Editorial: Prediction, pattern recognition and machine learning in agricultural economics

1. Statistics and econometrics facing the challenges of big data and machine learning

The rise of machine learning and big data is reshaping society and research paradigms. Recent technical breakthroughs in natural language processing (NLP) models, such as ChatGPT, have garnered a lot of attention. Traditional research paradigms in economics and statistics are facing strong headwinds. On the one hand, big data nullifies the theoretical foundation of econometrics and statistics. With big data, the central limit theorem and the concept of sampling seem outdated.

Firstly, big data implies that the research object is not a sample, but rather a population. Traditional sampling techniques are costly and often impossible due to the curse of "Dimensionality." However, big data techniques show that computation is not so costly and can often directly access the population's distribution.

Secondly, when the sample size or population is very large, the standard deviation in the central limit theorem collapses, making all tests for parameters in econometric or statistical models statistically significant, rendering statistical tests less useful.

On the other hand, traditional econometrics focuses on causal inference and model tests, while machine learning algorithms prioritize predictions and pattern recognition, which is crucial for policy-making and business practices. As a worldly subject, economic research should groundly help policy makers and businesses make decisions. With the rise of machine learning and big data, we can expect it to compete for gaining the traditional land of econometrics and statistics.

Agricultural economics is a sub-field of applied economics that deals with agricultural and rural issues. In this changing era, many touchy issues such as aging, rural-urban inequality, inflation, poverty, trade and environmental pollution require new tools, insights and perspectives. Machine learning is a timely and powerful tool to meet these requirements. Although a few studies have applied machine learning algorithms in energy and agricultural economics, the general applications are still limited.

To address this gap, we organized a special section on "Machine Learning in Agricultural Economics," attempting to apply machine learning in agricultural economics. After the peer-review process, this special section accepted four papers with different applications in agricultural economics.

2. Machine learning and its application in agricultural economics

In addition to traditional econometric tools that are mainly used to test hypotheses and identify causal relations between variables, machine learning algorithms provide a much broader toolbox. Broadly speaking, machine learning can provide five types of services. This special section particularly pays attention the first three services.

2.1 Identification and prediction

Machine learning can help identify, predict and classify objectives, which is often called "supervised machine learning." In this category, the outcome variables are labeled, while a number of features are used to predict the outcomes. There are many algorithms in

China Agricultural Economic Review
Vol. 15 No. 2, 2023
pp. 375-378

1756-137X
DOI 10.1108/CAER-05-2023-307

predictions, including logit models, support vector machines, classification trees, gradient descent boost, random forests, neural networks, etc. For example, Wang *et al.* (2021) used random forest and remote sensing data to predict energy poverty in India.

In this special section, Maruejols *et al.* (2023) used different machine learning algorithms to predict subjective poverty in rural China. Among three algorithms—random forest, support vector machine and logit with LASSO—the performance of logit with LASSO is better in fitting than the other two for the middle-income group. Indeed, the no free lunch theory in machine learning states that no specific algorithm is superior to others in all possible data and distribution, and theoretically, they shall have similar predictive power on average (Wolpert and Macready, 1997). Empirical results also show that income, education, health and gift expenses are key predictors of subjective poverty, and the predictive accuracy is above 85% with the test set of cross-validation.

2.2 Clustering

Data also have a lot of unobserved heterogeneities that need careful scrutiny. We can detect different patterns in data, cluster the data into a few groups, and make precise policies targeting each group. In a broad definition, this is called "unsupervised machine learning." For instance, Graskemper *et al.* (2021) used a clustering algorithm and found three types of farms in Germany, and suggested different policies that should be made toward each group. In this special section, two papers (Zeng and Chen, 2023; Liu *et al.*, 2023) used different clustering algorithms to study two important policy issues: rural-urban integration and egg price clustering at the provincial level of China. Given the sheer size of China and the heterogeneous development levels between regions, such studies could help make more precise policies in these fields.

Zeng and Chen (2023) used the partitioning around medoids (PAM) algorithm to reveal four different types of rural-urban integration across provinces: high-level urban-rural integration, urban-rural integration in transition, low-level urban-rural integration and early urban-rural integration in the backward stage. In contrast to the classical clustering algorithm of K-means, which has a restricting application to continuous variables, PAM can take into account discrete variables.

Liu *et al.* (2023) used the dynamic time warping (DTW) method to cluster time-series data of egg prices across provinces in China and identified three different clusters. Within each cluster, the trends of food price dynamics are very similar between these provinces. Compared with PAM and K-means, the DTW method takes into account time-series properties. The first cluster includes main egg production provinces, the second cluster is the neighbor of the first cluster and the third cluster is mainly egg-importing regions. Due to transaction costs, the importing areas may have less price volatility.

2.3 Feature engineering

Feature engineering is a machine learning method that automates the process of selecting independent variables, also known as features. In traditional econometric and statistical models, features are subjectively selected by researchers, which can lead to the "cherry-picking" problem. Machine learning provides a few methods to select features automatically, especially in high-dimensional data where the number of independent variables is higher than the number of observations.

In regularized regression models, parameters can be penalized to identify the importance of different variables. The first-order penalty is called "LASSO," and the second-order penalty is called "ridge."

In this special section, Mao (2022) used the LASSO algorithm to identify international relation coalition partnerships and used network analysis to characterize the clustering

pattern of coalitions with high-frequency records of global event data. The author then constructed a monthly dataset of agricultural non-tariff measures (NTMs) against China and international relations with China of each importer and its coalition partners and estimated impulse response functions of agricultural NTMs with regard to international relation shocks. Mao (2023) also identified two major clusters of coalitions, one composed of coalitions primarily among "North" countries and the other of coalitions among "South" countries. The USA is found to play a pivotal role by connecting the two clusters.

In addition, Maruejols *et al.* (2023) used random forest and LASSO to identify the importance of variables in predicting poverty in China.

Apart from supervised and unsupervised machine learning and feature engineering, there are two other types of applications: reinforcement machine learning and the generation of variables and content, which have not been included in this special section. For instance, ChatGPT, which made a breakthrough in NLP, has widely applied both reinforcement and generation of contents. With these machine learning tools, researchers can generate new variables from pictures, videos, texts, or other non-stylized sources for economic research. For example, in Wang *et al.* (2021) a machine learning algorithm selected variables from remote sensing data to predict energy poverty in India.

3. Concluding remarks and future researches

In conclusion, this collection of four papers presents a small sample of the applications of different machine learning algorithms to agricultural economic research. By using supervised and unsupervised machine learning and feature engineering, the papers shed light on poverty, rural-urban integration, food prices and agricultural trade, which are traditional but important research questions. The results could offer better and more precise policy insights for policymakers.

Furthermore, the papers show that machine learning can be empirically applied to a broad range of research questions in agricultural economics, compared with econometrics. Traditional agricultural economic analyses are mainly top-down, and econometric tools are used for testing predesigned economic hypotheses. However, pattern recognition using unsupervised machine learning could help automatically recognize patterns and regularities in the data. Clustering, principal component analysis, market basket analysis, recommendation engine, text mining and visual recognition have provided effective tools for dimension reduction and pattern recognition.

After this special section, we look forward to more applications of machine learning that can further enhance the field of agricultural economics.

Xiaohua Yu and Lucie Maruejols

References

Graskemper, V.S., Yu, X. and Feil, J.-H. (2021), "Farmer typology and implications for policy design – an unsupervised machine learning approach", *Land Use Policy*, Vol. 103 April 2021, 105328.

Liu, C., Zhou, L., Höschle, L. and Yu, X. (2023), "Food price dynamics and regional clusters: machine learning analysis of egg prices in China", *China Agricultural Economic Review*, Vol. 15 No. 2, pp. 416-432, doi: 10.1108/CAER-01-2022-0003.

Mao, R. (2023), "Coalitions in international relations and coordination of agricultural trade policies", *China Agricultural Economic Review*, Vol. 15 No. 2, pp. 433-449, doi: 10.1108/CAER-01-2022-0011.

Maruejols, L., Wang, H., Zhao, Q., Bai, Y. and Zhang, L. (2023), "Comparison of machine learning predictions of subjective poverty in rural China", *China Agricultural Economic Review*, Vol. 15 No. 2, pp. 379-399, doi: 10.1108/CAER-03-2022-0051.

Wang, H., Maruejols, L. and Yu, X. (2021), "Predicting energy poverty with combinations of remote-sensing and socioeconomic survey data in India: evidence from machine learning", *Energy Economics*, Vol. 102, 105510, doi: 10.1016/j.eneco.2021.105510.

Wolpert, D.H. and Macready, W.G. (1997), "No free Lunch theorems for optimization", *IEEE Transactions on Evolutionary Computation*, Vol. 1, p. 67.

Zeng, Q. and Chen, X. (2023), "Identification of urban-rural integration types in China – an unsupervised machine learning approach", *China Agricultural Economic Review*, Vol. 15 No. 2, pp. 400-415, doi: 10.1108/CAER-03-2022-0045.

Further reading

Graskemper, V.Y.X. and Feil, J.-H. (2022), "Values of farmers-evidence from Germany", *Journal of Rural Studies*, Vol. 89, pp. 13-24.

Maruejols, L., Hoeschle, L. and Yu, X. (2022), "Vietnam between economic growth and ethnic divergence: a LASSO examination of income-mediated energy consumption", *Energy Economics*.

Wang, H. and Yu, X. (2023), "Carbon dioxide emission typology and policy implications: evidence from machine learning", *Forthcoming in China Economic Review*.