

REVIEW

# Analysis of the influence factors of global film box office based on a log-linear model

**Xinyu Xie\***

School of Management, Tianjin University of Technology, Tianjin, China

**\*Correspondence:**Xinyu Xie,  
xiexinyu666@foxmail.com**Received:** 24 February 2022; **Accepted:** 18 March 2022; **Published:** 01 April 2022

In this paper, python is used to obtain the data of the top 100 global movie ticket rooms, and through data visualization, it is concluded that the global movie box office will peak in the next 10 years. Science fiction action movies are more popular. r-language was used to construct log-linear models, respectively. The results of the random unit group design model showed that science fiction action movies were more favorable to the box office. Through k-means clustering analysis, it was concluded that the guidance of well-known directors had more positive effects on the box office. According to the above conclusions, the major cinemas and film and television companies make relevant suggestions.

**Keywords:** r-language, log-linear model, random unit group design model, k-means clustering

## 1. Introduction

As global economy continues to grow, the potential value of the film market continues to increase and the film industry is growing rapidly. However, this has given rise to many problems in the film industry. More and more scholars have begun to pay attention to research on the film industry, and more and more quantitative studies have been conducted on the factors influencing the movie box office. Focusing on the domestic market, China's box office has only emerged as a global player since 2002 when the movie theater system was reformed and the movie market was activated, but it is still at an extreme disadvantage. Since 2002, Chinese movies have embarked on an industrialized path of development, with the total annual box office volume growing from around 1 billion RMB to 63 billion RMB in 2019. However, from 1982 to 2016, China's box office remained at the bottom of the global ranking. It was not until 2017 that the domestic movie "War Wolf 2" took a place in the global movie box office, which can be far behind European and American movies. This paper examines the data of the top 100 global movie box office from 1982 to 2022. Through quantitative analysis of different kinds of movies, we can determine which kinds of movies are more

popular with the public. In turn, relevant conclusions are drawn to encourage the domestic film industry to produce films of that type. In addition, it provides a reference for the types of movies released by domestic cinemas and provides a basis for arranging the frequency of movies. It visualizes the time dimension of movie box office data, analyzes the trend of movie box office, and predicts future changes. It provides a basis for decision-making in the movie industry.

## 2. Literature review

Since the 1960s, there has been a growing body of research on the factors influencing movie box office.

Liao Lin et al. studied the influence of social media marketing channels and events on movie box office through the likelihood model and concluded that social media can improve customers' purchase intention through relevant routes (1). Lee Young-Jin et al. used the two-stage instrumental variable method with fixed effect to explain the heterogeneity of films and the simultaneous relationship between user reviews, advertising expenditure, and sales through the box office data of American films and evaluated the impact of advertising expenditure on

sales after the release of films (2). Wei lu determined the main index system affecting film box office based on China's film box office, combined with a questionnaire survey and expert interview, and provided a valuable reference for future risk control and film investment decision through a neural network prediction model (3). Ya-Han Hu et al. used emotional tools to quantify film reviews. They combined the quantified data with basic film information and external environmental factors and predicted the performance of film box office through the model tree (M5P) model, linear regression model, and support vector regression model (4). Oh Chong et al. used the ordinary least squares (OLS) regression model and found that consumer participation behavior on Facebook and YouTube was positively correlated with the total box office revenue. However, the same effect was not observed on Twitter. The results show the importance of investing in social media communication in multiple channels (5). Yong Lou et al. studied the dynamic patterns of word of mouth and how it helps explain box office receipts by using actual word-of-mouth (word of mouth) information. The results show that word-of-mouth information has an important explanatory power for both gross and weekly box office receipts, especially in the first few weeks after a film's release. However, this explanatory power comes mainly from the amount of word of mouth, not its value, which is measured by the percentage of positive and negative information (6). Tingting Song et al. built a prediction model of movie box office revenue by studying the complex relationship between movie box office revenue and user-generated content (UGC) on the Weibo platform, marketer-generated content (MGC), and UGC on third-party platforms. It is found through research that the volume of corporate microblogging (MGC) can not only directly predict the box office revenue but also indirectly predict the box office revenue through MUGC. Therefore, MUGC plays a partial intermediary role in predicting the relationship between the volume of corporate microblogging and the box office revenue (7).

Elliot Mbunge et al. applied the PRISMA model to review published papers from 2010 to 2019 extracted from Google Scholar, Science Direct, IEEE Xplore Digital Library, ACM Digital Library, and Springer Link. The study shows that support vector machine has the highest frequency of predicting box office success, accounting for 21.74%, followed by linear regression, accounting for 17.39% of the total frequency contribution. This study also provides some valuable references for this paper (8). U. Ahmed et al. predicted the box office before film production based on actors' experience, journalists' comments, media reports, user ratings, and income generated by associated films and other information (9). L. Kang et al. examined the role of film quality signals (such as star power, Internet media reviews, and industry recognition) in an empirical analysis of the relationship between family friendly content and box office receipts in films by analyzing Chinese film data

from 2009 to 2018. The results show that explicit sex and profanity in films have a negative and statistically significant effect on box office takings, confirming the role of cultural values in the economic success of Chinese films. However, in big-budget films involving superstars, the violent and bloody (graphic violence) content attracts the audience, so the box office revenue increases (10). Based on the Chinese film market, this paper considers the influencing factors of film box office from multiple dimensions, uses the joint analysis method of questionnaire survey and expert interview to determine the main index system affecting MBO, and then establishes the MBO prediction model through the neural network BRP method to predict the electronic music box office (11). B. S. Wibowo et al. studied the key factors of the box office failure and success of Indonesian local films, and the results showed that the success of local films was mainly driven by the popularity of actors and the existence of foreign films at the box office (12). L. A. B. S. Ghazal et al. studied the data of 361 movies, and the results showed that the two factors affecting box office success were the number of days the movie was released and the number of theaters. In addition, the proximity of film release dates to seasonal holidays in Malaysia is quite consistent in producing a positive correlation (3). L. I. Pei-zhi et al. established a hybrid prediction model based on web search data. First, the optimal training set (OTS) was constructed by matching the training data most similar to the test set. Second, the Empire competition algorithm (ICA) is used to select the best parameter combination of least square support vector machine (LSSVM). Finally, the optimization model is used for prediction (13).

Although many domestic and foreign scholars have studied the factors influencing movie box office, few experts and scholars have conducted multivariate statistical analysis studies using the top 100 global movie box office data from 1982 to 2021. The act of using movie categories as influencing factors to study movie box office is even rarer. In addition, this paper will use cluster analysis methods to classify different box office different categories of box office and then draw correlation conclusions.

### 3. Data description and processing

#### 3.1. Data mining

Using Python's requests function to crawl the <http://www.piaofang.biz/> website data, we obtained the top 100 global movie box office data from 1982 to 2022, which contains information such as "movie name," "release date," "movie type," "movie box office," "director name," and so on. The data include "movie name," "release date," "movie type," "movie box office," "director name," and so on.

### 3.2. Digital description of the data

The resulting data were digitally described, and the digitized description table is shown in [Table 1](#).

### 3.3. Data cleaning and missing value processing

Using Python to clean and process the noisy data, duplicate data, and so on after screening, we found that the data of the movie “War Wolf 2” category were missing in these data, so we used the random forest model to fill in the missing values in [Table 2](#).

#### 3.3.1. Random forest model

The random forest algorithm is used to build a decision tree with the movie category 2 as the label, the release time, category 1, and the movie director’s name as the feature

**TABLE 1** | Table describing the digitization of variables.

Category name	Quantifying the numbers
Science fiction	1
Disaster	2
Action	3
Love	4
Adventure	5
Animation	6
Fantasy	7
Comedy	8
Gun battle	9
Thriller	10
Crime	11
War	12
Music	13
Biography	14
Military	15
Intimacy	16
Family	17

**TABLE 2** | Table of variables of the random forest model.

Model variables	Variable meaning
P	Probability of each species
H	Current leaf node entropy value
ID	GINI factor
n	Total data
C	Rate the value
V	Current node data volume

values, and the information entropy and information gain rate of each node are calculated to determine the root node and child nodes of the decision tree. Multiple decision trees are built to form a random forest, and each decision tree randomly selects 60% of the data in Form 1 for training and finally performs the prediction of missing values of category 2. The effect of the prediction is evaluated by calculating the evaluation function for each decision tree, and the final filled kind 2 missing value is 15 (i.e., military kinds).

The random forest computational model is as follows.

$$H(x) = - \sum_{i=1}^n P_i \ln(P_i)$$

$$ID = \frac{H(x)}{\sum_{i=1}^n -\frac{1}{n} \lg n}$$

Evaluation function:

$$C(T) = \sum_{t \in leaf} V_t H(t)$$

### 3.4. Preliminary analysis of data visualization

A time series plot of these data with the time variable as the horizontal axis and the box office as the vertical axis is shown in [Figure 1](#).

As shown in the figure from 1982 to around 1998, the movie box office showed an increasing trend, and in 2009, the movie box office reached its peak, and thereafter until 2022, the highest value of the movie box office still did not exceed 2009. Through further research, it can be seen that although the box office in 2009 was high for the movie “Avatar” whose movie special effects were perfect, the plot was appealing, and the director was the experienced James Cameron, so not only that but many of James Cameron’s films are listed in the top 100 films at the global box office. In addition, the sequence shows an up-and-down trend, with peaks occurring every 10 years or so. The graph inferred that there is a high probability that the peak will still occur in the next few years, but whether the box office will surpass the movie Avatar still needs further research and analysis.

Each genre is disaggregated and combined vertically, and the box office of each genre is averaged. A pie chart of the average box office of different genres is drawn. This is shown in [Figure 2](#).

As shown in [Figure 2](#), the average box office of romance movies is higher, but further research shows that there are fewer movies in this category, so the average is higher, which does not mean that the box office of all these movies is higher, followed by the second highest percentage of the average

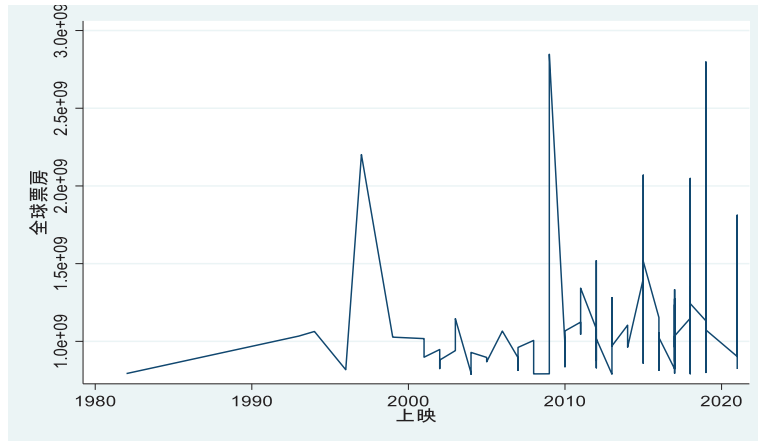


FIGURE 1 | Time series chart.

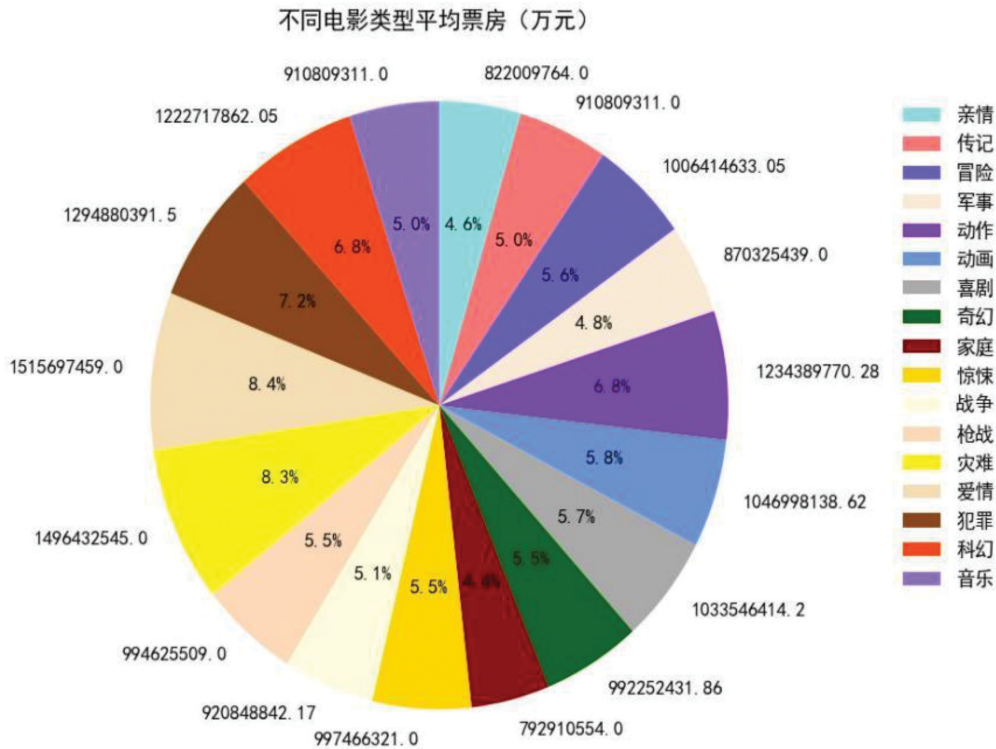


FIGURE 2 | Average box office of different movie types.

box office of disaster movies. The average box office share of family movies and military movies is lower.

Therefore, film companies should carefully consider shooting romance movies, although the average box office of romance movies is high according to Figure 2. The average box office of such movies is high due to the small number of data, which is not necessarily suitable for the company's profitability. In addition, from the perspective of profitability, they should shoot less number of military, family, and affection movies because the box office revenue of such movies is generally not high. They should shoot more action, science fiction, comedy, and disaster movies. The reason for

this is that these films account for the majority of the data, and the average box office is still in the middle to upper range.

## 4. Model construction

### 4.1. Log-linear model

The mathematical model was constructed with box office as the dependent variable and factors such as category 1, category 2, time, and director as the independent variables. The independent variables were analyzed as categorical

variables, while the dependent variables could be considered categorical variables or continuous variables, so different mathematical models were used to fit the obtained data and compare which model was more suitable.

First, the dependent variable was considered a multicategorical variable, so a log-linear model was used to fit the data. Type 1 and type 2 are two qualitative variables, and the director can be considered a continuous variable. Therefore, there is the formula:

$$\ln(\lambda) = \mu + \alpha_i + \beta_j + \gamma x + \varepsilon_{ij}$$

where  $\mu$  is the constant term,  $\alpha_i$  and  $\beta_j$  are the main effects of the two qualitative variables type 1 and type 2,  $x$  is the director continuous variable, while  $\gamma$  is its coefficient and  $\varepsilon_{ij}$  is the residual term. The positive parameter  $\lambda$  for the Poisson distribution is taken logarithmically in order to be the left side of the model that takes the entire range of values of the real axis.

## 4.2. Stochastic unit group design modeling

Considering the dependent variable as a continuous variable and the kind1, director as a qualitative factor, a randomized unit group design model can be used to fit the data to the model. After treating factor  $x_2$  with Glevels and unit group  $x_4$  with  $n$ , which can be viewed as  $n$  levels, and generating  $x_2$  dummy variables for  $G$  and  $x_4$  dummy variables for unit group  $n$ , respectively, the box office results  $y_{ij}$  are expressed as follows:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} (i = 1, 2, \dots, G; j = 1, 2, \dots, n)$$

After further processing, it is obtained that:

$$Y = X\beta + e$$

## 4.3. k-means clustering

The k-means clustering of movie title numbers was performed by using factors such as category 1, category 2, director, and worldwide box office as indicators. The 100 objects are divided into three classes, and the distance between cluster and cluster clustering centers is calculated continuously and iteratively until the criterion function converges. The squared error criterion is usually used:

$$E = \sum_{i=1}^k \sum_{p=c_i} (p - m_i)^2$$

where  $E$  is the sum of the mean squared differences of all objects in the data with the corresponding cluster centers,  $p$  represents a point in the object space, and  $m_i$  is the mean value of class  $c_i$ .

The k-means algorithm for clustering based on the mean value in the class is as follows.

1. Collect the sample set  $\{x_1, x_2, x_3, \dots, x_n\}$ , where  $n$  is the total number of samples is 100, and each sample vector is  $x_j = \{x_{j1}, x_{j2}, x_{j3}, \dots, x_{jm}\}$ ,  $x_{jt}$  is the  $j$ th sample  $t$ th attribute, total  $m$ -dimensional attributes.
2. Loop through 3 to 4 below until each cluster no longer changes.
3. Calculate the distance of each object from these central objects based on the mean value of the objects in each cluster (central objects) and re-divide the corresponding objects according to the minimum distance.
4. Recalculate the mean value of each (changed) cluster.

## 5. Model solving and result analysis

### 5.1. Log-linear model solving and analysis

The processed experimental data are imported into the r-language software, and then the log-linear model is constructed and solved for the data. The solution results are shown in [Figure 3](#).

Solving the results shows that the regression model can be expressed as follows:

$$y = -1.539 \times 10^{-2} x_2 - 6.350 \times 10^{-3} x_3 - 1.144 \times 10^{-2} x_4 + 21.21$$

As shown in the model solution results,  $p_1 < 0.01$  indicates that  $x_2$  (type 1) has a significant impact on box office, and the solved coefficients show that genre 1 has a significant negative correlation with box office. It further shows that science fiction, action, and adventure movies are more popular and have higher box office. Therefore, major cinemas can consider further increasing the number of science fiction and action movies introduced. Major film companies can also produce more high-quality science fiction and action movies to increase box office revenue and gain more profits.

$p_2 < 0.01$ , indicating that  $x_3$  (type 2) also has a significant impact on box office, and there is also a significant negative correlation between the two. This indicates that family films are not very popular among the general public, and further analysis shows that not many family films are listed in the top 100 global box office. It also shows that family movies have a low probability of making theaters more profitable. In addition, such films do not provide a good guarantee for the profitability of film companies. Therefore, major cinemas can selectively introduce family movies. Major film companies can also appropriately reduce the output of family movies to avoid the wastage of funds.

$p_3 < 0.01$  also shows that  $x_4$  (director) also has a significant effect on the box office, and the two become a significant



```

Call:
glm(formula = y ~ x2 + x3 + x4, family = poisson(link = log),
     data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-17389.7  -3942.1   439.2   2698.9  29357.2

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.121e+01  7.165e-06 2960051  <2e-16 ***
x2          -1.539e-02  1.083e-06 -14216   <2e-16 ***
x3          -6.350e-03  1.061e-06  -5986   <2e-16 ***
x4          -1.144e-02  1.838e-07 -62263   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1.0415e+10  on 98  degrees of freedom
Residual deviance: 4.9594e+09  on 95  degrees of freedom
AIC: 4959381641

Number of Fisher Scoring iterations: 4

```

FIGURE 3 | Results of the log-linear model solution.

negative correlation. Through the data, it can be illustrated that if the director James Cameron directed the movie, box office can be guaranteed. Therefore, major cinemas can further increase the number of films directed by James Cameron when introducing new films. In addition, major film companies can also invite the director to be a director or ask him to be a director.

## 5.2. Stochastic unit group design model solving and analysis

The processed data are imported into the r-language software, and the model is constructed and solved for these data. The solution is shown in Figure 4.

As shown in the figure,  $P_{x_2} < 0.05$ , indicating that type 1 has a significant effect on box office, and  $P_{x_4} < 0.05$ , indicating that the director factor also has a significant effect on box office. This is consistent with the conclusion obtained from the linear logit model. However, its  $P$  results are not as significant as the log-linear model. Therefore, it is more appropriate to consider the dependent variable box office as a categorical variable.

## 5.3. K-means clustering results and analysis

The required clustering data are brought into the r-language program, and category 1, category 2, director, and worldwide box office are used as clustering indicators for k-means

clustering, and the data are first normalized before clustering.

$$x_0 = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

where  $x_0$  denotes the resulting new data,  $x$  denotes the original data,  $x_{\min}$  denotes the minimum data value, and  $x_{\max}$  denotes the maximum data value. The data were clustered after the standardization process, and the clustering results were as follows.

The top movies in the clustering results are in the same category, and the analysis shows that most of the top movies in the box office are science fiction action movies, and the directors are internationally famous directors, such as James Cameron. Therefore, when introducing movies, cinemas should increase the number of movies introduced for science fiction action movies and movies directed by famous directors, and cooperate with film and television companies to earn higher profits. Film and TV companies can also invest more money in such movies to improve the quality of the films to tap the consumer surplus to earn high money. The films ranked 6–21 have a high proportion of science fiction and adventure films, so major cinemas can invest appropriately in science fiction and adventure films to earn more profits. Love, family, and military films are relatively low in the ranking, and the number of films of various categories is low. The high proportion is still science fiction and action films, so it is difficult for these films to enter the top 100 list at the global box office and difficult to hit the top 20 of box office. Therefore, major cinemas should consider carefully when introducing such films, and film companies should make comprehensive

```

Analysis of Variance Table

Response: y
      Df    Sum Sq   Mean Sq F value    Pr(>F)
factor(x2)  8 1.6527e+18 2.0658e+17  2.3035 0.04481 *
factor(x4) 58 9.3962e+18 1.6200e+17  1.8064 0.03635 *
Residuals 32 2.8698e+18 8.9681e+16
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

FIGURE 4 | Results of solving the random unit group design model.

```

1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
2  2  2  2  2  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  3  3  3  3  3  3  3  3  3
32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62
3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3
63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93
3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3  3
94 95 96 97 98 99
3  3  3  3  3  3

```

FIGURE 5 | K-means clustering results.

consideration for the investment and production of the film in shooting or investment.

## 6. Summary and prospect

This study of the top 100 global movie box office data visualizes the data and shows that the box office peaks every 10 years and is expected to peak again in the future through a time series chart. By fitting the data through a log-linear model and randomized unit group design model, it can be seen that science fiction and action movies are mostly high-grossing movies. In addition, family and biography movies occupy a smaller proportion in the top 100 box office list. Therefore, it is suggested that movie companies should invest more money in science fiction and action movies and less money in family movies. Through the cluster analysis, we can see that the directors of high-grossing movies are all internationally renowned directors, and the top-ranking movies at the box office are mostly science fiction action movies. Therefore, the Chinese film and television industry can consider producing high-quality science fiction action movies to impact the box office.

This study still has many shortcomings. The data can be considered to choose a larger amount of data in different dimensions for analysis. In addition, neural networks and other optimization intelligence algorithms can be introduced into it to obtain more profound conclusions. The time column of the data can also be fully utilized to build ar, ma, and other models through time series analysis to predict future data and obtain more accurate conclusions.

## References

- Liao L, Huang T. The effect of different social media marketing channels and events on movie box office: an elaboration likelihood model perspective. *Inf Manag.* (2021) 58:103481.
- Lee Y, Keeling K, Urbaczewski A. The economic value of online userreviews with ad spending on movie box-office sales. *Inf Syst Front.* (2019) 21:829–44.
- Lu W, Xing R. Research on movie box office prediction model with conjoint analysis. *Int J Inf Syst Supp Chain Manag.* (2019) 12:72–84.
- Hu Y, Shiau W, Shih S, Chen C. Considering online consumerreviews to predict movie box-office performance between the years 2009 and 2014 in the US. *Electron Lib.* (2018) 36:1010–26.
- Oh C, Roumani Y, Nwankpa J, Hue H. Beyond likes and tweets: Consumer engagement behavior and movie box office in social media. *Inf Manag.* (2017) 54:25–37.
- Zhou C, Leng M, Liu Z, Cui X, Yu J. The impact of recommender systems and pricing strategies on brand competition and consumer search. *Electron Commer Res Appl.* (2022) 53:1–15.
- Liu Y. Word of mouth for movies: its dynamics and impact on box office revenue. *J Market.* (2006) 70:74–89.
- Song T, Huang J, Tan T, Yu Y. Using user-and marketer-generated content for box office revenue prediction: Differences between microblogging and third-party platforms. *Inf Syst Res.* (2019) 30:191–203.
- Mbunge E, Fashoto S, Bimha H. Prediction of box-office success: a review of trends and machine learning computational models. *Int J Bus Intell Data Min.* (2022) 20:192–207.
- Zhou C, Ma N, Cui X, Liu Z. The impact of online referral on brand market strategies with consumer search and spillover effect. *Soft Comput.* (2020) 24:2551–65.
- Ahmed U, Waqas H, Afzal MT. Pre-production box-office success quotient forecasting. *Soft Comput.* (2020) 24:6635–53.
- Kang L, Peng F, Anwar S. All that glitters is not gold: do movie quality and contents influence box-office revenues in China? *J Policy Model.* (2022) 44:492–510.
- Zhou C, Tang W, Zhao R. Optimal consumption with reference-dependent preferences in on-the-job search and savings. *J Ind Manag Opt.* (2017) 13(1):503–27.
- Wibowo BS, Rubiana F, Hartono B. A data-driven investigation of successful local film profiles in the Indonesian box office. *J Manajemen Indones.* (2022) 22:333–44.
- Ghazali L, Islam R. Critical determinants of box office success for the Malaysian film industry. *Int J Bus Syst Res.* (2021) 15:491–509.
- Liu Z. Impact of cost uncertainty on supply chain competition under different confidence levels. *Int Trans Operat Res.* (2021) 28:1465–504.

17. Chen P, Zhao R, Yan Y, Zhou C. Promoting end-of-season product through online channel in an uncertain market. *Eur J Operat Res.* (2021) 295:935–48.
18. Li P, Dong Q. Box office prediction model based on web search data and machine learning. *Operat Res Manag Sci.* (2021) 30:168.
19. Oomiya N, Nakamura Y. Proposal of an estimate of box office revenues using a movie scripts—case of romance movies in Japan. *J Jap Soc Fuzzy Theor Intell Inf.* (2020) 32:935–43.
20. Boisvert S. Les Hommes du box-office québécois: la construction sérielle du genre dans les sequels Nitro Rush et Les 3 p'tits cochons 2. *Nouvelles Études Francophones.* (2020) 35:101–16.
21. Barnett VL. Super Fly (1972), Coffy (1973) and The Mack (1973): under- and over-estimating blaxploitation box office. *Hist J Film Radio Telev.* (2020) 40:373–88.
22. Koo HY, Lee HJ, Lee G. Influence of movie success factors including holdbacks in box office and VOD Market. *Korean Manag Sci Rev.* (2021) 38:47–61.
23. Chu M. The impact of online referral services on cooperation modes between brander and platform\*. *J Ind Manag Optim.* (2022). doi: 10.3934/jimo.2022174
24. Yu J, Zhao J, Zhou C, Ren Y. Strategic business mode choices for e-commerce platforms under brand competition. *J Theor Appl Electron Comm Res.* (2022) 17:1769–90.
25. Yu J, Song Z. Self-supporting or third-party? The optimal delivery strategy selection decision for e-tailers under competition\*. *Kybernetes.* (2022). doi: 10.1108/K-02-2022-0216
26. Wang X, Pan HR, Zhu N, Cai S. East Asian films in the European market: the roles of cultural distance and cultural specificity. *Int Market Rev.* (2021) 38:717–35.
27. Zhou C, Tang W, Zhao R. Optimal consumer search with prospect utility in hybrid uncertain environment. *J Uncertainty Anal Appl.* (2015) 3:1–20.
28. Maulud D, Abdulazeez AM. A review on linear regression comprehensive in machine learning. *J Appl Sci Technol Trends.* (2020) 1:140–7.
29. Filzmoser P, Nordhausen K. Robust linear regression for high-dimensional data: an overview. *Wiley Interdisc Rev Comput Stat.* (2021) 13:e1524.
30. Yuan C, Yang H. Research on K-value selection method of K-means clustering algorithm. *Multidisc Sci J.* (2019) 2:226–35.