Advances in Social Science, Education and Humanities Research, volume 580
Proceedings of the 2021 International Conference on Social Science:Public Administration, Law and International Relations (SSPALIR 2021)

Research on Characteristic Identification of Relatively Poor Migrant Workers Based on Random Forest Model

Lilu Sun[1,*] Xin Ma[2] Yuxin Ding[1]

[1] *School of Management, Chongqing University of Technology, Chongqing 400054, China*
[2] *School of Economics, Chongqing University of Technology, Chongqing 400054, China*
[*] *Corresponding author. Email:15990021646@163.com*

ABSTRACT
After China has eliminated absolute poverty and entered the post-poverty era, the anti-poverty cause has evolved to focus on relative poverty, which has the characteristics of relativity, multidimensionality, and dynamics. In particular, migrant workers have become the main group of relative poverty. Its characteristics such as strong mobility, difficulty in continuously increasing income, and lack of endogenous motivation make it more difficult to identify and manage relative poverty. Therefore, this paper uses the double-boundary method to identify and analyze the relatively poor migrant workers. It is found that: (1) Compared with the one-dimensional income level as the standard, the multi-dimensional poverty identification system can better reflect the poverty status of migrant workers; (2) The migrant workers showed a higher incidence of poverty in the highest education level, children's school attendance, five social insurance and one housing fund and other indicators; (3) The random forest algorithm was used to construct the relative poverty identification model of migrant workers, with an accuracy rate of 98.8%. Meanwhile, the model reflected that education and employment dimensions should be emphasized in the identification process. Based on the research conclusions, this paper puts forward corresponding policy suggestions for China to deal with the relative poverty of migrant workers and puts forward further research directions.

Keywords: *Migrant workers, Relative poverty, Double bounds method, Random forest.*

1. INTRODUCTION

By November 2020, China had completely eliminated absolute and regional poverty, lifting nearly 100 million people out of poverty. However, the data show that there are still risks of returning to poverty among the people who have been lifted out of poverty and the risk of poverty among the marginal population. Therefore, the anti-poverty cause has evolved to focus on relative poverty. Relative poverty governance reflects a wider range of social equity, justice, and public demands. Therefore, relative poverty needs to present a variety of aspects such as interest coordinator, multi-dimensional view, and social security mechanism to carry out method transformation, reflecting the relationship between technical rationality and cultural values. According to the 2018 Monitoring and Investigation Report on Migrant Workers by the National Bureau of Statistics, the total number of migrant workers in 2018 was 288 million. They are also the focus of combating relative poverty. Their multidimensional poverty is related to their liquidity and lack of endogenous power, which also resulted in the difficulties in identification and risks to fall in poverty. This brings a significant challenge to establish a long-term mechanism to solve relative poverty. Therefore, the identification of the relative poverty of migrant workers is the main connotation of the future poverty alleviation work mechanism, and it is also an important direction for the systematic thinking and integrated deployment of relative poverty management.

Townsend (1979) believes that poverty is not only the lack of the most basic necessities for life but also the relative state of being excluded from the life and activities of normal social groups due to

the lack of development resources [1]. The "relativity" of relative poverty refers to the perception of relative poverty by members of society through comparison with others (Cao, 2020) [2]. From a multi-dimensional perspective, relative poverty not only focuses on the income gap but also needs to use multi-dimensional identification criteria to identify poor groups in terms of living standards, education, and health care (Seth and Santos, 2018; Bourquin, 2019) [3] [4]. At present, the most widely used multidimensional poverty research framework is the "double boundary method" (Alkire and Foster, 2007) [5]. Firstly, the poverty line of each dimension is selected to determine the poverty status of individuals in each dimension, and then the poverty dimension critical value is determined. Individuals are defined as poor when they are poor in a certain number of dimensions or more. From the perspective of dynamics, the measurement standard of relative poverty changes with economic changes and its measurement method also needs to be adjusted appropriately according to different regions in different periods (Joyce, Ziliak, 2019) [6]. Traditional methods tend to ignore implicit characteristics, resulting in the lack of comprehensive and objective evaluation criteria. In the previous phase of the battle against poverty, China used Internet big data technology to register and accurately identify poor families, which played an effective role. Relative poverty has more complex indicators and difficult problems. It is necessary to optimize Internet big data management technology, set up relative poverty population information database, and ensure scientific management and effective monitoring relative poverty data. The random forest model is very suitable for dealing with the relative poverty problem with large number of indicators and large sample size.

Therefore, this study aims to combine relative poverty theory and practical experience at home and abroad. Combing and extracting the characteristics of China's relatively poor groups with the "double boundary method", a multi-dimensional relatively poor identification index system is constructed. And random forest model is used to make a systematic test on the characteristics of migrant workers' relatively poor groups. Theoretically, the study expands relative poverty governance theory logic space. Practically, it can provide a targeted relative poverty management plan and path reference for related research.

2. EMPIRICAL RESEARCH

With China Family Panel Studies of 2018 (2018CFPS) data [1], this paper takes migrant workers' families as the research object, and after screening out sample data such as missing and inapplicable key index data, 2001 sample data of families were finally retained.

2.1 Index System Setting

In terms of the construction of the relative poverty indicator system, it is necessary to standardize the multi-dimensional relative poverty indicator system. In addition to the income poverty line, the identification standards should also include multi-dimensional indicators such as education, employment, and health. This is the continuation of the policy of "two no worries and three guarantees". It is also a practical response to the people's growing needs for a better life. Therefore, in terms of indicator selection, this article combines the three dimensions of education, health and living standards in the MPI announced by UNDP, and increases the two dimensions of income and employment according to the actual situation in China, and adjusts the indicators and thresholds of each dimension. The determination of weights is an important part of the A-F method. At present, there is no unified weight determination standard in the research of the multi-dimensional relative poverty index system. The equal-weight method (that is, the weight of each dimension is equal, and the weight of each indicator within the same dimension is equal) is relatively simple and easy to implement and is widely used (Datt G, 2019) [7], so this method is also adopted in this article. The specific index system and weight design are shown in "Table 1" below.

1. The data used in this paper are from the results of the China Family Panel Studies (CFPS) conducted by the Institute of Social Science Survey, (ISSS) of Peking University in 2018. The indicators cover a number of research topics, including economic behavior, educational access, family relationships and family dynamics, population migration, and physical and mental health

Table 1. Multi-dimensional relative poverty identification index system of urban households

Dimensions	Weight of dimensions	Indicators	Indicator interpretation and assignment	Weight of indicators
Income	1/5	Family income per capita	If family income per capita is less than 3535 yuan, assign a value of 1, otherwise it is 0;	1/10
		Proportion of transfer income	If the proportion of transfer income is over 60%, assign a value of 1, otherwise it is 0;	1/10
Education	1/5	Education level	If the highest level of education is or under middle school, assign a value of 1, otherwise it is 0;	1/10
		Children's school enrollment	If there are School-age children drop out, assign a value of 1, otherwise it is 0;	1/10
Employment	1/5	Work status	If there is labor force could get a job, assign a value of 1, otherwise it is 0;	1/15
		Purchase situation of five social insurances and one housing fund	If there is one employee who do not get the insurances and housing fund in accordance with the regulations, assign a value of 1, otherwise it is 0;	1/15
		Participation in endowment insurance	If there is a retiree without endowment insurance, assign a value of 1, otherwise it is 0;	1/15
Health	1/5	Health	If there is a family member who is labeled as "unhealthy" or has chronic disease in the past six months, assign a value of 1, otherwise it is 0;	1/10
		Participation in medical insurance	If there is a family member without any medical insurance, assign a value of 1, otherwise it is 0;	1/10
		Drinking water	If the drinking water for the family is not clean water, assign a value of 1, otherwise it is 0;	1/20
		Cooking fuel	If the cooking fuel for the family is not clean fuel, assign a value of 1, otherwise it is 0;	1/20
		House	If the family do not have ownership of any house or apartment, assign a value of 1, otherwise it is 0;	1/20
		Living standard	1/5	1/20

2.2 Measuring the Incidence of Relative Poverty

Based on the weight of the indicators selected in this paper, the relative poverty situation of the sample is estimated. The incidence of one-dimensional poverty of each indicator is shown in "Table 2". The multidimensional poverty conditions under different relative poverty thresholds K are shown in "Table 3".

Table 2. Poverty incidence rate of various indicators

Indicators	Incidence of poverty
Family income per capita	0.010
Proportion of transfer income	0.022
Education level	1.000
Children's school enrollment	0.628
Work status	0.366
Purchase situation of five social insurances and one housing fund	0.864
Participation in endowment insurance	0.486
Health	0.377
Participation in medical insurance	0.160
Drinking water	0.128
Indicators	Incidence of poverty
Cooking fuel	0.071
House	0.209
Daily traffic	0.422

There are only two poverty rates below 5%, which are in the income dimension, among them, the lowest incidence of average per capita income poverty families, only 1.0%, followed by metastatic income proportion is 2.2%. It can be seen that the migrant worker group does not appear to be in

poverty in terms of income. Although their income level is low, it is slightly above the absolute poverty line, which means they belong to the edge of the poverty population. Therefore, in the previous process of targeted poverty alleviation, it has not received preferential policies and widespread attention. There is an indicator, education level, that the incidence of poverty is above 90%, which has reached 100%. It reflects that the migrant workers are generally not well educated, and the low level of human capital makes the migrant workers have to engage in manual labor with low added value, so there is no channel for them to get rich through labor. In addition, the incidence of poverty in the purchase of five social insurances and one housing fund is as high as 86.4%. The labor form of migrant workers is inherently more dangerous. The lack of insurance rights greatly increases the risk of falling into poverty. It also shows that the current social security for labor needs to be strengthened.

Among the remaining indicators, in terms of health, there is a certain gap between the health status of the family and the poverty incidence of the medical insurance participation index. The health status is 37.7%, and the medical insurance is 16.0%. The majority of migrant workers are engaged in jobs with high physical strength, which brings many risks to their health. Therefore, health problems are still an important factor causing family poverty, and the medical burden significantly increases the risk of family poverty. But the universality of medical insurance is helpful. With the promotion and help of medical insurance, many families have the awareness of prevention and purchase medical insurance in advance, which reduces the poverty problem caused by the disease to a certain extent. In terms of living standards, the incidence of poverty in drinking water is 12.8%, which is more than 10%, reflecting the need to improve the living environment of migrant workers. The incidence of poverty in the housing situation is not high, at 20.9%. The main reason is that although migrant workers only have temporary housing in cities, they own their own houses in rural areas, which to some extent provides a guarantee for them to resist the risk of poverty.

Table 3. Multi-dimensional relative poverty measurement results

k	H	A	MPI
0.1	1.000	0.376	0.376
0.2	0.938	0.391	0.367
0.3	0.741	0.429	0.318
0.4	0.413	0.496	0.205
0.5	0.180	0.566	0.102
0.6	0.058	0.643	0.037
0.7	0.012	0.718	0.009
0.8	0.001	0.800	0.001

①H: The proportion of poor people in the total sample

②A: The proportion of the average deprivation dimension of multi-dimensional relative poor families in the total number of dimensions

③MPI: Equals to H×A, is a comprehensive indicator of poverty

It can be seen that as the multidimensional relative poverty threshold k increases, the incidence of multidimensional relative poverty (H) and the multidimensional relative poverty index (MPI) show a downward trend. On the other hand, the average deprivation share (A) is on the rise. When k=0.1, because the sample group shows poverty in the indicator of the highest level of education, the multi-dimensional relative poverty incidence rate of the sample family at this time is 100%; when k=0.2, the multi-dimensional relative poverty incidence rate (H) When k=0.3, the multi-dimensional relative poverty incidence (H) is 74.1%, which is 25.9% lower than the multi-dimensional relative poverty incidence (H) when k=0.1. The multi-dimensional relative poverty index (MPI) dropped to 0.318, and the average deprivation share (A) rose to 0.429; when k=0.4, the multi-dimensional relative poverty incidence (H) dropped to 41.3%, which was a drop of 32.8%, and a significant drop occurred. The multidimensional relative poverty index (MPI) dropped to 0.205, and the average deprivation share (A) rose to 0.496; when k=0.5, the multidimensional relative poverty incidence (H) continued to drop to 18%, and the average deprivation share (A) increased to 0.566. With the continuous increase of the K value, when k=0.6, k=0.7, and k=0.8, the multi-dimensional relative poverty incidence (H) at this time is 5.8%, 1.2%, and 0.1%, respectively, which all drop below 10%, That is, only a small number of families fall into multi-dimensional relative poverty under this critical value. Compared with other multi-dimensional relative poor families, these families have more serious poverty, and their multi-dimensional relative poverty is deeper.

According to the Human Development Report, multidimensional poverty is defined as when a person is deprived of at least 30% of the weighted indicators. Based on this, this study believes that when the total deprivation score of the sample k ≥0.3, the family is considered as a multidimensional relative poverty family, and the multidimensional relative poverty incidence (H) at this time is 74.1%, that is, 1,483

families are multidimensional relative poverty and 518 families are non-multidimensional relative poverty.

2.3 Analysis with Random Forest Model

Random Forest is a combined supervised learning method. In the random forest, multiple prediction models are generated at the same time, and then the results of the models are summarized to improve the accuracy of the models and thus the classification accuracy. The "double boundary method" has been used to decomposition the dimensions of the multi-dimensional relative poverty. Then the machine learning algorithm — random forest will be used to analyze and rank the characteristic importance of the multi-dimensional relative poverty indicators and analyze the contribution rate of the indicators.

The data set is divided into two parts at a ratio of 3:1, which are respectively used as the training data set and the test data set. The test set data has a total of 501 families, including 360 poor families and 141 non-poor families. Among the 360 poor families, 6 families are predicted to be non-poor families, and 354 families are predicted to be poor families; among the 141 non-poor families, 141 families are predicted to be non-poor families, and 0 families are predicted to be Poor families. From this, the overall correct rate is calculated: (141+354)/501=98.8%. The results show that the model can be used for accurate estimation.

The random forest can measure the importance of variables, which is obtained and output by the importance function. The impurity of the node is defined by the Gini coefficient. The specific values are shown in "Table 4".

Table 4. Model feature importance

Feature	Mean Decrease Gini
1) Children's school enrollment	0.207
2) Participation in endowment insurance	0.173
3) Health	0.167
4) Daily traffic	0.098
5) Work status	0.093
6) Purchase situation of five social insurances and one housing fund	0.078
7) Participation in medical insurance	0.066
8) Drinking water	0.046
9) House	0.046
10) Cooking fuel	0.017
11) Proportion of transfer income	0.007
12) Family income per capita	0.002
13) Education level	0.000

This parameter reflects the importance of these indicators in the classification process. Indexes with a higher degree of importance can be regarded as the objects that should be focused on in the practice of managing relative poverty. Among all the features, children's school enrollment is the most important feature in the classification process, with an importance weight of 20.7%, followed by pension insurance, with an importance weight of 17.3%. Summarize according to its dimension division, and get the order of dimension importance, as shown in "Table 5". The results show that the total importance of the employment dimension is the highest at 34.4%, followed by the health dimension (23.3%), the living standard dimension (20.7%), the education dimension (20.7%), and the final income dimension (0.9%). The above analysis shows that, when analyzing and identifying the relatively poor families among migrant workers, we should pay attention to the schooling status of the children in the family, and at the dimension level, we should focus on the performance of the employment dimension of the family.

Table 5. Dimension importance

Dimension	Mean Decrease Gini
Employment	0.344
Health	0.233

Living standard	0.207
Education	0.207
Income	0.009

3. CONCLUSION

3.1 Conclusions and Recommendations

It is not difficult to draw the following conclusions by making multi-dimensional identification and analysis of the relative poverty of migrant workers with the double-boundary method.

Firstly, according to international practice, when the weight index of 30% is deprived, it is regarded as the standard of multidimensional poverty. The multidimensional relative poverty index is 0.318, that is, 31.8% of migrant worker families are relatively poor under this standard. However, it's worth noting that this index is much bigger than the 3.9% incidence of tridimensional poverty in China measured by Alkire (2019) [8]. This is mainly ascribed to two aspects. On the one hand, the dimensional system is expanded in this paper. On the other hand, the results indicated that the relative poverty of migrant workers is far higher than the average level. In general, migrant workers have a large population and strong mobility, and their income levels are usually on the verge of poverty. Therefore, when establishing a long-term mechanism for relative poverty governance, research should pay more attention to migrant workers.

Secondly, from the perspective of the single-dimensional poverty incidence of various indicators, the indicators with higher poverty incidence are the education level and the enrollment of children in the education dimension. The second is the purchase of five social insurances and one housing fund in the work dimension. From the perspective of education, the level of education of migrant workers' families is generally low. This has no obvious restriction when migrant workers are engaged in agricultural work. However, when agricultural income cannot meet family needs and farmers choose to work in cities, the level of education becomes an obvious disadvantage. It forces them to choose low-value-added manual labor, and often high-intensity labor-intensive work, which limits their opportunities for education and training and thus cannot achieve human capital accumulation through learning, so there is room for income improvement. At the same time, with the increase of age and decline in physical strength, the income level of migrant workers may further decline, so their poverty risk will continue to increase. On the other side of the education dimension, the incidence of one-dimensional poverty in the enrollment of migrant children in this sample is as high as 62.8%, which reflects the plight of children's education in migrant families and the risk of intergenerational transmission of poverty. When the children of migrant workers cannot complete the accumulation of human capital, they can easily fall into the same predicament as the previous generation. Therefore, in terms of education, on the one hand, it is necessary to increase government investment in adult vocational training and encourage social organizations to participate and provide subsidies for relatively poor marginal populations such as migrant workers to make up for the decline in income caused by their reduced working hours and participation in training. On the other hand, in cities, it is necessary to continue to provide convenience and policy support for children of migrant workers, and at the same time increase investment in teachers and social workers in township schools to support the education of "left-behind children".

From the perspective of employment, the one-dimensional poverty incidence rate for the purchase of five social insurances and one housing fund is 0.864. Five insurances and one housing fund are a kind of welfare benefits for employees, involving the legal rights and interests of the labor force such as a pension, housing, illness, and casualties. For the group of migrant workers engaged in high-intensity manual labor, the lack of insurance means a sharp increase in the risk of poverty. Migrant workers are facing an unstable job market and troubled by illegal employment and even malicious delays in wages. While vigorously advancing supervision measures, local governments must actively explore and improve the supervision mechanism to protect the labor market for migrant workers and establish a law popularization agency for migrant workers to encourage migrant workers to protect their legitimate rights and interests.

Thirdly, the proportions of per capita annual household income and transfer income under the income dimension, are 1% and 2.2% respectively. Due to the particularity of the migrant worker group, this article selects the rural poverty standard to measure. However, the daily living cost of migrant workers is benchmarked against urban residents. While working in cities, migrant workers face the situation that their income and living standards are generally lower than urban residents, so their sense of deprivation and poverty is far more than farmer groups that have the same income in the countryside. In addition, the low incidence of poverty as a

proportion of transfer income also reflects a problem. That is migrant workers have a low-income level but are slightly above the poverty line. This situation has excluded them from the scope of the previous poverty alleviation policies. This also shows that in the previous process of absolute poverty governance, the poverty identification method that mainly relied on income levels to delineate the poverty line will lag behind the multi-dimensional needs of relative poverty governance. With the dynamic adjustment of the relative poverty standard, a large number of migrant workers will be included in the category of relative poverty, and the particularity of migrant workers' strong mobility and unstable income has brought challenges to the establishment of the poverty reduction mechanism, that's why urgently needs to be targeted the design of the scheme. In the health dimension, the incidence of poverty in terms of health status is relatively high, but the incidence of poverty in medical insurance participation is low, indicating that with China's vigorous efforts to promote medical insurance for urban and rural residents and the new rural cooperative medical system, the basic medical care of migrant workers has been guaranteed to a considerable extent.

Finally, the random forest algorithm is used to establish the relative poverty identification model of the migrant worker group, and the prediction accuracy of the test set is 98.8%, which reflects the applicability of the random forest algorithm. This article gives the importance of indicators in the model classification process, that is, those indicators can effectively distinguish between non-relatively poor families and relatively poor families of migrant workers. The results show that children's school enrollment is the most important feature. The logic is that in a country like China that values children's education, only families with really difficult conditions will give up supporting children's schooling. At the dimension level, the focus is on the employment dimension. To have a stable job and to obtain employment security such as five social insurance and one housing fund has also become an important direction for investigating the risk of a family falling into relative poverty. With the development of information technology and the process of digitalization and electronification of social work, big data algorithms such as random forest will play an increasingly important role in public management affairs such as the identification of relative poverty. At the theoretical level, interdisciplinary research should be carried out and effective analytical tools and poverty reduction programs should be actively explored. In practical social public management, it is necessary to further promote the coverage of information collection and input to migrant workers. This will, on the one hand, promote the standardization of migrant workers' employment market, and on the other hand, facilitate community management and big data analysis.

3.2 Research Prospect

The eradication of absolute poverty has been historically achieved in China, then the governance of relative poverty is on the agenda. Thereinto, the problem of relative poverty bears the brunt, and its governance difficulties are particularly prominent. This paper employed the double-boundary method to identify and analyze the relative poverty families of migrant workers, finds and summarizes the law of relative poverty of migrant workers, and puts forward corresponding suggestions. Moreover, the random forest algorithm is used to construct the recognition model, with a high accuracy rate, which verifies the applicability of the big data algorithm in the research of poverty problems. Nevertheless, this article still has two shortcomings. For one thing, according to the continuous changes in poverty patterns, the design of the indicator system should be improved following the principle of "a reasonable reflection of the moderate difference between the relatively poor group and the social average level", which can better identify the relatively poor and guide the practice of poverty governance. In addition, the flexibility of the random forest algorithm provides convenience for further expansion of the index system, and subsequent research will further expand its application scenarios. For another thing, the relative poverty of migrant workers should be solved in combination with rural revitalization and urbanization. It's noteworthy that increasing public investment in rural areas can reduce the gap between urban and rural areas, and policy support and institutional space should be provided for migrant workers when they return to hometowns for employment or entrepreneurship. These problems may become the focus of research in the future.

AUTHORS' CONTRIBUTIONS

Lilu Sun is responsible for the design of article frame and writing of introduction. Xin Ma writes the empirical research analysis and conclusion. Yuxin Ding processes the data.

REFERENCES

[1] Townsend Peter. Poverty in the United Kingdom: A Survey of Household Resources and Standards of Living [M]. University of California Press: 1979-12-31.

[2] Cao Yongping. Choice of Paths to Solve Relative Poverty in the Post-Poverty Era [J]. Gansu Agriculture, 2020, 520(10): 38-40.

[3] Seth. S, M. E. Santos, Multidimensional Inequality and Human Development [C]. OPHI Working Paper, No. 114, 2018

[4] Bourquin P, J Cribb, T Waters, and X Xu. Living Standards, Poverty and Inequality in the Uk: 2019 [J]. Institute for Fiscal Studies, 2019.

[5] Alkire S, Foster J. Counting and multidimensional poverty measurement [J]. Journal of Public Economics, 2007, 95(7-8):476-487.

[6] Robert Joyce, James P. Ziliak. Relative Poverty in Great Britain and the United States, 1979–2017 [J]. Fiscal Studies. 2019, 40(4):485-518.

[7] Datt G. Multidimensional poverty in the Philippines, 2004–2013: How much do choices for weighting, identification and aggregation matter? [J]. Empirical Economics. 2019;57(4):1103-1128.

[8] Alkire S, et al., Global Multidimensional Poverty Index 2019: Illuminating Inequalities [R]. New York: UNDP and OPHI,2019.