형태의 클러스터 모습을 보여준다. 와드측정법 모델을 이용하여 데이터를 분석한 결과, 8가지 지표가 4개 클러스터 사용을 제안하였고, 4개 지표가 3개 클러스터 사용을 제안하였다(〈그림 4〉 참조).

와드측정법 모델의 덴드로그램과 클러스터 수를 확인한 결과, 가장 많은 지표가 제안한 클러스터 4개를 만든 결과는 완전측정법 모델과는 전혀 다른 결과를 보여주었다. 4개의 클러스터는 클러스터 1에 974개 도서관, 클러스터 2에 115개 도서관, 클러스터 3에 2개 도서관, 클러스터 4에 5개 도서관이 포함되었다(〈그림 5〉 참조). 크게 본다면 2개의 클러스터로 구분되었다고 볼 수 있으며, 각 클러스터에 포함된 도서관 수는 9:1의 비율이었다.

3.1.4 k-평균 군집분석

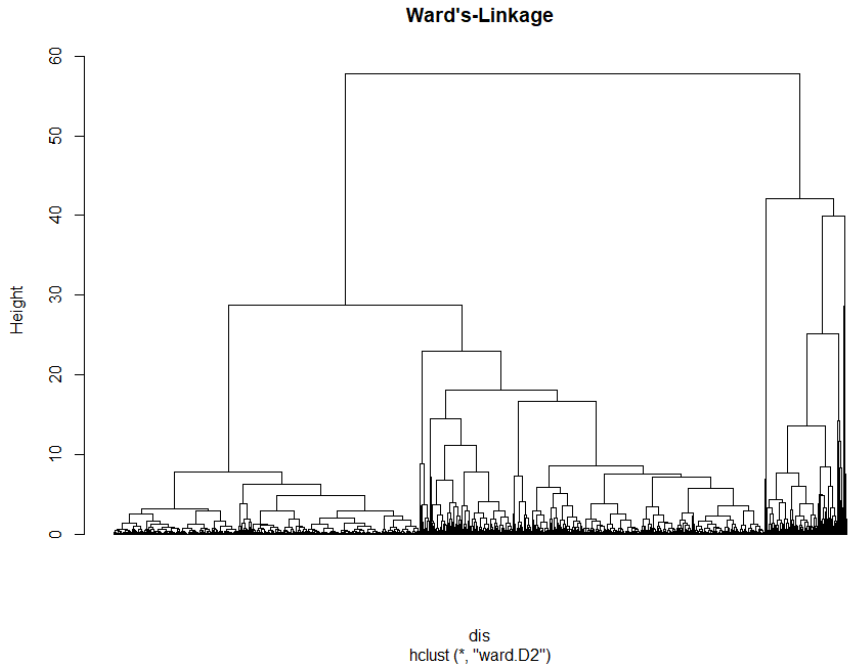
k-평균 군집분석은 계층적 군집분석과는 다른 방법이다. k-평균 군집분석에서 가장 많은 지표가 제안한 클러스터 수는 3개였다. 이 모델에는 클러스터 1에 170개 도서관, 클러스터 2에 2개 도서관, 클러스터 3에 924개 도서관이 포함

```
> nymWard <- NbClust(librarydata, diss=NULL, distance = "euclidean", min.nc = 2,
                     max.nc = 20, method = "ward.D2", index = "all")
----- 중략 ------
* 3 proposed 2 as the best number of clusters
* 4 proposed 3 as the best number of clusters
* 8 proposed 4 as the best number of clusters
```

〈그림 4〉 계층적 군집분석 와드측정법 모델의 R 코드 및 결과

```
> hcWard <- hclust(dis, method="ward.D2")
> plot(hc, hang=-1, labels = FALSE, main="Complete-Linkage")
> ward4 <- cutree(hcWard, 4)
> table(ward4)
ward4
  1   2   3   4
974 115   2   5
```

〈그림 5〉 계층적 군집분석 와드측정법 모델의 군집 테이블 작성 R 코드 및 결과



〈그림 6〉 계층적 군집분석 와드측정법 모델의 덴드로그램

되었다. 이 모델이 만든 클러스터는 크게 2개 클러스터로 구분할 수 있는데 각 클러스터에 포함된 도서관 수의 비율은 앞의 와드측정법 모델과 유사하게 9:1 비율이었다.

3.1.5 군집분석 모델 비교

공공도서관 통계 데이터를 이용한 도서관 그룹화에 적합한 모델은 k-평균 군집분석과 계층적 군집분석의 와드측정법 모델이라 판단된다. 이들 두 모델은 클러스터에 포함된 도서관 수가 매우 편향되기는 하지만, 대규모 도서관이 포함된 클러스터와 그 이외의 도서관 집단이 포함된 클러스터로 구분하였다. 반면에, 계층적 군집분석의 완전측정법 모델은 그룹화라 표현하기 어려운 1개 도서관과 1,095개 도서관으로 클러스터를 구분하였다. 그러므로 공공도서관 통계 데이터를 이용한 그룹화 방법으로 완전측정법 모델은 적합하지 않다고 판단할 수 있다.

k-평균 군집분석과 계층적 군집분석의 와드측정법 모델 사이에서 클러스터 구분의 정확도를 비교하기 위해 각 변수들을 박스플롯(box-plot) 그림을 사용하였다(〈그림 9〉 참조). 도서관 규모와 관련된 변수로 사용한 국내서 장서량의 경우, k-평균 군집분석과 와드측정법 모델의 결과는 이상치를 제외하고, 비교적 잘 구분되고 있는 모습을 보여준다. 〈그림 10〉을 보면 k-평균

```
> numKMeans <- NbClust(librarydata, diss=NULL, distance = "euclidean", min.nc = 2,
                       max.nc = 20, method = "kmeans", index = "all")
------ 중략 ------
* 7 proposed 2 as the best number of clusters
* 8 proposed 3 as the best number of clusters
```

〈그림 7〉 k-평균 군집분석 모델의 R 코드 및 결과

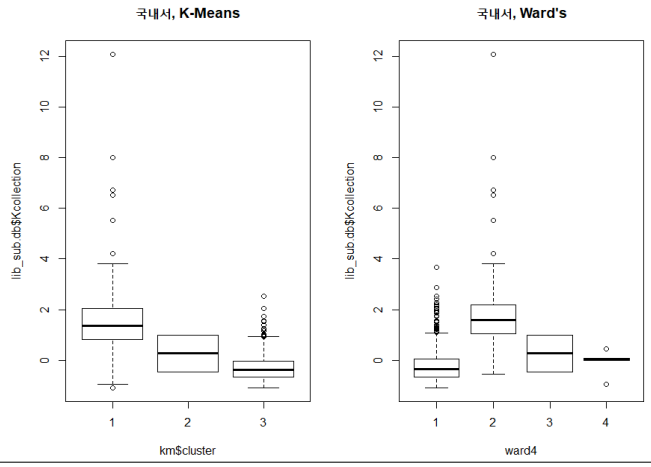
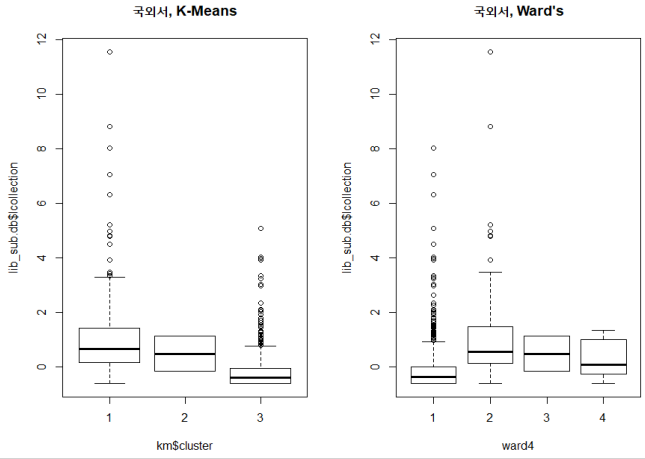
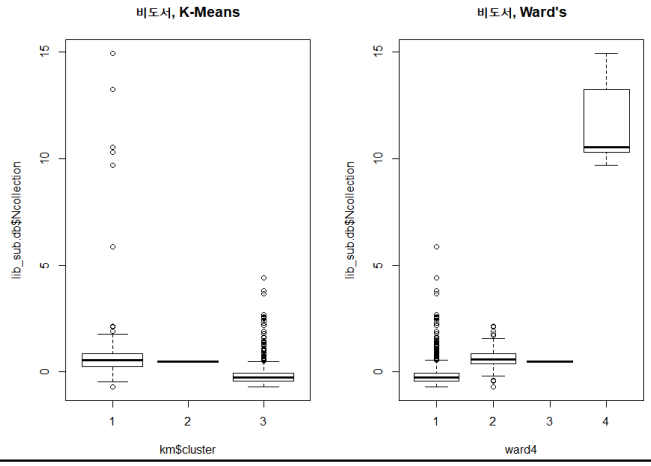
```
> set.seed(1234)
> km <- kmeans(librarydata_matrix, 3, nstart=25)
> table(km$cluster)

  1   2   3
170   2 924
```

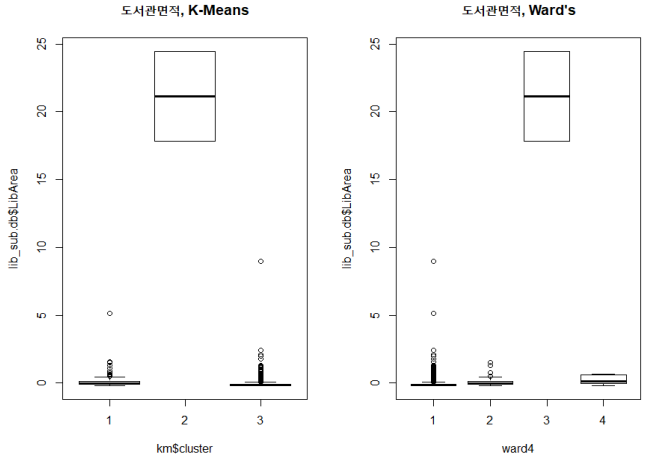
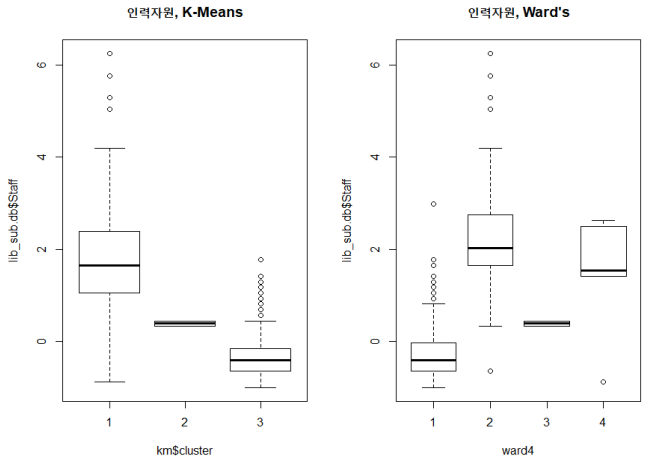
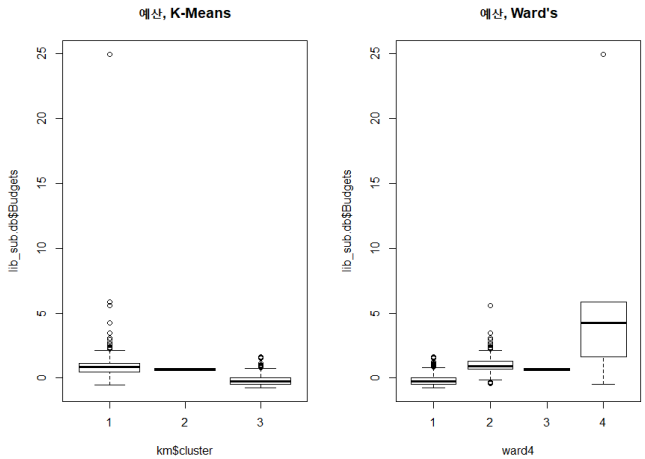
〈그림 8〉 k-평균 군집분석 군집 모델의 테이블 작성 R 코드 및 결과

```
> boxplot(librarydata$Kcollection ~ km$cluster, data=lib_sub.db, main = "국내서, K-Means")
> boxplot(librarydata$Kcollection ~ ward4, data = lib_sub.db, main = "국내서, Ward's")
```

〈그림 9〉 박스블록 그리기 R 코드

변수	클러스터 비교
국내서	
국외서	
비도서	

〈그림 10〉 변수별로 k-평균 군집분석과 계층적 군집분석 와드측정법 모델의 박스블록 비교(계속)

변수	클러스터 비교
도서관 면적	
인력 자원	
예산	

〈그림 10〉 변수별로 k-평균 군집분석과 계층적 군집분석 와드측정법 모델의 박스블록 비교

군집분석의 클러스터 1의 상위 75%는 다른 클러스터와 중복되지 않고 있다(단, 클러스터 2는 2개 도서관만 포함하고 있어서 비교에서 제외한다). 와드측정법 모델을 이용한 결과에서도 클러스터 2의 상위 75%는 다른 클러스터와 중복되지 않고 있다.

이러한 방법으로 다른 변수들을 비교하면, 도서관 면적을 제외하고는 k-평균 군집분석과 와드측정법 모델에서 만들어진 클러스터는 도서관들을 상당히 잘 구분하고 있다고 판단할 수 있다. 도서관 면적의 경우는 k-평균 군집분석이나 와드측정법 모델에서 구분한 클러스터들이 거의 서로 구분되지 않아 이들 군집분석 모델이 클러스터를 구분할 때, 도서관 면적은 거의 영향을 미치지 못했다고 판단할 수 있다.

변수별로 특이한 사항을 살펴보면, 국외서 장서량의 경우 규모가 큰 클러스터에 속한 도서관의 경우, 장서량 규모가 다양하지만, 규모가 작은 클러스터에 속한 도서관의 경우 이상치를 제외하고는 대부분 비슷한 양의 장서량을 보여주고 있다. 예산의 경우, 규모가 큰 클러스터에 속한 도서관이나 규모가 작은 클러스터에 속한 도서관 모두가 비슷한 수준의 예산을 가지고 있는 것으로 나타나 있다.

이처럼 군집분석 결과를 박스플롯 함수를 이용하여 시각화하면, 클러스터가 얼마나 정확하게 구분되었는지 확인할 수 있다. 동시에 클러스터 별로 데이터의 범위를 최솟값에서 1사분위수(25%), 중앙값(50%), 3사분위수(75%) 최댓값까지 확인할 수 있다. 결과적으로 k-평균 군집분석과 와드측정법 모델은 규모를 기준으로 하는 도서관 그룹화에 있어서 거의 유사한 정확도를 보이고 있다.

3.2 그룹 세분화를 위한 군집분석

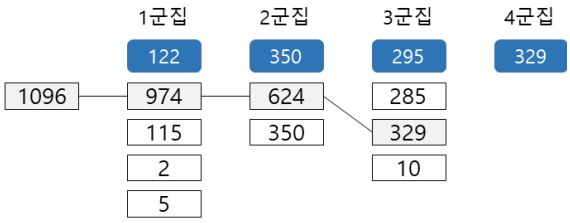
3.2.1 군집화 방법

일반적인 데이터 세트를 대상으로 군집분석을 실시하는 경우 1회의 군집분석으로 데이터 분석이 종료된다. 하지만, 공공도서관 통계 데이터를 이용한 군집분석 결과는 도서관 간의 규모가 9:1 비율로 양분되어 한쪽으로 치우치는 결과를 보여주고 있다. 그러므로 도서관 그룹의 크기를 유사하게 만들기 위해 반복적인 군집분석이 가능한지 실험하였고, 만들어진 클러스터가 서로 명확하게 구분되는지 분석하였다.

이를 위해 첫 번째 군집분석으로 구분된 클러스터 2에 포함된 974개 도서관을 대상으로 다시 군집분석을 실시하였다. 하지만, 두 번째 군집분석 역시 한쪽으로 편향된 2개의 클러스터로 구분되었다. 그러므로 구분된 클러스터가 비슷한 규모로 나누어질 때까지 반복적으로 군집분석을 실시하였다. 그리고 군집분석 결과로 매번 만들어진 클러스터 사이의 구분이 명확하게 이루어졌는지 비교하였다.

k-평균 군집분석 모델과 계층적 군집 분석의 와드측정법 모델의 결과로 만들어진 클러스터는 유사하지만, k-평균 군집분석 모델이 이상치에 민감하게 반응하므로, 와드측정법 모델을 적용하였다. 공공도서관 규모를 기준으로 총 3회의 군집분석을 반복적으로 실시하여 최종적으로 4개의 클러스터로 공공도서관을 그룹화 하였다.

첫 번째 군집분석 결과 4개 클러스터로 나뉘었는데, 974개 도서관이 포함된 클러스터를 제외하고 나머지 3개 클러스터가 유사하여 이들 3개 클러스터를 하나로 묶어 1군집으로 구분하고, 974개 도서관이 포함된 클러스터를 대상으



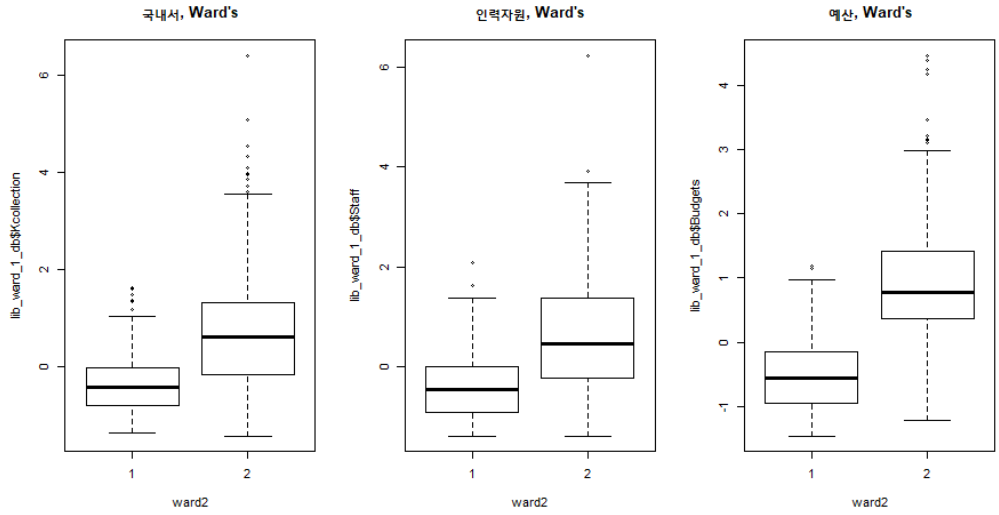
〈그림 11〉 그룹 세분화에 따른 군집별 포함 도서관 수

로 다시 군집분석을 실시하였다. 군집분석 결과 2개의 클러스터로 나뉘었고, 이때 규모가 큰 클러스터를 대상으로 다시 군집분석을 실시하는 방법으로 최종적으로 4개의 군집을 만들었다.

그리고 각 단계별 군집분석의 정확도를 측정하기 위해 각각의 군집분석 결과를 박스플롯 그림을 이용하여 분석하였다. 명확한 비교를 위해 군집분석 결과 차이가 있었던 3가지 변수, 즉 국내서, 인력자원, 예산을 대상으로 삼았다.

3.2.2 2단계 군집분석

2단계 군집 분석은 974개 도서관이 포함된 군집을 대상으로 와드측정법 모델을 이용하여 군집분석을 실시하였다. 분석 결과 2개 클러스터로 구분되었는데, 클러스터 1은 624개 도서관을 포함하고 있었고, 클러스터 2는 350개 도서관을 포함하고 있었다. 군집분석의 정확도를 측정하기 위해 박스플롯 그림을 이용하여 국내서, 인력자원, 예산 변수를 비교하였다. 비교 결과 두 클러스터는 부분적으로 차이점을 보여주고 있었다(〈그림 12〉 참조). 이는 첫 번째 군집분석에서 동일한 클러스터로 구분한 도서관들을 대상으로 다시 군집분석한 결과이기 때문이다. 변수별 클러스터 구분에서 예산 변수는 두



〈그림 12〉 2단계 군집분석 결과 국내서, 인력자원, 예산의 박스플롯 비교

클러스터에서 도서관의 75%가 명확하게 구분되는 모습을 보여주고 있다. 값의 차이를 살펴보면 클러스터 2가 3개 변수를 기준으로 더 높은 값을 가지고 있었다.

3.2.3 3단계 군집분석

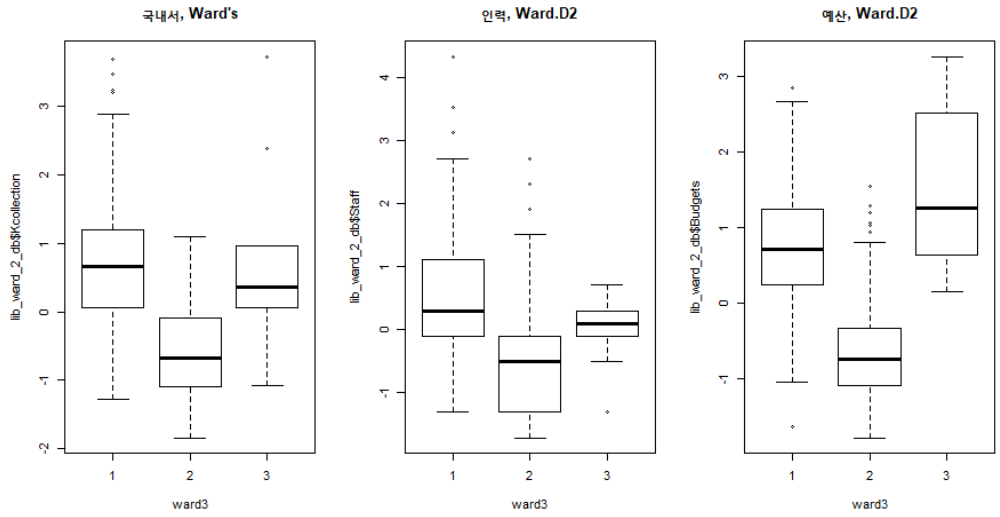
3단계 군집분석은 2단계 군집분석에서 624개 도서관을 포함하고 있는 클러스터를 대상으로 실시하였다. 군집분석 결과 3개 클러스터로 구분되었다. 클러스터 1은 285개 도서관이 포함되었고, 클러스터 2는 329개 도서관, 클러스터 3은 10개 도서관이 포함되었다. 클러스터 1과 클러스터 3은 서로 유사하므로 이를 합하여 군집 3으로 정하고, 클러스터 2를 군집 4로 정하였다. 클러스터의 구분 정도를 비교하기 위해, 3 가지 변수를 기준으로 박스플롯 그림을 이용하였다. 클러스터 1과 클러스터 3의 75%가 클러스터 2와 명확하게 구분되었다(〈그림 13〉 참조).

3.2.4 군집별 지역 적용

군집분석 결과 만들어진 4개 군집은 도서관 규모별로 1군집이 가장 크고, 4군집이 가장 작았다. 이들 군집을 지역별로 정리하면 다음의 〈표 3〉과 같다. 전체적으로 보았을 때, 주로 광역시에 위치한 도서관들이 1군집에 포함된 비율이 상대적으로 많았고, 부산, 대전, 세종, 경기, 전남, 경남에 위치한 도서관들이 2군집에 포함된 비율이 평균보다 많았다. 다시 말하면 이들 지역의 도서관 규모는 타 지역에 비해 상대적으로 크다고 이야기할 수 있다. 반면에 서울, 대구, 인천, 울산, 강원, 전남에는 소규모 도서관 비율이 타 지역보다 높다는 것을 보여주고 있다.

3.3 도서관 서비스를 기준으로 한 군집화

도서관을 그룹으로 구분하는 기준은 도서관 규모 외에도 도서관 서비스를 기준으로 적용할



〈그림 13〉 3단계 군집분석 결과 국내서, 인력자원, 예산의 박스플롯 비교

〈표 3〉 지역에 따른 군집별 도서관 수(비율)

	1군집	2군집	3군집	4군집	계
서울	29(16.8%)	34(19.7%)	31(17.9%)	79(45.7%)	173(100%)
부산	6(14.0%)	25(58.1%)	4(9.3%)	8(18.7%)	43(100%)
대구	8(19.5%)	12(29.3%)	7(17.1%)	14(34.2%)	41(100%)
인천	11(22.0%)	12(24.0%)	8(16.0%)	19(38.0%)	50(100%)
광주	3(13.0%)	7(30.4%)	6(26.1%)	7(30.4%)	23(100%)
대전	2(8.3%)	8(33.3%)	9(37.5%)	5(20.8%)	24(100%)
울산	5(26.3%)	2(10.5%)	4(21.1%)	8(42.1%)	19(100%)
세종	0(0.0%)	5(50.0%)	0(0.0%)	5(50.0%)	10(100%)
강원	4(7.0%)	8(14.0%)	23(40.4%)	12(38.6%)	57(100%)
경기	26(9.8%)	137(51.9%)	55(20.8%)	46(17.5%)	264(100%)
충북	4(8.9%)	7(15.6%)	13(28.9%)	21(46.7%)	45(100%)
충남	7(11.3%)	13(21.0%)	26(41.9%)	16(25.8%)	62(100%)
전북	1(1.7%)	14(21.1%)	16(27.6%)	27(46.5%)	58(100%)
전남	3(4.3%)	24(34.8%)	24(34.8%)	18(26.0%)	69(100%)
경북	5(7.7%)	9(13.8%)	32(49.2%)	19(29.3%)	65(100%)
경남	4(5.6%)	31(43.7%)	23(32.4%)	13(18.4%)	71(100%)
제주	4(18.2%)	2(9.1%)	14(63.6%)	2(9.1%)	22(100%)
계	122(11.1%)	350(31.9%)	295(26.9%)	329(30.1%)	1,096(100%)

수 있다. 대규모 군집(1군집)에 포함된 122개 도서관을 대상으로, 20개의 도서관 서비스 변수를 사용하여 군집분석을 실시하였다. 와드측정법 모델을 이용한 군집분석 결과 클러스터 1에 60개 도서관이 포함되었고, 클러스터 2에 62개 도서관이 포함되었다. 각 변수별로 클러스터 구분의 정확도를 측정하기 위해 박스플롯 그림을 통하여 비교하였다([부록 1] 참조).

비교 결과, '직원 전문교육', '이용자 교육 횟수', '이용자 교육시간', '독서프로그램 강좌 횟수', '독서프로그램 참가자 수', '지식취약대상 봉사' 변수는 클러스터 사이에서 75% 이상이 구분되었다. 전체 20개 변수 중 이들 변수들이 클러스터를 나누는데 중요한 역할을 담당하고 있다고 판단된다.

반면에 도서관 서비스 변수 20가지 중 클러스터 사이에 중복 비율이 높은 변수는 '어린이 회원 수', '청소년 회원 수', '성인 회원 수', '대출자 수', '어린이 대출 책 수', '청소년 대출 책 수', '성인 대출 책 수'의 7가지로 도서관 이용자와 대출에 관련된 변수이다. 이들 변수들 특징을 파악하기 위해 7가지 변수만을 추출하여 군집분석을 실시하였다.

군집분석 결과 2개의 클러스터로 구분되었는데, 클러스터 1에는 51개 도서관이 포함되고 클러스터 2에는 71개 도서관이 포함되었다. 7가지 변수에서 클러스터 사이의 구분은 '어린이 회원 수', '청소년 회원 수', '어린이 대출 책 수', '청소년 대출 책 수'에서 잘 구분이 되었다([부록 2] 참조). 이 결과는 군집변수를 실시할 때 사용하는 변수에 따라 클러스터 구분이 차이가 있다는 것을 보여준다.

도서관 서비스와 관련된 20개 변수를 사용한 군집분석과 이중 이용자와 대출에 관련된 7가지 변수만 사용한 군집분석 결과를 비교하면 매우 흥미로운 결과를 발견할 수 있다. 2가지 군집분석 결과 동일한 클러스터에 포함된 도서관 수의 비율이 54.9%((28+32)/(60+62) × 100)에 불과하였다. 20개 변수를 사용한 결과에서 클러스터 1에 포함된 도서관 수는 60개관이었으나, 이들 60개관은 7개 변수만 사용한 결과에서 동일한 클러스터에 포함된 비율은 46.7%(28/60 × 100)에 불과하였다. 또한 20개 변수를 사용한 결과에서 클러스터 2에 포함된 도서관의 62.9% (39/62 × 100)만이 7개 변수만 사용한 경우에도 동일한 클러스터에 포함되었다.

이 결과는 공공도서관 통계 데이터를 사용하여 도서관을 군집분석 할 때, 도서관의 핵심 서비스인 이용자와 대출 관련 변수가 군집분석에 큰 영향을 미치지 못하고 있음을 보여준다. 다양한 도서관 서비스 관련 변수를 이용하여 군집분석을 하였을 때, 이용자와 대출 관련 변수는 클러스터를 구분하는데 중심이 되지 못하고, 다른 변수들 중심으로 클러스터가 만들어 질 수 있다. 그러므로 공공도서관 통계 데이터를 이용하여 공공도서관을 서비스 기준으로 그룹으로 구분할 때, 다양한 서비스 관련 많은 변수를 사용하는 대신, 도서관 그룹화 목적에 관련이 있는 소수의 변수를 사용하는 것이 매우 중요하다고 판단된다.

4. 공공도서관 그룹화 결과의 시사점

4.1 공공도서관 서비스에 관한 데이터

공공도서관 그룹화를 위해 통계 데이터를 사용하였을 때, 통계에 포함된 대부분의 데이터는 양적 데이터로 도서관 규모와 밀접한 관련이 있는 것들이었다. 비록 서비스와 관련된 내용의 변수를 기준으로 삼아 그룹을 만든다 하더라도 도서관 규모가 영향을 끼칠 수 밖에 없다. 이는 공공도서관 통계의 한계라 생각할 수 있다.

도서관 통계 역시 공공도서관 서비스에 관련된 모든 내용을 다 포함할 수 없다. 예를 들면, 공공도서관 통계에는 이용 및 이용자 분야에 정보서비스 의뢰와 정보서비스 제공 필드가 있는데, 단순히 의뢰 건수와 제공 건수만 포함되어 있다. 공공도서관에서 제공하는 정보서비스의 범위가 어느 정도인지 혹은 정보서비스 수준이 어느 정도인지 알 수 없다. 대출에 관련된 내용도 보다 세부적으로 파악한다면 1인당 주

〈표 4〉 도서관 서비스 변수와 이용자 및 대출 책 수 변수 결과 비교

		이용자 및 대출 책 수 변수(7개 변수) 기준		
		1	2	합계
서비스 전체 변수 (20개 변수) 기준	1	28	32	60
	2	23	39	62
	합계	51	71	

제별 대출 책 수가 몇 책이며, 도서를 많이 대출하는 이용자와 적게 대출하는 이용자 수는 몇 명인지 알 수 없다.

예를 들어, 두 도서관이 도서관 통계에서 대출 책 수가 같다고 해서 두 도서관이 대출에 있어서 유사하다고 이야기할 수 없을 것이다. 한 도서관은 소수의 이용자가 많은 도서를 대출하지만, 다른 도서관은 많은 이용자가 한 책 혹은 두 책의 도서만 대출하였다면, 두 도서관은 전혀 다른 특성을 가지고 있고 서비스 방향도 달라져야 할 것이다. 그러므로 보다 정확하게 유사한 도서관을 그룹화하기 위해서는 공공도서관 통계 데이터만으로는 충분하지 않고, 도서관 서비스에 관련된 다양한 데이터가 체계적으로 수집되어야 할 것이다.

4.2 데이터 세트에 적합한 그룹화 모델

공공도서관 통계 데이터를 이용하여 군집분석을 한 이유는 공공도서관을 규모에 따라 유사한 집단으로 구분하기 위함이었다. 이 연구에서는 계층적 군집분석의 완전측정법, 계층적 군집분석의 와드측정법, k-평균 군집분석법의 3가지 모델을 사용하였는데, 모델에 따라 도서관 클러스터가 다르게 만들어졌다. 공공도서관 통계 데이터를 이용하여 1,096개 도서관을 그룹으로 구분하기 위한 목적이라면, 계층적 군집분석의 와드측정법과 k-평균 군집분석법 모델이 적합할 것으로 판단된다. 하지만, 공공도서관 통계 데이터가 아닌 다른 유형의 공공도서관 데이터를 이용하여 군집분석을 한다면 다른 모델이 더 적합할 수도 있다.

공공도서관 통계 데이터를 이용하여 군집분석을 실시한 결과 크게 두 가지 클러스터로 구분되었다. 이 결과를 본다면 문화체육관광부에서 실시하는 전국 도서관 운영평가에서 공공도서관을 2개 집단으로 구분하는 방안을 고려할 필요가 있을 것 같다. 대규모의 도서관 그룹에 포함되지 않는 약 90% 공공도서관은 적용하는 변수에 따라 다른 그룹에 포함될 가능성이 크다. 그러므로 특정 도서관이 유사한 도서관 그룹임에도 불구하고 다른 그룹으로 세분화될 가능성이 있어서, 도서관 서비스 향상을 위해 그룹을 구분하는 목적이 훼손될 수 있을 수 있다.

공공도서관을 여러 그룹으로 세분화해야만 한다면 군집분석을 반복적으로 실시하는 것도 고려해 볼 만하다. 일반적으로 군집분석 기법은 여러 번 반복적으로 사용하기 보다는 1회 사용하여 군집분석을 실시한다. 하지만 공공도서관 통계 데이터처럼 한쪽으로 치우친 형태의 데이터인 경우, 반복적인 군집분석을 사용할 필요가 있다. 이 경우는 군집분석 결과의 정확도를 판단하기 위해 만들어진 클러스터가 서로 명백하게 분리되어 있다는 것을 증명할 필요가 있다.

4.3 목적에 따른 군집분석의 활용

공공도서관 통계 데이터를 이용하여 군집분석을 실시한 결과, 공공도서관 서비스와 관련된 특정 변수의 포함 여부에 따라 소속되는 도서관 그룹이 달라지는 것을 확인할 수 있었다. 공공도서관 규모를 기준으로 그룹을 만든 후에 목적에 적합한 변수를 사용하여 그룹을 세분화할 수 있다. 이 방법은 단순히 유사한 규모의 도서관을 그룹으로 모아주는 역할 외에도 다른 목적에 따른 그룹을 구성하는데 활용할 수 있을 것이다.

예를 들면, 군집분석을 통하여 대규모 그룹에 속하는 도서관들을 구분한 후에 이들 도서관들을 대상으로 도서관 서비스와 관련된 변수를 사용하여 군집분석을 실시하였다. 그 결과 값이 큰 클러스터에 속한 도서관들을 대상으로 2019년도 전국 도서관 운영평가 우수 공공도서관 리스트와 비교했을 때, 다수의 도서관이 일치하였다. 우수 공공도서관이 단순히 양적인 지표를 통해서만 결정되는 것이 아니기 때문에 군집분석 결과와 우수도서관 선정 결과를 비교할 수 없지만 관련성은 파악할 수 있다.

군집분석은 공공도서관을 단지 그룹으로 구분 해주는 기능만을 가지고 있지 않다. 구분된 클러스터를 비교해보면 어떤 변수가 핵심적인 역할을 했는지 파악할 수 있다. 겉으로 들어나지 않아서 알 수 없었던 변수를 파악할 수 있는 장점도 가지고 있다. 공공도서관 규모를 기준으로 군집분석을 하였을 때, 도서관 면적은 영향이 미비하고, 대신 소장 장서, 인력자원, 예산 등이 그룹으로 구분하는데 큰 영향을 미친 것을 파악할 수 있다. 이러한 측면에서 군집분석 기법은 도서관 분야에서 다양하게 활용할 수 있을 것으로 판단된다.

5. 결론 및 제언

이 연구는 공공도서관 그룹화를 위해 적합한 군집분석 모델을 파악하고 그 특징을 분석하는데 목적이 있다. 공공도서관 통계 데이터 중 규모와 관련된 데이터를 이용하여 세 가지 군집분석 모델을 적용한 결과, 계층적 군집분석의 와드측정법과 k-평균 군집분석 모델이 서로 비슷한 형태의 군집을 만들어 규모를 기준으로 공공도서관 그룹화에 적합하였다. 이들 모델을 이용한 군집분석 결과 2개 군집으로 구분되었지만, 9:1로 한쪽으로 치우친 형태였다.

공공도서관 서비스와 관련된 변수를 기준으로 군집분석을 실시한 결과, 이용자와 대출 관련 변수는 군집분석에 영향을 미치지 못하고 다른 변수들 중심으로 군집이 만들어졌다. 그러므로 공공도서관 서비스 관련 통계 데이터를 이용하여 군집분석을 실시할 경우, 많은 변수를 포함시키는 대신, 군집분석 목적에 적합한 소수의 변수를 사용할 필요가 있다.

시사점으로 첫째 공공도서관 서비스에 관련된 데이터로 공공도서관 통계 데이터는 한계가 있으므로 서비스에 관한 다양한 데이터가 체계적으로 수집하여 사용한다면 공공도서관에 대한 보다 체계적인 그룹화가 가능할 것이다. 둘째, 군집분석 모델에 대한 부분으로 만약 공공도서관에 대한 다른 데이터 세트를 사용한다면 군집분석 기법의 다른 모델도 검토되어야 할 것이다. 셋째, 군집분석의 활용 부분으로 공공도서관 분야에서 군집분석은 도서관을 유사한 규모로 모아주는 역할 외에도 특정 변수의 역할도 파악할 수 있어서 공공도서관 서비스의 다양한 분야에 활용할 수 있다.

제언으로 공공도서관 통계 데이터가 연도별로 축적되어 있다. 연도별로 공공도서관이 어떤 군집에 포함되는지 그리고 어떤 변수로 인하여 해당 군집에 포함되는지 조사할 필요가 있다. 이 결과는 공공도서관의 발달을 파악하는데 중요한 데이터가 될 수 있을 것이다. 또한 공공도서관 통계 데이터와 우수도서관 사례 관계에 대한 체계적인 조사도 필요할 것이다.

참 고 문 헌

국가도서관통계시스템. 2020a. 열린마당. 2020년 전국 도서관 운영평가 공공도서관 그룹핑 변경 안내. [online]. [cited 2020.8.10]. 〈https://www.libsta.go.kr/libportal/openMdg/notice/getNoticeList.do〉.

국가도서관통계시스템. 2020b. 공공도서관 통계보기. [online]. [cited 2020.8.10]. 〈https://www.libsta.go.kr/libportal/libStats/publicLib/unitStats/getUnitStatsPop.do?gubun=STEP0000000001&libGubun=LIBTYPE002〉.

김재희, 고윤실. 2009. 군집분석 비교 및 한우 관능평가데이터 군집화. 『응용통계연구』, 22(4): 745-758.

장철호. 2009. Clustering DEA/AHP 모형을 이용한 전국 공공도서관 효율성 평가. 『한국도서관·정보학회지』, 40(2): 491-514.

조민호. 2019. 『R 데이터 분석: 데이터 분석 전문가를 위한』. 서울: 정보문화사.

조용준. 2009. 2단계 군집분석을 통한 해구별 조업정보의 유사성 분석. 『한국데이터정보과학회지』, 20(3): 551-562.

차미경, 표순희. 2015. 전국 공공도서관 운영평가의 성과에 관한 연구: 2010년~2013년도를 중심으로. 『한국비블리아학회지』, 26(2): 241-268.

한국직업능력개발원, 명지대학교 산학협력단. [2017]. 07 머신러닝 기반 데이터 분석. [online]. [cited 2020.2.10]. 〈https://www.ncs.go.kr/unity/th03/ncsResultSearch.do〉.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

Cha, Mikyeong and Soon Hee Pyo. 2015. "A Study on the Results of the National Evaluation on Public Library Management during 2010~2013." *Journal of the Korean Biblia Society for Library and Information Science*, 26(2): 241-268.

Cho, Min Ho. 2019. *R Data Analysis for Data Scientists*. Seoul: Information Publishing Group.

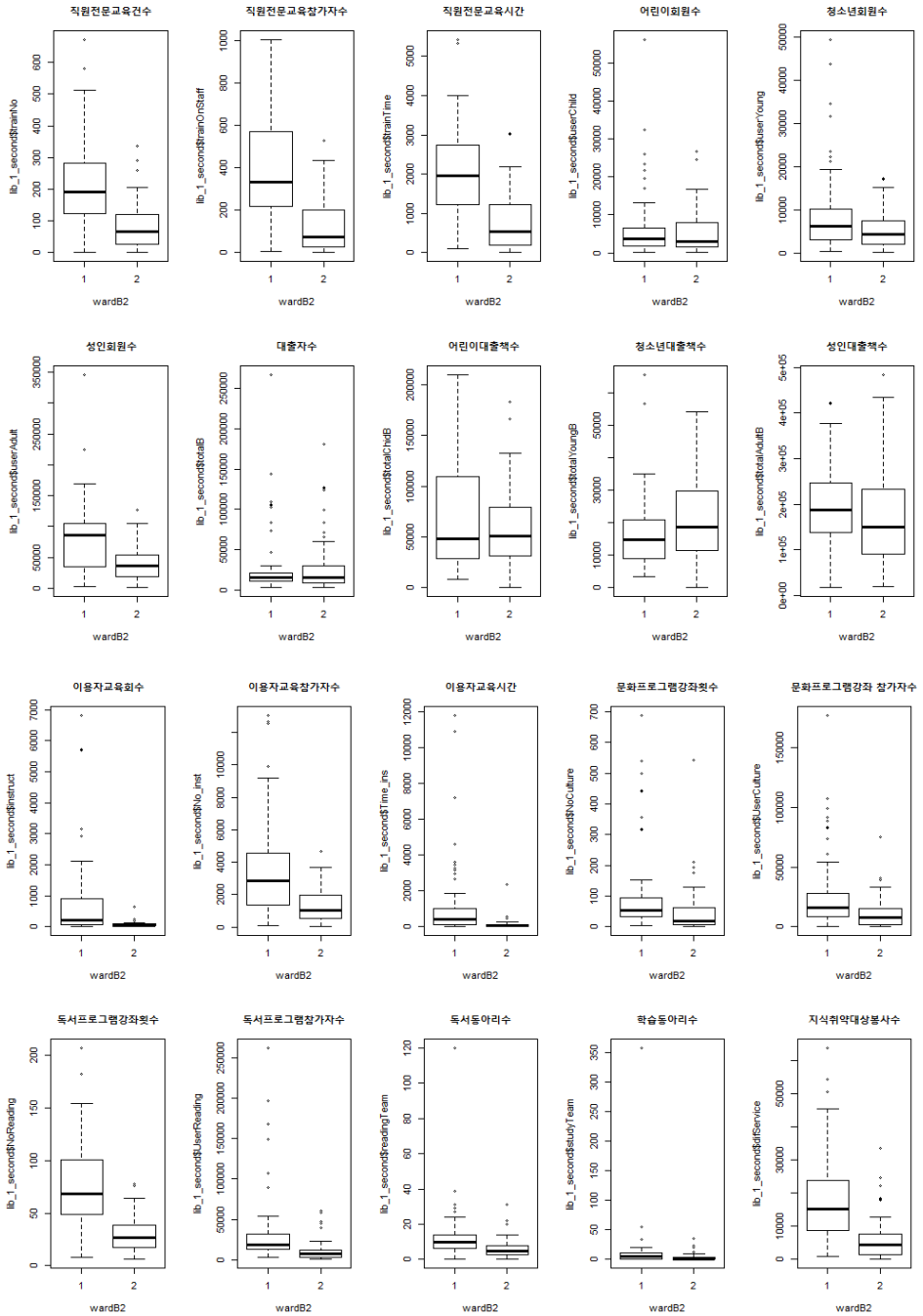
Cho, Yongjoon. 2009. "The Similarities Analysis of Location Fishing Information through 2 Step Clustering." *Journal of the Korean Data & Information Science Society*, 20(3): 551-562.

Jang, Chul-Ho. 2009. "A Study on Efficiency Analysis about the Public Libraries Using Clustering DEA/AHP Model." *Journal of Korean Library and Information Science Society*, 40(2): 491-514.

Kim, Jaehee and Yoon Sil Ko. 2009. "A Comparison of Cluster Analyses and Clustering of Sensory Data on Hanwoo Bulls." *Korean Journal of Applied Statistics*, 22(4): 745-758.

Korea Research Institute for Vocational Education & Training and Myongji University. Industry and Academia Cooperation Foundation. [2017]. *07 Data Analysis based upon Machine Learning*. [online]. [cited 2020.2.10].
〈https://www.ncs.go.kr/unity/th03/ncsResultSearch.do〉.

[부록 1] 대규모 도서관 군집의 와드측정법 모델을 이용한 군집분석 결과 군집간 비교



[부록 2] 대규모 도서관 군집의 와드측정법 모델을 이용한 이용자 및 대출책수 관련 7가지 변수를 이용한 군집분석 결과 군집간 비교