Journal of Autonomous Intelligence (2024) Volume 7 Issue 1
doi: 10.32629/jai.v7i1.900

ORIGINAL RESEARCH ARTICLE

A pricing model for agricultural insurance based on big data and machine learning

Yu Wang, Muhammad Asraf bin Abdullah*

Faculty of Economics and Business, Universiti Malaysia Sarawak (UNIMAS), Kota Samarahan Sarawak 94300, Malaysia

*** Corresponding author:** Muhammad Asraf bin Abdullah, amasraf@unimas.my

ABSTRACT

Agricultural insurance is a crucial element of policies that promote and protect agriculture. It protects agriculture from risk and distributes agricultural hazards. The rural economy's stabilization has been a significant stabilizer function. But as agriculture insurance has quickly advanced, a number of issues have unavoidably come to light. Agricultural insurance still offers a wide range of products and services available today. Big data will play a significant supporting role in the pressing need to innovate and improve goods and services. Other information supporting agricultural insurance includes agricultural data connected to it. The two previously most often utilized agricultural index insurances are regional yield insurance and weather index insurance. They struggle with risk pricing mostly due to a lack of appropriate empirical data, complicated dependence linkages between various hazards, and the prevalence of basis risk. A comprehensive study and review of pertinent research findings are carried out by modelling regional yield risk, building weather indicators and their distribution fitting, modelling agricultural dependence risk, and measuring and reducing basis risk. This article highlights the flaws in the current pricing models as well as the problems that need to be addressed in future studies. The need to further develop agricultural index insurance's risk modelling techniques and increase the objectivity and precision of the pricing outcomes cannot be overstated in terms of their practical importance.

Keywords: agricultural insurance; basis risk; big data; machine learning

ARTICLE INFO

Received: 5 July 2023
Accepted: 22 August 2023
Available online: 14 November 2023

1. Introduction

In recent years, with the deepening of big data technology, a large number of commercialized big data companies in agricultural related fields have emerged in China[1]. These big data companies, relying on their own advantages, not only carry out applications and services in their respective fields, but also collect big data in related fields. These enterprise big data are mostly distributed among various companies and serve the main business of the enterprise, with few integrated applications with agricultural insurance[2]. Agricultural insurance is the "stabilizer" and "safety valve" of agricultural production and operation, and artificial intelligence, as the core technology of insurance technology, is changing the business model of agricultural insurance companies[3]. First, determine the internal and external drivers of agriculture insurance businesses' digital operations and the rationale behind their shift to an AI approach; Second, examine the route of artificial intelligence enabling the whole business process of agricultural insurance firms by dividing the front-end, mid-range, and back-end business processes of these organizations[4]; examining the difficulties that arise while implementing technology, changing

business strategy, and processing data in the context of the agriculture insurance business system; The article concludes by proposing remedies for AI to continually support agricultural insurance operation, including optimizing the rational design of AI strategy[5], enhancing data governance competence, concentrating on AI talent development, and technology innovation. The core of scientific pricing in agricultural insurance is the achievement of equivalence between insurance premiums and insurance liabilities, and the key to effectively evaluating agricultural production risks is the scientific determination of premiums[6]. The techniques for assessing agricultural production risk may be divided into three groups depending on the various risk components: methods based on risk factors, methods based on risk mechanisms, and methods based on risk losses. In the insurance sector, the loss-based evaluation approach is presently employed extensively[7].

2. The application of big data and machine learning in insurance risk management

In fact, traditional risk management methods are no longer effective in analyzing, preventing, and supervising various risks faced by insurance companies, including insurance risks. New cognitive technologies such as big data, machine learning and natural language processing are replacing traditional analysis methods to quantitatively analyze and process the increasingly large data sets generated in the insurance market[8]. Ultimately, it helps to identify various risk indicators and achieve more effective risk management. With the development of modern technology, especially the emergence of internet technology, massive data on human consumption, entertainment, credit and other behaviors has also emerged[9]. At this point, traditional software tools cannot obtain, manage, and analyze data sets of this size within a certain time frame. New processing ideas and technological advancements are therefore required. The insurance industry has a lot of information about policyholders, and several types of information about policyholders may be obtained online. using identification data, interpersonal connections, consumer behavior, credit ratings, etc., and assessing policyholders' credit based on big data technologies. A massive data modelling and analysis technology called machine learning is used to forecast corresponding insurance risks. By continuously choosing data, establishing model data, establishing model data, validating data, readjusting models, etc., it extracts usable data from a vast quantity of data and regularly modifies training samples[10]. In an effort to provide the best result, look for internal patterns and patterns in the data itself. Machine learning includes supervised learning, unsupervised learning, semi supervised learning, deep learning, etc. Different learning method systems correspond to different algorithms. It includes BP neural network, support vector machine, random forest, clustering analysis algorithm, etc. Different algorithms may provide inconsistent prediction results for different application methods, datasets, and prediction targets.

3. Construction of a system for evaluating the performance of agricultural insurance subsidies

3.1. Application framework and evaluation system

The application framework of big data analysis not only needs to master the research content of the previous evaluation system, but also needs to conduct systematic research on financial business. The evaluation system business framework contains core data information, which can serve as a key part of the application framework[11,12]. According to the characteristics of different stages of the big data lifecycle, the collection data, analysis indicators, and calculation models of evaluation indicators are stored in corresponding positions[13]. Transform the three forms using the platform domain to meet the research needs. Evaluate various businesses through big data analysis application domains and interact with existing "Jincai" business application systems. **Figure 1** shows a big data analysis application framework.

Based on the research on the current development status of big data in agriculture, insurance, and finance, it is concluded that the performance evaluation big data of agricultural insurance subsidies collected and

organized through diversified channels plays a role in data management. Achieving the integrity and scientificity of the big data field plays an important role in improving the performance rating system[14,15]. Data sources mainly rely on agricultural production statistics and monitoring data, agricultural insurance and risk monitoring data, financial data, etc. Agricultural production statistics and monitoring are an important foundation for big data analysis, and also serve as a guiding basis for various policies and systems of agricultural insurance subsidies. It plays an important role in agricultural production and economic development in rural areas. Agricultural insurance and risk monitoring are important supplements to big data evaluation, and agricultural insurance plays a protective role in the healthy development of agriculture. Risk monitoring is the basis for evaluating the scientific nature of agricultural insurance policies, and bears the important responsibility of preventing agricultural production risks and improving risk control measures. Financial data is an integral part of the evaluation system, recording information on financial subsidies and fund allocation[16,17]. With the support of the above system, effective implementation of fiscal policies can be achieved. The scientific allocation of subsidy funds plays an important role in realizing the value of the big data domain.

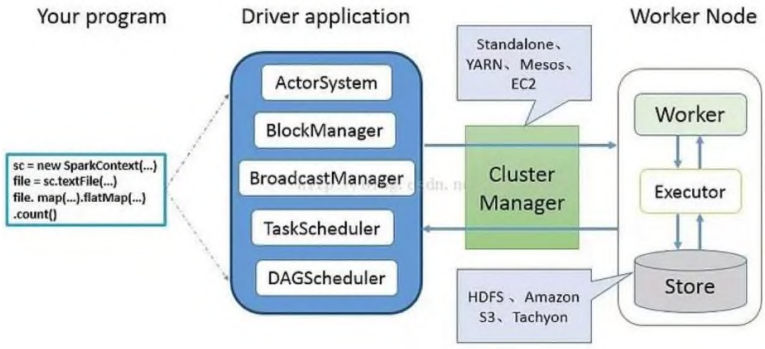

Figure 1. Big data analysis application framework.

3.2. Static panel data model for machine learning algorithm based on machine learning algorithm

Panel data refers to the data that takes multiple sections on the time series, that is, the data that combines the time series data and section data. The difference between panel data, time series data and section data is that the variables of panel data have double subscripts. The general expression is as follows:

$$y_{it} = \alpha + X_{it}^{'}\kappa + \varepsilon_{it}, i = 1,2,\cdots,N; t = 1,2,\cdots,T \tag{1}$$

where i represents an individual, t represents time, x is a constant, K is a constant vector of K × 1 order, and X_{it} is the i-th observation value. Most panel data applications use one-way error component models, namely:

$$\varepsilon_{it} = \mu_i + v_{it} \tag{2}$$

among μ_i represents unobservable individual effects, and vit is a random disturbance term. Equation (1) can be written in the following vector form:

$$y = \alpha t_{NT} + X\kappa + \varepsilon = Z\delta + \varepsilon \tag{3}$$

where y is a vector of NT × 1 order, X is a matrix of NT × K order, Z = [Lnt, X], δ' = (at', K'), IvT is a vector of NT order whose elements are all 1. Equation (2) can be written in the following vector form:

$$\varepsilon = Z_{\mu}\mu + v \tag{4}$$

of which

$$u^{'} = (u_{11},\cdots,u_{1T},u_{21},\cdots,u_{2T},\cdots,u_{N1},\cdots,u_{NT}) \tag{5}$$

In the fixed effect model, μ_i is a fixed parameter, v_{it} is a random perturbation term, and v_{it} is independent

and identically distributed, For all i and t, X_{it} and v_{it} are independent. Substitute Equation (4) into Equation (3) to get:

$$y = \alpha v_{NT} + X\kappa + Z_\mu \mu + v = Z\delta + Z_\mu \mu + v \tag{6}$$

For Equation (6), the least squares (OLS) method can be used to estimate a, β and μ. Z is $NT \times (K+1)$ order matrix, z is $(NT \times N)$ order individual dummy variable matrix. When N is large, Equation (6) contains too many dummy variables. Because the matrix dimension of $(N+K)$ dimension is too large, its inverse matrix is difficult to solve, so using OLS will lead to large deviation. At this time, just use Q to multiply Equation (6) left and then apply OLS to the converted model to obtain the least squares dummy variable (LSDV) estimate of the parameter. The converted model is:

$$Qy = \alpha Q t_{NT} + QX\kappa + QZ_\mu \mu + Qv \tag{7}$$

Matrix Q eliminates the individual effect. At this time, let:

$$\tilde{y} = Qy, \tilde{X} = QX \tag{8}$$

The OLS estimator of Equation (7) is:

$$\tilde{\kappa} = \left(X'QX\right)^{-1} X'Qy \tag{9}$$

and

$$var(\tilde{\kappa}) = \sigma_v^2 \left(X'QX\right)^{-1} = \sigma_v^2 \left(\tilde{X}'\tilde{X}\right)^{-1} \tag{10}$$

For regression model:

$$y_{it} = \alpha + \kappa x_{it} + \mu_i + v_{it} \tag{11}$$

Calculate the mean value of all observations according to Equation (11):

$$\bar{y}_{..} = \alpha + \kappa \bar{x} + \bar{v}_{...} \tag{12}$$

F test was used to test the significance of fixed effects, and the original hypothesis was used. OLS is used to regress the mixed model to obtain the constrained residual sum of squares (RRSS), and LSDV regression is used to obtain the unconstrained residual sum of squares (URSS). When N is large, the sum of squares of residuals can be used as URSS by means of intra group mean conversion. At this time, the test statistic is:

$$F_0 = \frac{(RRSS - URSS)/(N-1)}{URSS/(NT - N - K)} \sim F_{(N-1),N(T-1)-K} \tag{13}$$

4. Results and discussion

4.1. Agricultural index insurance pricing model

The actuarial pricing method, which builds an actuarial model for a specific type of insurance product, is the most popular approach for creating insurance products. Calculate the insurance product's premium amount using the actuarial assumptions that have been established. Modelling the relationship between weather index and crop production per unit area is crucial for the development of agricultural index insurance. primarily by using a linear regression model to describe the relationship between the yield and the weather index. On the basis of this, a piecewise linear compensation function for the weather index is built. The greatest high-quality japonica rice production region in China is Heilongjiang Province, which serves as the "ballast stone" of the country's food security. The yield changes of single season rice in Heilongjiang Province are related to national food security. Studying the determination of rice index insurance premium rates in Heilongjiang Province can provide certain reference ideas for the further promotion and application of weather index insurance products.

Taking single season rice in Heilongjiang Province as an example, pure rate pricing is used for single season rice weather index insurance. Based on the complete data that can be collected, this article has compiled the yield data and related weather indicator data of single season rice in Heilongjiang Province from 1999 to 2019, all of which are sourced from the "Heilongjiang Statistical Yearbook" and "China Statistical Yearbook". Single season rice in Heilongjiang Province is harvested once a year, with a growth cycle of five months from

April to August each year. Six meteorological parameters, including temperature, precipitation, sunshine, humidity, wind speed, and light temperature ratio, have a major influence on the growth of single-season rice. The yield of rice grown in a single season is significantly influenced by the growth of crops during the first three months. Since the efficacy of rice tillering is directly correlated with the number of ears per unit area, different weather elements at different growth stages have varying effects on crop output, such as the large influence of precipitation on rice during the tillering stage. In order to create a total of 18 weather elements, the six dimensional weather factors are further divided into several months. This article uses the relaxed Lasso dimensionality reduction model in machine learning to screen weather factors that are correlated with yield.

By constructing a relaxed Lasso in machine learning to screen the correlation between weather factors and single cropping rice yield, a sparse model is generated that only involves a subset of the original variable set. The Relaxed Lasso algorithm identifies a set of non-zero parameters by running Lasso, and then fits an unconstrained linear model to the selected feature set. The optimal selection of penalty parameters can be obtained through cross validation methods.

Using a stepwise regression model, select the weather factors that have the most significant impact. The stepwise regression analysis method combines the characteristics of one by one introduction method and one by one elimination method, and is one of the extended methods of multiple linear regression analysis. The basic idea is to perform regression analysis on all explanatory variables and the dependent variable, then hypothesis test the sum of squares of partial regression, and eliminate one explanatory variable at a time that has the smallest sum of squares of partial regression and is not significantly correlated with the dependent variable. Then, on this basis, regression analysis and partial regression sum of squares test are repeated until all independent variables included in the model have a significant impact on the dependent variable.

4.2. Result discussion

The overall fitting effect of the model is good, and at the 95% significance level, the June average light temperature ratio index (G6) has the greatest impact. Therefore, this article selects this weather factor as the insurable risk of the weather index, and takes the average value of the light temperature ratio index in June during the sample period as the trigger value, resulting in a trigger value of 10.501. According to the Lasso model estimation, the average light temperature ratio parameter in June was estimated to be negative, indicating that the larger the light temperature ratio index in June, the greater the impact on yield reduction. The historical average light temperature ratio data for June fluctuates significantly, which is an insurable risk. Therefore, it is determined to price the product at a pure rate for the June light temperature ratio index insurance of Heilongjiang single crop rice.

In countries and regions where the development of index insurance has been relatively successful, the application of actuarial pricing method is the most widespread. Especially the classic combustion analysis method, which can perform relatively simple calculations based on limited historical data and has a lower cost, is suitable for promotion in developing countries. This article takes the tariff determination of rice index insurance in Heilongjiang Province, China as an example, and selects 18 weather factor covariates from six dimensions of climate factors: precipitation, temperature, sunlight, humidity, wind speed, and light temperature ratio index. The Relaxed Lasso algorithm is utilized to diminish the dimensionality of the multivariate framework, identifying five weather factors with significant correlations from a pool of high-dimensional variables. Subsequently, a stepwise regression evaluation is executed. The findings highlight that the light-temperature ratio for June possesses a notable correlation with the yield of single season rice per unit. By employing the traditional combustion model, we can efficiently and swiftly determine the pricing for associated index insurance products. The outcome indicates that in Heilongjiang Province, the net premium for the insurance product related to the light temperature ratio index of single season rice in June stands at 786.376 yuan/ha.

Weather monitoring is fundamental for implementing weather index insurance. It's vital to meticulously discern the relationship between the weather index and per unit yield loss by evaluating historical meteorological and agricultural yield data. Establishing credible compensation benchmarks and product rates remains at the heart of weather index insurance pricing. A conventional meteorological observation post can encompass several square kilometers of risk-prone areas. However, by this criterion, the existing station count in China is insufficient for the prevailing demands. While China's meteorological sector is fundamentally supported by governmental finances, it also has the capacity to offer specialized, fee-based services, catering to the pricing requisites of insurance entities. Offering granular data to insurance agencies not only aids in setting apt rates but also fosters enhancements in meteorological station observational capabilities. This ensures the delivery of more precise, rapid, and encompassing meteorological data, bolstering the forecasting and alert mechanisms for extreme weather phenomena, thus instituting a beneficial feedback loop.

5. Conclusions

For most farmers, agricultural index insurance is a new type of thing. In addition, most farmers themselves have a lack of trust in insurance products. To address this issue, the government should promote knowledge related to weather index insurance among farmers and introduce the advantages of this new type of insurance compared to traditional agricultural insurance. And enhance farmers' trust in insurance companies and encourage them to actively participate in the pilot work of agricultural weather index insurance. Given the limited insurance knowledge and level of farmers, if the acceptance of the insurance product cannot be significantly improved, the government should further adopt a combination of voluntary and mandatory measures. In addition, government data and technical support are also extremely important factors. The immature and complete weather index database is a weakness in the development of agricultural weather index insurance in China. The government should establish and improve the climate database for agricultural production, such as strengthening the construction of weather data infrastructure, establishing an independent third-party agency to monitor and publish meteorological data, and regularly adjust the index value according to the actual situation. At the same time, the government should carry out the construction of laws and regulations related to weather index insurance, standardize the pilot process, provide legal support for weather index insurance pilot projects, and provide intellectual property protection mechanisms for insurance product design, in order to promote the exploration and practice of weather index insurance.

Author contributions

Conceptualization, MAbA; methodology, YW; write original drafts, organize data, MAbA and YW. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. Benami E, Jin Z, Carter MR, et al. Uniting remote sensing, crop modelling and economics for agricultural risk management. *Nature Reviews Earth & Environment* 2021; 2(2): 140–159. doi: 10.1038/s43017-020-00122-y
2. Hill RV, Kumar N, Magnan N, et al. Ex ante and ex post effects of hybrid index insurance in Bangladesh. *Journal of Development Economics* 2019; 136: 1–17. doi: 10.1016/j.jdeveco.2018.09.003
3. Smith VH, Watts M. Index based agricultural insurance in developing countries: Feasibility, scalability and sustainability. *Gates Open Research* 2019; 3(65): 65. doi: 10.21955/GATESOPENRES.1114971.1
4. Budhathoki NK, Lassa JA, Pun S, Zander KK. Farmers' interest and willingness-to-pay for index-based crop insurance in the lowlands of Nepal. *Land Use Policy* 2019; 85: 1–10. doi: 10.1016/j.landusepol.2019.03.029
5. Takahashi K, Barrett CB, Ikegami M. Does index insurance crowd in or crowd out informal risk sharing? Evidence from rural Ethiopia. *American Journal of Agricultural Economics* 2019; 101(3): 672–691.
6. de Janvry A, Sadoulet E. Using agriculture for development: Supply- and demand-side approaches. *World*

Development 2020; 133: 105003. doi: 10.1016/j.worlddev.2020.105003
7. Cariappa AGA, Acharya KK, Adhav CA, et al. Impact of COVID-19 on the Indian agricultural system: A 10-point strategy for post-pandemic recovery. *Outlook on Agriculture* 2021; 50(1): 26–33. doi: 10.1177/0030727021989060
8. Ceballos F, Kramer B, Robles M. The feasibility of picture-based insurance (PBI): Smartphone pictures for affordable crop insurance. *Development Engineering* 2019; 4: 100042. doi: 10.1016/j.deveng.2019.100042
9. Jung J, Maeda M, Chang A, et al. The potential of remote sensing and artificial intelligence as tools to improve the resilience of agriculture production systems. *Current Opinion in Biotechnology* 2021; 70: 15–22. doi: 10.1016/j.copbio.2020.09.003
10. Möhring N, Dalhaus T, Enjolras G, et al. Crop insurance and pesticide use in European agriculture. *Agricultural Systems* 2020; 184: 102902. doi: 10.1016/j.agsy.2020.102902
11. Wiener M, Saunders C, Marabelli M. Big-data business models: A critical literature review and multiperspective research framework. *Journal of Information Technology* 2020; 35(1): 66–91. doi: 10.1177/0268396219896811
12. Roetzel PG. Information overload in the information age: A review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Business Research* 2019; 12(2): 479–522. doi: 10.1007/s40685-018-0069-z
13. Grover V, Chiang RHL, Liang TP, Zhang D. Creating strategic business value from big data analytics: A research framework. *Journal of Management Information Systems* 2018; 35(2): 388–423. doi: 10.1080/07421222.2018.1451951
14. Lim C, Kim KH, Kim MJ, et al. From data to value: A nine-factor framework for data-based value creation in information-intensive services. *International Journal of Information Management* 2018; 39: 121–135. doi: 10.1016/j.ijinfomgt.2017.12.007
15. Niño HAC, Niño JPC, Ortega RM. Business intelligence governance framework in a university: Universidad de la costa case study. *International Journal of Information Management* 2020; 50: 405–412. doi: 10.1016/j.ijinfomgt.2018.11.012
16. Jean RJ, Sinkovics RR, Kim D. Information technology and organizational performance within international business to business relationships: A review and an integrated conceptual framework. *International Marketing Review* 2008; 25(5): 563–583. doi: 10.1108/02651330810904099
17. Rahardja U. Using Highchart to implement business intelligence on attendance assessment system based on YII framework. *International Transactions on Education Technology* 2022; 1(1): 19–28.