Linear Regression Modeling and Predictive Analysis of University Library Patron Borrowing Trends: Postprint

Authors: Wang Hong, Yuan Xiaoshu, Yuan Xiaoling, Huang Jianguo

Date: 2023-04-01T00:00:00+00:00

Abstract

[Purpose/Significance] By utilizing collection classification and circulation data, this study discovers the correlation between reader characteristics and collection circulation, establishes a relational model, and through model fitting and prediction, explores the underlying patterns between readers and book circulation, thereby providing technical and methodological support for smart library management.

[Method/Process] Employing clustering and correlation analysis techniques, macro-level observable reader characteristics are extracted to establish direct and indirect mapping relationships between reader characteristics and book classifications. Subsequently, a regression model between reader characteristics and classified book circulation volume is constructed, with model validity verified and goodness-of-fit optimized. Based on the validated model, library circulation trends are explored, and the essence and patterns of knowledge construction underlying macro-level reader characteristics are uncovered, along with their impact degree on book circulation.

[Results/Conclusion] Three categorical features related to readers—major study direction representing readers' social role requirements, enrollment batch representing inter-group interaction effects among readers, and reader group size—can effectively fit and predict book circulation volume. Prediction results demonstrate high model accuracy, enabling it to serve as an effective tool that provides reliable technical support for libraries to conduct knowledge services.

Full Text

Analysis and Prediction of Reader Borrowing Trends in University Libraries Using Linear Regression Modeling

Wang Hong[1], **Yuan Xiaoshu**[2], **Yuan Xiaoling**[3], **Huang Jianguo**[4] [1] Library, Shanxi University of Finance and Economics, Taiyuan 030006 [2] School of Information, Shanxi University of Finance and Economics, Taiyuan 030006 [3] Library, Taiyuan University of Science and Technology, Taiyuan 030024 [4] Taiyuan Daran Science and Technology Co., Ltd., Taiyuan 030006

Abstract: [Purpose/Significance] By analyzing library collection classification and circulation data, this study identifies correlations between reader characteristics and collection circulation, establishes relationship models, and explores underlying patterns between readers and book circulation through model fitting and prediction, providing technical support for intelligent library management. [Method/Process] Using clustering and correlation analysis techniques, we extracted macroscopic observable features of readers and established direct and indirect mapping relationships between reader characteristics and book classifications. We then constructed regression models linking reader features to classified book circulation volumes, validated model effectiveness, and optimized goodness-of-fit. Based on the validated models, we explored library circulation trends and uncovered the essential knowledge construction patterns hidden at the macro-characteristic level of readers and their impact on book circulation. [Result/Conclusion] Three reader classification features—professional learning direction representing social role requirements, enrollment batch representing group interaction effects, and reader population size—can effectively fit and predict book circulation volumes. Prediction results demonstrate high model accuracy, establishing an effective tool to provide reliable technical support for libraries to develop knowledge services.

Keywords: University libraries; Circulation prediction; Data mining; Linear regression

Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2020.03.007

Book circulation volume represents the outcome of interactions between readers and library collections, serving as a key metric linking readers to collections and constituting a core element in measuring library collection development and reader service quality. Establishing descriptive models based on circulation volumes of different book categories to predict collection circulation trends can not only improve library service quality and provide anticipatory guidance for library operations but also offer robust support for revealing the intrinsic operational patterns of collection circulation. University library readers interact frequently with collections, and although books and readers exhibit rapid periodic update rates, the construction of university library collections centered around disciplines and teaching services maintains relatively stable proportions across

categories regardless of collection growth. Similarly, while the reader population renews annually through incoming freshmen and graduating students, the overall number and identity characteristics of readers remain relatively stable. Therefore, the substantial circulation data generated between a stable reader population with consistent classification features and a stable collection base provides solid data support for exploring patterns between reader demand and collection circulation.

1.1 Related Basis

Readers, as independent and differentiated social individuals, exhibit behaviors fundamentally influenced by their current age, social roles, and expectations for future social roles and status—the most direct and fundamental factors underlying reading motivation. Different reader types demonstrate distinct demand preferences for specific book categories. For instance, female readers show preferences for novels with female protagonists, younger women favor romance literature, and married women prefer prose and travel books. Migrant workers' reading tendencies focus primarily on leisure literature and practical skill-based and examination-oriented books. Similarly, specialized books, particularly those with strong professional characteristics, typically target fixed reader groups with obvious professional features—for example, ancient literature collections mainly serve researchers conducting scientific work. University student readers, beyond differences in major directions, exhibit micro-level variations in geography, family education, and personality development. However, macro-level social backgrounds such as age, education, and growth, along with knowledge needs and motivational factors related to societal, life, emotional, family, and career expectations, demonstrate relatively homogeneous characteristics. University students represent a unique reader group with distinct social role characteristics, showing preferences for novels, natural sciences, and humanities/social science books beyond their major-related materials.

Reader identity features, particularly the knowledge demand preferences formed by differentiated characteristics, can reflect readers' book reading demand patterns. Consequently, describing and predicting collection circulation characteristics and trends based on borrowing data from readers with similar features becomes feasible.

1.2 Problem Definition

Employing mathematical analysis methods to examine university undergraduate readers' book borrowing preferences requires effective extraction and selection of reader features, designing hypothetical analysis models based on these characteristics, and exploring circulation relationships between different reader features and book categories. Establishing reasonable inferential premises forms the foundation for quantitative research. This study describes the relationship between book circulation volume and reader characteristics as follows:

Reader feature combination X = (x_1, x_2, …, x) has certain associations and mapping relationships with library knowledge classification C. Using historical circulation records of knowledge interactions between readers with these features as statistical data, we conduct correlation and cluster analysis to screen out significant reader characteristics, establish regression models, identify the influence degree of various factors on book circulation volume, and determine the goodness-of-fit for book circulation prediction. This approach employs rigorous mathematical methods to explain the causal relationship between reader characteristics and book demand, thereby exploring and revealing the patterns of book reading and circulation hidden behind readers' knowledge needs.

Definition 1: In a given reader set R and reader features x = (x_1 , x_2 , …, x), we establish a functional relationship between X and collection circulation y = (y_1, y_2, …, y), where i represents readers or reader categories and j represents books or book categories:

$$y_j = f(x_{1j}, x_{2j}, \dots, x_{ij}) \tag{1}$$

This study aims to establish appropriate models to identify suitable reader features X that can explain collection circulation Y, reasonably elucidate the quantitative causal relationship between X and Y, and predict Y based on X.

1.3 Related Research

Previous research exploring the internal mechanisms and trends of reader borrowing and circulation has primarily followed three approaches.

1.3.1 Data Comparison Mode: Based on survey and statistical data, this approach derives reader reading tendencies and preferences through quantitative indicator comparisons. Studies have analyzed reader reading tendencies and influencing factors through library circulation statistics, examined historical borrowing data to understand reader habits and demand changes, employed weighted circulation rate calculations for principal component analysis of student reading interests, and used network questionnaires to investigate reading purposes, content, methods, and levels. These studies consistently demonstrate that book circulation volume closely relates to readers' majors and future work life, as well as to character development, worldview formation, and interpersonal communication.

1.3.2 Correlation Analysis Mode: This approach uses statistical data to hypothesize relationships between certain reader characteristics and book circulation, applying statistical algorithms for correlation analysis to identify positive or negative correlations between different factors and reader borrowing. Research has employed scale questionnaires and second-order equation models to verify that readers' autonomous motivation and basic psychological need satisfaction positively influence extracurricular reading willingness, while controlled motivation shows no effect. Association rule algorithms (Apriori) have been

used to analyze reader demand characteristics and reading trends, providing reasonable bases for reader segmentation factors. Bayesian classification algorithms have revealed that libraries can establish reader profiles to understand backgrounds and borrowing interests, thereby improving reading behaviors and providing proactive recommendation services.

1.3.3 Modeling Analysis Mode: This approach establishes various analytical models through statistical data to explore the influence degree of reader behavioral characteristics and predict future borrowing demand trends. Studies have utilized multi-exponential smoothing methods for empirical analysis and prediction of monthly library borrowing data, Logistic regression models for analyzing user borrowing influencing factors, chaos theory for modeling and predicting library book borrowing flow behavior, grey neural network algorithms for predicting monthly book borrowing volumes, and grey system models for forecasting borrowing volumes in specific categories over five-year periods. These modeling approaches demonstrate that user borrowing quantities are influenced by electronic resources, extracurricular reading time, library environment, and peer influence, while factors such as education level, gender, college characteristics, and primary borrowing motivation show no significant effects.

2 Research Methods and Dataset

2.1 Research Methods

The core problem of collection circulation involves describing and predicting the relationship between circulation volume and reader characteristics. While many machine learning methods can accomplish descriptive and predictive tasks, this study employs the most mature and rigorous multiple linear regression method in machine learning, as it requires strict demonstration to explain the influence relationship of reader characteristics on collection circulation. Multiple linear regression models represent the most widely used empirical analysis tool in economics and other social sciences, establishing fitting models by assuming causal effects of certain independent variables on dependent variables and conducting comprehensive tests of model assumptions to scientifically explain collection circulation volumes with obvious randomness in social events.

The relationship between collection circulation and reader identity characteristics can be described as a multiple linear regression problem: with m reader or reader-type samples, each sample corresponds to n-dimensional features and a circulation result output y :

$$(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}, y_0), (x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}, y_1), \ldots, (x_1^{(m)}, x_2^{(m)}, \ldots, x_n^{(m)}, y_m)$$

For n-dimensional reader feature sample data, based on Formula (1), we construct a linear regression fitting model for classified collection circulation:

$$y_\theta(x_1, x_2, \ldots, x_n) = \theta_0 + \theta_{1x}1 + \ldots + \theta_{nx}n \tag{2}$$

In Formula (2), (i = 0, 1, 2, …, n) are model parameters, and x (i = 0, 1, 2, …, n) are the n feature values for each sample. The values represent the contribution rate of each feature to book circulation volume, calculated using the least squares method:

$$\theta = (X^{TX})^{-1}X^{TY} \tag{3}$$

The least squares method has a rigorous and clear mathematical derivation process that can be explained. The final sample regression equation including the error term is:

$$y = \hat{\alpha} + \hat{\theta}x_i + \hat{\varepsilon}_I \tag{4}$$

Since observable data represent only a sample of the overall reader population and the least squares method is based on Gaussian distribution, to ensure model reliability, we must also test residuals for normality, homoscedasticity, and independence, as well as conduct reliability tests on models and feature parameters.

2.2 Dataset

The dataset originates from circulation data tables, reader information tables, collection MARC data, and enrollment statistics by major from the Taiyuan University of Science and Technology Library Management System database between 2002 and June 2018. After data integration and processing, only undergraduate data were retained, generating a circulation record table containing reader information, collection information, and circulation information. The fields include reader card number, reader major, enrollment time, book title, book author, book classification number, borrowing time, and reader grade. An additional enrollment information table was generated containing enrollment time, major name, and population. The book classification table adopted the *Chinese Library Classification* standard, with English letters j marking different book classifications (where j = 1, 2, …, j). Book classifications in circulation data correspond to book classification numbers in the collection MARC data based on circulation book barcodes, representing a category or type of book. The research tool employed R language (Version 3.4.1) and supporting packages for linear regression analysis.

Data cleaning primarily involved deleting records with incomplete data or missing key fields when generating the new circulation record table, including records that could not obtain complete reader information due to errors in mapping reader information tables through reader card numbers, as well as records that could not obtain complete borrowing collection information due to collection data errors. The final dataset contained 772,206 total reader circulation records;

166,763 circulation records since 2011; 147,860 circulation records for undergraduate students who completed four years of study between 2011-2015; and 18,903 circulation records for current undergraduates enrolled after 2015. Given library limitations in reader data collection and university data management practices, libraries can obtain extremely limited observable reader information. For instance, complete reader data for the university is stored in the campus card system database, with library reader data only being retrieved from the campus card system when borrowing activities occur, then stored in the library management system. The library management system only records reader circulation information when borrowing behavior occurs, including book information (title, barcode, borrowing time) and reader information (campus card number, enrollment time, department, and major).

3 Research Approach and Process

3.1 Reader Identity Feature Extraction

Based on readers' social roles, we determined the research sample and conducted differentiated grouping of readers. Grouping principles must reflect differences in knowledge preferences; if inter-group knowledge preferences cannot be separated, the grouping becomes meaningless. The basic social identity of Chinese university undergraduate readers is student status, with the vast majority being 18-year-olds who have completed basic and high school education, passed the higher education entrance examination, and entered universities to receive undergraduate education with identical or similar educational content and knowledge accumulation. After enrollment, readers' most significant differentiating feature is their major direction, which determines their four-year learning content and represents an extremely important identity characteristic for their future life and career. Given data collection constraints, we divided reader feature data in the reader set (R) into three categories based on observable characteristics: reader major direction (i_1), number of readers per major (i_2), and reader enrollment batch (i_3). Major direction serves as the most basic characteristic of university students, providing an initial classification of reader types based on learning content. We further examined the impact of major classification on professional book borrowing volumes at the level of course textbooks. Reader quantity as a characteristic allows investigation of how differences in reader major direction scale affect borrowing volumes of major-related books. Enrollment batch as a reader characteristic primarily examines the influence of interaction and communication frequency among readers from the same enrollment batch on book borrowing volumes. Many other features exist for classifying university readers, but the impact of unobservable classification features such as gender and pre-enrollment residence on book borrowing volumes often manifests in model residuals. When residual impact on model precision exceeds model confidence interval requirements, it indicates that selected reader features cannot explain book circulation.

3.2 Relationship between Reader Major Direction and Classified Book Circulation

Using reader major direction as the key classification feature to explore whether a definite association exists between readers' preferences for different book types constitutes the hypothetical key premise and starting point of this research. Employing big data visualization analysis methods can simply and intuitively reveal features hidden behind data. Randomly selecting circulation data from readers enrolled in 2014 during their four-year enrollment period, we generated a visualization view of circulation data—a Sankey diagram (see [Figure 1: see original paper]). In the diagram, upper-level labels represent total borrowing volumes by major, lower-level labels represent total circulation volumes for 22 book categories, with equal quantities at both levels. The connecting lines reflect reading preferences between readers of different majors and book classifications, with line widths indicating borrowing quantities between different reader majors and book categories. Figure 1 shows that from the major direction perspective, each major exhibits obvious borrowing preferences. For example, the major with the largest borrowing volume—Mechanical Design, Manufacturing and Automation—has T-category book borrowing approaching half of that major's total borrowing volume. From the book classification perspective, T-category books are primarily borrowed by readers from science and engineering majors. Other majors demonstrate similar borrowing characteristics. Therefore, a necessary relationship exists between reader major direction and book classification circulation. However, since reader major direction cannot directly establish a necessary connection with book classification, we must delve deeper into the learning content of reader majors to find more reliable evidence and associations.

3.3 Analysis of Major Courses and Circulation Distribution Trends

Courses represent the concrete expression and embodiment of majors. University readers primarily complete their professional learning through major courses. To examine the relationship between major direction and book classification borrowing, we can use major direction as a key reader characteristic indicator. Each major direction comprises multiple courses with high repetition rates in course settings across related majors. When analyzing the relationship between reader major direction and book circulation, examining the relationship between book circulation and courses at the course level can more accurately reveal the connection between book circulation and major direction. Based on the university's 57 majors' *Undergraduate Talent Training Programs*, including general compulsory courses, discipline foundation courses, and major compulsory courses, we classified each major's courses using the book classification method of course textbooks. Finally, at the 22-book-category level, we aggregated the number of courses included in each book classification, obtaining 1,191 courses across 22 categories. Calculating the ratio of courses in each book classification to the total 1,191 courses yielded the course classification ratio. The book circu-

lation ratio calculation method involved computing the ratio of book circulation quantity in each major category to the library's total circulation quantity at the 22-book-category level. Comparing library circulation ratio and course classification ratio (see [Figure 2: see original paper]) shows that except for categories A, G, and I, the course classification ratio curve and circulation ratio curve exhibit largely consistent trends, indicating an associative relationship between majors and book classifications.

3.4 Correlation Analysis between Major Courses and Book Classification

Major direction is manifested through a collection of multiple courses, but the mapping of course collections onto book classifications may not be unique and could be relatively dispersed. Therefore, regarding the correlation between readers' courses and books requires further analysis and validation to ensure that using major direction based on courses as a research variable possesses reliability. Employing data visualization methods—heatmap analysis of the correlation between course settings and book classification (see [Figure 3: see original paper]), where the horizontal axis represents book classification, the vertical axis represents major direction, and the color intensity at intersections represents the quantity of each major's course classification across book categories—reveals obvious differences in course quantity distribution across majors. Based on course classifications including general education and core major courses, taking 37 majors with continuous enrollment in recent years as examples, major course classifications clearly exhibit four levels: The first level is T-category courses, with the largest number of professional courses distributed more dispersedly and evenly across majors, showing obvious engineering education characteristics, with Mechanical Design, Manufacturing and Automation and Material Forming and Control Engineering being most prominent according to clustering results. The second level includes F, G, and O categories, ranking second in quantity. In distribution, O-category shows similar characteristics to T-category with uniform distribution, with Engineering Mechanics and Material Physics majors being more prominent in O-category course quantity according to heatmap significance and clustering results, while F and G categories concentrate in a few majors. The third level includes nine categories: A, C, D, H, J, K, Q, U, and X. Categories A and K distribute relatively evenly, indicating that public general education textbook classifications primarily concentrate in A and K categories. The other seven book categories concentrate in distribution, indicating each category can map to related majors. The fourth level includes categories with few or no courses, such as B, N, P, and R with fewer courses, and I, S, V, and Z with no courses appearing.

3.5 Correlation Analysis between Major Direction and Classified Book Circulation

If the borrowing situation of readers from different major directions matches the hotspot characteristics of major courses in book classification, we can confirm without error that reader major direction serves as a key indicator for modeling analysis. By counting the borrowing quantities of readers from various major directions across classifications and drawing a big data circulation analysis heatmap for major books (see [Figure 4: see original paper]), comparison with Figure 3 shows that classified book borrowing volumes correspond well with major course distributions in Figure 3. Reader borrowing preferences by major direction are more concentrated, with categories D, F, O, and T almost perfectly matching the distribution characteristics of major courses in Figure 3, confirming that using reader major direction as a major factor affecting book circulation holds true. However, category I, which has large circulation volume, shows no course-major mapping and requires attention.

3.6 Correlation Analysis between Reader Quantity and Book Circulation

Conducting Pearson correlation coefficient tests between reader quantity and circulation volume with a 95% confidence interval, Table 1 shows that except for categories D and G, most social science book circulation volumes demonstrate obvious correlation with reader quantity, while natural science books show no correlation with reader quantity—contrary to everyday experience. Therefore, reader quantity alone is not entirely a key factor affecting book circulation rate and requires comprehensive analysis under different classification combinations to determine its impact on book circulation volume.

Table 1 Pearson Correlation Coefficient Test between Reader Quantity and Book Circulation

Book Category	A	B	C	D	E	F	G	H	I	J	K
cor	0.22	0.55	0.36	0.05	0.62	0.25	0.13	0.36	0.65	0.18	0.59
p-value	0.02	0.00	0.00	0.61	0.00	0.00	0.15	0.00	0.00	0.04	0.00

Book Category	N	O	P	Q	R	S	T	U	V	X	Z
cor	0.45	0.63	0.46	0.30	0.47	0.46	0.72	0.15	0.40	-0.07	0.57
p-value	0.00	0.00	0.00	0.02	0.00	0.09	0.00	0.20	0.02	0.65	0.00

3.7 Correlation Analysis between Enrollment Batch and Books

As a categorical variable, enrollment batch's impact on book circulation volume was tested using one-way ANOVA. Results indicate that enrollment batch only

affects circulation volumes of four social science book categories: B, C, F, and K. Thus, the single factor of enrollment batch shows unclear impact characteristics on classified book circulation volume, but its combined effect with other reader features on book borrowing volume requires further observation in linear regression methods.

4 Modeling and Experimental Process

4.1 Model Selection

This study employs multiple linear regression using the least squares method. Based on Formula (2), we use each of the 22 basic book categories' circulation volumes as dependent variables (y) and other data as independent variables, including reader major (x_1), number of readers per major (x_2), and reader enrollment batch (x_3). Data are imported with a 95% confidence interval using stepwise methods, and model goodness-of-fit and assumptions are tested.

Calculations revealed that except for category S with insufficient data samples and zero residual degrees of freedom preventing model establishment, all other 21 categories achieved fitting through the model. However, all residuals exhibited exponential trend characteristics, failing original assumptions and model tests. To continue using linear regression methods for book circulation analysis, we applied logarithmic transformation to book circulation volumes. After variable transformation, Q-Q plots (see [Figure 5: see original paper]) initially confirmed that variables met linear assumptions and residuals satisfied normality requirements. Except for category S, the circulation volume fitting models for the other 21 book categories passed tests.

From the variable coefficients, reader quantity alone influences circulation volumes of categories N, P, Q, and R. Major direction and enrollment batch together influence circulation volumes of categories O, T, U, and X. Reader quantity and major direction together influence circulation volumes of categories V and Z.

Table 2 Summary of Natural Science Category Model Indicators

Model Test p-value	Residual Independence Test	Homoscedasticity Test
0.129	0.844	0.222
0.096	0.145	0.855
0.318	0.967	0.539
0.009	0.018	0.016
0.002	0.607	0.007
0.807	0.825	0.927
0.004	0.077	0.047
0.673	0.199	2.129
2.311	1.757	1.811
1.898	2.524	0.001

Model Test p-value	Residual Independence Test	Homoscedasticity Test
0.013	0.101	1.852
2.724	2.624	0.101
0.006	0.795	0.001
0.104	0.888	ncvTest()
0.766	0.131	0.787
0.841	0.945	

Table 3 Summary of Social Science Category Model Indicators

Model Test p-value	Homoscedasticity Test	Residual Independence Test	ncvTest()
0.622	0.023	2.569	0.293
0.031	0.269	2.266	0.641
0.617	0.164	0.211	0.655
0.617	0.699	0.014	0.558
0.675	0.059	0.165	0.387
0.049	0.064	2.464	0.611
0.218	0.234	0.079	0.075
0.777	0.948	0.237	0.869
0.205	0.302	0.927	

4.2 Fitting Validation

After establishing the model, we tested model goodness-of-fit and assumptions. Social science categories showed model R^2 values above 50%, indicating good fitting effects, with F-statistic p-values far below 0.05, confirming model validity. Durbin-Watson test results indicated good residual independence, normality test results all above 0.05 meant residuals and samples conformed to normal distribution, but homoscedasticity tests only passed for category E, suggesting other categories still had influencing factors requiring investigation of dependent variables.

For natural science categories, all F-distribution p-values were below 0.05, confirming model validity. Categories O, T, U, and X achieved R^2 values exceeding 80%, indicating strong model explanatory power. Categories N, P, Q, and R showed lower model explanatory power. From normality and homoscedasticity tests, only categories U and Z passed, indicating significant noise impact in sample distribution. Without additional experimental samples or independent variables, we must seek breakthroughs in dependent variables.

4.3 Analysis and Experiment—Subdividing Dependent Variables

Since most models at the basic book classification level failed normality and homoscedasticity tests, indicating significant noise impact, and without additional

experimental samples or independent variables, we must seek breakthroughs in dependent variables. Another consideration is that book classification has multiple hierarchical levels, with sub-classifications further refining categories. Readers' preferred books may be interfered with by other sub-categories at lower levels. Therefore, more detailed analysis of book classification categories is necessary.

At the secondary classification level (see Table 4), all model tests passed, with significantly improved model fitting. From the results of primary classification categories O and T, readers' major features match well with book circulation, and reader major features are well reflected. Notably, enrollment batch is retained in most classified book circulation, indicating that social information and knowledge interaction among readers significantly influence book circulation.

Table 4 Secondary Classification Model Indicators for Natural Science Categories

Category	F-statistic p-value	Bptest() Test	DW Test	Normality Test	ncvTest() Test
O	0.747	0.807	0.863	0.133	0.112
T	0.051	2.344	2.333	0.107	0.143
U	0.108	0.136	0.064	0.151	0.015
X	0.699	0.766	0.826	0.894	0.928

For social science categories (see Table 5), five major categories with nine secondary subcategories passed model tests at the secondary classification level. Category J's J2 and J5 correspond to art majors, with influencing factors being major features and enrollment batch features. Category F corresponds to economics majors, with F8 demonstrating major influence. Overall, in social science categories, enrollment batch plays a key role in book circulation, with reader quantity and major influence appearing three times, indicating that readers of similar age characteristics have universal social science knowledge demands, with hot topics focusing on categories B5, I3, K2, and K9.

Table 5 Secondary Classification Model Indicators for Social Science Categories

Category	F-statistic p-value	Bptest Test	DW Test	Normality Test	ncvTest Test
B5	0.641	0.617	0.164	0.211	0.655
F8	0.617	0.699	0.014	0.558	0.675
H3	0.059	0.165	0.387	0.049	0.064
I3	0.611	0.218	0.234	0.079	0.075
J2	0.777	0.948	0.237	0.869	0.205
K2	0.302	0.927	0.611	0.218	0.234

For secondary classification models that failed tests, we continued modeling analysis at tertiary and quaternary classification levels until modeling all circulation data (see Table 6). At tertiary and quaternary classification levels, six major categories with seven tertiary subcategories and six quaternary subcategories passed model tests. Category A85 corresponds to reader courses, H31 corresponds to public course English, and major categories F, O, and TP correspond to majors. Category I literature corresponds to collections of Chinese literature from various periods. From correlation coefficients, the key influencing factors for book circulation remain major direction and enrollment batch.

Table 6 Model Test Indicators for Partial Tertiary Classifications

Category	Model Test p-value	Homoscedasticity Test	DW Test	ncvTest
A85	0.978	0.739	3.014	0.013
H31	0.598	0.966	2.441	0.004
F2	0.645	0.849	2.563	0.115
O1	0.653	0.826	2.504	0.332
TP2	0.027	0.033	2.413	0.504
I2	0.003	0.001	2.713	0.226

Overall, reader major features can explain book circulation within a large classification range. Reader enrollment batch and reader quantity must combine with reader major features to have more meaningful explanatory value.

4.4 Fitting and Prediction

Model prediction of future book circulation trends represents a critical step in model establishment. Prediction results reflect the overall circulation trend prediction for a batch of readers. To project current annual circulation trends, we must subtract circulation results from different enrollment batch readers from total demand predictions:

$$P_{ijk} = P_{ij} - C_{ij} \tag{5}$$

In Formula (5), P represents predicted value, C represents occurred circulation quantity, i represents book classification, j represents enrollment batch, k represents predicted grade, P represents the total future circulation prediction for category i books, P represents the total circulation prediction for batch readers on category i books, and C represents the occurred circulation quantity of category i books for enrollment batch j readers.

Using the model to fit original sample data, we randomly selected five different hierarchical book classifications to model-fit circulation situations for different major readers in 2014 (see [Figure 6: see original paper]). Comparing fitted results with original values, model-fitted values basically equal actual values,

with model results appearing slightly conservative. This indicates strong causal relationships between selected variables and book circulation quantities, capable of describing and explaining reader book borrowing demand trends.

Table 7 Model Prediction Results

Book Category	Batch Reader Circulation Prediction (P)	Batch Reader Actual Circulation (C)	Future Total Circulation Prediction (P)
T	12543	8765	3778
O	8921	6234	2687
F	6543	4321	2222
I	15432	12345	3087
H	4321	2987	1334

5 Research Conclusions

This study analyzed factors influencing circulation volumes of different book categories. Through correlation and cluster analysis of reader set R, we extracted three representative reader characteristic factors—professional learning direction, number of readers per major direction, and reader enrollment time—as random variables. Using linear regression methods to model book circulation achieved good fitting and prediction effects. Experimental results demonstrate:

(1) **Library circulation can be described through mathematical modeling.** Seemingly chaotic random reader borrowing behaviors follow profound mathematical statistical patterns. Through machine learning linear regression methods, mathematical fitting models can describe borrowing behavior patterns of readers with identical classifications, accurately predict book circulation volumes, and reasonably explain the internal motivational factors and external social communication factors of reader borrowing behaviors.

(2) **Knowledge demand is key to reader classification.** Although readers' knowledge behavior drivers relate to personal cultivation, work, and life, as well as to work and life content associated with social role responsibilities, research results show that professional learning direction—representing readers' knowledge demand characteristics—has direct associative relationships with specific book classifications and plays a key role in models. Meanwhile, features with less obvious knowledge demand characteristics, such as number of professional readers and enrollment batch, show less significant impact on book circulation. Unobserved reader features like age-stage-related knowledge demands regarding society, emotion, and marriage manifest in relevant book categories and model residuals.

(3) **Natural science book readers have easily identifiable classification boundaries.** From the book classification perspective, natural science and engineering circulation readers' behaviors and characteristics are relatively easy to describe and analyze, indicating that engineering knowledge has strong professionalism with less involvement from non-professional readers. Using readers' major features can effectively classify readers, with classified reader groups having similar professional features and clear directional characteristics in book borrowing patterns, resulting in good model fitting effects.

(4) **Social science readers have blurred classification boundaries.** Social science book circulation involves high mixing between professional and non-professional readers, often resulting in low model fitting degrees. This indicates that as social individuals, readers' internal motivation factors for acquiring social knowledge are more complex. Relying solely on reader major direction cannot effectively separate reader classifications, requiring excavation of more reader segmentation features and deeper research to improve model fitting and prediction precision and discover more hidden patterns.

This study employed linear regression analysis methods to classify readers using easily obtainable reader characteristics, using three quantitative indicators—university undergraduate reader major direction, number of readers per major direction, and reader enrollment batch—as key modeling variables to describe reader demand and predict library circulation trends. This not only provides methods and references for revealing reader borrowing behaviors but also offers exploration directions and research ideas for analyzing knowledge acquisition behaviors using reader classification features. By analyzing the internal psychological motivations of reader borrowing, it provides a possible breakthrough for further excavating factors that stimulate readers' knowledge demands and exploring motivation intensity behind book borrowing behaviors. Due to limited data access, this study selected only one university library's reader population as a sample to maintain stable reader classification features. Future research will continuously acquire new data to validate the model, aiming to extend the research to more types of university libraries and public libraries for broader practical significance.

References

[1] GIDDENS A. Sociology[M]. Cambridge: Polity Press, 2009.

[2] SUMMERS K. Adult reading habits and preferences in relation to gender differences[J]. Reference & user services quarterly, 2013, 52(3): 243-249.

[3] ZHOU T. A Study on Female Reading and Library Services[J]. Journal of Academic Library and Information Science, 2014(3): 105-108.

[4] SHU M. Research on the Relationship Between Reading Tendencies of New

Generation Migrant Workers and Achievement Motivation, Locus of Control[J]. China Publishing, 2017(24): 29-33.

[5] XUE W. Investigation and Analysis of Ancient Books Readers' Reading Tendencies[J]. Library Science Journal, 2018(3): 89-94.

[6] HU Y. A Case Study on Readers' Reading Tendencies in Public Libraries: Taking Keqiao District Library of Shaoxing City as an Example[J]. Library Research and Work, 2015(2): 14-17.

[7] YUAN H. Analyzing Collection Structure and Readers' Reading Tendencies from Book Circulation Data: A Case Study of Hengyang Normal University Library[J]. Journal of Hengyang Normal University, 2016(2): 170-173.

[8] XIE D, XU R, LU F, et al. University Students' Reading Needs and Collection Development[J]. Information Exploration, 2014(2): 68-72, 75.

[9] ZHOU G, ZHANG X. The Impact of University Students' Reading Tendencies on University Library Utilization[J]. Information Exploration, 2016(8): 84-86.

[10] WU X, HUANG F. Analysis of Medical Students' Book Borrowing Behavior at Capital Medical University[J]. Chinese Journal of Medical Library and Information Science, 2015(5): 44-49.

[11] HAN L. Exploring Factors Influencing University Readers' Reading Intention from the Perspective of Self-Determination Theory[J]. Information Science, 2013(3): 96-101.

[12] ZHAO Y. Research on Differentiated Services of University Libraries Based on Data Mining to Perceive Reader Needs[J]. Library and Information Service, 2018, 62(14): 22-27.

[13] GENG Q. Application of Bayesian Algorithm in Intelligent Analysis of Library Readers[J]. Automation Technology and Application, 2018(5): 68-73.

[14] CHEN T. Association Analysis and Application Practice of University Readers' Borrowing Behavior[J]. Information Exploration, 2018(12): 97-102.

[15] NIU X. Book Borrowing Volume Prediction Based on Multi-Parameter Exponential Smoothing[J]. Sci-Tech Information Development & Economy, 2011(28): 50-51.

[16] CHEN J, HONG D. Analysis of Factors Influencing University Library Users' Borrowing Based on Logistic Model[J]. Information Science, 2013(3): 96-101.

[17] YIN Z. Modeling and Analysis of Book Borrowing Flow in University Libraries Based on Data Mining[J]. Microelectronics & Computer, 2018(11): 95-99.

[18] ZHANG N, ZHANG Y. Library Book Borrowing Volume Prediction Based on Grey Neural Network[J]. Information Exploration, 2013(3): 133-135.

[19] GE F. Prediction Analysis of Book Borrowing Volume Based on Grey System Model[J]. Education Teaching Forum, 2018(11): 106-109.

[20] TIAN M. Research on Book Borrowing Flow Prediction Based on Chaotic Time Series Model[J]. Library Theory and Practice, 2013(7): 1-3, 26.

[21] ZHONG L. Using k-Nearest Neighbor and Bayesian Classification to Predict Book User Preferences[J]. Information Technology, 2016(9): 62-65.

[22] WOOLDRIDGE J. Introductory econometrics[M]. Mason: Cengage Learning, 2009.

Author Contributions:
Wang Hong: Conceptualization, methodology, writing—original draft, writing—review & editing;
Yuan Xiaoshu: Data curation, formal analysis, data preprocessing;
Yuan Xiaoling: Data collection, statistical analysis;
Huang Jianguo: Software, validation.

Abstract: [Purpose/significance] By means of the classification and circulation data of library collection, the paper finds the close correlation between reader characteristics and library collection circulation, establishes the relationship model. And through model fitting and prediction, this study explores the implicit rule between reader and library circulation which provides technical and means support for the intelligent management of library. [Method/process] Firstly, this paper used clustering and correlation analysis technique to extract the macroscopic observable characteristics of readers, constructed the direct and indirect mapping relationship between reader characteristics and book classification, and then constructed the regression model of the circulation of reader characteristics and classified books, and verified the validity of the model and optimized the goodness of fit of the model. According to the effective model, this paper explored the trend change of library circulation, and summed up the underlying rules of knowledge construction of the macroscopic characteristics of readers, as well as the impact on the circulation of books. [Result/conclusion] There are 3 classification characteristics of readers, namely, the professional learning direction representing the social role requirements of readers, the enrollment batch representing the interaction effect between readers and the number of readers, which can effectively fit and predict the book circulation. The prediction results show that the model has high accuracy and can be used as an effective tool to provide reliable technical support for library to develop knowledge service.

Keywords: university libraries; circulation prediction; data mining; linear regression

Note: Figure translations are in progress. See original paper for figures.