Exploring the Impact of Pesticide Usage on Crop Condition: A Causal Analysis of Agricultural Factors

Mee Qi Siow, Yang Sok Kim, Mi Jin Noh, Mu Moung Cho Han

Abstract

Human lifestyle is affected by the agricultural development in the last 12,000 years ago. The development of agriculture is one of the reasons that global population surged. To ensure sufficient food production for supporting human life, pesticides as a more effective and economical tools, are extensively used to enhance the yield quality and boost crop production. This study investigated the factors that affect crop production and whether the factors of pesticide usage are the most important factors in crop production using the dataset from Kaggle that provides information based on crops harvested by various farmers. Logistic regression is used to investigate the relationship between various factors and crop production. However, the logistic regression is unable to deal with predictors that are related to each other and identifying the greatest impact factor. Therefore, causal discovery is applied to address the above limitations. The result of causal discovery showed that crop condition is greatly impacted by the estimated insects count, where estimated insects count is affected by the factors of pesticide usage. This study enhances our understanding of the influence of pesticide usage on crop production and contributes to the progress of agricultural practices.

Keywords: Pesticide Usage| Crop Condition| Logistic Regression| Causal Discovery

I. INTRODUCTION

The impact of agricultural development, which began 12,000 years ago, on human lifestyle is quite profound, leading to a significant shift from traditional hunter-gatherer ways of life to settled living and a preference for a reliable food supply [1]. Since the crops and animals could be farmed to fulfill the demand, the global population has surged to eight billion today [1]. However, this population increase necessitates a greater food supply, and to meet this demand, pesticides are extensively used in modern agriculture as a more effective and economical approach [2]. According to Mahmood et.al, a pesticide is a toxic chemical substance or a mixture of substances or biological agents that are intentionally released into the environment in order to avert, deter, control, or kill and destroy the population of insects, weeds, rodents, fungi or other harmful pests [3].

Although there are other influences that might affect crop production, the use of crop protection chemicals seems a simple way to obtain better crop yields [4]. The study of Abhilash and Singh disclosed that approximately 45% of the annual food production is lost due to pest infestation [5]. Hence, the extensive use

Manuscript : 2023.10.31
Confirmation of Publication: 2023.11.20

Corresponding Author : Yang Sok Kim, e-mail : yangsok.kim@kmu.ac.kr

of pesticides is required to confront pests and to increase the yield quantity [5,6]. Based on the study of Rahman, farmers seem to treat pesticides as substitutes for fertilizers, indicated by the positive influence of fertilizer prices on pesticide use [7]. The usage of pesticides has become a significant factor in the development of agricultural activity.

The objective of this study is to investigate the factors influencing crop production, with a specific emphasis on the role of pesticide usage. The overarching goal is to enhance our understanding of how agricultural practices, particularly the use of pesticides, contribute to crop production outcomes. The study aims to identify and analyze the key factors affecting crop yields, assess the significance of pesticide usage among these factors, and employ logistic regression and causal discovery methods to uncover relationships and potential causal links within the dataset obtained from Kaggle.

Logistic regression is used to identify the relationship between the factors of pesticide usage and crop condition. Logistic regression is a statistical analysis that describe the relationship between a binary dependent variable and a set of independent variables. However, there is a limitation in logistic regression that assumes the independent variables do not influence each other. Hence, causal discovery is applied in this study to look for the structural relationship that appears between the predictors and outcomes in crop production.

From a perspective of the contribution of this study, it will enhance our understanding of the influence of pesticide usage on crop production, thereby making a valuable contribution to the progress of agricultural practices.

II. RELATED WORK

The study of Liliane and Charles revealed that the factors affecting crop production can be categorized into three groups: technological, biological, and environmental [8]. However, despite other factors, the biological factor is being focused, indicated by the extensively used pesticides that helped in confronting pests. Pesticides are important for ensuring sufficient food production for the global population [3]. The role of pesticides is not only preventing food shortages but also indirectly keeping food prices under control [3]. Meanwhile, pesticides also contribute to improving human health by preventing disease outbreaks through the control of rodents and insect vectors [3].

However, studies have revealed the non-negligible bad sides of pesticides [2,3,9]. The toxic chemical substances in pesticides have drastic effects on non-target species, including the natural predators of pests and parasites, affecting the biodiversity of animal and plant and hence harming the ecosystems [3]. Even though the bad sides of the usage of pesticides were disclosed, pesticides are still continuously being used in agriculture to maximize the yield quantity.

The relationship between crop production and the factors affecting it can be demonstrated by logistic regression. Logistic regression is a statistical analysis that describe the relationship

between a binary or dichotomous outcome and a set of independent predictors or explanatory variables [10]. However, the statistically significant relationship between the independent and dependent variables does not indicate the existence of a causal relationship [11]. In addition, logistic regression assumes that the effect of one predictor is independent of the value of other predictors, which is often unrealistic in real-world data [12]. Hence, studies nowadays have been paying attention to the structural relationship between variables [13,14]. The structural relationship allows researchers to incorporate causal assumptions into the model for causal relationship analysis. [14].

In the preceding literature, there has been a focus on the role of pesticides and their adverse effects. However, a noticeable gap exists in the investigation of the relationship between crop condition and pesticide usage. Recognizing this gap, our study proposes to bridge it by introducing a novel methodology. This methodology aims to unravel the indirect relationship between crop condition and pesticide usage, offering a fresh perspective on the intricate interplay between these factors. Our study seeks to contribute to a more comprehensive understanding of the impact of pesticide usage on crop conditions, thereby addressing a crucial gap in the existing literature.

III. METHODOLOGY

1. Dataset

This study used the agriculture data collected from Kaggle that provided information based on crops harvested by various farmers at the end of the harvest season [15]. This dataset contained 10 attributes that collected from 88,858 records and the detail information of the dataset is shown in Table 1.

Table 1. Data Description

Variables	Definition
ID	Unique ID
Estimated_Insects_Count	Estimated insects count per square meter
Crop_Type	Category of Crop (0,1)
Soil_Type	Category of Soil (0,1)
Pesticide_Use_Category	Type of pesticides uses (1-Never, 2- Previously Used, 3- Currently Using)
Number_Doses_Week	Number of doses of pesticides used per week
Number_Weeks_Used	Number of weeks used pesticides
Number_Weeks_Quit	Number of weeks quit pesticides
Season	Season Category (1,2,3)
Crop_Damage	Crop condition/production (0-Alive, 1-Damage due to other causes, 2-Damage due to pesticides)

The dependent variable of this study is crop damage, and the independent variables are variables that related to pesticide usage, such as pesticide use category, number doses week, number weeks used, number weeks quit, and estimated insects count. Meanwhile, the other variables such as crop type, soil type and season are the control variables in this study.

2. Data-Preprocessing

The attribute of ID, which is the unique value for each data entry, is excluded from this study. Furthermore, the condition of crops (labeled as crop damage) is categorized into two groups: 0

the group that crops are alive and 1 the group that crops are damaged. Both categories that labelled as 1−crops that damaged due to other causes and 2−crops that damaged due to pesticides in the initial dataset are grouped as damaged crops in this study.

Moreover, as the values of estimated insects count are larger than the other variables' values, the large value will cause the other variables with smaller values to be overwhelmed in the regression. Hence, the variable is log−transformed in this model [16].

There are 9000 missing values in the variable of the number of weeks pesticides used, and these data rows have been dropped from this study. After removing the missing values, there are 79,858 data in the final dataset.

3. Logistic Regression

Logistic regression is an analysis that enables us to estimate the relationship between binary dependent variable and independent variables [10]. According to Cokluk, the main purpose of logistic regression is to classify individuals into different groups, where logistic regression reveals the possibility of particular consequences for each subject [17]. The study of Cokluk revealed that logistic regression produces a regression equation that enables us to make an accurate estimation of the possibility that an individual falls into one of the categories [17]. Logistic regression analysis is unlike the discriminant analysis and multiple regression analysis, where it does not require assumptions met concerning the distribution of independent variables [17]. Hence, the study suggested that logistic regression analysis is much more flexible than the two techniques [17]. However, there are several research problems that are difficult to handle by logistic regression [13]. The most notable problem is when an outcome is determined not only by direct influences of the predictor variables but also by the unobserved common cause, and the unavailability to determine the structural relationship between variables [13].

4. Causal Discovery

To address the limitations of logistic regression mentioned above, causal discovery is used in this study. Causal discovery is a method that is able to reveal causal information by analysing purely observational data [18]. A traditional way to discover causal relations is to use intervention or randomized experiments [18]. However, in most cases, this way might be too expensive, time−consuming, unethical or even impossible [18]. Hence, using pure observation data for obtaining the causal relations between variables has drawn much attention. This study will use Direct LiNGAM for conducting causal discovery. This algorithm is one of the algorithms of LiNGAM, a family of causal discovery algorithms used in machine learning to estimate causal relations between variables [19]. This algorithm assumes that the causal relations between the variables are linear and have non−Gaussian distributions [19]. The algorithm seeks to identify a causal ordering of the variables by exploiting the

fact that the residuals of the linear regression models of each variable on the others should be independent and non-Gaussian [19]. Direct LiNGAM can estimate the causal relations between the variables by iteratively fitting these regression models and testing for the residuals' independence [20]. According to Shimizu et.al, the basic LiNGAM model makes the following assumptions when estimating the causal relationship among variables [20]:

1. Linearity
2. Non-Gaussian continuous error variables (except at most one)
3. Acyclicity
4. No hidden common causes

In this study, the relationship of crop production and factors affected it is estimated by logistic regression and causal discovery using the attributes of crop condition and other variables in the data collected.

IV. RESULT

Table 2 shows the result of logistic regression model for identifying the crop condition under the influence of various factors. In the logistics model, season and number of doses of pesticides used per week are factors that are not significantly impacting the crop condition, proved by the p-values that are larger than 0.05. Meanwhile, crop type, soil type, pesticide use category, estimated insects count, number of weeks quit pesticide, and number of weeks used pesticide all have a significant relationship with crop condition as the p-values of these factors are all less than 0.05.

Table 2. Logistic regression result.

Variables	Coef	P-value
Intercept	-2.10	0.00
Crop_Type[T.1]	-0.50	0.00
Soil_Type[T.1]	-0.18	0.00
Pesticide_Use_Category [T.2]	-5.99	0.00
Pesticide_Use_Category [T.3]	-4.35	0.00
Season [T.2]	-0.01	0.69
Season [T.3]	0.00	0.96
Estimated_Insects_Count	0.61	0.00
Number_Doses_Week	0.00	0.70
Number_Weeks_Quit	0.03	0.00
Number_Weeks_Used	0.04	0.00

However, the limitation of logistic regression that assumes that the effect of one independent variable is not influenced by another independent variable is reflected in this model. The pesticide use category, number of doses of pesticide used per week, number of weeks quite pesticide, and number of weeks used pesticide are factors of pesticide usage that highly influenced the estimated insects count, and the model above is unable to address this problem. In addition, the model above only shows the flat relationship between variables and is unable to show the factor that has the greatest impact on crop condition.

Therefore, causal discovery is applied to the data to address the limitations of logistic regression. Figure 1 is a DAG (Directed Acyclic Graph) labeled with the value of the adjacency matrix, which denotes the connection strength of a variable to another variable. The result shows us the structural causal relationship among the variables.

According to DAG, the crop condition is directly impacted by the factors of crop type, soil type, pesticide use category,

and estimated insects count. Meanwhile, the number of doses of pesticides used per week, number of weeks quit pesticide, and number of weeks used pesticide affect the estimated insects count and do not have a direct impact on crop condition.

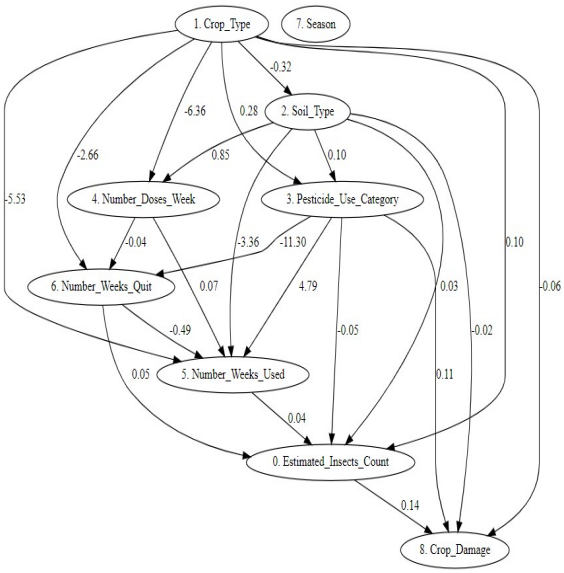

Fig. 1. DAG estimated by Causal Discovery

The value of the intervention effect will be estimated by the formula of $E[Y \mid do(Xi=mean)] - E[Y \mid do(Xi=mean \pm std)]$ [21]. The maximum estimated value means the feature is having the greatest intervention effect [21]. Table 3 shows the estimated values. According to the estimated values of intervention effect, the attribute of estimated insects count has the highest intervention effect towards the crop condition, followed by number of weeks used pesticide, pesticide use category, number of doses of pesticide used per week, number of weeks quit pesticide, crop type, soil type, and season. As the existence of a strong dependence of estimated insects count on pesticide usage is believed, the intervention effect on estimated insects count is estimated and the result is showed in Table 4.

Table 3. Intervention effect estimated for crop condition.

Features	effect_plus	effect_minus
Estimated_Insects_Count	0.229	0.133
Crop_Type	0.021	0.023
Soil_Type	0.021	0.022
Pesticide_Use_Category	0.131	0.093
Number_Doses_Week	0.026	0.028
Number_Weeks_Used	0.168	0.110
Number_Weeks_Quit	0.025	0.027
Season	0.002	0.002
Crop_Damage	0.000	0.000

Table 4. Intervention effect estimated for estimated insects count.

Features	effect_plus	effect_minus
Estimated_Insects_Count	0.000	0.000
Crop_Type	0.074	0.074
Soil_Type	0.056	0.056
Pesticide_Use_Category	0.065	0.065
Number_Doses_Week	0.064	0.064
Number_Weeks_Used	0.528	0.528
Number_Weeks_Quit	0.240	0.240
Season	0.002	0.002

The number of weeks used pesticide is the attribute with the greatest intervention effect (0.528), followed by the number of weeks quit pesticide (0.240). This also means that the factors of pesticide usage are important factors affecting crop conditions indirectly.

V. CONCLUSION

This study shed some light on the method to identify the factors that affect the crop conditions. This study suggested a new methodology to be used to identify the structural relationship between predictors and dependent variables.

Causal discovery is also able to help in discovering the attribute that has the greatest impact on the dependent variable. The conclusion that pesticide usage is a critical factor indirectly affecting crop conditions signifies a noteworthy outcome. This implies that interventions and strategies aimed at optimizing pesticide application, in terms of frequency and duration, can have a profound impact on overall crop health and yield.

The valuable insights into the complex interplay of factors influencing crop conditions, particularly by introducing and successfully applying causal discovery to overcome the limitations of logistic regression is provided in this study. The identification of pesticide usage as a crucial factor indirectly impacting crop conditions highlights the practical significance of optimizing pesticide application strategies. The research contributes methodologically by proposing a new approach for uncovering structural relationships between predictors and dependent variables. Practically, these findings suggest that interventions aimed at enhancing pesticide usage practices could significantly contribute to improved crop health and yield, providing a pathway for more informed and targeted agricultural practices.

While this study sheds light on important factors influencing crop conditions, it is essential to acknowledge its limitations. The analysis does not encompass various factors, such as technological variables, specific pesticide types, and climate factors, which are known to impact crop production. These omissions restrict the generalizability of the findings and underscore the need for more comprehensive investigations in future studies.

Future research should delve into the impact of technological factors on crop conditions. Assessing the role of advanced agricultural technologies, machinery, and precision farming methods can provide a more comprehensive understanding of their influence on overall crop health. Also, a more granular examination of different pesticide types and their specific effects on crop conditions is warranted. Understanding the impacts of various pesticides can inform targeted strategies for optimizing pest control while minimizing potential negative consequences. Investigating the interaction between climate factors and crop conditions is crucial for developing resilient agricultural practices. Future studies should explore how changing climate patterns, temperature variations, and precipitation levels influence crop health and production.

REFERENCES

[1] The development of Agriculture (2023).https://education.nationalgeographic.org/resource/development-agriculture/ (Accessed: Oct., 2, 2023).

[2] A. Sharma, V. Kumar, B. Shahzad, M. Tanveer, G.P.S. Sidhu, N. Handa, S.K. Kohli, P. Yadav, A.S. Bali, R.D. Parihar, and O.I. Dar, "Worldwide pesticide usage and its impacts on ecosystem," *SN Applied Sciences*, vol.1, pp.1-16, Oct. 2019

[3] I. Mahmood, S.R. Imadi, K. Shazadi, A. Gul, and K.R. Hakeem,

"Effects of pesticides on environment," *Plant, soil and microbes: Implications in Crop Science*, vol.1, pp.253-269, Mar. 2016

[4] F.P. Carvalho, "Agriculture, pesticides, food security and food safety," *Environmental Science & Policy*, vol.9, no.7-8, pp.685-692, Nov. 2006

[5] P.C. Abhilash, and N. Singh, "Pesticide use and application: an Indian scenario," *Journal of Hazardous Materials*, vol.165, no.1-3, pp.1-12, Jun. 2009

[6] J.P.G. Webster, R.G. Bowles, and N.T. Williams, "Estimating the economic benefits of alternative pesticide usage scenarios: wheat production in the United Kingdom," *Crop Protection,* vol.18, no.2, pp.83-89, Mar. 1999

[7] S. Rahman, "Farm-level pesticide use in Bangladesh: determinants and awareness," *Agriculture, Ecosystems & Environment*, vol.95, no.1, pp.241-252, Apr. 2003

[8] T.N. Liliane, and M.S. Charles, "Factors affecting yield of crops," *Agronomy-climate change & food security*, pp.9, Jul, 2020

[9] M. Lykogianni, E. Bempelou, F. Karamaouna, and K.A. Aliferis, "Do pesticides promote or hinder sustainability in agriculture? The challenge of sustainable use of pesticides in modern agriculture," *Science of the Total Environment*, vol.795, pp.148625, Nov. 2021

[10] A. Das, "Logistic regression," *In Encyclopedia of Quality of Life and Well-Being Research*, pp.1-2, Jan. 2021

[11] A. Worster, J. Fan, and A. Ismaila, "Understanding linear and logistic regression analyses," *Canadian Journal of Emergency Medicine*, vol.9, no.2, pp.111-113, Mar. 2007

[12] J. Tolles, and W.J. Meurer, "Logistic regression: relating patient characteristics to outcomes," *Jama*, vol.316, no.5, pp.533-534, Aug. 2016

[13] E. Kupek, "Beyond logistic regression: structural equations modelling for binary variables and its application to investigating unobserved confounders," *BMC Medical Research Methodology*, vol.6, no.1, pp.1-10, Mar. 2006

[14] K.A. Bollen, and J. Pearl, *Handbook of causal analysis for social research*, Springer Dordrecht, pp. 301-328, 2013

[15] AV JantaHack: Machine Learning in Agriculture. (2020). https://www.kaggle.com/datasets/sumeetsawant/av-jantahack-machine-learning-in-agriculture/data?select=test.csv (Accessed: Sep., 4, 2023).

[16] L. Metcalf, and W. Casey, *Cybersecurity and applied mathematics*, Syngress, pp.43-65, 2016

[17] O. Cokluk, "Logistic Regression: Concept and Application," *Educational Sciences: Theory and Practice*, vol.10, no.3, pp.1397-1407, 2010

[18] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in Genetics*, vol.10, pp.524, Jun. 2019

[19] S. Shimizu, P.O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan, "A linear non-Gaussian acyclic model for causal discovery," *Journal of Machine Learning Research*, vol.7, no.10, Oct. 2006

[20] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvarinen, Y. Kawahara, T. Washio, P.O. Hoyer, K. Bollen, and P. Hoyer, "DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model," *Journal of Machine Learning Research-JMLR*, pp.1225-1248, Apr. 2011

[21] Causal Effect. (2023). https://lingam.readthedocs.io/en/latest/reference/causal_effect.html (Accessed: Sep., 10, 2023).

Authors

Mee Qi Siow

She received her B.S. degree at the Department of Economics, University of Malaya, Malaysia, in year 2020. She is currently pursuing her Master studies at the Department of Management Information Systems in Keimyung University, South Korea. Her research interests include Big Data Analytics, Text Mining and Data Visualization.

Yang Sok Kim

He has been serving as an Associate Professor at the department of Management Information Systems, Keimyung University, South Korea. He received his Ph.D. from University of Tasmania (UTAS), Australia. His research interests are Machine Learning, Web Search/Mining, Social Network, and Recommenders Systems. He has published papers in Electronic Commerce Research and Applications, Expert Systems with Applications, Mobile Information Systems, International Journal of Human-Computer Studies, Sustainability, and other reputed journals.

Mi Jin Noh

She received her M.S. and Ph.D. degree in Management Information Systems from Kyungpook National University, Korea in 2001 and 2006, respectively. Since 2022, she has been an assistant professor in Department of Business Big Data, Keimyung University, Korea. Her research interests are Big Data Analysis, Smart City, Text Mining and Mobile Services.

Mu Moung Cho Han

She received her B.S. degree in Computer Science from Korea National Open University in 2006. Subsequently, she received her M.Ed. and Ph.D. degrees in Computer Education and Management Information Systems from Keimyung University in Korea in 2009 and 2016, respectively. Since 2021, she has been serving as an Assistant Professor in the Institute of General Convergence Education at Dongguk University WISE Campus in Korea. Her research interests are Machine Learning, Text Mining, Information Technology, Knowledge Management, and Education Technology.