Crowdsourcing Logistics Pricing Optimization Model Based on DBSCAN Clustering Algorithm

ZHICHAO LI¹, YILIN LI¹, WANCHUN LU¹, AND JILIN HUANG²
¹School of Political Science and Public Administration, East China University of Political Science and Law, Shanghai 201620, China
²Department of Environment Science, College of Environmental Sciences and Engineering, Peking University, Beijing 100871, China

Corresponding author: Jilin Huang (jilinhuang@pku.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 71974057.

ABSTRACT
From the perspective of platform economics, crowdsourcing is a very efficient business model, and the pricing of crowdsourcing tasks is a key factor for the sustainable development of the crowdsourcing model. In the logistics industry, crowdsourcing provides a new idea of sustainable development for logistics enterprises, and reasonable distribution pricing is the key to achieving sustainable development. This paper innovatively adds dynamic and decentralized characteristics of logistics on the basis of a detailed analysis of pricing methods and uses this as a basis to build a pricing model. First, based on existing crowdsourced photography task pricing data, this paper establishes a project-centric domain and builds metrics into the attributes of each project based on the data in that domain. Then, a regression model is used to fit the completion rate of previous projects, and a multiple linear regression and optimal pricing mechanism are established. Finally, the DBSCAN algorithm is used to cluster areas with a high project density, and a pricing optimization model based on polynomial Logit (MNL) is established. We found through the model analysis that the optimized pricing strategy of crowdsourcing logistics services has a better packaging completion rate based on a combination of complex factors including bundling and outliers. In short, the main contributions of this paper are to build a complex mathematical model for crowdsourcing tasks, improve the algorithmic deficiencies of the previous crowdsourcing task pricing methods, and provide a reference for further research on crowdsourcing tasks.

INDEX TERMS
Crowdsourcing logistics, pricing optimization model, DBSCAN clustering algorithm, sustainable development.

I. INTRODUCTION
In the logistics industry, crowdsourcing provides a new idea for the sustainable development of logistics enterprises, and reasonable distribution pricing is the key to the sustainable development of logistics enterprises [1]. Therefore, this paper attempts to establish an optimal pricing model for logistics crowdsourcing tasks.

The platform economy is becoming an important driver to promote the upgrading and transformation of industrial structures. Different from the traditional linear channel value chain model, platform enterprises try to create a circular-driven process to maximize the overall value of the business ecosystem. In digital business, many platform enterprises continue to practice sustainable business models [2]. For example, in many businesses, platform enterprises can use the crowdsourcing model, using network resources, group power and collective wisdom to improve their operating efficiency, allocate resources reasonably and achieve sustainable development. Crowdsourcing is a business model based on outsourcing that transforms external resources into sustainable internal resources. Although it has a broad application prospect, it faces a problem that must be solved, that is, the price of each task must be reasonably determined [3]. If the price is not reasonable, some tasks will be ignored, leading to the failure of the business plan. A reasonable pricing scheme is directly related to the sustainable development of platform enterprises using crowdsourcing. Therefore, the pricing scheme must conform to reality to ensure the stability of the supply and demand balance.

The research content of this paper is crowdsourcing task pricing in the logistics industry. How to price tasks on...
a crowdsourcing logistics platform is an important topic in platform economic theory. Crowdsourcing logistics is an Internet platform that uses idle transportation and human resources to provide distribution services. The crowdsourcing model, which greatly reduces the cost of human resources, is very suitable for the labor-intensive logistics enterprises, and its flexible working hours are also suitable for food delivery services, which have obvious peaks and valleys. In two-sided market theory, platform pricing is the core issue and has important value in realizing the sustainable development of crowdsourcing logistics platforms [4]. The contribution of this paper is to discuss the network externality characteristics and pricing model of crowdsourcing logistics in the platform and to establish a reasonable pricing model based on the practical characteristics of crowdsourcing logistics to make it sustainable. Since models are supposed to reflect reality, we consider the common problem of packaging sales. After establishing the new pricing model, we need to simulate and predict the implementation effect of the scheme. Then, a logistics classifier is selected to train the implementation of the original pricing scheme (project completion rate) to predict the completion rate of the new scheme.

The structure of this paper as follows: The first section of the article is an introduction, and the second section is a literature review. The third section introduces the clustering algorithm and index selection, and the fourth section presents the DBSCAN clustering algorithm and multicluster linear regression model pricing scheme. The fifth section considers a package model that uses bundling to optimize the model established in the fourth part, and the sixth section includes the summary and conclusions. This paper attempts to establish a sustainable pricing model for crowdsourcing logistics.

II. LITERATURE REVIEW
In recent years, quite a number of papers have studied the platform economy. Among them, the pricing model of platform competition and the factors influencing project pricing are the issues that have most concerned scholars [5]. The pricing structure of the platform will have an important impact on the transaction scale, market structure and interest pattern of the transaction subjects and can even affect the success or failure of the operation of the platform enterprises [6]. Reasonable pricing is crucial to the development and operation of crowdsourcing logistics platforms [7]. In this section, the relevant literature is divided into three parts: the pattern and mechanism of crowdsourcing, the price structure of logistics platforms in bilateral markets, and the applicability of the DBSCAN clustering algorithm. We will collect views from different reports and further study the theoretical significance of the pricing optimization model.

A. THE MODE AND MECHANISM OF CROWDSOURCING
Crowdsourcing is the action of a requester subcontracting a human intelligence task (HIT) to workers on a network platform. This concept was originally proposed by Jeff Howe [8], and there have been many attempts of crowdsourcing research and practice. Estelle is-Arolas thought of crowdsourcing as a distributed problem-solving mode and production mode. The response of an individual is a solution that is difficult to handle online [9]. Based on the practice of crowdsourcing, the crowdsourcing modes are elaborated upon according to different standards, which are used as the basis for the classification of the crowdsourcing mode [10]. Among them, the purpose-based crowdsourcing mode emphasizes the purpose of crowdsourcing, with the focus on “what to do”. Saxton et al. summarized 9 crowdsourcing patterns based on 103 crowdsourcing websites [11]. On the other hand, the process-based crowdsourcing pattern classification emphasizes the specific process of carrying out crowdsourcing activities, focusing on the “how to do” problem. Afuah et al. divided crowdsourcing into competitive crowdsourcing and collaborative crowdsourcing. The main difference between the two models is whether there is collaboration between the participants [12]. In addition, in the study of crowdsourcing mechanisms, Kittur studied the problem of complex task decomposition and solution integration, proposed an analysis framework based on MapReduce, and found that the crowdsourcing mechanism has good collaboration ability and creativity [13]. Zhao et al. proposed an effective task allocation method based on the ability of crowdsourcing participants to complete tasks, which can effectively help enterprises assign tasks to the most appropriate participants [14]. Stieger et al. found in their research on the design of crowdsourcing incentive mechanisms that the key to motivating employees to complete crowdsourcing tasks is to create a suitable process incentive to guide users to participate [15]. Giret, aiming at the quality assessment of crowdsourcing results, proposed that the results could be evaluated from the four dimensions of novelty, flexibility, relevance and comprehensiveness [1].

In the broad sense, crowdsourcing is a distributed digital business model. It helps enterprises to explore public creativity, deal with problems in organization, production and R&D, and solve logistics problems between products and distribution [16]. For example, crowdsourcing logistics platforms outsource tasks previously done by full-time deliverymen to nonspecific groups through the Internet in a voluntary and paid way [17]. This is what we are interested in. In terms of logistics and distribution, crowdsourcing provides a new way of thinking and approach. This type of breakthrough innovation can more effectively integrate internal and external resources and help logistics enterprises achieve sustainable development with the help of crowdsourcing [16].

B. PRICE STRUCTURE OF A LOGISTICS PLATFORM IN BILATERAL MARKET
A platform is a real or virtual space that can guide or facilitate transactions between two or more customers, with the core being connection, bridging or matchmaking. Since the platform has the typical characteristics of a two-sided market, many scholars have studied the price strategy of the
platform by using two-sided market theory (Caillaud & Jullien, 2003; Rochet & Tirole, 2003; Armstrong, 2006; Hagiu, 2006; Economides & Tag, 2012). The literature presents a variety of pricing strategies to attract bilateral users to a platform. Among them, Caillaud & Jullien solved the problem of platform competition through pricing decisions [18]. Rochet and Tirole respectively discussed the platform benefit maximization equilibrium price and social welfare maximization Ramsey pricing in the case of a monopoly platform and competitive platform and concluded that the price distribution is proportional to the elasticity of the bilateral price demand [19]. Armstrong studied the price decisions of a monopoly platform, bilateral single-owned competition platform and unilateral single-owned competition platform and drew the conclusion that the equilibrium price is related to the size of the cross-network effect, charging mode and market structure [20]. Economides & Tag analyzed the net neutrality of the Internet by using the price tool of two-sided market theory and found that the net neutrality platform increased the total surplus compared with the private platform in a certain parameter range of the cross-network externality while reducing the total surplus in other parameter ranges [21].

Different from the traditional network externalities, the formation of the platform economy is based on cross-network externalities [22]. Katz & Shapiro argued that network externalities can be divided into direct and indirect cases [23]. Direct network externalities are the values associated with the number of users using a product or service, while indirect network externalities mainly refer to the externalities caused by complementary products and services. The platform marketplace consists of two users who interact with each other to gain value through a common platform. The crowdsourcing logistics platform enterprises studied in this paper have the characteristics of cross-network externalities. Their participants are merchants on one side and free couriers on the other. The two parties involved in platform transactions often have complementary needs. Only when bilateral users participate in the platform and have demand for the products or services provided by the platform can the platform realize its own value and earn profits [24].

From the perspective of crowdsourcing logistics, this two-sided market does not always exist independently. In general, many merchants connect to the platform through the application programming interface (API) of the open logistics platform and distribute it in the form of crowdsourcing. Eichhoff and Vries proposed to evaluate the quality of crowdsourcing results from the four dimensions of novelty, flexibility, relevance and comprehensiveness to meet the needs of bilateral complementarity [25]. The pricing of the distribution tasks released by logistics enterprises on the platform will affect the changes in the supply and demand of the merchants and express personnel. Compared with products and services such as credit cards that can realize transactions online or offline, logistics platforms must combine online and offline functions to realize transactions. Logistics products are generally nonstandardized and nonstorable, so pricing studies on logistics platforms are rare. Based on the research results of Armstrong (2006), Wang analyzed the operation strategy of the fourth-party logistics platform in China and conducted a case study on Chuanhua logistics [26]. Kung & Zhong studied the optimal pricing strategy of a distribution platform, but this paper considered the cross-network effect only in the consumer market, not in the distributor market [27]. Punel & Stathopulos empirically analyzed how the degree of acceptance and preference of crowdsourcing logistics affect the task pricing [28] and found that consumers of local short-distance delivery preferred to pay for faster delivery while consumers of long-distance delivery preferred to pay for deliverymen’s expertise and experience. In the study of platform economic theory, the pricing model is one of the core issues [11]. There are some influencing factors to be taken into account in the pricing decision of the main body of the platform. Filistrucchi & Geradin pointed out that even in the case of bilateral information symmetry, platform pricing asymmetry and other asymmetries would occur [29]. Caillaud & Jullien found that the demand elasticity of both sides of the market is an important factor to be considered, and they obtained an intuitive conclusion based on a static pricing model [18]: the size of a given market is the influencing factor of the demand elasticity of the other side. Dou et al. showed that when intrinsic motivation is ineffective, a reward is a prerequisite for users to participate in crowdsourcing [30].

Although the above research provides a general introduction to the pricing structure of the crowdsourcing distribution tasks published by a logistics platform, it also enables a preliminary study on the selection and use of algorithms. However, there is still a lack of more optimized pricing models that can better utilize the crowdsourcing model of logistics platforms. With the accelerated development of the Internet era, crowdsourcing platforms have developed rapidly, playing a positive role in improving the logistics efficiency and reducing logistics costs. Crowdsourcing logistics platforms have broad development prospects and exhibit a strong sustainable development momentum. The theory of the platform economy provides a new perspective for the development of crowdsourcing platforms. Therefore, the objective of this paper is to build a model that can optimize the pricing of crowdsourcing logistics platforms.

### III. CLUSTERING ALGORITHM AND INDEX SELECTION

In this section, we have reviewed and categorized the relevant literature [31–33]; our article belongs to the third type of study described in Table 1). We use a mathematical model to improve the pricing mechanism of the crowdsourcing task using actual data from Guangdong Province, China. We take part of the data as the training set, and the other part (containing only geographic information with all other information hidden) is used as the test set on which recognition and classification is performed to make it accurate.
TABLE 1. The basic paper types.

| Type | Concrete content | Main representative scholars |
|------|------------------|-----------------------------|
| 1. The rise, development and expansion of crowdsourcing | The definition of crowdsourcing, a theoretical discussion of crowdsourcing task sources, the exploration of crowdsourcing’s legitimacy, and the business model formation of the crowdsourcing task. | Chris Y; Marcus A; Estelles; Arolas E |
| 2. Pricing and incentive mechanism of the crowdsourcing task (theoretical discussion) | How to make appropriate pricing strategies for crowdsourcing tasks to attract more riders to take orders and reduce costs. The incentive mechanism of crowdsourcing tasks, including the novelty and flexibility of dispatching tasks. | Saxton D; Steiger; Afuah A |
| 3. Improvement of the mathematical model for the crowdsourcing task and the cost. | The optimization of the pricing mechanism by an algorithm to achieve a comprehensive order quality optimization and the cost. | Zhang Q; Yang L; Zhuo G; Jia; Q; Li M; Tong Y; Chen L; Zhou; Z |

A. CLUSTERING ALGORITHM SELECTION

Clustering is an important unsupervised algorithm process in machine learning; it can combine data points into a series of specific combinations [34]. In theory, data classified as a group have the same characteristics, while data of different types have different attributes. The goal of the clustering algorithm is to obtain a compact subset of multiple classes by clustering a series of data. There are currently four different types of clustering algorithms, including partition-based, hierarchical, density-based, and grid-based.

1) PARTITION-BASED CLUSTERING ALGORITHM

Given a data set with N tuples, the splitting method will construct K groups, each representing a cluster, K < N. The partition-based algorithm needs a large amount of computation but is very suitable for discovering spherical clusters in small- and medium-sized databases (K-MEANS, K-MEDOIDS, CLARANS).

2) HIERARCHICAL CLUSTERING ALGORITHM

The given data set is decomposed hierarchically until some condition is satisfied. This type can be divided into “bottom-up” and “top-down” schemes. This algorithm is characterized by a small amount of computation, but it cannot correct wrong decisions (BIRCH, CURE, CHAMELEON).

3) DENSITY-BASED CLUSTERING ALGORITHM

As long as the density of a point in a region is greater than a certain threshold, it is added to the cluster that is closest to it. This algorithm can overcome the shortcomings of the distance-based algorithms, which only find clusters that are “circular-like” (DBSCAN, OPTICS, DENCLUE).

4) GRID-BASED CLUSTERING ALGORITHM

The data space is divided into a grid structure with a finite number of cells, and all processing is performed on a single unit. The grid-based algorithm has a fast processing speed that is only dependent on the number of cells in the data space (STING, CLIQUE, WAVE-CLUSTER).

Of the above algorithms, only the K-MEANS algorithm and DBSCAN algorithm have strong applicability. Hierarchical clustering is too inefficient and is not suitable for large-scale dataset analysis, so it is excluded. The input parameters of the grid-based algorithm have a great influence on the clustering results, and these parameters are difficult to set. When there is noise in the data, the clustering quality of the algorithm will be poor if no special treatment is applied. Moreover, the grid-based algorithm is less scalable for the data dimensions. There are many data dimensions in this paper, and the input parameters are difficult to determine, so the grid-based algorithm is not considered.

Figure 1 below is a brief overview of this paper, which shows theoretically that this paper is a step-by-step improvement of the pricing model, so it is reasonable.

The k-means algorithm is fast, and its calculation process is simple. However, it is insufficient to judge the number of classifications. Additionally, because k-means is characterized by the random selection of center points, the process...
cannot be repeated, and the randomness is relatively large. The following Figure 2 shows different clustering results using the K-MEANS clustering algorithm for the same set of four-dimensional data. As seen from the graph, the results from the repeated application of the K-MEANS clustering algorithm are quite different.

The DBSCAN algorithm is a density clustering algorithm based on high-density connected regions [35]. The algorithm has the advantages of a fast clustering speed, good noise processing effect and insensitivity to cluster shape. It can find spatial clusters of arbitrary shapes [36] without affecting the data input sequence. In addition, the DBSCAN algorithm can filter high-density areas during operation. It does not require the number of clusters to be known, and the clustering is not biased. Therefore, this paper chooses the DBSCAN algorithm for clustering, and to evaluate the efficiency of the clustering, K-MEANS clustering is used for comparative testing.

**B. INDEX SELECTION**

The pricing mechanism of a task varies with the region, and each geographical block requires different pricing strategies because of its particular location advantages and traffic conditions. When exploring the project pricing mode, if it is not broken down, the conclusion is very likely to be biased. This paper takes a task-based approach. Based on the latitude and longitude of a task, with each independent task as the center, and the dimensions being extended 0.2 units in the surrounding (southeast northwest) directions, the area that each task contains is called the field. The field is a sociological concept [37], specifically referring to the relationships within the independent space. This paper distinguishes the concept from the grid. Each field corresponds to a number of indicators in their respective regions. The field’s advantages include that, based on the task, it can not only take into account the refinement of each task’s indicators but can also better analyze the reasons why a task is not completed. Additionally, the geographical location and other factors do not cause any deviation.

With the development of mobile networks and the emergence of the sharing economy, there are many data sets describing crowdsourcing. The data set we analyzed includes the data of a completed project (which can be regarded as training data). The training data includes the location, pricing and completion of each task (where “1” means completed, and “0” means not completed). Another data set is additional information, which contains the personal information of the rider (order executor), including his location, reputation value and reservation limit. In principle, the higher the rider’s reputation is, the higher the priority of starting to select tasks and the larger the quota is (in fact, the allocation is based on the proportion of the reservation limit reached during the task allocation) - this data set can be regarded as an additional data set that is a supplementary part of how to correctly carry out pattern recognition. The third data set is a new check item task data set that contains only the location information of the tasks. Our aim is to recognize and classify the new data set correctly to achieve the highest accuracy.

The data in this paper comes from the Higher Education Press. The original data includes the following indicators: the contractor’s number, scheduled task quota, scheduled task start time, reputation value, task number, task location (GPS), task price, and task execution (binary data). In principle, the higher the reputation of the contractor, the higher the priority of the task selection and the larger the quota (task allocation is actually based on the proportion of the reserved quota occupied).

Indicators of the original data cannot be directly substituted into the model for pricing deduction, so we regenerate some variables according to the characteristics of these indicators and eliminate unnecessary variables by a stepwise regression method, making the pricing model more rigorous and concise. The following is an introduction to the newly generated variables.

In this paper, we select seven indexes: 1. the supply and demand density of the field, 2. the average distance from the mission point, 3. the actual task quota in the field, 4. the average reputation value of the field, 5. the time covariance, 6. the task cutting rate, and 7. the minimum distance from the mission point.

**FIGURE 2.** Comparison of results of repeated runs of k-means algorithm under the same conditions.
supply and demand density is set as \( X_1 \). That is, the demand and supply density of the field is:

\[
X_1 = \frac{\sigma_1 \times S_{per}}{\sigma_2}
\]  

(1)

The average distance from the task point represents the average distance of each member from the mission in the field. Each individual distance is calculated as \( l_{per} \), the sum of the individual distances is \( \sum_{i=1}^{n} l_{per} \), and the average distance between range points is \( X_2 \). That is, the average distance from the mission point is:

\[
X_2 = \frac{\sum_{i=1}^{n} l_{per}}{\sigma_2}
\]  

(2)

The actual task quota of the field reflects the ratio of the total number of tasks in the field to the actual number of tasks in the field summed for all members and is expressed as \( X_3 \). Within the field, each member has a predetermined quota of tasks \( U_1 \), and the total number of tasks a member has to order from is a computable constant of 12,830 (known data). Currently, the number of tasks supplied is a constant of 835 pieces (known data). Set the actual task quota for the field, and the actual field quota is calculated as:

\[
X_3 = \frac{\sigma_1}{\sum_{i=1}^{n} \frac{835 \times U_1}{12830}}
\]  

(3)

The credit value of each member is calculated as \( M \). Set the field average reputation value, \( X_4 \), defined as:

\[
X_4 = \frac{\sum_{i=1}^{n} M}{\sigma_2}
\]  

(4)

The time covariance describes the change in the time difference, and the task of the price will be subject to the corresponding changes. Set the time difference as \( t \), and the time covariance is \( X_5 \). Using the block growth model (Logistic), its formula is:

\[
X_5 = \frac{1}{1 + e^{-t}}
\]  

(5)

The task cut rate is the ratio of the scheduled task quota \( U_1 \) to each member’s reputation value \( M \). It is defined as \( X_6 \), with a formula:

\[
X_6 = \frac{U_1}{M}
\]  

(6)

Set the minimum distance for each member to the task point as \( l_{ave} \), and \( n \) is the number of members. The minimum distance from the task point is \( X_7 \), with a formula:

\[
X_7 = \frac{\sum_{i=1}^{10} l_{ave}}{10}
\]  

(7)

Then, the data processed by the above indexes are processed in a dimensionless manner, and the Z-score method is adopted. To avoid the co-occurrence of the above indicators, a stepwise regression method is adopted to eliminate the variables that are not significant. The basic idea of stepwise regression is to introduce the variables one by one into the model. Each time an independent variable is introduced, the F test is carried out, and the independent variables are selected for the t test. When an original independent variable is determined to no longer be significant, it is deleted. This is an iterative process that is performed until either no significant independent variables are selected in the regression equation or no significant independent variables are removed from the regression equation. Some of the variables \( X_2, X_5, X_6, X_7 \) were eliminated. The reason for the elimination of the second variables \( X_2 \) may be that the average distance attribute within each task is not obvious due to the establishment of the field. This reflects the pricing model that can be obtained according to the above formula. Its constant term is \( X_0 = 69.113 \), and the values of \( X_1, X_3, \) and \( X_4 \) are 0.021, -0.027, and -0.036. The pricing pattern is expressed as:

\[
y = 69.113 + 0.021X_1 - 0.027X_3 - 0.036X_4
\]  

(8)

IV. DBSCAN CLUSTERING ALGORITHM AND MULTICLUSTER LINEAR REGRESSION MODEL PRICING SCHEME

This paper is based on the data background of Guangdong, China, but the original data is not classified, irregular and intricate. To design an optimal pricing model, a cluster analysis of the original data is needed. The original data is classified into multiple clusters, with data in the same cluster having high similarity and the similarity between data in different clusters being low [38].

In this paper, the density-based DBSCAN clustering algorithm is used to cluster the original data. DBSCAN is a density-based clustering algorithm. It has the advantage of fast clustering and can effectively deal with noise points (outliers) and discover spatial clustering of arbitrary shapes. Its principle is to find high-density regions separated by low-density regions. To start the calculation, we first need to divide the task points into three categories according to the definition of the density, that is, the points (core points) in the dense area, the points on the edge of the dense area (the boundary points), and the points in the sparse area (noise or Background). The neighborhood within the radius of a given object Eps is called the Eps neighborhood of the target. In this paper, we use \( N_{Eps}(p) \) to represent the set of points within the Eps radius of point \( p \):

\[
N_{Eps}(p) = \{ q | \text{distance}(p, q) \leq Eps \}
\]  

(9)

In addition, if an object’s Eps neighborhood contains at least minpts objects, the object is called a core object. The algorithm is as follows:

1) DBSCAN searches the clusters by checking the Eps neighborhood of each point in the dataset and creates a cluster that considers the core object if the Eps neighborhood of the point contains more than minpts.

2) Then, DBSCAN iteratively gathers the core object and the objects that are reachable; this process involves the merging of some density reachable clusters.
3) When no new points are added to any cluster, the process ends.

At the same time, in order to test the clustering effect of DBSCAN, we also conducted K-means clustering in the first clustering. Figures 3 and 4 show a comparison between DBSCAN clustering and K-means clustering. As shown in Figures 5 and 6, in the second and third clusterings, we use the DBSCAN clustering method to recluster Cluster 1 and Cluster 2.

The new optimization scheme needs to be based on certain rules. According to the original data, we study the characteristics of the indexes in the tasks that have been successfully executed and obtain a new pricing scheme based on feature fitting. Generally, the data generated during the operation is discrete, that is, pure regression or fitting methods cannot obtain good results. Therefore, this paper uses a multiple linear regression model based on multiple clusters to establish a pricing scheme. Multiple linear regression [39] has a good effect in solving multivariable problems. The algorithm is based on least squares (OLS), and the calculation process is as follows: Suppose that a dependent variable, \( y \), is affected by the \( k \) independent variables \( T_1, T_2, \ldots, T_k \), and the \( n \) group observations are \( (y_a, T_{1a}, T_{2a}, \ldots, T_{ka}), a = 1, 2, \ldots, n \). The structure of the multiple linear regression model is:

\[
y_a = x_0 + x_1T_{1a} + x_2T_{2a} + \ldots + x_kT_{ka} + \epsilon_a \quad (10)
\]

Among them, \( x_1, x_2, \ldots, x_k \) are the undetermined coefficients; and \( \epsilon_a \) is a random variable. If \( b_0, b_1, \ldots, b_k \) are the fitting values of \( x_1, x_2, \ldots, x_k \), respectively, the regression equation is:

\[
\hat{y} = b_0 + b_1T_1 + b_2T_2 + \ldots + b_kT_k \quad (11)
\]

where \( b_0 \) is a constant, \( T \) is an independent variable, and \( b_0, b_1, \ldots, b_k \) are the partial regression coefficients.

Subsequently, for raw data that has been successfully executed, the price of the task is the dependent variable, and \( X_1, X_3, X_4 \) were used as independent variables to fit the multiple linear regression. The fitting process divides the
original data into clusters, where the first two clusters occupy 73.3% and 21.1% and the remaining clusters occupy 5.4%. According to its characteristics, the first cluster is selected, and the second clusters and the remaining clusters are fitted. We can obtain the results of the first, second and remaining clusters, where \( p_i \) represents the price:

\[
\begin{align*}
    p_1 & = 71.232 + 13.0514X_1 - 0.0153X_3 - 0.0026X_4 \\
    p_2 & = 68.847 - 0.08323X_1 - 0.0015X_3 - 0.0073X_4 \\
    p_3 & = 85.6811 - 14.6013X_1 - 0.0001X_3 - 0.0121X_4
\end{align*}
\] (12)

After the establishment of a new pricing model, it is necessary to simulate the effect of the implementation of the proposed scheme. In this paper, we select a binary logistic classifier to study the implementation rate (task completion rate) after the original pricing scheme to predict the completion rate of the new scheme. The logistic classifier [40] is based on the delivery of Bernoulli (Bernoulli), specifically to solve the training sample data in the dichotomous variable (0 or 1) of the relevant issues [41]. The calculation process is as follows: Let \( \eta = (\eta_1, \eta_2, \ldots, \eta_{p-1})^T \) be a factor that affects the probability of occurrence of event \( A \). \( P(\eta) \) indicates the probability of event \( A \) occurring, and the probability that event \( A \) does not occur is \( 1 - P(\eta) \).

\[
P(\eta) = \int (\eta_1, \eta_2, \ldots, \eta_{p-1})
\] (13)

It can be determined from the above that:

\[
\emptyset [P(\eta)] = \ln \left( \frac{P(\eta)}{1 - P(\eta)} \right)
\] (14)

When \( 0 < P(\eta) < 1 \) is satisfied, the following formula is applicable, \( -\infty < \emptyset [P(\eta)] < \infty \)

\[
\ln \left( \frac{P(\eta)}{1 - P(\eta)} \right) = F(\eta_1, \eta_2, \ldots, \eta_{p-1})
\] (15)

Consider \( F \) a linear function:

\[
F(\eta_1, \eta_2, \ldots, \eta_{p-1}) = \beta_0 + \ldots + \beta_{p-1}\eta_{p-1} = \beta_0 + \sum_{k=1}^{p-1} \beta_k \eta_k
\] (16)

Thus, the probability of event occurrence can be calculated directly, and the coefficients in the model are estimated by iterative calculations using the maximum likelihood parameter.

To evaluate the rationality of the pricing model proposed in this paper, the least squares linear regression model, multivariate linear regression and multicluster linear regression are applied in comparative experiments. As shown in Table 2, the accuracy of the multicluster linear regression model established in this paper is relatively high, up to 0.80357, with strong applicability.

### V. A PACKAGING MODEL THAT TAKES BUNDLING INTO ACCOUNT

In this section, a packaging model is presented that takes bundling into account. In addition to introducing how to use the DBSCAN clustering algorithm to package tasks, the pricing optimization model based on minimal residual tasks, and the discount rate optimization model based on the multinominal logit model, we also discuss the validity of the logistics classification model and perform an error analysis.

#### A. USING DBSCAN CLUSTERING ALGORITHM TO PACKAGE TASKS

In real life, task publishers may use packaged sales in order to make tasks more quickly available. Packaging can be considered as a bundle of tasks that focus on location and can also be packaged in accordance with the principle of distance. This paper considers the task of focusing on the location of the package. In the case of guaranteeing the completion rate of a task, the price is converted, and the cost is reduced. The impact of packaging on pricing is not a simple addition process and should take into account the more complex real situation. Based on this, this paper first uses the DBSCAN clustering algorithm to cluster the more intensive task points. Using MATLAB, through the density-based DBSCAN clustering, we can obtain the cluster diagram and the number of clusters diagram, as shown in Figure 7 and Figure 8. From the clustering diagram, many of the denser points are packaged together. By clustering the number of packages, we find that the number of clusters containing one task is 109, the number of clusters with 2 tasks is 42, and so on. A block contains 20 tasks, and according to the data, the clustering term is 266.

#### B. PRICING OPTIMIZATION MODEL BASED ON MINIMAL RESIDUAL TASKS

After a cluster is complete, it needs to be packaged. This paper builds an optimization model of the packaging method based on the smallest remaining tasks. The purpose of this model is to consider the situation of the package after the task of clustering in the event that all phenomena cannot be received. The existence of single tasks is bound to cause the rate of completion of the tasks to decline. Therefore, the selection of this package should be aligned as much as possible with the intensive task of all packages being completed.

First, select a cluster that has more than two tasks and package it. Consider that the binding method is not always going to be the same one but different circumstances should be discussed. We cannot oversimplify the number of packages in each group, so we have developed three different packaging methods, namely, with 3, 4, and 5 tasks in a set. According to the above idea, the optimization model is established.
MATLAB is used in accordance with the above principles to achieve a bundled package.

To make the selected packaging method pack all the intensive tasks as far as possible, the minimum packing method and optimization model for the remaining tasks are established. Take the minimum tasks as the objective function, and mark it as $\min \beta$. The number of package strategies in the first group is $\alpha_1$, the number of packaging strategies in the second group is $\alpha_2$, and the number of packaging strategies in the third groups is $\alpha_3$. $Q_{\text{ove}}$ is the total number of tasks. The following optimization formulas are given:

$$\min \beta = Q_{\text{ove}} - 3\alpha_1 - 4\alpha_2 - 5\alpha_3$$

$$0 \leq \alpha_1 \leq \frac{Q_{\text{ove}}}{3}$$

$$0 \leq \alpha_2 \leq \frac{Q_{\text{ove}}}{4}$$

$$0 \leq \alpha_3 \leq \frac{Q_{\text{ove}}}{5}$$  \hspace{1cm} (17)

They can then be solved by MATLAB. After clustering, the packaging results are as shown in Figure 9. Among them, the red indicates successful packaging, and the black the unpackaged.

C. DISCOUNT RATE OPTIMIZATION MODEL BASED ON MULTINOMIAL LOGIT MODEL

After considering the packaging problem, many parameter values are bound to change. At the same time, we need to revise the previous pricing model and measure the task completion rate at this time. Once a package is taken, the package price exhibits a downward trend, that is, a common discount. The discount rate is expressed as $\theta$. The idea of this paper is to use the discount rate to reflect the modified pricing model and to find the model with the lowest cost or the highest discount rate when the completion rate is guaranteed. A multinomial logit model (MNL) is used in this paper.

Based on the current research literature, a suitable method to describe utility is the MNL model based on random utility theory [42]. Because of the completion of the task after binding and packaging, it is necessary to consider the subjective psychology of the member. Therefore, the MNL model ideally includes the subjective factors of the member, which is reasonable and feasible. The basic idea is that a member faces different choices when choosing tasks. These choices constitute a choice set, and a member chooses only the most useful one. The hypothesis of the MNL model is derived:

1) For a given member, the utility of a task can be divided into two parts:

$$U_j = G_j + \varepsilon_j$$  \hspace{1cm} (18)

in which $G_j$ is the determining part and $\varepsilon_j$ is the stochastic part, which considers unobservable variables.

2) According to the random utility model, members will choose their most useful task. In the selection set $C$, the probability that a given member will select task $j$ is:

$$q_j = q(\varepsilon_i - \varepsilon_j < G_j - G_i, i, j \in C, i \neq j)$$  \hspace{1cm} (19)

3) For the random part $\varepsilon$ of the utility, there are two conditions: one is an independent random variable, and the second
is a probability variable subject to the double power function probability delivery:

\[ q_j = q(e_{\theta} \leq e) = e^{-e}, \quad -\infty < e < \infty \] (20)

Based on the formula, the task completion rate can be represented by the following basic principles:

\[ q_j = \frac{e^{G_j}}{\sum_{j \in C} e^{G_j}}, \quad j \in C \] (21)

The numerator is the power function of the decision part of the task \(j\), and the denominator is the sum of all the utility functions in the selection collection. In this way, the stochastic part of the utility can be quantified, and the computation process of the membership selection probability is simplified. As a result, the rate of completion of the task based on the discount rate can be derived:

\[ q_j(\theta) = \frac{e^{a_{\theta}I_{sk} - \beta p(\sum_{i=1}^{k} p_0)}}{e^{a_{\theta}I_{sk} - \beta p(a_{\theta} + e^{a_{\theta} + e^{I_{sk} - \beta p_0}})}} \] (22)

\(\theta\) represents the discount rate, \(a_{\theta}\) stands for the number of \(S_k\) tasks in combination \(n\). \(I_{sk}\) is the additional task value in the bundled \(S_k\), \(\beta\) indicates the influence coefficient of the price \(p_0\) on the member selection utility \(p\left(\sum_{i=1}^{k} p_0\right)\) represents the sum of the \(k\) tasks in a package, which is a constant, \(r\) stands for the retention value of a member when a task is not selected, and \(p_0\) is the price tag for a single random task. Using MATLAB, the corresponding discount rate can be obtained. The newly assembled task set is priced by the total prepackaged price multiplied by the discount rate, the number of tasks changes from the total number \(n\) before the package to 1, and the membership quota is changed.

\[ P = p\left(\sum_{i=1}^{k} p_0\right) \times \theta \] (23)

For the calculation of the degree of completion, the DBSCAN clustering algorithm is used to package and price the tasks in clusters according to the density in accordance with the DBSCAN clustering algorithm. Pricing is performed according to the model pricing model in Sections 4 and 5 of these papers, and the model is modified after the pricing into the Logistics classifier to predict the completion of the situation. By the MATLAB calculation, the new result is 84.9%, which is better than that of the original model, 62.5%.

**D. THE TEST OF THE LOGISTIC CLASSIFICATION MODEL VALIDITY**

To test the effect of the optimized pricing, this paper establishes a logistic classification model to predict the results. The accuracy of this paper is measured by MSE. MSE is usually used to measure the deviation between the predicted value of a model and the real value. The real value is recorded as \(y\), and the predicted value is recorded as \(\hat{y}\):

\[ MSE = \frac{1}{M} \sum_{m=1}^{M} (y_m - \hat{y}_m)^2 \] (24)

The original task data is brought into the logistic classification model to predict the error contrast. The comparison between the predicted results and the original results is given below. The accuracy rate is up to 0.838, as shown in Figure 10. Therefore, the accuracy of the model is higher, and the results of the model are more reliable.

**E. ERROR ANALYSIS**

1) **SYSTEMATIC ERROR**

The original data only cover the task number, task location (GPS), task execution price, membership number, membership position (GPS), booking, booking limit of a task, start time of a task, and the reputation value. According to these nine factors, we study the pricing rule in reverse, but in fact, the actual pricing model may be related to other factors, such as the task difficulty, competition price, and so on. At the same time, some process indicators, such as the scheduled process and scheduled time, will also occupy a larger role. Therefore, the model is affected by the confidence of the data, and inevitably there is a certain system error.

2) **RANDOM ERROR**

In this paper, the crowdsourcing logistics pricing model uses the idea of machine learning, so the data quality of the training set is relatively high, although the small sample size in some intervals will lead to errors caused by random factors. Moreover, the user orders will have a certain subjective consideration, which will lead to a certain random error.

**F. DISCUSSION**

Different from the existing research, this paper innovatively adds dynamic and decentralized characteristics of logistics on the basis of a detailed analysis of pricing methods and takes them as a sample to try to build a pricing model that can make the crowdsourcing logistics model a sustainable business model. Therefore, during the analysis, the objects of the first cluster cover all the data, and on the basis of the first cluster, two clusters are clustered again. After two clusters, the coordinate system is established and divided into
four quadrants, which serve as the basis for classification and discussion, making the conclusion more specific and more logical. To make the model more scientific and eliminate collinearity, the method of stepwise regression is used to simplify 7 variables into 4 variables. The contribution of this paper is that we establish a new pricing scheme by using a two-point logic model and multiple regression function of clustering. When analyzing bundles, we considered the combination of the number of bundles in detail and used DBSCAN clustering to cluster and then package the denser data sets. Finally, an optimization model based on the MNL project completion rate is established. The achievement of the model lies in the fact that this paper takes into account relatively complex factors, such as the selection of data outliers, indicators and regression equations, as well as the reality of bundling and model testing. However, in order to make the model more accurate, we conducted two clustering analyses on the data. These are the prediction of the completion rate of the system by using the simulation method based on the logic classifier. In addition, we conducted reliability tests and an error analysis to make the model more rigorous.

VI. CONCLUSIONS

As crowdsourcing tasks become more and more popular, there is still no complex system model considering the pricing of crowdsourcing tasks. The main feature of this paper is to train a model with real data and test a training set separated from real data. Experimental results show that the classification accuracy of the model is high. Based on the data characteristics, this paper conducts clustering and develops a subregional research task pricing model and establishes an optimized package pricing scheme and package pricing model. When the new pricing model is designed, a multilinear regression pricing model based on clustering is established. To explain the difference between the completion rate of the new scheme and that of the original scheme, a binary logic classifier was used to predict the task completion rate, and the results showed that the completion rate increased from 62.5% to 70.3%. At the same time, considering the different combinations of packaging, the density-based DBSCAN clustering algorithm is used for the regions with a high cluster task density, and then a single-objective optimization model is used to package the cluster task set. Finally, the MNL pricing optimization model is established, and the results are compared with those of the original scheme. The results showed that the task publisher’s expenditure decreased, and the task completion rate increased from 62.5% to 84.9%. We then grouped them by task density and packaged the denser parts. For the packaged task set, we use the MNL model to solve the pricing problem. For unpackaged data, we perform a multiple linear regression of the price by clustering and use a binary logic classifier to predict the corresponding task completion rate, which is 71.1%. The main contribution of this paper is to establish a complex mathematical model of crowdsourcing tasks, mitigating the shortcomings of previous crowdsourcing task pricing algorithms and enriching the research content of crowdsourcing enterprise pricing in the logistics industry.

The research content of this paper also provides some reference for further research on crowdsourcing tasks. In network platforms, intergroup network externalities also have an impact on platform pricing, which is a key to follow-up research. This is because a buyer will observe the product sales, after-sales evaluation and other information and then decide whether to buy. In addition, the research can be expanded to include the case of multiple purchases by users and the impact of flow nodes on the sustainability of crowdsourcing logistics. These are the discussions that will unfold in future studies.

ACKNOWLEDGMENT

The authors would like to thank the comments from the reviewers and the editor. They would also like to thank Yinye Hua for the manuscript discussion and Zhimin Li and Rui Xiao for their assistance in collecting data.

REFERENCES

[1] A. Giret, C. Carrascosa, V. Julian, M. Rebollo, and V. Botti, “A crowdsourcing approach for sustainable last mile delivery,” Sustainability, vol. 10, no. 12, p. 4563, 2018.
[2] K. Täuscher and N. Abdelkafi, “Scalability and robustness of business models for sustainability: A simulation experiment,” J. Cleaner Prod., vol. 170, pp. 654–664, Jan. 2018.
[3] R. T. Nakatsu, E. B. Grossman, and C. L. Iacovou, “A taxonomy of crowdsourcing based on task complexity,” J. Inf. Sci., vol. 40, no. 6, pp. 823–834, Dec. 2014.
[4] K. Chatterjee and K. Das, “Bilateral trading and incomplete information: Price convergence in a small market,” Games Econ. Behav., vol. 106, pp. 89–113, Nov. 2017.
[5] W. Sutherland and M. H. Jarrahi, “The sharing economy and digital platforms: A review and research agenda,” Int. J. Inf. Manage., vol. 43, pp. 328–341, Dec. 2018.
[6] A. C. Garavelli, A. M. Petruzzielli, A. Natalechio, and W. Vanhaverbeke, “Benefiting from markets for ideas—An investigation across different typologies,” Int. J. Int. Manag., vol. 17, no. 6, 2013, Art. no. 1340017.
[7] A. Natalechio, A. M. Petruzzielli, and A. C. Garavelli, “A literature review on markets for ideas: Emerging characteristics and unanswered questions,” Technovation, vol. 34, pp. 65–76, Feb. 2014.
[8] Z. Y. Chris and Q. Zhu, “Effects of Extrinsic and Intrinsic Motivation on Participation in Crowdsourcing,” Contest. Onl. Inf. Rev., vol. 38, pp. 896–917, 2014.
[9] E. Estellés-Arolas and F. González-Ladrón-de-Guevara, “Towards an integrated crowdsourcing definition,” J. Inf. Sci., vol. 38, no. 2, pp. 189–200, Apr. 2012.
[10] A. Marcus and A. Parameswaran, “Crowdsourced data management: Industry and academic perspectives,” Found. Trends Databases, vol. 6, nos. 1–2, pp. 1–161, 2015.
[11] G. D. Saxton, O. Oh, and R. Kishore, “Rules of crowdsourcing: Models, issues, and systems of control,” Inf. Syst. Manage., vol. 30, no. 1, pp. 2–20, Jan. 2013.
[12] A. Aftab and C. L. Tucci, “Crowdsourcing as a solution to distant search,” Acad. Manage. Rev., vol. 37, no. 3, pp. 355–375, Jul. 2012.
[13] A. Kittur, B. Smus, S. Khamkar, and R. E. Kraut, “Crowdforging: Crowdsourcing complex work,” in Proc. 24th Annu. ACM Symp. User Interface Softw. Technol., Santa Barbara, CA, USA, Oct. 2011, pp. 43–52.
[14] Y. Zhao and Q. Zhu, “Evaluation on crowdsourcing research: Current status and future direction,” Inf. Syst. Frontiers, vol. 16, no. 3, pp. 417–434, Jul. 2014.
[15] D. Steiger, K. Matzler, S. Chatterjee, and F. Ladstaetter-Fussenegger, “Democratizing strategy: How crowdsourcing can be used for strategy dialogues,” California Manage. Rev., vol. 54, no. 4, pp. 44–68, Jul. 2012.
[16] D. Schlagwein and N. Bjorn-Andersen, “Organizational learning with crowdsourcing: The revelatory case of LEGO,” J. Assoc. Inf. Syst., vol. 15, no. 11, pp. 754–778, 2014.

[17] M. Klumpp, “Crowdsourcing in Logistics: An Evaluation Scheme,” in Proc. 5th Int. Conferences Dyn. Logist., Bremen, Germany, 2016, pp. 401–411.

[18] B. Caillaud and B. Jullien, “Chicken & egg: Competition among intermediation service providers,” RAND J. Econ., vol. 34, pp. 309–328, Jul. 2003.

[19] J.-C. Rochet and J. Tirole, “Platform competition in two-sided markets,” J. Eur. Econ. Assoc., vol. 1, no. 4, pp. 990–1029, Jun. 2003.

[20] M. Armstrong, “Competition in two-sided markets,” RAND J. Econ., vol. 37, no. 3, pp. 668–691, 2006.

[21] N. Economides and J. Tag, “Network neutrality on the Internet: A two-sided market analysis,” Inf. Econ. Policy, vol. 24, no. 2, pp. 91–104, Jun. 2012.

[22] E. Minton, C. Lee, U. Orth, C. H. Kim, and L. Kahle, “Sustainable marketing and social media: A cross-country analysis of motives for sustainable behaviors,” J. Advertising, vol. 41, no. 4, pp. 69–84, 2012.

[23] D. Eickhoff and A. P. de Vries, “Increasing cheat robustness of crowdsourcing tasks,” Inf. Retr., vol. 16, no. 2, pp. 121–137, Apr. 2013.

[24] Y. Wang, D. Zhang, Q. Liu, F. Shen, and L. H. Lee, “Towards enhancing the last-mile delivery: An effective crowd-tasking model with scalable solutions,” Transp. Res. E, Logistics Transp. Rev., vol. 93, pp. 279–293, Sep. 2016.

[25] L.-C. Kung and G.-Y. Zhong, “The optimal pricing strategy for two-sided platform delivery in the sharing economy,” Transp. Res. E, Logistics Transp. Rev., vol. 101, pp. 1–12, May 2017.

[26] A. Purnel and A. Stathopoulos, “Modeling the acceptability of crowdsourced goods deliveries: Role of context and experience effects,” Transp. Res. E, Logistics Transp. Rev., vol. 105, pp. 18–38, Sep. 2017.

[27] C. Li and Z. Li, “A text classification algorithm based on improved multidimensional-multiresolution topological pattern recognition,” Int. J. Pattern Recognit. Artif. Intell., vol. 2, pp. 10021–10032, Sep. 2018.

[28] Z. Li, R. Jia, and J. Huang, “A text classification algorithm based on improved multidimensional-multiresolution topological pattern recognition,” J. Assoc. Inf. Syst., vol. 15, no. 11, pp. 5919–5929, May 2019.

[29] F. Zhu, “Modeling overdispersed or underdispersed count data with generalized Poisson integer-valued GARCH models,” J. Math. Anal. Appl., vol. 389, no. 1, pp. 58–71, May 2012.

[30] J. A. Adell and A. Lekuona, “Binomial convolution and transformations of appel polynomials,” J. Math. Anal. Appl., vol. 456, no. 1, pp. 16–33, Dec. 2017.

[31] S. Huda, J. Abawajy, M. Abdullahian, R. Islam, and J. L. Yearwood, “A fast malware feature selection approach using a hybrid of multi-linear and stepwise binary logistic regression,” Concurrency Comput., Pract. Exper., vol. 29, no. 23, 2016, Art. no. e3912.

[32] P. G. Ipeirotis, “Analyzing the amazon mechanical turk marketplace,” XRDS: Crossroads, ACM Mag. for Students, vol. 17, no. 2, pp. 16–21, Dec. 2010.

[33] S. M. Laudien and B. Daxbäck, “Business model innovation processes of average market players: A qualitative-empirical analysis,” R&D Manage., vol. 47, no. 3, pp. 420–430, Jun. 2017.

[34] N. Saeed, M. Amir, and S. Shahaboddin, “Sustainable business models: A review,” Sustainability, vol. 11, no. 6, pp. 1663–1693, 2018.