A Comparative Study on Agent Based Decision Making Models: A Proof of Concept Focused on Farmers' Decisions Regarding Best Management Practices

By

Duo Zhang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Geography

Waterloo, Ontario, Canada, 2020

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In recent times, with the increasing availability of large datasets, applications of machine learning techniques have grown at a rapid speed. However, due to the black-box nature of these tools, it can be hard for model builders to understand the detailed structure of the system that machine learning models simulate. Agent-based modelling (ABM) is a popular approach to studying complex systems., One of the challenges for this technique is to design the decision making processes of the agents in the model. As machine learning tools have a strong ability to transform the information from the raw data into a functional model as the decision making processes for agents in ABMs. Because an ABM can provide a detailed structure for the system that the machine learning model simulates, it is reasonable to combine the two kinds of models. However, although in previous studies, some researchers combine the two models, most of them use one of the two models as a validation tool for the other, rather than to integrate the machine learning model into the decision making processes of agents in ABMs. Therefore, this thesis focuses on integrating a machine learning model into the ABM, and contrast it with the ABMs with two traditional decision making models, including an optimal model and a stochastic model.

To compare the three decision making models, we use farmers' BMP adoption case in the Upper Medway subwatershed, and contrast the three models through three metrics, including the percentage of BMP adoption, size of agricultural land of BMP adoption, and the correlation between BMP adoption and landuse types. As a result, the ABM with the machine learning model presents a high level of accuracy compared with the other two traditional models, but its adaptability to other cases and the robustness to uncertainties still require a further study.

Acknowledgements

I would first like to express my sincere appreciation to my supervisor, Professor Peter Deadman for his expert guidance and patience for the two years during my master’s studies. I am truly grateful for his continuous support and guidance on my research and writing of this thesis, especially in this tough time of COVID-19.

Then, I would also like to thank my committee members Dr. Derek Robinson and Dr. Rob Feick for their insightful comments and time. Their questions and immense ideas are extremely valuable.

My sincere gratitude also goes to my friends and fellow researchers. Particularly, I would like to thank my friend Jiaxin Zhang for discussing with me the valuable literature, design of the ABMs and possible problems and limitations. I would also like to thank the writers whose books encouraged me in this difficult time of COVID-19.

Last but not least, I would like to express my appreciation to my family members for their love and support all the time. This accomplishment will not be possible without them.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

The study of complex systems has been applied in various fields like global climate change, transportation planning, and market dynamics. These complex systems are comprised of interacting parts that generate collective behavior at the system level (Commendatore et al., 2018; Robinson & Brown, 2016; Chen et al., 2013; Gilli & Rossier, 1979). Many studies of complex systems have employed agent-based modelling (ABM) to describe the system and explore the dynamics of its behavior (Miller & Page, 2009; Heckbert et al., 2010). ABM is a computational modelling method that describes the entities or individual elements of a system as agents that take decentralized actions to interact and communicate with each other and their environment (Wilensky & Rand, 2015). With these actions, the model of a whole system can generate some global properties and patterns that could be difficult to discover by studying the system at the individual level (Lai & Liao, 2013). Therefore, the ABM method is an ideal tool for simulating the complex system, since the computational model consists of heterogenous and interactive agents, and can generate nonlinear dynamics, especially compared with traditional mathematical and statistical methods (Parker & Robinson, 2017).

In an ABM, it is important to define an agent and its possible actions. The agents can be heterogenous representatives of some abstract entities with microscopic motives. The agent has the ability to make decisions autonomously, interact with each other and the environment, and respond to them. The actions of agents will lead to a macroscopic pattern (Handel, 2016), which is also how the results of a model are presented. Therefore, it is significant for a model builder to design the actions of agents, and hence, the decision models or structures can be introduced into ABMs to achieve this. According to Groeneveld et al. (2017), these decision making models are commonly implemented by optimization, heuristics, and the process with stochastics parts, which means a stochastic modification of the optimal models. For all these methods, it is usually difficult to determine the optimal initial settings for model parameters, and sometimes it is strongly influenced by personal perceptions (Hayashi et al., 2016). As a result of the personal perceptions, sometimes the information and relationship in the dataset can be easily ignored, e.g. the first version of the model that simulates the European beech forest ignores the vertical patterns, which was questioned by beech forest scientists, for it is a significant factor in forest structure, but it satisfied the demands of landscape ecologists (Railsback & Grimm, 2019). One possible approach to addressing this problem could involve the introduction of machine learning methods into ABMs.

Machine learning in recent times grows in a rapid speed due to the large amount of online data collected (Jordan & Mitchell, 2015). It is a study of algorithms that has the ability to automatically learn and improve from experience and data (Mitchell, 1997). As we discussed above, for ABM, it is not easy to design the decision making processes of agents, and personal perceptions have a strong influence. Compared with ABM, Machine learning techniques can take advantage of the data collected and use the experience derived from data to make decisions, but it can hardly present the heterogeneity among agents or parts of systems and predict the dynamic influence of actions of them as it is sometimes a black-box method (Hayashi et al., 2016), while ABM can provide an opportunity for model builders to check the detailed structures inside the whole complex system. As we discussed above, in the machine learning model, there is a difficulty with explaining the detailed structure of the simulated system, and in ABMs, it is hard to design the actions of agents. Therefore, an integration of ABM and machine learning can be an alternative to solve the problems in both of ABMs and machine learning, because machine learning model can use the data more efficiently to get the action rules of agents, while ABMs can provide a clear structure of the system that the machine learning method sometimes has difficulty with.

However, the integration of the two models is not widely discussed. Although some researchers provide some examples (Wojtusiak et al., 2012; Lamperti, 2018; Torrens et al., 2011), most of their research are not published in a peer-reviewed journal (Hayashi et al., 2016; Raman & Leidner, 2019; Furtado, 2017; Rand & Stonedahl, 2007; Sheikhha, 2009; Zhou, 2018; Rand, 2006). Furthermore, these research findings mainly only focus on the specific case of the integration of the two models, instead of the comparison of a machine learning model and other decision models in ABM. For those research findings that focus on different performances of decision models in ABM, the involvement of machine learning is not usually mentioned (Groeneveld et al., 2017). Consequently, comparing the decision models in ABMs, including a machine learning model based one, remains a research gap. To fill the research gap, this project attempts to answer the primary research question: To what degree do the performance of ABMs with different types of decision making models, especially for the machine learning based heuristic one, differ by patterns in a practical case, and what is the main characteristic of the involvement of machine learning models in ABMs?

In order to answer these research questions, this project focuses on a case study that explores farmer's adoption of agricultural Best Management Practices (BMP) designed to improve water quality in agricultural watersheds. These BMPs include several effective, practical and affordable approaches to prevent or reduce the pollution on farms' soil and water resources, like building a windbreak or retiring fragile land. Since the BMP adoption is determined by farmers, and each farmer that is affected by others or the environment can make their own decision, it can be generalized as a complex system. Hence, ABM can be implemented in this farmers' BMP adoption case. Guo (2018) built an ABM for this case with an optimized decision making structure, and this provides a general structure for this project to build other comparable ABM with different decision making models.

Additionally, this project can also be a part of the Agricultural Water Futures (AWF). The AWF is a seven-year and pan-Canadian project funded by a $77.8-million grant from the Canada First Research Excellence Fund. The goal of this project is to investigate how agriculture will change in response to climate factors or socio-economic drivers and help to improve the current and future water sustainability (University of Waterloo, n.d.). Especially, this study is a part of Work Package 3 in

AWF, strengthening capacity for adaptation in agricultural water decision making, as through the comparison of these decision making structures, future model builders can also get some insights about which decision model is suitable for their purpose.

In conclusion, this project aims to achieve the following goals and objectives to answer the research question above:

1. Examining the previous ABMs with different decision making structures that simulates farmers' BMP adoption by conducting a literature review and searching related projects and papers in the journal database including Scopus, Google scholar and the University library database.
2. Building a general framework of ABMs that make the decision making models of the previous ABMs possible to be implemented in a practical case and making them comparable.
3. Building a machine learning model based on empirical data and integrating it as heuristic rules of agents in the general structure.
4. Contrasting the results of the ABM with these decision making models and analyzing the characteristics of them.

During the processes of this project, there are limitations caused by time and limitations in available empirical data. This project only tries to focus on filling the research gap and answering the questions mentioned previously by achieving the research objectives above, so other objectives, particularly related to the practical case like making the results of models more realistic and analyzing the impacts of factors on farmers are not considered here.

There are 4 other chapters after the introduction in this thesis. In the literature review section, we will review the background knowledge about the decision making structure in ABMs, machine learning models, and BMP in the previous related literature, In the methodology section, we will introduce the main ABM structure used in this project, and the details about the practical case. The results chapter will introduce some metrics to evaluate and compare the ABMs with different decision making structures, and try to analyze the characteristics of them. The last chapter is the discussion, and we will discuss the achievements, limitations, and future possible work about this project.

Chapter 2 Literature review

This section will start by introducing the definition and importance of complex systems, and then discuss the background of ABM. Next, pattern oriented modelling will be introduced to examine whether the ABM is acceptable. Since one challenge for ABM is to design the actions of its agents, machine learning is introduced as a tool to inform the design of agent decision making algorithms. After discussing the general design of ABMs, we introduce the background of the specific case study used here, which serves as the proof-of-concept, in this thesis. Therefore, the knowledge about BMP and their previous implementation are also provided. Particularly, in one of these previous projects, the integration of farmers' typology is highlighted, for in this thesis, it is also a significant difference in decision making models, so the relevant information is also included.

2.1 Complex systems

Complexity exists in a variety of research fields, including computer design, social organization, ecology research, and even astronomy (Booch, 2008, pp. 8-10). However, the concise definition of a complex system remains a problem (Ladyman & Lambert, 2011). Since the concept of a complex system is widely used in multiple fields, social scientists and philosophers utilize various definitions. For example, Arthur (1999) claimed that the complex system is a system "with multiple elements adapting or reacting to the pattern these elements create", while Rind (1999) defines a complex system as a system "in which there are multiple interactions between many different components”. Because researchers in multiple fields have different approaches to define this concept, it is hard to explain a complex system in a definition form. However, these researchers still provided some ideas about the characteristics of this concept. According to Ladyman & Lambert (2011), there are five main features of the complex system: nonlinearity, feedback, not centrally controlled, emergence, and hierarchy.

Nonlinearity is a corresponding concept to linearity. In a linear system, researchers can add any two solutions or multiply factors to get another, while nonlinearity indicates that this principle cannot be applied to the kind of nonlinear system. The attribute of feedback means that the actions of a member of the system respond to the earlier decisions made by its neighbors, and this requires a way to simulate the interactions among parts of the system. For most complex systems, they are not controlled by a central entity, which means that in the system there are no elements that can obtain all the information and guide or restrict all the other parts of the system. Simon (1991) suggests that the complex systems usually have several levels of organizations like some ecological systems, in which each level of structures can have its own functions and interact with the upper and lower structures. Additionally, as described by Anderson (1972), “more is different”, and this points out the limitation of reductionism, which means a complex system can usually generate some emergence, which means that the whole system has the properties that its parts do not own. Therefore, only focusing on the microlevel entities or studying on the macrolevel system can neither capture these emergent outcomes. For example, only focusing on individual farmers' decision making will make it hard to analyze the whole pattern, while only considering the change in the whole study area will lack the decision making processes of each farmer.

According to Shalizi (2006), complex systems can be studied by some modelling techniques, including cellular automata and agent based models (ABM), to make the systems easier to be understood. In these techniques, Shalizi (2006) believes that the ABM is the most frequently associated modelling techniques with complex systems. For example, in the research of coupled human and natural systems (CHANS), which focus on the interconnection of human and natural systems, the ABM is the major tool (An, 2012). Therefore, since a complex system exists in many fields and the features of complex systems require researchers to use some special methods to study it, rather than the traditional linear model, the agent based model can be a significant alternative.

2.2 Agent based models

ABM is a computational modelling tool in which the key actors or individual components of the system being studied are directly simulated (Parker & Robinson, 2017). These components are referred to as unique and autonomous agents that can interact with each other or the environment created within the model. The uniqueness indicates that each agent can be specified with its own characteristics, while being autonomous implies that agents can make decisions and take actions independently with the aim of achieving their own goals (Railsback & Grimm, 2019).

The capability of ABMs to respond to the features of complex systems makes them as a useful tool in understanding the complex systems. According to Dong et al. (2010), the ABM structure can represent the nonlinearity dynamics, which respond to the nonlinearity feature of a complex system as we mentioned above. In addition, the interactions between agents and the environment allow the model to represent the feedback and decentralization attributes of complex systems (Robinson & Brown, 2016). Furthermore, studying the connection between the systematic behaviors and characteristics of individual agents provides the opportunity to get some emergent results for researchers (Railsback & Grimm, 2019). Since the ability of ABM structure can satisfy the demands arise from the features of a complex system as we mentioned in the last section, including nonlinearity, feedback, not centrally controlled, emergence, and hierarchy, it is a powerful tool to simulate the complexity and multi-level problems in models (Van Dam et al., 2012).

ABMs are widely used for analyzing complex human decision making and behavior. Klabunde & Wilekens (2016) used the ABM structure to simulate the decision-making rules in migration. Miksch et al. (2019) applied the ABM to simulate the impacts of human interactions on the spread of infectious diseases. Fabian Adelt et al. (2014) involved the ABM structure to simulate different modes of governance. However, identifying the decision making structure of individual agents is still a challenge for designing the ABMs (Zeman, 2019). According to Groeneveld et al. (2017), three main decision-making techniques are used in the ABM structure: optimization, stochastics, and heuristics. Some researchers have studied the comparison of these models, especially for optimization and heuristic models. For example, Grovermann et al. (2017) and Schreinemachers & Berger (2006) discussed the advantages and disadvantages of optimization and heuristic decision making structures. Additionally, researchers like Cabrera et al. (2010) provided some case studies for presenting the performance of these decision making models. However, as described by Robinson et al. (2007), heuristics models are often described as "IF…THEN" structures, while some statistic models and the black-box approaches are ignored, because it does not present a clear "IF... THEN" conditional relationship.

The agents in optimization models usually rely on solving mathematical programming models to make their decisions, which is based on the idea that agents are always rational actors, able to evaluate all possible alternatives and find the choice that can maximize their utility (Schreinemachers and Berger, 2006). This is the most commonly used decision structure in ABMs (Groeneveld et al., 2017). For instance, Miller et al. (2010) used an optimization-based ABM, in which the agents, farmers, in this model use a calculated net income as the utility value to drive their decisions, to simulate the livelihoods of the population and the resulting landcover dynamics in the Galápagos Islands. This optimization-based structure is usually criticized as being unrealistic as choices are often affected by imperfect resources and bounded rationality (Simon, 1997). However, this model also has advantages, in that it is clear to monitor the economic trade-offs, and the factors driving the decision making processes can be easily tracked to inform policy making (Schreinemachers and Berger, 2006).

The stochastic structure, as the name suggests, implies that there are stochastic elements in the decision making processes. According to Groeneveld et al. (2017), models with stochastic elements are classified as an independent category of optimal and heuristics models. For example, Hoertel et al. (2020) developed a stochastic ABM to study the COVID-19 epidemic in France. In this model, the stochastic elements allow it to simplify the simulation of population social contacts and virus propagation dynamics. Additionally, the stochastic elements can also be used to investigate heterogeneity among agents, as the agents have the ability to have different preferences or follow divergent decision making rules (Ligmann-Zielinska, 2009).

If in a model, agents make decisions based on some rules that are derived from empirical data without a strong theoretical basis, we can define this type of model as the heuristic model (An, 2012). Compared with the optimization-based models, this approach allows one to simulate the bounded rationality of individual agents (Schreinemachers and Berger, 2006). According to Groeneveld et al. (2017), it is the second most popular decision-making structure used in ABMs, and it is very close to the optimization-based models (77 cases of heuristic rules vs 91 cases of optimal models). For example, Cabrera et al. (2010) implemented a heuristic-based structure, which is performed by a decision tree, ABM to simulate farming household behavior in the Brazilian Amazon. According to An (2012), statistical models like logistic regression or some black-box machine learning models like neural networks can be regarded as a tool to derive rules from empirical data. Therefore, although these black-box approaches are different from some theoretically guided methods, they can be also used to derive heuristic rules for ABMs, and this provides an opportunity to introduce some machine learning models in the ABM decision making structure. Since the heuristic structure usually uses empirical data to derive a rule for model design, the models may be regarded as more realistic compared with the optimal structure, because in the optimal structure, agents are usually assumed to be rational while the heuristic structure can present the bounded rationality of agents (Schreinemachers and Berger, 2006). However, it has a disadvantage that this structure can hardly explain the reasons or mechanism of decision making processes (Evans et al., 2006). Nevertheless, even though the reasons for decision making remains an unanswered question, these derived rules can explore some knowledge or information underlying the empirical data (An, 2012).

2.3 Pattern Oriented Model

Models can be regarded as a simplified representation to answer some research questions (Starfield et al., 1990), so there is still a significant problem that is necessary to be highlighted: how can a model both provide a response to the research questions and keep its characteristics as simple as possible, since the higher complexity will lead to increased difficulty in understanding the behavior of the model and reduced the generalizability of the implementation of models? For example, the additional information and submodels can make the models focus more on a specific case, and it may not be suitable for a different case. Therefore, a concept called structurally realistic is used to describe models that satisfy this requirement that means the model can present the essential elements to answer the research question, but it is not necessary to include everything we know to make it “realistic” (Railsback & Grimm, 2019). To decide whether the ABM is structurally realistic, an approach named pattern oriented model (POM) is an alternative.

The POM idea is mainly to design and analyze models based on multiple patterns from the real systems (Wiegand et al., 2003), in which patterns are “small numbers of weak and qualitative but diverse” stylized facts (Railsback & Grimm, 2019). Each pattern can work as a filter, and usually three to five patterns can help the models to become structurally realistic. It can mainly be used for designing the model structure, testing the agent behavior, and determining the appropriate parameter values (Railsback & Grimm, 2019).

Especially, as it is usually hard to determine the agent behaviors, POM can be very helpful for theory development, which means the agent behaviors serve as a theory described by submodels and researchers can examine possible theories by contrasting the levels of pattern-matching of the reproductions of these submodels (Railsback & Grimm, 2019), where ABMs work as virtual laboratories. For example, Huth & Wissel (1992) used an ABM structure to simulate and compare two theories for how fish moves in a school. One is to react with the nearest other fish, and the other one is to move based on the average direction of several neighbors. They used the mean distances between fish and the nearest neighborhood, the mean angle between each fish and the school, and the root mean square distance from individuals to the centroid of the school as patterns (Grimm et al., 2005). With this comparison, the second theory performed better when it came to reproducing fish schooling patterns. Therefore, POM can serve as a tool to compare different agent behaviors or decision making structures.

2.4 Machine learning models

Machine learning, or statistical learning, is an algorithm that has the ability to learn to solve a particular problem like driving or chess, from data automatically, and it usually relies on the sample data to train and build the model, and aims to make decisions or predictions (Goodfellow et al., 2016). As heuristics models require one to derive rules from empirical data, a machine learning model can be a good tool for turning the empirical data into useful rules or information (Mitchell, 1997), e.g. the machine learning model can be used to derive a rule for detecting the spam email based on the empirical records of spam email (Dada et al., 2019).

Based on the difference that whether the datasets have both inputs and outputs or only have the inputs, the machine learning model can be divided into supervised learning and unsupervised learning. Specifically, for the supervised learning models, in the raw dataset, there are model builders' expected values or classes, and the models are expected to produce results with these determined classes, while in the unsupervised learning, this expected information is not in the raw datasets, and model builders often seek the similarity or clusters in the datasets. As in the ABM, the machine learning model is expected to be used for decision making, and this means that a determined decision is necessary, only the supervised learning is related, since it can produce a determined result.

Moreover, supervised learning consists of two categories: classification for classifying the samples into discrete categories and regression for generating continuous numerical results, and since the decisions or choices can be regarded as several discrete values, classifiers, which means the models are used for classification, are more important here.

Specifically, logistic regression, k-nearest neighbors (KNN), naïve Bayes, decision trees and random forests, support vector machines (SVMs), and the neural networks are the most used classifiers, and the comparison of these classifiers are conducted in some past studies. Manzouri et al. (2018), Lorena et al. (2011), and Ul Hassan et al. (2018) used different case studies to examine the performance of these machine learning classifiers, and among them, random forest neural network, and SVM have the highest accuracy, while the random forest can also have a quick training speed and requires a smaller amount of data than neural networks, although the overfitting problem is necessary to be highlighted (Hastie et al., 2009).

Random forest is an ensemble learning method based on decision trees (Ho, 1998). As its name represents, it is just like a “forest” of decision trees. It combines multiple decision trees to predict the results. In this model, multiple decision trees are created by the bootstrap aggregating technique. Specifically, the algorithm chooses several samples in the training data without replacement to build decision trees. In these decision trees, at each split only a random subset of features in the training data are selected for enhancing the influence of strong predictors. Finally, the random forest model will use the average of predictions as the results (Hastie et al., 2009). Therefore, three parameters are the most important to be tuned, including the number of decision trees, the number of the random subsets of features at each split, and the minimum terminal node size (Koehrsen, 2018).

The overfitting problem is another significant challenge for machine learning models. It implies that the statistical model is too close to fit the training data, and therefore with the additional data, the model may not predict the results successfully (Santos et al, 2018). Due to the errors and noises in the training dataset, the ability of generalization will be reduced and leads to a big problem if the model is aimed to be used in another dataset, especially for the low-quality training data. To solve this problem, cross-validation can be a good technique. This technique chooses to split the input training dataset several times, and each time it uses one split for validation, the other splits are used for training the model. Then, it uses the averaging results as the output of this statistical learning model. Hence, this technique reduces the impacts of training data, and further reduces the overfitting effects caused by them.

There are some previous studies in which the machine learning models are integrated into ABMs. Most of them used the ABMs for the validation of the machine learning models, or used machine learning models to evaluate the performance of ABMs. Rand (2006) provided an integrated structure for both ABM and machine learning models, applied to the El Farol Bar Problem. In this structure, the results of machine learning models are used for updating the decision making processes of the ABM, while the running results are used as the observations to train the machine learning models. Lamperti et al. (2018) also introduced machine learning for ABM calibration to the real-world data. Wojtusiak et al. (2012) provided another perspective that regards the ABMs as an ideal platform for intelligent agents to simulate learning procedures. Each time the model runs, agents can use the results of machine learning models to learn, or in other words, to update its original behavior rules. Additionally, Hayashi et al. (2016) used the customer churning case for the comparison of the ABM and machine learning model, and the results of machine learning models were used for improving the prediction accuracy of ABM. However, to my knowledge, these projects usually focus on calibrating the results of ABMs or machine learning models by each other, instead of integrating the machine learning model into the decision making procedures of ABMs.

Therefore, the fact that few of these studies focus on integrating the machine learning model into an ABM for defining the decision making processes provides an opportunity to examine the effects of this kind of model, for one of the big challenges for ABM building is to design the action of agents. One of the examples is a framework provided by Jäger (2019). In this project, neural networks were used to replace manual rules in the case of reproducing Schelling's model, although the data train the machine learning model are collected from randomized ABM first. Since few researchers study the ABM with a machine learning model as the behavior structure, to my knowledge, the comparison of it and other decision making models are hardly performed.

As described, few researche focus on integrating machine learning models into

ABM as submodels to design actions of agents, which is a challenge in ABM building. Although An (2012) mentioned this technique as a subgroup in heuristic models, the practical implementation and its characteristics still remain a gap. In this thesis, we choose BMP adoption as a proof-of-concept to examine the difference between this type of ABM and ABM with other decision making structures. Particularly, since BMP adoption can be influenced by interactions of other agents, and can also interact with the environment, it is a good case for ABM application, and can also be used for contrasting ABMs with different decision making models.

2.5 Best management practices in agriculture

Best management practices (BMPs) are the descriptions of approaches that are devoted to managing human activities for reducing the pollution of surface and groundwater. These approaches can be implemented in agricultural, urban and forestry land (Utah State University, n.d.). Specifically, for the agricultural BMP, it implies a practical and affordable approach for farmers to conserve their soil and water with the guarantee of the productivity (Ontario Ministry of Agriculture, Food and Rural Affairs [OMAFRA], n.d.). The Clean Water Program is a program that aims to provide technical and financial assistance to protect the rural water quality. According to Clean Water Program (n.d.), there are totally 12 eligible BMP projects for its funding in the installation dataset, which is also the training dataset used for the machine learning model in this thesis. However, most of them cannot be easily adopted by farmers like septic systems, which require special eligibility and the assessment of experts. Additionally, there are other classification methods for BMP types as described by Upper Thames River Conservation Authority [UTRCA] (n.d. a), and OMAFRA (n.d.), but due to the limit from the availability of datasets, this thesis will only focus on the following four types of BMPs.

Grassed waterways are broad, shallow channels that are designed to convey local runoff without soil erosion (OMAFRA, 2009). The vegetation cover can serve as a protector to retard the flow of surface water and reduce the erosion of soils (Alberta Government, 2015). A number of factors should be considered before designing the grassed waterways, including the slope of the waterway, the vegetation suitable, and the size of the waterways like length and width (United States Department of Agriculture [USDA], 2007). According to Schroter and Kansas (n.d.), the grassed waterway has a minimum lifespan of ten years.

Water and Sediment Control Basin (WASCob) refers to an embankment built across a depression area to store the runoff water (Natural Resources Conservation Service [NRCS], n.d. a). It can improve the water quality by trapping and collecting sediment. Since it can also control the flow within a drainage area, the gully erosion can be further reduced. Usually, the slope, the area watershed, and the soil characteristics should be considered before designing the WASCob (NRCS, 2010).

According to NRCS (n.d. b), “buffer strips” can refer to riparian buffers, filter strips, grassed waterways, and windbreaks. However, since these approaches are quite different in terms of installation and life span, here the buffer strip only implies the riparian buffers. The riparian buffer strips are the area of permanent vegetation that can serve as the barriers between water bodies and farming areas (UTRCA, n.d. b). As a result, it can usually protect the water bodies from the influence of land use in the neighborhood, and thus protect and improve the water quality. Slope, soil texture, vegetation types can all have an impact on the effectiveness of the buffer strips (Hawes & Smiths, 2005).

Field windbreaks are a linear vegetative barrier in one or two rows for reducing the impacts of excessive wind speed in the fields (OMAFRA, n.d.). They can reduce the erosion caused by the wind, improve the crop quality by reducing the wind damage (UTRCA, n.d. c), and since the decreased wind speed can enhance the pollutant filtering capacity of soils, it can also be used to protect the water quality (Lower Thames Valley Conservation Authority [LTVCA], 2015). Additionally, the windbreak can serve to moderate the soil temperature, increase the soil moisture, and change the distributions of snow. According to Essex Region Conservation (ERCA) (n.d.) and Clean Water Program (n.d.), both projects aim to improve the water quality have listed the windbreak installation as an alternative. Before the installation of windbreaks, the density, height and species are all necessary factors, which can have impacts on their effectiveness, to be determined based on the local conditions (Brandle, n.d.).

There are also some other types of agricultural BMP used widely, such as changes in tillage systems, manure management, and the pesticide storage. The model in this project provides a possible interface for various BMP types if the model users can assess the adequate local observations and the cost data about these BMP types.

Nevertheless, in the proof of concept, only the above four BMP types can be used, in order to make the decision making models comparable with the limited available data. The model does not prohibit other BMP cases used for the proof of concept. The reason for the BMP choices in this thesis is mainly two-fold. One main reason is the limitation of raw datasets, and even the four BMP types will be recategorized as two large groups for the comparability of the decision making structures in ABMs. The other reason is that since the goal of this project is to provide a practical case to examine the performance of decision making structures in ABMs, especially for the machine learning model, some model structures and elements mainly inherits from the previous work that focuses on farmers' BMP adoptions. Hence, in the following section, we will try to introduce some previous work using ABM to simulate this case, which serves as an important basis for the practical implementation in this project.

2.6 ABMs to simulate the BMP adoption of farmers

In this project, the main modelling structure inherits from two ABMs that simulate the adoption of farmers, Guo (2018)'s optimizing ABM in the Upper Medway Creek and Zeman (2019)'s typology-based ABM in an Iowa watershed.

Guo (2018) aims to analyze the different impacts of socio-economic conditions on farmers' BMP adoption processes in the Medway Creek subwatershed, and provide some insights for policymakers to make strategies. In Guo (2018)'s project, there are totally six BMP types used as the potential outcomes, while these BMPs are organized as eleven BMP scenarios, and these scenarios serve as the alternatives for farmer agents. For the decision making processes, this model introduced a weighted sum equation to evaluate farmers' preferences for economic, environmental, and social factors based on Brown and Robinson's (2006) work. Especially, the cost and effectiveness of each BMP type, which is an important factor of the generation of economic and environmental scores, is calculated in detail based on previous surveys and literature. Additionally, farmers' knowledge levels and the subsidy rates of BMPs are involved in the decision making processes and examined by the sensitivity analysis. However, this model assumes that all the agents follow the same decision making processes, and the heterogeneity among farmer groups is not illustrated.

Therefore, Zeman (2019)'s model can be a good supplement for this. Similar to Guo (2018), this project also focuses on building an ABM to examine the influences of different factors, but it introduces farmers with different attitudes and preferences to perform the heterogeneity in farmers' decision making processes. In this model, only three BMP types are used, including cover cropping, nutrient management, and drainage water management, and the cost and effectiveness of these BMPs are collected from literature, which is relatively simple compared with Guo's (2018) calculation. However, although both the two models used the weighted sum function to calculate the utility value, besides economic, social and environmental factors, this model especially introduces other two factors, risk aversion and interest in innovation. Particularly, the risk aversion is evaluated by the income growth rate every year, and the innovation is presented by the amount of work of different BMPs, which is assigned by the modeler. Unlike Guo (2018)'s model, the impacts of subsidy rates and knowledge levels of farmers on BMP are not the concerns of this model. Instead, the sensitivity analysis used in this model is aimed at studying the influence of farmer demographics, which are represented by the typology of farmers, change in BMP adoption. In this model, the farmers are divided into six groups that have different preferences to social, economic, and environmental concerns, risk aversion, and interests in innovation these five factors, including business, conventional, environmental, innovative, supplemental, and investor farmers. This group categorization can introduce the heterogeneity among farmer agents and serve as an alternative way to simulate the decision making processes.

In this thesis, the main ABM structure is the integration of the simplified two models in order to make the decision making models comparable. Specifically, since the study area of the proof of concept is the same as Guo (2018)'s model, the BMP categorizations and economic score calculation mainly inherit from it. In addition, to make the decision making model comparable, the utility function only takes the economic, environmental, and social factors into considerations. Correspondingly, the farmers' typology is simplified into four types: business, environmental, conventional, and hobby farmers. According to Railsback & Grimm (2019), the sensitivity analysis is aimed at examining whether the results are sensitive to parameters and structure change. As in both of the two models, the sensitivity analysis has already been performed, and the objectives of this thesis are to contrast the decision making processes, the sensitivity analysis will not be a major concern in this project.

As mentioned before, the main difference of Zeman's work from Guo (2018)'s project is to introduce farmers' typology, which is a way to creating distinct groups of farmers that have different characteristics, to perform different decision making processes and reactions, and this can present the heterogeneity among the agents' group. Therefore, it is necessary to provide some information and background about farmers' heterogeneity and typology.

2.7 Farmers' typology

Heterogeneity among agents is still a significant concern in decision making processes of ABMs (Brown & Robinson, 2006), since it can have a marked impact on the agent behaviors and the model output (Uchmański, 2000). Therefore, it is important to figure out an appropriate approach to represent the heterogeneous agents. As many researchers have studied the typology of farmers, which divides farmers into different types with distinct attitudes towards their land, it provides an opportunity to involve it into the decision making processes to represent the heterogeneity among the farmer agents. For different types of farmers, they can have distinct preferences towards factors.

Based on the analysis of data and the Zeman's(2019) previous work about farmers' typology, to make it appropriate with the possible BMP choices, there are four groups of farmers that will be simulated in the model: business, conventional, environmental, and the hobby farmers.

Business farmers are those who mainly tend to maximize their farm economic profits, and these farmers are also called "profit maximizer" farmers (Malawska &

Topping, 2016). They are the most common groups of farmers, and according to Guillem et al. (2012), they usually have a negative attitude towards ecological farming. Therefore, it is reasonable to assign this group of farmers the highest economic preference values, while they have the lowest environmental preferences.

Conventional farmers, or traditional farmers, are the second large group. These people have a higher concern about the social and environmental aspects than the business farmers. They tend to maximize their yield and use more fertilizer and preserve their rural lifestyle (Malawska & Topping, 2016; Daloğlu et al., 2014). Additionally, compared with the environmental factors, they are more concerned about the social factors (Guillem et al., 2012). Therefore, they have the highest social preference, high environmental importance, and low economic concern.

The environmental farmers, as its name shows, have the highest concern about the environmental aspects. Compared with the above two groups, these people have a smaller proportion. (Guillem et al., 2012) In addition, this group has a strong willingness to reducing the harmful impacts of agricultural activities on the environment (Malawska & Topping, 2016).

Hobby farmers are a small proportion of the farmer population (Guillem et al., 2012). The main income resource for them is off-farm, and this means they show a relatively low interest to maximize their economic farm profit. They usually own a small size farm and have an interest in improving the environmental conditions. Since their main income is not from farming activities, they tend to use less labor on farming (Daloğlu et al., 2014).

2.8 Summary

This section starts from introducing the features of complex systems and their importance. Due to the special features, ABMs can respond to them and work as an alternative to study the systems with complexity. Additionally, three decision making structures of agents in ABM are introduced and will be further discussed in this thesis. However, it is still a challenge to decide whether the ABM is acceptable, the POM can a useful tool for this. Especially for examining agents' behavior based theory, the POM can work as a virtual laboratory, and this can be helpful to contrast decision making structures of agents in ABMs. The most special decision making structures examined in this thesis is a machine learning model. Based on its strong ability to get the underlying information for raw data, it can satisfy the requirements of heuristic models, whose characteristic is also to make decisions based on empirical data. However, there are few examples that integrate machine learning models into ABM, and this thesis can try to fill this gap.

After we introduce the background of modelling, the information of BMP and its adoption, which is the proof-of-concept case used for our model to implement, is introduced. Because the BMP has different kinds of categorizations, only the BMP used in this thesis is described. Next, two previous projects that simulate farmers' BMP adoption developed by Guo (2018) and Zeman (2019), are introduced in the following part, and this thesis mainly inherits from these two projects, especially for Guo (2018)'s work, including the study area, environment design, and value calculations. However, Zeman (2019)'s work is also important since it introduces farmers' typology to present the heterogeneity among farmers, and this is what the stochastic model in this thesis relies on. To provide some information and backgrounds about this, farmers' typology is also introduced, while since there are many methods to conduct the categorization, we only introduced the classification method used in this thesis like in the BMP part.

Chapter 3 Methodology

In this section, the main method used in this thesis will be introduced. There are two topics that will be discussed in this chapter, including the model structure and the practical case works as the proof-of-concept to examine and compare the three decision making structures, optimal model, stochastic model, and the heuristic model. The main reason to divide the methodology into the two parts is to emphasize the main goal of this thesis is not to analyze a specific case, but to contrast the three different decision making models that can guide the actions of agents. Hence, the model structure is a major part of this section, and it will describe the ABM with the Overview, Design Concepts, and Details protocol (ODD). Next, this section will introduce the information about the practical case used in this thesis. In addition, with the model documented, we will discuss the approaches that can describe and evaluate the results of our models.

To make the three decision making models comparable, we used the same ABM framework to implement them. The main difference among the three ABMs is the decision-making part. To achieve that, the ABM to simulate farmers' decision-making process on BMP selection and adoption is used based on Guo (2018) and Zeman (2019)'s work. Some sensitivity analysis and optimizations of variables are conducted in their paper, which provides an additional reference for this model to focus on the comparison of the decision-making methods.

The computational model is implemented in the Repast Simphony 2.7 platform.

Repast Simphony is a Java-based agent based modelling system on Windows, Mac OS X, and Linux. The various and large amounts of Java libraries make it convenient to implement complex sub-model, import different kinds of data, and other features, especially for the machine-learning model in this project. The performance of machine learning models relies on the Weka 3 library. Weka is an open source machine learning software with a Java API. The import of Weka library helps to integrate the machine learning model in the Repast Simphony platform. The optimization of variables in the model inherits from historical data and ABMs of Farmers' adoption of BMP designed by Guo (2018) and Zeman (2019). The machine learning model used in this project is the random forest, and the dataset on which random forest is based is from the Clean Water Program (n.d.).

3.1 The overview of the model process

Before we start to go into the detailed stage documented by the ODD protocol of this ABM, to have a better understanding of the model process in this project, we can first introduce the overall workflow of this ABM as in Figure 3-1.

As shown in Figure 3-1, the geography context which contains the environment of this model and agents is created in the first stage. There are two types of agents in this thesis work, including the farmer agent and the fieldcell agent. The fieldcell agents represent the agricultural land parcels, while the farmer agents represent the owner of these fieldcells. Both farmers and fieldcells have their own attributes.

With the generated agents, the initialization of the landcover and BMP adoption will be started in the model. Since the fieldcells are agricultural land parcels, the landcover here represents the crop types in the fieldcells, and the BMP adoption on this fieldcell will also be initialized randomly. After these initialization processes, we can start the model runs.

The time step in this ABM is one single year, and in each time step, the landcover of fieldcells will be updated following a given landcover scenario. The landcover scenario is based on the landcover change patterns in the Upper Medway subwatershed analyzed by Guo (2018). There are three major land cover change patterns in this area, including the corn-soybean rotation, corn-soybean-wheat rotation, and hay with no change. The proportions of the three landcover change patterns are respectively about 30.1%, 29.1%, and 40.8%. Hence, we build the landcover scenarios based on these three landcover change patterns, and the landcover scenarios also follow the proportions of these landcover change patterns.

In each time step, after the landcovers of the fieldcells are updated, the current BMP adoption of fieldcells will be examined. If the current BMP adoption reaches its lifespan, the agent will start to use the decision making models to decide the BMP adoption, while if the current BMP adoption does not reach its lifespan, we can examine whether the model reaches its time span. If so, the model will be stopped, while if the model does not reach its time span, another time step will be run.

There are three decision making models in this ABM, including the optimal model, stochastic model, and the heuristic model as demonstrated by Table 3-1. As shown in the table, both the optimal model and stochastic model rely on the score calculation of the economic, environmental, and social factors, while the machine learning model depends on the training dataset. Compared with the optimal model, the main difference of the stochastic model is to introduce a stochastic element to represents farmers categorizations, including the business, conventional, environmental, and hobby farmers. The hobby farmers' income is mainly composited by the off-farm income, while the others rely on the on-farm income. Since the decision making models are the main objectives to be compared, in this overview we will not go into the detailed descriptions of them. The detailed description of the decision making models will be described in the submodel part of the ODD protocol. In the next section, we will start to describe the details and elements of the ABM by the ODD protocol.

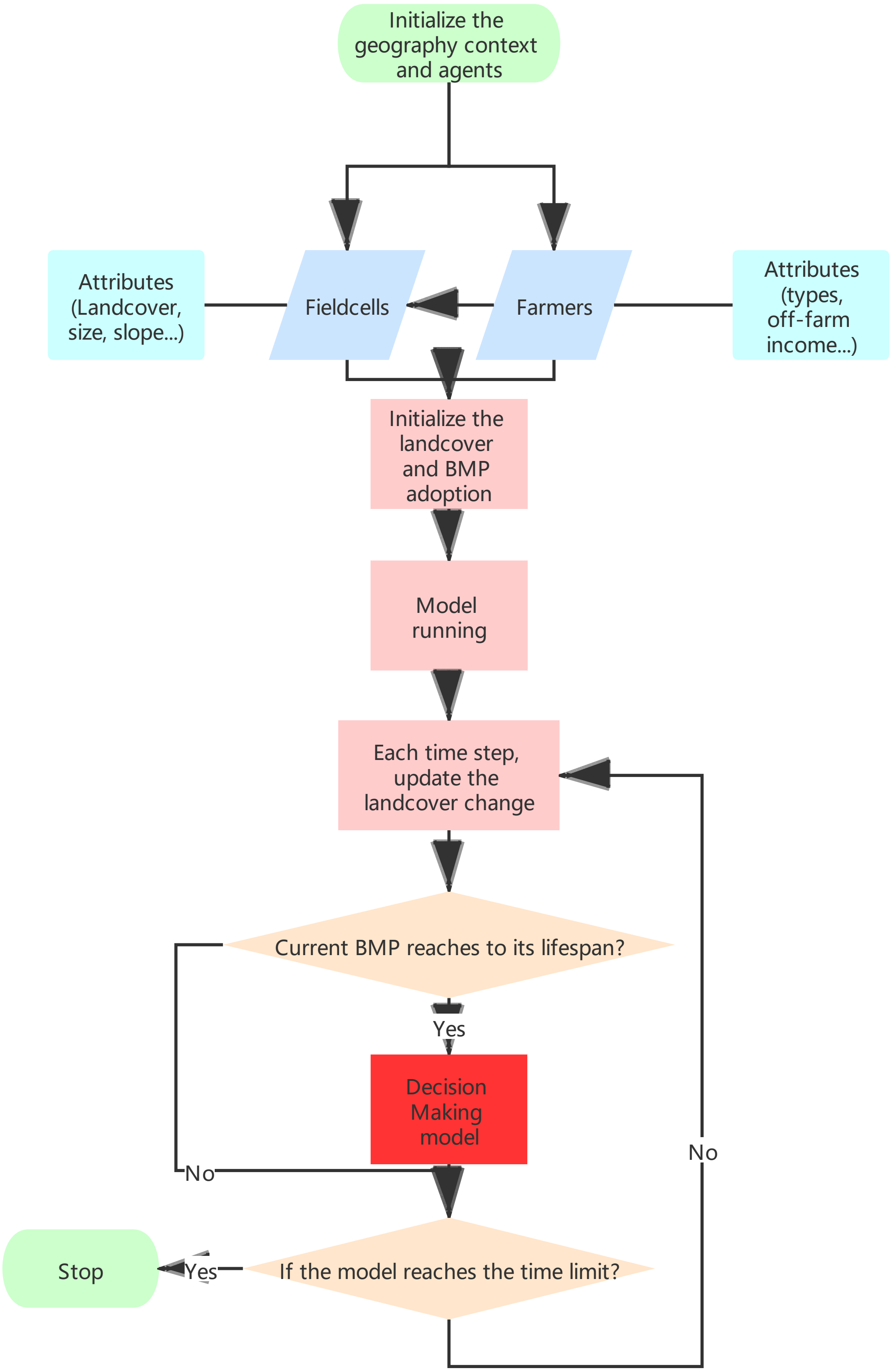
Initialize the geography context and agents
Attributes (Landcover, size, slope...)
Fieldcells
Farmers
Attributes (types, off-farm income...)
Initialize the landcover and BMP adoption
Model running
Each time step, update the landcover change
Current BMP reaches to its lifespan?
Yes
Decision Making model
No
No
Stop
Yes
If the model reaches the time limit?

Figure 3-1 The overall workflow of the ABM in this project

Table 3-1 The characteristic of the three decision making models

Decision making models	Main evaluation approaches of possible choices of agents	Factors/data used for the evaluation
Optimal model Stochastic model	Utility function based on score calculation	Economic scores
		Environment scores
		Society scores
Heuristic model	Machine learning model (Random Forest)	Training dataset

3.2 ODD protocol

The ODD protocol is a widely used and standard protocol to describe and document the ABM model, and it leads to more efficiency for readers to understand the structure and the details of models (Grimm et al., 2020). Although there are also some updated versions of this protocol like ODD+D or ODD+2D, the examples of these documentation methods are not so adequate, and the ODD protocol can satisfy the need of this model. Hence, we choose to use the ODD protocol, and describe the decision making structures in the submodel part. Generally, it consists of three main sections: overview, design concepts, and details. Particularly, the ODD protocol in this section mainly describes the ABM structure that aims to simulate farmers' BMP adoption, while the goal of the whole thesis project is to contrast the decision making structures through this ABM documented by the ODD protocol.

3.2.1 Purpose

The model is designed to provide an opportunity to simulate the BMP adoption decisions of farmers with three alternate decision making models, including the optimal model, the stochastic model, and the machine learning based heuristic model. All these three decision making models are documented in the submodel part of this ODD protocol, as they are used for simulating the agents' actions and decisions. The optimal model depends on a weighted sum utility function based on the economic, environmental, and social factor scores, and the stochastic model only adds a stochastic element to simulate farmers' typology to the optimal model. As both of the two models are based on the utility function, and the utility function is calculated by three factor scores, including the economic, environmental, and social factors, both of the two utility-based models depend on the calculation of these three scores, as shown in Table 3-1. For the machine learning based heuristic model, it works in a different way by introducing a random forest model for agents' decision making. To examine the difference among these three models, we set four different scenarios, including the agents with the three decision making structures, and a baseline case in which agents make decisions randomly. Specifically, what types of BMP will farmers choose to adopt in their farmland? What patterns will be generated with three types of decision making models, in terms of spatial analysis and categorical statistics? With the comparison of the pattern generated, the project tries to explore the suitable cases for the implementation of three decision making models, and the appropriate model to be used for studying the farmers' dynamics of BMP adoption.

3.2.2 Entities, state variables, and scales

Entities

The entities in this model consist of two types: farmers and fieldcells. Fieldcells are agents that represent agricultural land parcels, while farmers are their managers. Every farmer and their fieldcell follows a one-to-one correspondence to simplify the model structure and reduce the complexity in model running.

The fieldcell has the following attributes: current year, area, length, current BMP adoption, corresponding farmer and availability for grassed waterway, buffer strip, Wascob, and the windbreak, which means whether a certain BMP can be adopted on the fieldcell, which is calculated before model building based on the standard provided by Guo (2018). Since this model has three types of decision making structures as the submodels, and among them, the machine learning model depends on a training dataset that includes some attributes that are not important to the other two utility-based models, to make these models comparable, both these data and the attributes which the utility calculation relies on are integrated into the fieldcell.

For the first two decision making models, the optimal model and the stochastic model, the current crop type on this land, and the land cover change scenario are represented as two attributes in the fieldcell agents. Moreover, to get the social information, the neighbor fieldcell list and the corresponding BMP adoption are also necessary. The neighbor fieldcell stores the fieldcell ID in the neighborhood, and the neighborhood is defined as the fieldcell within the 30m buffer of the fieldcell agents. For further analysis, the maximum utility score is also stored in the fieldcell agent. As for the machine learning model, since the fieldcell agent has all the attributes that are used for training the random forest, the details of these attributes will be described in the submodel part.

Farmer agents are relatively less complicated than the fieldcell, only the farmer type, their off-farm income, and the economic, environmental, and social preference weights. For this type of agent, the difference between the optimal model and the stochastic model has the most significant impacts. When the optimal model randomly assigns the preference weights, the stochastic one generates the preference weights based on the categorization of farmers.

State variable

There are two kinds of agents in this project, including farmer agents and the fieldcell agents. For the farmer agents, the agent state variables include farmer ID and its preference type. Both of the two variables are represented by integers. For the fieldcell agents, there are multiple state variables as shown in Table 3-2.

Table 3-2 The state variable of the current BMP

State variable	Description	Type
ID	The id of the fieldcell	Integer
BMP availability list	What kind of BMP can be applied to this land	List
Farmer	The landowner of this fieldcell	Farmer agent
Landcover scenario	The landcover scenario of the fieldcell	List
Neighbor list	The lists of neighbor fieldcells of this agent, and neighbor is defined as other fieldcells within 30m distance of the agent	List
BMP adoption	The current BMP adoption	BMP
Year	How many time steps after the model running	Integer

Scale

The temporal resolution of this model is one year. Every year the crop type of the fieldcell will change following the corresponding landcover scenario. Moreover, since the lifespan of most BMP is 10 years long or its multiples, every ten years the farmer will choose to adopt a BMP based on the decision making model. The temporal extent is not necessarily restricted, but in this model, we choose 100 years as the whole time period, although 100 years are too long and unrealistic for the practical case. The reason for choosing 100 years is that the model starts from a random state, while 50 years is the longest lifespan for BMPs, so 100 years can fully capture the 2 cycles, including the initial choice of BMP adoption and the decision made after an initial choice. Therefore, we have to assume that farmers or the family of farmers always work on their fields during the model running. Fieldcells in this model are GIS vector data, and as a result, no minimum cell size is necessarily defined.

3.2.3 Process overview and scheduling

The processes in the model mainly include two stages: the change of crop type landcover and the adoption of BMP. Every time step the crop type will change based on the corresponding landcover scenario, while every time the current BMP reaches its lifespan, the farmers will change their BMP adoption based on the decision making models. Although for the Fieldcells that has no BMP employed, farmers can choose to implement BMPs, in this thesis, for the reason that there are not so many No BMP cases and to simplify the complexity in terms of computation, this case is not taken into consideration. Particularly, because the machine learning models depend on its training dataset, and in that training dataset, there are not No BMP cases, it would be more complicated to simulate this case in this decision making structures. The decision making models will be described more clearly and in detail in the submodel section below. The change of crop type is ahead of the BMP adoption here, as the farmers usually have a plan that what crop will be planted in the next years.

3.2.4 Design concepts

Basic principles

As described in the literature review, the ABM can be divided into three categories in terms of decision making structure, including optimal model, stochastic model, and the heuristic model. Because the farmer agents will select the BMP to adopt in their farmland in the model, economic concern, social influence and environmental factors can drive farmers' decision making, which can support us to involve a utility function to calculate the utility value for farmers based on these factors as an optimal model (Guo, 2018). However, farmers can have various categorizations as discussed in the farmers' typology topic of literature review, and this can have an impact on their perspectives on BMP adoption. As a result, a stochastic element can be introduced in this project to simulate the heterogeneity among these categories. Moreover, the real-world BMP installation of farmers dataset can provide an opportunity to involve some data-based machine learning model to predict the BMP adoption, and use the machine learning model as a heuristic rule for ABMs. Particularly, in this thesis, a random forest model is used for this type of decision making.

Adaptation

For the optimal and the stochastic model, there are five types of BMP as the alternatives of decisions: grassed waterway, buffer strip, WASCoBs, windbreak, and no BMP adoption. In these two models, the farmers' decision making is driven by the change of crop type landcover and adoption of BMP in the neighborhood every time the current BMP lifespan is reached. The machine learning model can have two types of decisions, erosion control structures and fragile land retirement, which inherits from the dataset (Clean Water Program, n.d.), because there are no detailed categories like the five types listed above in this training dataset. According to Clean Water Program(n.d.), erosion control structures consist of the grassed waterway and buffer strip, while WASCoBs and windbreak belong to the fragile land retirement category, for it is listed in the description of fragile land retirement of Clean Water Program, even though it may not be so reasonable. There are only few No BMP cases, so it does not make a strong difference in the comparison. To make the three decision making models comparable, we regroup the four BMP into the two classes as mentioned above. Particularly, here the fragile land retirement type actually does not include the land retirement, and this results from the difficulty to simulate the economic revenue to make it as a possible alternative for agents. All the decision making models will be described in detail in the submodel section.

Objectives

The farmers will tend to maximize their utility-based on the function, in which economic, environmental, and the social factors are taken into consideration, in the optimal and stochastic models. This allows farmers to evaluate driving factors from economic, environmental, and social aspects and make a decision. However, in the heuristic model, the result is generated by the trained random forest model based on the installation dataset, and the data mining method can take an advantage of using some information underlying the data. As mentioned in the adaption section, the results of the two utility-based models will be regrouped into two classes as same as the results of the heuristic model. Therefore, through the three decision making models, we can compare the results of the ABMs with them.

Sensing

In this model, all the decision making structures rely on sensing, which means farmer agents will make their decision based on the attributes of the fieldcells they manage and their preference weights or characteristics. Even for the machine learning model, it is necessary to get the attributes of fieldcells to predict the result of the trained model. This means that the farmer agents are supposed to get the information from their own Fieldcells. Additionally, the optimal model and stochastic model also sense the number of same BMP adoption in their neighborhood as a social factor as shown in Table 3-1, and this means that in these two submodels, farmer agents should also collect information from the neighbor Fieldcells.

Interaction

The interactions of agents are represented in different ways for the three decision making models. For all kinds of the model, since the landcover scenarios are distributed proportionally, mediated interactions among are represented. Since if one agent is assigned with a particular landcover scenario, other agents will have a lower possibility to be assigned with the same landcover scenario.

For the optimal model, the interactions among agents are mainly represented by the social factor. As the social factor is represented by the percentage of neighborhood similarity, the actions of agents will have a direct impact on the decision making processes of the neighbor agents.

For the stochastic model, besides the social factor, farmers' typology can also represent the mediated interactions among agents. Similar to the distribution of landcover scenarios, if one farmer agent is assigned with a particular type, other farmers will have a lower possibility to be assigned as this type.

Therefore, in terms of the interactions, all the three models can represent the interactions among agents, while compared with the heuristic model, the interactions of the optimal model and the stochastic model are represented directly, and the interactions of the stochastic model are stronger than the optimal model due to the introduction of farmers' typology.

Stochasticity

The stochasticity used in this model mainly is for two reasons, to simulate the variability and the proportion. Most randomness in this model is used to simulate the variability, including the initialization part, scores of factors calculation as shown in Table 3-1, and the preference weight assignment for each farmer.

In all the decision making models, the randomness of landcover scenario distribution represents the proportion. In addition, in the stochastic model, to assign the categorization of farmers, the randomness that represents the proportion is also used. This stochastic element can reflect the proportion of different types of farmers, and further becomes the main difference of the stochastic model from the optimal one.

3.2.5 Details

Initialization

In this model, firstly, a geography context is built for model graphical display, and farmers are created in the geographical context for fieldcell initialization. Each farmer is assigned with a categorization based on the proportion of the typology of farmers as described in the stochasticity section. Although many farmers have off-farm jobs, in this thesis, to present the difference among types of farmers, only every hobby farmer will get an off-farm income, while other types of farmers will have a 0 off-farm income. For the optimal model and stochastic model, farmers' preference weights for factors will be generated in different ways. Next, the model starts to read the vector files of agricultural parcels in the study area. Each agricultural parcel, whose features will be imported as attributes of fieldcells, in the file is loaded as a fieldcell agent, and it will be assigned with a farmer agent as the manager from the farmer created earlier. To initialize the land cover information, the crop type scenario file is loaded into the model, and the fieldcell randomly picks a scenario to guide the landcover change during the model processes. Although some crop type inventories can be collected year by year, the actual temporal information about the observation data from Clean Water Program (n.d.) is not clear, and the actual date for these BMP adoptions are not fully recorded, and not at the same year. Hence, we have to use a random selection to initialize land cover. In the input vector data, the neighbor fieldcells and the availability of BMP types have been already calculated and recorded, so the fieldcell saves these available BMPs and the neighbor fieldcells id into two lists. Based on the availability of the BMP, the initialized BMP adoption is randomly generated from these available BMP in the list of fieldcell. After that, the fieldcells get all the current BMP adoption information in the neighborhood based on the neighbor cell list. After all the 167 fieldcell go through all these steps, the initialization process is finished.

Input data

The input data can be categorized into three main types: the data used for creating the environment, the data used for training the machine learning based heuristic model, and the data used for score calculation in the optimal and stochastic models. Hence, this part will introduce these three types respectively. Additionally, because we want to keep the adaptability of this model structure to different cases, the study site, the practical data used in the proof-of-concept will not be introduced here, but in the proof-of-concept section after the ODD protocol.

The data used for compositing the environment mainly refer to the GIS land parcel data and the corresponding attributes to the parcels. The GIS parcel data represent the modelling world and entities of field cells. Since these data are GIS polygon data, parcel ID, the area and perimeter of each polygon will be recorded. To reduce the running time of computational models, the availability of each BMP is defined for each parcel. For simulating landcover change in the model processes, model builders can introduce some crop type change scenarios, and assign these scenarios to the land parcels. In addition, since the model introduces the machine learning model as one of the submodels, all the attributes within the training dataset are supposed to be also included in the parcels.

The data used in machine learning submodels are actually the attributes of training datasets. Since the agents in this project are expected to choose a BMP type, which requires the machine learning submodels to produce a specific type as the results, these training data have to include the BMP types for supervised learning. Additionally, other information about the records in the datasets that can be used for training, but in this project, due to the limitation of lack of other information in the dataset, we try to join other datasets to introduce enough information for training models, and these datasets will be introduced in the proof-of-concept part as they are used in the practical case. The other information is not limited, but in this project what we used for training the models will be included in the submodel part. As mentioned previously, the GIS parcel data should also include these attributes, for the trained machine learning model also requires the corresponding information to generate the results.

The data used for score calculation have three parts corresponding to the three factors in the score calculation. For the calculation of economic scores, the cost of each BMP and the cost, yield and price data of different crops per area unit are supposed to be included. For the environmental factors, the effectiveness of each

BMP is required. Since most of these kinds of data are collected from surveys and statistics, these data are described as mean and standard deviation values. In addition, for the social factors, it is necessary to get the information about the neighbors of each parcel.

Submodels

As shown in Table 3-1, both of the optimal model and the stochastic model require the score calculations of the three factors, including the economic, environmental and social factors. Therefore, before we describe the details of the three decision making models, we will first introduce the way of the score calculations.

Score calculation

Economic score

In both the optimal model and the stochastic model, the model needs to quantify the economic, environmental, and social factors. Therefore, it is necessary to involve a submodel to calculate these three scores.

To calculate the economic score S_{Eco}, the following equation is used:

$$\boldsymbol{S_{Eco} = I_O + I_F - C_F * A - C_B * (1 - R_s)}$$

In this equation, the income is summed by I_o, the off-farm income collected from the farmer agent, and I_F, the farm income. The cost includes the cost in the farm, which is calculated by multiplying the unit cost of the crop type, C_F, and the area of the parcel A, and the cost for BMP adoption, C_B. Since some institutions can provide some subsidies for farmers, the R_S is used to describe the share rate for the BMP adoption. The economic score can be calculated by the total income minus the cost finally like previous farmers' BMP adoption simulation work (Guo, 2018; Zeman, 2019).

To get the farm income I_F for each parcel, the equation is used:

$$\boldsymbol{I_F = A * Normal\left(\mu_p, \delta_p^2\right) * \ Normal\left(\mu_y, \delta_y^2\right)}$$

The farm income is calculated by multiplying A, which is the area of fieldcell, price, and the yield. Price and yield are generated by normal distributions based on their average μ and standard error δ.

For grassed waterway, buffer strip, and WASCoBs, the BMP adoption cost C_B is calculated in a relatively direct way, multiplying the area A of fieldcells, the adoption cost per unit C_{BU}, and LS, the lifespan of the BMP:

$$\boldsymbol{C_B = A * \ C_{BU} * LS}$$

The cost per unit C_{BU} can be calculated by the sum of lower ranges of the unit cost C_L and the difference between C_L and the upper range C_U. To involve some variance, a uniform distribution random generator is used for the difference:

$$\boldsymbol{C_{BU} = Random(0, C_U - \ C_L) + \ C_L}$$

Although the C_{BU} calculation for the windbreak is the same as other BMP types, the total C_B is generated in a different way:

$$\boldsymbol{C_B = L * \ C_{BU} + \ L * \ C_{BU} * R_M * LS}$$

In this equation, the length L of fieldcells works as the unit to calculate the cost. C_{BU} means the cost per unit like other BMP types. The main difference between windbreak and other BMP types is that the maintenance fees are much less than the installation fees. Therefore, the rate of maintenance R_m is used here to represent this, and the result of multiplying of this and the lifespan LS denotes the maintenance fees. As a result, C_B is generated by the sum of installation fees and the maintenance fees.

Environmental score

The environmental score calculation three aspects, the phosphorus loss reduction, sediment removal, and the soil erosion control effectiveness, according to Guo (2018). Similar to the generation of costs per unit C_{BU} for BMP adoptions, these three factors are quantified by these equations:

$$\boldsymbol{P = Random(0, P_U - P_L) + P_L}$$

$$\boldsymbol{Sed = Random(0, Sed_U - Sed_L) + Sed_L}$$

$$\boldsymbol{Ero = Random(0, Ero_U - Ero_L) + Ero_L}$$

In these equations, P, Sed, Ero means the phosphorus loss reduction, sediment removal, and the soil erosion control effectiveness, while the subscript U represents the upper level, and L means the lower level. Hence, the environmental score is directly calculated by the following equation (Guo, 2018):

$$\boldsymbol{S_{Env} = (P + Sed + Ero) * Sc1}$$

In this formula, S_{Env} represents the environmental score for fieldcells, and the Sc1 means the scaling factor. The scaling factor is used for solving the problems of unit. Since the unit of economic scores is a dollar, this can be a large number and dominate the whole decision making processes. Therefore, if the preference weights to the economic, environmental, and social indicators are the same, the involvement of scaling factors can rebalance the value of the three scores to make them balanced.

Social Score

The social score is mainly presented as the proportion of neighbors that adopt the same kind of BMP in all the neighbors of fieldcells suggested by Zeman (2019). To make the results of the optimization models and the machine learning model comparable, as mentioned in the adaptation section, the grassed waterway and buffer strips are reorganized as the erosion control structure, and the WASCoBs and windbreak are regarded as the fragile land retirement type. Therefore, the social score S_{soc} can be generated by this equation:

$$\boldsymbol{S_{soc} = \frac{N_s}{N_t} * Sc2}$$

As described, the proportion is calculated by N_s, the number of neighbors that adopt the same type of BMP, divided by the number of neighbors, N_t. For the same reason as the environmental indicator, a scaling factor Sc2 is used to balance the magnitude of all the three factors.

With the economic, environmental, and social scores, the utility value for different BMP types can be calculated as suggested both by Guo (2018) and Zeman (2019):

$$\boldsymbol{u = S_{Eco} * W_{Eco} + S_{Env} * W_{Env} + S_{soc} * W_{soc}}$$

In this function, $W_{Eco,}$ W_{Env}, and W_{Soc} represent the farmers' preference weights of economic, environmental, and social aspects. Additionally, the farmers will choose the one with the highest utility value from the available BMPs or no BMP, if the utility value is negative. However, the determination of preference weights is still a problem, while it is also the main difference between the optimization model and the stochastic model.

Decision making model 1: the optimal model

This model is generally inherited from Guo's (2018) decision making structure. In this model, the farmer agents have the ability to evaluate the three main factors including economy, environment and society, and use a weighted sum model to optimize the best actions for agents. However, to make these submodels comparable, it is still necessary to change some elements.

In this model, the only impact of the typology of farmers is the determination of off-farm income, which means the preference weights to three factors are chosen in a more homogenous way, and the reason for this assumption is to highlight the difference among groups of farmers. In this model, the preference weights of all the farmers are generated following one rule: the economic preference weights are always the highest one among the three factors (Guo, 2018). Specifically, the economic weight W_{Eco} is generated by uniformly distributed random numbers in the region [0,10] firstly, and then the environmental W_{Env} and social weights W_{Soc} are generated by a uniform distribution in [0, W_{Eco}]. After three weights are determined, a normalization method is used to restrict the sum of three weights to 1 by dividing each weight by their sum.

Every time when the BMP of the fieldcell reaches its lifespan, the fieldcell will evaluate the three scores and calculate the utility values for each type in their available BMP list. If the highest utility value is positive, the agent will choose the corresponding BMP. While the utility value is negative, agents will not adopt any BMP type, which is performed by choosing the “No BMP” type.

Decision model 2: the stochastic model

This model is inspired by Zeman (2019), and it is also an improvement to the optimal model. Although this model also relies on the utility calculation to adopt the most appropriate BMP as described in Table 3-1, the typology of farmers works as a stochastic element to determine the preference weights. Hence, in the optimal model, the preference weights for different factors of all the agents are generated by the single method described in the previous part, even though the involvement of randomness will increase some variations, while in the stochastic model, all the agents will be firstly divided into several types, and for each type, they will have different priority to factors, so the preference weights are generated in different ways.

Before generating the preference weights for different categories of farmers, it is necessary to get the categorizations and their proportions. The typology of farmers is generally inherited from Zeman’s (2019) work, although similarly, to make these models comparable, there are some changes made in this model. For example, in Zeman (2019)’s work, besides economic, environment, and society, there are still two kinds of factors that are supposed to be evaluated, including risk aversion and innovation, and the BMP used and the way of calculating the scores are also different from what in this project. As the risk aversion and innovation factors require some data to evaluate the risks and the innovation of BMPs, and we do not have this kind of data, these two factors are not concerned in this model. In this model, there are four categories of farmers: business farmers, conventional farmers, environment farmers, and hobby farmers. Farmers from different categories have divergent perspectives to these three factors. The preference rank suggested by Zeman (2019) is presented in Table 3-3, and in this table, the highest rank is 1, and the lowest rank is 3.

Table 3-3 The preference rank for different types of farmers

	Economics	*Environment*	*Social*
business farmers	1	3	2
conventional farmers	2	3	1
environment farmers	2	1	3
hobby farmers	2	2	2

With the rank of preference for different categories of farmers, the preference weights of three factors can be generated by uniformly random distribution. To restrict the differential degree of these weights, the uniform distribution is performed in the fixed region for divergent levels of preference. Similarly, the utility will be calculated based on the preference weights and factor scores, and the BMP with the highest utility value can be picked from the available BMP list every time when the current BMP reaches its lifespan.

Decision model 3: the machine learning based heuristic model

The third decision making model is quite different from the first two models in input data, model structures, and adaptive choices. As mentioned previously, although the statistical method like machine learning model is a possible alternative used as a heuristic rule in ABMs, few researchers practically implemented it with ABMs in this way, and its performance compared with other decision making structures is still not clearly examined, which composite the gaps that this thesis tends to fill, so this decision model is the most important one in this thesis. The input data of this model are collected from the real-world case rather than the calculation results from literatures. Since the objective of this ABM is to predict the adoption of farmers on BMP types, it can be transformed into a classification problem: with the given information in the fieldcells, how can we classify the land into several BMP types? Therefore, some machine learning models that solve the classification problems can be introduced into the ABM as submodels. Among a variety of machine learning classifiers, the random forest is chosen for this. The reasons for choosing the random forest can be explained in four reasons:

(1) Since there are 733 records in the dataset, and the data quality is not at a high level, a classifier that has a high fitting ability is supposed to be used for higher accuracy. Although this may also lead to high overfitting, a cross validation can be used to solve this.

(2) There are only two categories as results of the classification, as described in the adaption part of the ODD protocol previously, and this low complexity allows the

implantation of classifiers with a relatively low generalization capability.

(3) The types of attributes to train the model are various, including the numerical, ordinary, and string types, and this makes it hard to use some classifiers like logit regression, since they cannot directly deal with the categorical data like the string types.

(4) Compared with other classifiers with similar high accuracy like SVMs, the execution speed of random forests is significantly rapid, and its parameter tuning processes are relatively simple. Especially, since all the 167 agents have to use it every time when they make decisions, the time has been attached to great importance.

Hence, the random forest is a good choice here as it can satisfy these requirements. In the random forest model, there are 16 attributes used as explanatory variables, and the attribute of type is the response variable. The name and descriptions of these attributes are shown in Table 3-4. The reason for choosing these attributes is based on the availability of BMP as described by Guo (2018). The detailed source of these attributes will be described in the proof of concept section, as they are practically used in this case.

Table 3-4 The descriptions and types of the training data for the random forest

ATTRIBUTES	DESCRIPTION	TYPE
PRECITIPATIONINTER	The indicator of precipitation collected by spatial interpolation	Numeric
TEMPINTER	The indicator of temperature collected by spatial interpolation	Numeric
SPIINTER	The SPI collected by spatial interpolation	Numeric
SLOPE1	The slope collected from DEM	Numeric
CROPVALUE	The crop type collected from the annual crop inventory	String
SOIL_COMPLEXITY	The number of associated soil components	Numeric
SOIL_PERCENT	The percentage of the polygon that the dominant components represent	Numeric
CLI1	The soil capability for the production of crops	String
DRAINAGE	How wet the soil drains	String
DR_DESIGN	The drainage design code reflects the soil drainage characteristics	String
HYDRO	Estimation of runoff from precipitation	String
ATEXTURE	Soil texture of the surface	String
K_FACTOR	Soil erodibility factor	Numeric
POLY_AREA_ACRE	The area of the land parcel	Numeric
WATER_AREA	The area of land influenced by water body or water courses	Numeric
WATER_PERCENTAGE	The percentage of WATER_AREA in the total area	Numeric
TYPE	The adopted BMP types, 1 means the erosion control structure; 2 means the fragile land retirement	1 or 2

As mentioned in the literature review part, there are three main parameters that need to be tuned: the maximum number of features that can be used in a single tree, the number of estimators in the forest, and the size of the terminal nodes. Particularly, increasing the number of estimators can improve the ability of prediction, while it also leads to a long running time of the model, while decreasing the size of terminal nodes will also increase the accuracy of the model, but is possible to cause the overfitting problem. As the size of the dataset is relatively small, the accuracy has a higher priority than the overfitting issue. Therefore, the size of terminal nodes is set to 1, and the number of estimators is set to 100 to get a more accurate result, while the maximum number of features is the logarithm of the amount of all the features to the base 2, which is the default and fixed option of Weka. Additionally, to reduce the complexity in terms of computation, the random forest model is trained only at the beginning, rather than for every agent, since we only have one training dataset, and this dataset is the same for all the agents, so it is not necessary to train the model for every individual agent.

To reduce the overfitting and test the accuracy of this model, the 10-fold cross validation is used. This method allows to divide the training data into 10 folds randomly, and totally build 10 random forest models. Each time when one model is trained, only 9 of 10 folds of data are used, and the other one is used to test the result. As a result, the average result of the 10 models is used as the output, and the accuracy of the prediction can also be collected.

Therefore, when the current BMP reaches its lifespan, the agent will make a decision based on the trained random forest model. To reduce the processing time, the random forest model is trained and used before the running of ABM, and when the decision making stage starts, the agent will just directly implement the random forest to classify their future BMP type.

3.3 Proof-of-concept

The proof-of-concept is mainly inherited from Guo (2018)'s work, and this means we adopt the same study area and input setting about the model environment for the evaluation of economic, environmental, and social factors and resources. However, the main modelling structure follows the one described in the ODD section. Therefore, this part will mainly introduce the study area and the practical input data in the proof-of-concept.

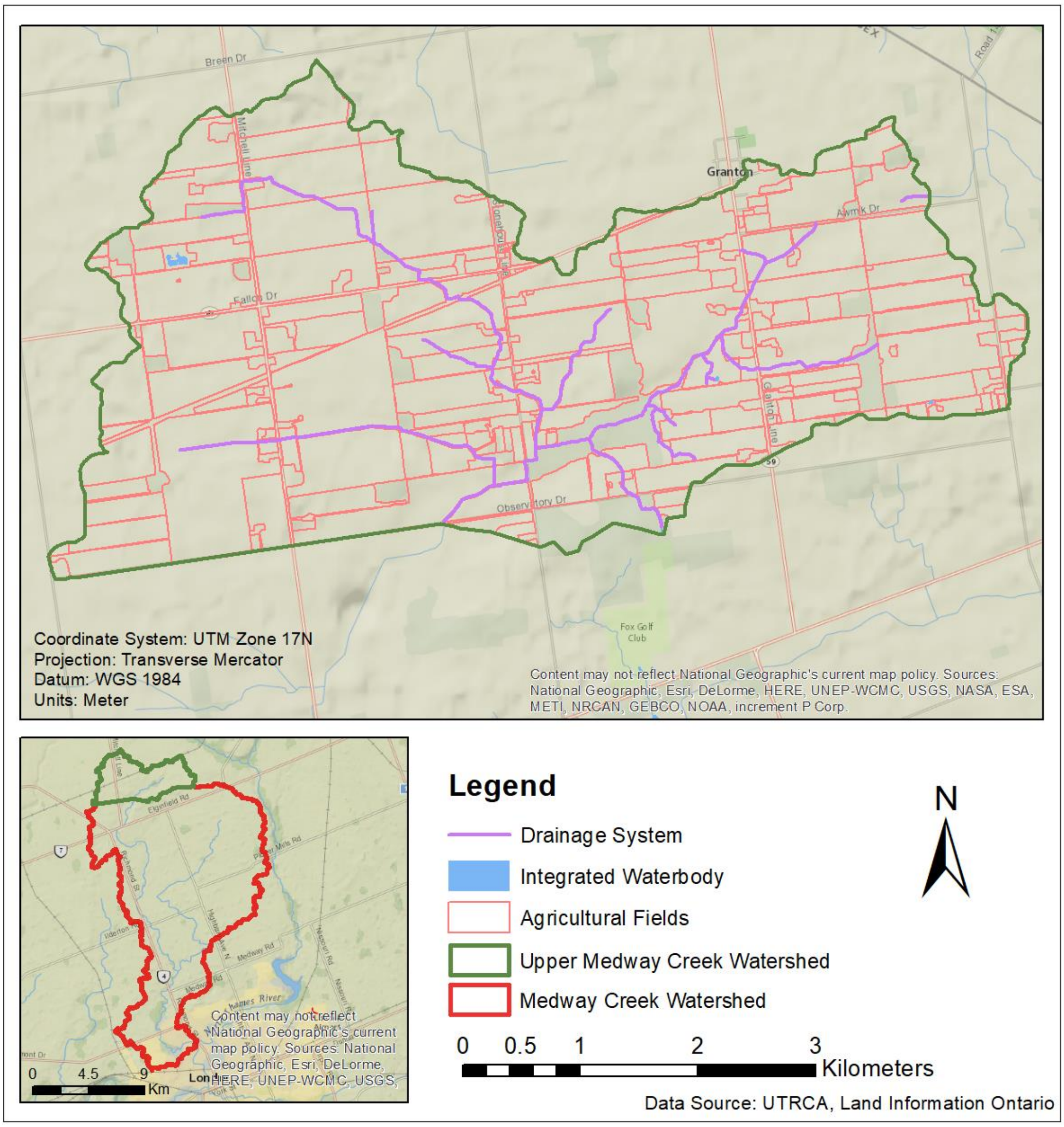
Breen Dr
Mitchell Line
Fallon Dr
Granton
Avonk Dr
Granton Line
Observatory Dr
Fox Golf Club
Coordinate System: UTM Zone 17N
Projection: Transverse Mercator
Datum: WGS 1984
Units: Meter
Content may not reflect National Geographic's current map policy. Sources: National Geographic, Esri, DeLorme, HERE, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, increment P Corp.
0
4.5
9
Km
Content may not reflect National Geographic's current map policy. Sources: National Geographic, Esri, DeLorme, HERE, UNEP-WCMC, USGS,
Legend
Drainage System
Integrated Waterbody
Agricultural Fields
Upper Medway Creek Watershed
Medway Creek Watershed
N
0
0.5
1
2
3
Kilometers
Data Source: UTRCA, Land Information Ontario

Figure 3-2 The boundary map of Upper Medway Creek subwatershed (Guo, 2018)

3.3.1 Study Area

The study area of the proof of concept is the same as Guo (2018)'s work, and it is presented in Figure 3-2. The Medway Creek watershed is used for the environment of model simulation. It is one of the subwatersheds in the Upper Thames Watershed in southwestern Ontario, Canada. The Medway Creek covers the area of 205 km^2, including parts of the City of London, the Municipalities of Middlesex Centre, Thames Centre, and the Township of Lucan Biddulph (UTRCA, n.d. d). It takes about 6% of the Upper Thames River watershed. According to 2017 Watershed Report Card (2017). The dominant land use in this area is the agricultural type, and it accounts for 82% of the whole land use type. Additionally, there are three main soil types in the Medway Creek watershed, including clay loam, silty loam, and silty clay loam.

According to UTRCA (n.d. e), the Medway Creek watershed is identified by UTRCA as a priority for environmental enhancement in 2007, and the surface water quality has remained steady at the grade of D (2017 Watershed Report Card, 2017). Phosphorus concentrations, which is one of the major factors influencing the surface water quality, have been improved over the long term, and reach a lower level than the Upper Thames average, but still remains at 2 times the provincial aquatic life guideline (2017 Watershed Report Card, 2017). To further improve the surface water quality, some projects have been implemented in this region. For example, the watershed owners have completed 34 Clean Water Program projects that help them to adopt BMPs, which are also the resources of the training data of the machine learning models. According to 2017 Watershed Report Card (2017), in these Clean Water

Program BMP projects, fragile land retirement and the erosion control measures are the main types.

Since we choose the Upper Medway Creek subwatershed as the study area of our proof-of-concept, to have a clear description of the case, next we will introduce the practical data used in this proof-of-concept to adapt the study area. Particularly, the structure to introduce the practical data will follow the input data part in the ODD section.

3.3.2 Practical Input Data

As described in the input data part of the ODD section previously, the input data can be divided into three categories: data used for compositing the environment, data used for training the machine learning submodel, and data used for score calculation.

The data used for compositing the environment mainly refers to the GIS parcel data. In this project, the GIS parcel data are mainly inherited from Guo (2018)'s work. There are 176 parcels digitized as agricultural land by this work, and the availability of each BMP is also determined. Additionally, Guo provided some land cover change scenarios in this region, and this makes it easier to simulate the change of crop types for each agricultural parcel. Nevertheless, since the machine learning submodel will also be examined in this project, other attributes in the training data will also be joined to the GIS parcels. As most of them come from the same resources, we will discuss them in the following part.

The data used for training the machine learning models are mainly based on the funded BMP implementation data collected by Clean Water Program(n.d.). In that raw dataset, there are totally 2910 records, and 17 types of BMP. However, due to the lack of data description, some of the BMP types are not clearly described like the BMP type called GPS, and some of them are strongly limited by the local conditions, such as the BMPs like wellhead protection and decommissioning unused wells. In addition, according to 2017 Watershed Report Card (2017), fragile land retirement and erosion control measures are the two main types in the Medway Creek watershed. As a result, we finally choose these two types as the possible choices for agents, and in the raw datasets, only records that adopt these two types are kept in the dataset. After removing some missing values, there are still 733 records for our training work. However, in the raw dataset, only geographical coordinates and adopted BMP types can be used for model building, and this is obviously not adequate to train a model. Hence, we used spatial join to link attributes from other sources with these records as shown in Table 3-5. Particularly, the description of all these attributes in Table 3-5 is recorded in Table 3-4.

Table 3-5 The sources of the attributes used in the training data of the random forest.

ATTRIBUTES	SOURCES
PRECITIPATIONINTER	Interpolated based on Weather Station data in southwestern Ontario
TEMPINTER	
SPIINTER	Interpolated based on Standardized Precipitation Index (Ontario Ministry of Natural Resources and Forestry [OMNRF],2016)
SLOPE1	Interpolated based on Ontario Digital Elevation Model (OMNRF, 2019)
CROPVALUE	Collected from Annual Crop Inventory 2011(Agriculture and Agri-Food Canada [AAFC], 2011)
SOIL_COMPLEXITY	Collected from Soil Survey Complex (OMNRF, 2015)
SOIL_PERCENT	
CLI1	
DRAINAGE	
DR_DESIGN	
HYDRO	
ATEXTURE	
K_FACTOR	
POLY_AREA_ACRE	Calculated from Regional Municipality of Waterloo Property Parcels (Teranet Incorporated, 2018)
WATER_AREA	Calculated from Ontario Hydro Network (OMNRF, 2020)
WATER_PERCENTAGE	

The data used for score calculation in optimal and stochastic models as described in the ODD sections are corresponding to three factors in the calculation. Therefore, the cost and effectiveness of BMP, yield, cost and price of crops, and the neighbors of each land parcel are required. In this project, due to the same study area, these data are mainly inherited from Guo (2018)'s work. For example, the neighbors of each parcel are collected from that work. Other sources of these data are recorded in Table 3-5.

Table 3-6 The sources of data used for score calculation in the proof-of-concept

Data	*Sources*
Cost of *BMPs*	
Grassed waterway	Kansas (1989)
Buffer strip	Mtibaa et al. (2018)
WASCoBs	Kansas (1989)
Windbreak	Kansas (1989)
Effectiveness of BMPs	
Grassed waterway	Kansas (1989)
Buffer strip	Hawes and Smith (2005)
WASCoBs	Kansas (1989)
Windbreak	López et al. (2017)
Yie*lds of Crops*	OMAFRA (2017)
Costs of Crops	OMAFRA (2016)

Price of Crops	OMAFRA (2018)

With the three types of data, we can input them into the ODD structure described previously and build the computational models to produce some results for the analysis of differences among optimal model, stochastic model, and the heuristic model in this proof-of-concept. In the next part, we will introduce the scenarios and metrics used in this thesis for comparison and the way to analyze the results.

3.4 The approaches to analyze results

In this part, we will firstly introduce the modelling scenario we run and what metrics used for evaluating the performance of BMPs with different decision making submodels. In addition, we will use the coefficient of variation to examine whether the model can produce a stable result in the given times of running.

To contrast the three decision making models as described in the submodel part of the ODD protocol, there are totally four scenarios used in this thesis, including the ABMs with the three decision making models, and the null model in which the decisions of agents are made by a uniform distribution to reject the null hypothesis that the agents made their decision randomly. For the three decision making model scenarios, the only difference among them is the decision making models that work as the action rules of agents, and the other parts of the models are kept the same. For the model in which decisions of agents are made randomly, it works as a baseline for the comparison, and in this scenario, it can illustrate the impacts of the involvement of decision making models.

Since to my knowledge, although there is some literature focusing on farmers' decision making processes on BMP adoption (Liu et al., 2018), most of them focus on the microlevel, which means the individual farmer, while it is not very common to have a clear description of the macrolevel, which refers to the pattern in the whole study area, especially for the case in which there are interactions among individual farmers. In addition, the contrast of decision making models are usually conducted in a theoretical framework, especially for a machine learning based heuristic model, (Schreinemachers and Berger, 2006; Groeneveld et al., 2017; An, 2012), and this leads to the challenge that few metrics can be considered appropriate to describe the comparison results of models, especially in a specific case of farmers' BMP adoption. As a result, the only reference for the model comparison is the original observations that trained the random forest model. However, due to the different scales in the study area and the uncertainties among the observations data from Clean Water Program (n.d.), it is hard to compare the results in the quantitative levels.

To solve this problem, as discussed in the literature review part, the POM can be an alternative for the theory test and further development. Therefore, in this project, the POM will be implemented as a tool for the analysis of results from the three decision making structures. According to Railsback and Grimms (2019), the patterns used in the POM usually are 3 to 5, qualitative and diverse characteristics. Ideally, if there is any material that focuses on the macrolevel patterns of farmers' BMP adoption, these results can be used as patterns that describe a system. However, as mentioned, most of the literature, to my knowledge, paid more attention to the individual factors, we have to analyze some metrics from the model results and regard them as the patterns in the POM. Specifically, the percentage of different BMP adoption, the size of agricultural land by different BMP, and the correlation between the land use type and the BMP adoption class are selected as the pattern to describe and contrast our results. Since the main question of this model is to predict farmers' BMP adoption, the percentage of different BMP adoption is a straightforward metric to describe the general pattern. Additionally, because some literature (Baumgart-Getz et al., 2012; Pannell et al., 2014) support that farm size has an impact on farmer's decision making on BMP adoption, the size of agricultural land by BMP class is an alternative to describe the pattern of results. Moreover, as the land use type, or the crop type, is the main change in each step of this model, the correlation between this and the BMP classes is very possible to help to analyze the dynamic mechanism during model running. Hence, the three metrics are selected for describing the characteristics of the three decision making structures.

However, before we analyze the results of the model running, it is important to make a decision on how many running times can provide us an acceptable and meaningful result. To solve this problem, we used the coefficients of variation. The coefficients of variation cv can be calculated by the following equation:

$$\boldsymbol{cv} = \frac{\boldsymbol{Sd.}}{\boldsymbol{Mean}} \times \mathbf{100}\ \%$$

In this equation, Sd. refers to standard deviation, and Mean represents the mean value of a series of data. Although its acceptable level differs in different fields, according to Gomez & Gomez (1984), in the agriculture field, a cv lower than 20% can be acceptable. Hence, we check the cv of three metrics, and all of them are lower than 20%, so it is reasonable to accept the 20 running times in this thesis.

Therefore, with the scenarios, metrics, and running times that can make sure these scenarios can produce stable results, we can start to compare the results of the ABMs with these decision making models. In the following section, we will use these metrics to examine and contrast the results, and expect to get some insights about them.

Chapter 4 Results

In this section, we will present the results of the scenarios mentioned previously in terms of three metrics we introduced, including the percentage of BMP adoption, size of the agricultural land, and the correlation between the landuse and BMP adoption type. We will first present the whole results of the scenarios by these metrics, and then contrast these models in two aspects. One is to examine the impacts of stochastic elements by comparing the optimal model and the stochastic model, and the other one is to examine the impacts of decision making structures by comparing the optimal model and the heuristic model. Therefore, in the part of presenting the whole results, we will not analyze the results in detail, and will leave this to the part in which we compare these models in the two aspects as described previously.

4.1 Percentage of BMP adoption

As shown in Table 4-1, similar to the observation data, all the models except the null model, provide a pattern that the fragile land retirement is the dominant class in the results. This can illustrate that all the models can represent this pattern. However, the variance within the results and the differential values between the two classes are not at the same level. Specifically, the random forest based heuristic model, have the highest contrast between the two classes, and the stochastic model, have the highest variance in model running processes. Therefore, through the percentage of BMP adoption, we can analyze that both the stochastic model and the heuristic model can produce a similar pattern as the observation, but as shown by the standard deviation, the stochastic model is more unstable even than the null random scenario. Relatively, the pattern produced by the optimal model although cannot be so similar to the observation, can still capture the main trend that the fragile land retirement is the dominant class, and as indicated by the standard deviation, the optimal model is stable as the same level as the heuristic model.

Table 4-1 The percentage of BMP adoption

Model		Optimal	Stochastic	Heuristic	Null Model	Observations
EC	Mean	30.03	21.20	19.67	49.34	21.68
	Std.	2.23	3.96	2.29	3.38	
FR	Mean	65.78	75.03	80.33	50.66	78.32
	Std.	2.24	3.89	2.29	3.38	
Diff.	Mean	35.75	53.83	60.66	1.32	56.64
	Std.	4.43	7.85	4.58	6.76	

In this table, Mean represents the mean percentage of the adoption of specific BMP types, while the std represents the standard deviations of the percentage in the 20 running times. The fragile land retirement class and the erosion control measures will be presented as FR and EC.

4.2 Size of agricultural land

The mean and standard deviation values are presented in Table 4-2. Each time the model runs, the average agricultural land size is firstly calculated by class. After all the running test is over, the class mean values and standard deviations are calculated. According to the results, it is clear that the optimal model and stochastic model have a similar pattern in terms of the size of agricultural land by classes, which means that farmers who adopt the fragile land retirement usually have a larger agricultural land than those who adopt the erosion control measures. Even at the quantitative level, the class distributions of the agricultural land size of the first two models are very similar. However, the heuristic model shows a very different pattern from the utility-based models. In the results of the heuristic model, the large agricultural land owners usually prefer to adopt the erosion control measures rather than the fragile land retirement, although the difference between the two classes is not so distinct as the utility-based models. Since the spatial scale of the observation is the southwestern Ontario, while the ABM simulates only the Upper Medway Subwatershed, they are not comparable quantitively. The results of the heuristic model, instead of the utility-based models, can reflect the general pattern in the observations, where farmers that adopt BMP in the erosion control measures usually have a larger agricultural land, and the ratio of the size of agricultural land of erosion control measures to the fragile land retirement type is at the similar level to the observation.

Table 4-2 The average size of agricultural land by BMP adoption classes

Model		Optimal	Stochastic	Heuristic	Null Model	Observations
EC	Mean	19.24	18.29	44.29	33.88	106.69
	Std.	3.08	3.21	4.98	3.34	
FR	Mean	42.71	40.17	31.82	34.61	86.14
	Std.	1.72	1.13	1.25	3.38	

In this table, Mean and std. represents the mean and standard deviation values of the 20 running times. The fragile land retirement class and the erosion control measures will be presented as FR and EC.

4.3 Correlation between landuse and BMP adoption type

There are usually three main approaches to evaluate the correlation between two or more variables, including Pearson, Kendal, and Spearman correlation tests. However, since these three correlation tests have some requirements on the types of variables, and the land use and BMP adoption types are all unordered categorical variables that do not meet the demands of the three tests, it is hard to introduce the correlation tests for the two attributes. Nevertheless, the chi-square test of independence can be used to measure the significant relationship between the two nominal variables as a metric to represent the correlation. Since the chi-square test of independence can only provide the significance that whether the correlation exists between the two variables, Cramer’s V is used for discussing the strength of the association.

As shown in Table 4-3, compared the baseline null model, in both of the utility-based models, it is not statistically significant to determine there is a correlation between the two attributes, although compared with the optimal model, the stochastic model shows a relatively higher correlation between land use and the BMP type. As for the machine learning based heuristic model, it is very clear that in all the running times, it shows a significance that the two attributes are correlated, and the strength of the correlation is at a high level. Compared with the observation data, it also shows a significance that the two attributes are not independent, but the strength of the correlation is not as high as the results in the heuristic model. Therefore, in the results of the heuristic model, it shows there is a correlation between the landuse type and BMP adoption choices as the observation data illustrates. However, optimal and stochastic models cannot present this relationship.

Table 4-3 The correlation between land use and BMP adoption type

Model	Optimal	Stochastic	Heuristic	Null Model	Observations
No. of significant tests	1	3	20	0	Yes
Cramer's V	0.198	0.217	0.689	null	0.232

No. of significant tests represents the amount of significant case ($p < 0.05$) in 20 running times. Cramer's V represents the average Cremer's V value in all the significant cases. The fragile land retirement class and the erosion control measures will be presented as FR and EC.

4.4 Impacts of the Stochastic elements: Comparison of the optimal and stochastic model

Since the optimal model and the stochastic model use the same utility function but different categorical groups of farmers, a comparison of the results of the two models will reflect the impacts of the stochastic elements. The first step of the comparison is a student t-test to check whether there is a significant difference between the two groups of results. Because the requirement of the t-test is that the samples should follow the normal distribution, for each metric, a Shapiro-Wilk test is performed, and all the three metrics of two results are normally distributed. The difference in the percentage of BMP types is significant, while for the other two metrics, the two models perform a similar pattern. For the percentage of the BMP types as illustrated in Table 4-1, even though the optimal model can also perform the general pattern, where the fragile land retirement is the dominant type, the difference between the two classes is much smaller than the results of the stochastic model. Compared with the observation data, the difference between the two types in the stochastic model is very close to the observation data, although for both of the utility-based models, the observation dataset is not used to train or build them. Additionally, the large standard deviations show that there is a large variance among the results of the stochastic model. For the other two attributes as shown in Table 4-2 and Table 4-3, the patterns are similar. Therefore, the main impacts of introducing the stochastic elements are illustrated in the percentage of the two BMP classes, and the similar pattern with the observation can be assumed as the influence of introducing some heterogeneity within the real-world data, especially when we consider that the observation data are not used in model building.

As a result, the involvement of the stochastic elements, which is the main difference of the stochastic model from the optimal model, can increase more similarity to the observation data in terms of the percentage of BMP adoption. However, the difference is not so strong in terms of other metrics. Hence, the stochastic elements can slightly improve the accuracy of the model results to the real-world case.

4.5 Impacts of the decision making structure: Comparison of the optimal and heuristic model

The main reason for comparing the heuristic model with the optimal model rather than the second model is in both of the heuristic and optimal models, farmers' typology structure is not implemented, and the main difference is the decision making approaches. Similarly, the normality and student t-tests are used to examine the significant differences between the two kinds of models, and all the three metrics of the two models are significantly different.

Similar to the comparison of the optimal and stochastic models, although the optimal model presents the pattern that the fragile land retirement is the dominant type, in the results of the heuristic model, the difference between the two classes is much larger than that in the results of the optimal model as illustrated in Table 4-1. Generally, like the results of the stochastic model, the percentage of BMP adoption types in the results of the heuristic model is quantitatively similar to the observations. Additionally, the variance of percentage in the heuristic model is much smaller than that in the stochastic model.

For the class average of agricultural land sizes, the two models have very different patterns. As shown in Table 4-2, in the optimal model, farmers that own larger land usually adopts the fragile land retirement, while in the heuristic model, those farmers usually adopt the erosion control measures BMP. However, the two classes average of agricultural land sizes are very close, and although due to differences in the spatial scale, the results of the heuristic model are quantitatively different from the observations, the general pattern is similar, which means that the larger land owners usually adopt the erosion control measures.

The most significant difference is in the correlation between land use type and the BMP adoption. As illustrated in Table 4-3, the results of the heuristic model are significantly different from the other two. In all the 20 tests, the p-value is lower than the threshold 0.05, and even very close to 0, and this means there is strong confidence to determine the two variables are dependent. With a further Cramer's V test, the average value of the 20 tests is 0.689, which shows a very strong correlation. Since the main dynamic mechanism for the three models is the change of crop type, and every year the land use is updated, it means the results of the heuristic model is very sensitive to the change during the model running stage. Compared with the observation data, although it is also significantly different, the strength of the correlation, described by the Cramer's V, is not so large. However, it still presents the pattern within the observation data, even in a different study area.

As a result, the stochastic elements, which refers to farmers' typology, introduce more heterogeneity among the agents by increasing the difference in the percentage of land parcels between the two BMP adoption classes. However, compared with the change in decision making structure, the impacts of the stochastic elements are not so strong. For the ABM that integrate the random forest as agents' decision making structure, it can successfully generate a similar macrolevel pattern with the microlevel training data, even not in the same study area and spatial scales. Nevertheless, the results of this kind of model tend to excessively follow the trend or patterns within the training datasets, and this can an interesting topic for further study.

In conclusion, the results of the machine learning based heuristic model show a high similarity to the observation data, and this similarity cannot be captured in the results of the optimal model. Therefore, we assume that the machine learning based heuristic model can capture the patterns in the observation data, and if the model builders aim to predict the future pattern of a system, and have a reliable dataset, the machine learning based heuristic model will be a satisfying alternative. However, due to the black box characteristics of the machine learning model, if the goal of the model is to examine the impacts of factors in agents' decision making, this type of heuristic model is not an acceptable choice. Under this circumstance, the stochastic model, which means an optimal model with the stochastic elements, can be a good choice, as it still slightly improves the optimal model. However, if the model builders have no knowledge about the stochastic elements like farmers' typology in this thesis, the optimal model is still acceptable as it can still capture some important features of the system, such as the percentage of BMP adoption in this project.

Chapter 5 Discussion

In this section, we will start to discuss what this project achieves and what it can provide for model designers and builders. Next, we will talk about the limitations and problems in this project, and in the future work, what is expected to be conducted and which direction can be followed further.

5.1 Achievement

This project tried to fill the gap that to my knowledge, there are not so many researches focusing on the comparison of decision making structures in ABMs, especially for the heuristic way in the form of machine learning models. Through this comparison, we can gain some understanding of the characteristics and differences among the decision making models. Furthermore, based on these characteristics and differences, we can provide some suggestions for general model builders and policy makers who want to develop the BMP adoptions.

Based on the previous analysis of results, it is clear that the design of decision making structure in terms of introducing the machine learning model has a stronger impact on the model patterns than the stochastic elements that represent farmers' typology in agents' design. However, stochastic elements of the farmers' typology can still increase the heterogeneity in the model, and make some patterns more similar to the observation dataset.

For the machine learning model, although restricted by the black-box design to some degrees, it is hard to explain why the decision made and the detailed decision making processes of agents, can represent the bounded rationality of agents, and predict the results at a better level of accuracy compared with the optimal and stochastic models. Specifically, in this project, even the machine learning model is trained to simulate the behavior of agents rather than the whole system, and used in different study areas and temporal-spatial scales, the ABM can still produce very similar patterns as the observation data in the macro level.

Nevertheless, there are still some disadvantages and risks of introducing the machine learning model. Since this kind of model will excessively present the trends in the training dataset, it may lead to some analogous problems to overfitting. However, because in this project we only have one source of data to train the model, it is hard for us to determine whether there is an overfitting problem, but as shown in the results, the machine learning model is too close to the observation data. Additionally, although introducing machine learning models to represent the decision making processes of the agents in ABMs rather than the whole system makes it easier to understand the elements and structures of the system, the interpretation of agents' behavior still remains a problem due to the black-box feature. Furthermore, the interactions among agents are also hard to be presented directly and explicitly in the machine learning based heuristic model, compared with the optimal model and the stochastic model.

Therefore, because this project provides a general idea about the difference of agents' action rules and with this, we can generalize a preliminary suggestion on the choices of these decision making models for model builders, as illustrated in Table 5-

1. If the aim of the ABMs is to examine the decision making processes and important factors to individual agents, it is better to use the optimal models to get some insights about them. In addition, the stochastic elements and detailed agents' typology will increase the heterogeneity and make a more similar pattern to the observation and calibration data. If the goal of ABMs is to predict some patterns or future change and the model builders have access to a reliable dataset, the machine learning model can be an appropriate alternative. However, the risk of overfitting problem, the interpretability problem, and the difficulty to represent the direct interactions should be also be taken into the consideration.

Table 5-1 The summary of characteristics of decision making models based on the results of this thesis

	Strengths	Weaknesses	Conditions that are suitable to use
Optimal model	Relatively easy to be implemented Producing an acceptable result Remaining opportunity to analyze the factors that drive agents' behavior	Results of the model may not meet the demand in accuracy compared with observations	Analyzing the factors that affect agents' behavior with no specific knowledge about the agent and the system
Stochastic model	Additional improvement on the optimal model More similar to the real-world circumstance than the optimal model Remaining opportunity to analyze the factors that drive agents' behavior	Still requires knowledge about the agent and the system that the model simulates Compared with the heuristic model, the similarity to the real-world observation is still lower	Analyzing the factors that affect agents' behavior
Heuristic model	High similarity to the observation data	Relying on an appropriate dataset Not suitable for interpreting the agents' decision The overfitting risks The difficulty to represent the interactions among agents	Predicting the future possible pattern of the system

In addition, the improvement of stochastic elements that represents farmers' typology to the optimal model allows us to provide some suggestions on the development and implementation of BMP. Since the introduction of farmers' typology makes the model has more similarity with the observations, it is very possible that different groups of farmers may have different preferences and priority for BMP adoption. Hence, if the policymakers expect to develop the BMP adoption, concerning about the heterogeneity among farmers, and providing different inspiring plans for them can be helpful.

5.2 Limitation

The limitation of this work can be discussed in two aspects, the data used in model building and the method to design and compare the models. For the data limitation, it comes from three main aspects: data dimension and description, data quality, and the spatial-temporal scales in the datasets. The limitation in the method mainly results from the two aspects, including the assumptions, and the stationary state of the decision making.

Specifically, the data dimension refers to the number of attributes and information within the dataset. In the observation dataset, only five categories of information are provided, including the BMP adoption type, geographical coordinates, project years, funding resources, and the final costs. Additionally, all these attributes lack detailed descriptions. For example, the locations represented by these geographical coordinates are still unclear, and it is still unknown for us that whether the points represent the centroid of land parcels or just a nearby place on the road. In addition, this dataset does not provide detailed information about the landowners' information and the characteristics of the agricultural land. To overcome the limitation of lacking data dimension, many external data are joined from other datasets based on the spatial location. However, the lack of the data dimension and data description leads to many uncertainties for this project. When we tried to join external information from other datasets for training a model, the lack of description of the features increases another level of uncertainties since what the geographical locations represent is not so clear.

Additionally, there are some errors in the datasets, and this indicates the low quality of the dataset. For instance, for some records in the data, there are attributes containing the missing information, and several geographical coordinates are incorrect, and some of the coordinates of raw data are just located in Peru or the Pacific Ocean. To overcome this challenge, we examined and cleaned the dataset, and removed these wrong records. However, because we want to deal with the data dimension problem by joining many attributes from other datasets by geographical locations, and the low data quality brings more uncertainties for the joining process.

Another problem in the dataset is the spatial-temporal scales. As mentioned, although there is an attribute called project year that provides the temporal information, the description of this attribute is not so clear and has some missing values, and this makes it hard to filter the records that may not meet the temporal demand. For the spatial scales, since the study area of the observation dataset is across the whole southwestern Ontario, while the spatial extent of the proof-of-concept is only the Upper Medway Creek subwatershed. Although the different spatial scales make it possible to examine the performance of the trained model, it is still hard to use it for quantitative comparison, for the different spatial scales make it hard to get evaluated metrics at the same level. To overcome the temporal limitation, we select the data records attached project year, and use the median of the total period as the representative. To overcome the limitation of the spatial scale, as we performed in the result section, we focused on the ratio among the two BMP adoption classes rather than the actual number.

For the method used in this model, the assumption is a significant limitation. As described in the methodology section, there are many assumptions introduced, e.g. the farmers and fieldcells follow a one-to-one correspondence relationship, and the impacts of the topology and non-contributing areas are not concerned in this model. These assumptions are helpful in the model design to reduce the complexity of the model designs and the computation, but sometimes these assumptions can be arguable that it cannot capture the characteristics of the system (Railsback& Grimm, 2019). Therefore, model builders can be aware of the problems that the assumptions may lead to. To overcome this problem, the POM, while it is also used in this project to evaluate the models, can be an alternative. Based on the goal of the model, model builders can define and examine whether the model is structurally realistic to check the suitability of assumptions.

Another limitation of the method resides in the stationary design of the agents. In this thesis, in all the decision making models, the agents will follow the same decision making structure during the model running, while in the real world, agents can change their decision making processes. To overcome this problem, introducing the learning mechanism to agents can be an alternative, e.g. Gimona & Polhill (2011) included a learning mechanism by storing the memory of agents' actions in the previous model running to make decisions. In addition, there are three decision making models in this thesis, and this can further provide an opportunity to use the results of other models as the observations to train or design the model. However, because the goal of this thesis is to compare, and introducing this will increase the difficulty to compare the decision making models, the learning mechanism is not included in this thesis. Furthermore, if the model builder thinks it is important to represent the change of decision making processes, the learning mechanism of agents can still be an alternative.

5.3 Future possible research

Due to the limitation in the datasets as mentioned, there is an opportunity to improve the project structure with a dataset that contains more information like the joined external data in the proof-of-concept. Additionally, the introduction of local data or working with experts in this field can provide a baseline for the validation. Moreover, since the comparison is not limited to a specific case, in future work, researchers can choose ABMs that have reliable calibration data or patterns that are cleared determined by previous experts to change the performance of different decision making models, and have more detailed or quantitative descriptions of the difference.

Additionally, this project also provides a simple structure for the future research. Specifically, since there is a distinct difference between the results of the optimal models and ABMs that integrate the machine learning model. It would be interesting to check the representativeness and whether the achievements of this project can be adapted in other model design cases. Due to the time limit in this project, we have to only compare the three ABMs in this BMP adoption case. With further research, we can contrast the characteristics of decision making models in different cases, such as market share forecasting. Introducing different cases would be useful to provide some insights about the optimal choice on decision making models for various purposes.

In conclusion, this project focuses on comparing and contrasting the differences of decision making models, especially for the machine learning based heuristic model, in a proof-of-concept of farmers' BMP adoption case. Next, we further analyzed the results of these decision making models based on the three metrics. According to the analysis of the results, we provide some suggestions on the model and policy choice based on the characteristics of the three models. Additionally, this project also provides a practical case of integrating machine learning model into ABMs as the decision rules of agents to fill the gap that few researches compare the decision making structures in ABMs, especially for the machine learning model integrated as a heuristic model.

References

2017 Watershed Report Card - Medway Creek. (2017). Retrieved from http://thamesriver.on.ca/wp-content/uploads//WatershedReportCards/RC_Medway.pdf

Agriculture and Agri-food Canada. (2013). Annual Crop Inventory 2011. Retrieved from https://open.canada.ca/data/en/dataset/58ca7629-4f6d-465a-88eb-ad7fd3a847e3

Alberta Government. (2015). Grassed waterway construction. Retrieved from https://open.alberta.ca/publications/2819583

An, L. (2012). Modeling human decisions in coupled human and natural systems: Review of agent-based models. Ecological Modelling, 229, 25–36. https://doi.org/10.1016/j.ecolmodel.2011.07.010

Anderson, P. W. (1972). More is different. Science, 177(4047), 393-396.

Arthur, W. B. (1999). Complexity and the economy. science, 284(5411), 107-109.

Baumgart-Getz, A., Prokopy, L. S., & Floress, K. (2012). Why farmers adopt best management practice in the United States: A meta-analysis of the adoption literature. Journal of environmental management, 96(1), 17-25.

Booch, G., Maksimchuk, R. A., Engle, M. W., Young, B. J., Connallen, J., & Houston, K. A. (2008). Object-oriented analysis and design with applications. *ACM SIGSOFT software engineering notes*, *33*(5), 29-29.

Brandle, R. J. (n.d.). How Windbreaks Work. Retrieved from https://www.fs.usda.gov/nac/documents/morepublications/ec1763.pdf

Brown, D. G., & Robinson, D. T. (2006). Effects of heterogeneity in residential preferences on an agent-based model of urban sprawl. Ecology and society, 11(1).

Cabrera, A. R., Deadman, P. J., Brondizio, E. S., & Pinedo-Vasquez, M. (2010). Exploring the choice of decision making method in an agent based model of land use change.

Chen, J. J., Zheng, B., & Tan, L. (2013). Agent-based model with asymmetric trading and herding for complex financial systems. PloS one, 8(11).

Clean Water Program. (n.d.). Eligible Projects. Retrieved August 18, 2020, from https://www.cleanwaterprogram.ca/eligible-projects/

Commendatore, P., Kubin, I., Bougheas, S., Kirman, A., Kopel, M., & Bischi, G. I. (2018). *The Economy as a Complex Spatial System: Macro, Meso and Micro Perspectives*. Cham: Springer International Publishing.

Dada, E. G., Bassi, J. S., Chiroma, H., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. Heliyon, 5(6), e01802.

Daloğlu, I., Nassauer, J. I., Riolo, R. L., & Scavia, D. (2014). Development of a farmer typology of agricultural conservation behavior in the American Corn Belt. Agricultural Systems, 129, 93-102.

Dong, X., Foteinou, P. T., Calvano, S. E., Lowry, S. F., & Androulakis, I. P. (2010). Agent-based modeling of endotoxin-induced acute inflammatory response in human blood leukocytes. *PloS one*, *5*(2), e9249.

Essex Region Conservation. (n.d.). Agricultural Stewardship. Retrieved from https://essexregionconservation.ca/stewardships-grants/agricultural-stewardship/

Evans, T. P., Sun, W., & Kelley, H. (2006). Spatially explicit experiments for the exploration of land‐use decision‐making dynamics. International Journal of Geographical Information Science, 20(9), 1013-1037.

Fabian Adelt, Johannes Weyer & Robin D. Fink (2014) Governance of complex systems: results of a sociological simulation experiment, Ergonomics, 57:3, 434-448, DOI: 10.1080/00140139.2013.877598

Furtado, B. A. (2017). Machine Learning simulates Agent-based Model. arXiv preprint arXiv:1712.04429.

Gilli, M., & Rossier, E. (1979). *Understanding complex systems*. Genève: Faculté des sciences économiques et sociales Département déconométrie.

Gimona, A., & Polhill, J. G. (2011). Exploring robustness of biodiversity policy with a coupled metacommunity and agent-based model. Journal of Land Use Science, 6(2-3), 175-193.

Gomez, K. A., & Gomez, A. A. (1984). Statistical procedures for agricultural research. John Wiley & Sons.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* . The MIT Press.

Grimm, V., Railsback, S. F., Vincenot, C. E., Berger, U., Gallagher, C., DeAngelis, D. L., ... & Johnston, A. S. (2020). The ODD protocol for describing agent-based and other simulation models: A second update to improve clarity, replication, and structural realism. Journal of Artificial Societies and Social Simulation, 23(2).

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., ... & DeAngelis, D. L. (2005). Pattern-oriented modeling of agent-based complex systems: lessons from ecology. science, 310(5750), 987-991.

Groeneveld, J., Müller, B., Buchmann, C. M., Dressler, G., Guo, C., Hase, N., ... & Liebelt, V. (2017). Theoretical foundations of human decision-making in agent-based land use models–A review. *Environmental modelling & software*, *87*, 39-48.

Grovermann, C., Schreinemachers, P., Riwthong, S., & Berger, T. (2017). *"Smart" policies to reduce pesticide use and avoid income trade-offs: An agent-based model applied to Thai agriculture. Ecological Economics, 132, 91–103.* doi:10.1016/j.ecolecon.2016.09.031

Guillem, E. E., Barnes, A. P., Rounsevell, M. D., & Renwick, A. (2012). Refining perception-based farmer typologies with the analysis of past census data. Journal of Environmental Management, 110, 226-235.

Guo, L. (2018). Simulating Farmer Adoption of Agricultural Best Management Practices in the Upper Medway Creek Subwatershed (Master's thesis, University of Waterloo).

Handel, O. (2016). Modeling dynamic decision-making of virtual humans. Systems, 4(1), 4.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning data mining, inference, and prediction (2nd ed.). Springer New York.

https://doi.org/10.1007/978-0-387-84858-7

Hawes, E., & Smith, M. (2005). Riparian buffer zones: functions and recommended widths. Eightmile River Wild and Scenic Study Committee, 15, 2005.

Hayashi, S., Prasasti, N., Kanamori, K., & Ohwada, H. (2016). Improving behavior prediction accuracy by using machine learning for agent-based simulation. In Asian Conference on Intelligent Information and Database Systems (pp. 280-289). Springer, Berlin, Heidelberg.

Heckbert, S., Baynes, T., & Reeson, A. (2010). Agent-based modeling in ecological economics. Annals of the New York Academy of Sciences, 1185(1), 39-53.

Ho, Tin Kam. (1998). The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8), 832–844. https://doi.org/10.1109/34.709601

Hoertel, N., Blachier, M., Blanco, C., Olfson, M., Massetti, M., Rico, M. S., ... & Leleu, H. (2020). A stochastic agent-based model of the SARS-CoV-2 epidemic in France. *Nature Medicine*, 1-5.

Huth, A., & Wissel, C. (1992). The simulation of the movement of fish schools. Journal of theoretical biology, 156(3), 365-385.

Jäger, G. (2019). Replacing Rules by Neural Networks A Framework for Agent-Based Modelling. Big Data and Cognitive Computing, 3(4), 51.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

Kansas, M. (1989). An Evaluation of the Cost Effectiveness of Agricultural Best

Management Practices and Publicly Owned Treatment Works in Controlling Phosphorus Pollution in the Great Lakes Basin. U.S. Environmental Protection Agency.

Klabunde, A., & Willekens, F. (2016). Decision-Making in Agent-Based Models of Migration: State of the Art and Challenges. *European Journal of Population*, *32*(1), 73–97. https://doi.org/10.1007/s10680-015-9362-0

Koehrsen, W. (2018, January 10). Hyperparameter Tuning the Random Forest in Python. Retrieved August 17, 2020, from https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

Ladyman, J., Lambert, J., & Wiesner, K. (2013). What is a complex system?. *European Journal for Philosophy of Science*, *3*(1), 33-67.

Lai, J. Z., & Liao, S. L. (2013). Agent-based simulation method of complex system. In Applied Mechanics and Materials (Vol. 380, pp. 1378-1381). Trans Tech Publications Ltd.

Lamperti, F., Roventini, A., & Sani, A. (2018). Agent-based model calibration using machine learning surrogates. Journal of Economic Dynamics and Control, 90(C), 366–389. https://doi.org/10.1016/j.jedc.2018.03.011

Ligmann-Zielinska, A. (2009). The impact of risk-taking attitudes on a land use pattern: an agent-based model of residential development. *Journal of Land Use Science*, *4*(4), 215-232.

Liu, T., Bruins, R. J., & Heberling, M. T. (2018). Factors influencing farmers’ adoption of best management practices: A review and synthesis. Sustainability, 10(2), 432.

López, A., Valera, D. L., Molina-Aiz, F. D., Lozano, F. J., & Asensio, C. (2017). Sonic anemometry and sediment traps to evaluate the effectiveness of windbreaks in preventing wind erosion. Scientia Agricola, 74(6), 425-435.

Lorena, A., Jacintho, L., Siqueira, M., Giovanni, R., Lohmann, L., de Carvalho, A., & Yamamoto, M. (2011). Comparing machine learning classifiers in potential distribution modelling. Expert Systems With Applications, 38(5), 5268–5275. https://doi.org/10.1016/j.eswa.2010.10.031

Lower Thames Valley Conservation Authority. (2015). Rural Best Management Practice Factsheet. Retrieved from https://www.lowerthames-conservation.on.ca/wp-content/uploads/2015/04/Fact-Sheet-Windbreak-LTVCA.pdf

Malawska, A., & Topping, C. J. (2016). Evaluating the role of behavioral factors and practical constraints in the performance of an agent-based model of farmer decision making. Agricultural Systems, 143, 136-146.

Manzouri, F., Heller, S., Dümpelmann, M., Woias, P., & Schulze-Bonhage, A. (2018). A Comparison of Machine Learning Classifiers for Energy-Efficient Implementation of Seizure Detection. Frontiers in Systems Neuroscience, 12, 43–. https://doi.org/10.3389/fnsys.2018.00043

Miksch, F., Jahn, B., Espinosa, K. J., Chhatwal, J., Siebert, U., & Popper, N. (2019). Why should we apply ABM for decision analysis for infectious diseases? —An example for dengue interventions. PloS one, 14(8), e0221564.

Miller, B. W., Breckheimer, I., McCleary, A. L., Guzmán-Ramirez, L., Caplow, S. C., Jones-Smith, J. C., & Walsh, S. J. (2010). Using stylized agent-based models for population–environment research: a case study from the Galápagos Islands. Population and environment, 31(6), 401-426.

Miller, J. H., & Page, S. E. (2009). Complex adaptive systems: An introduction to computational models of social life. Princeton university press.

Mitchell, T. (1997). *Machine learning* . McGraw-Hill.

Mtibaa, S., Hotta, N., & Irie, M. (2018). Analysis of the efficacy and cost-effectiveness of best management practices for controlling sediment yield: A case study of the Joumine watershed, Tunisia. Science of the Total Environment, 616, 1-16.

Natural Resources Conservation Service. (2010). WATER AND SEDIMENT CONTROL BASIN. Retrieved from https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs143_026238.pdf

Natural Resources Conservation Service. (n.d. a). Water and Sediment Control Basin. Retrieved from https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/water/?cid=nrcs142p2_044370

Natural Resources Conservation Service. (n.d. b). Buffer Strips: Common Sense Conservation. Retrieved from https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143_0

23568

Ontario Ministry of Agriculture, Food and Rural Affairs. (2009). Grassed Waterways. Retrieved from http://www.omafra.gov.on.ca/english/engineer/facts/09-021.htm

Ontario Ministry of Agriculture, Food and rural Affairs. (2016). Provincial Field Crop Production and Prices. Retrieved from http://www.omafra.gov.on.ca/english/stats/crops/index.html

Ontario Ministry of Agriculture, Food and rural Affairs. (2017). 2017 FIELD CROP BUDGETS. Retrieved from http://www.omafra.gov.on.ca/english/busdev/facts/pub60.htm

Ontario Ministry of Agriculture, Food and rural Affairs. (2018). Area, Yield, Production and Farm Value of Specified Field Crops, Ontario, 2012 - 2017 (Imperial and Metric Units).Retrieved from http://www.omafra.gov.on.ca/english/stats/crops/estimate_new.htm

Ontario Ministry of Agriculture, Food and Rural Affairs. (n.d.). Best Management Practices Series. Retrieved August 18, 2020, from http://www.omafra.gov.on.ca/english/environment/bmp/series.htm

Ontario Ministry of Natural Resources and Forestry. (2015). Soil Survey Complex. Retrieved from https://geohub.lio.gov.on.ca/datasets/ontarioca11::soil-survey-complex?geometry=-151.490%2C38.917%2C-17.984%2C58.786

Ontario Ministry of Natural Resources and Forestry. (2016). Standardized Precipitation Index. Retrieved from https://geohub.lio.gov.on.ca/datasets/ce2a10c9eea04a10bb514a303acc676b

Ontario Ministry of Natural Resources and Forestry. (2019). Ontario Digital Elevation Model (Imagery-Derived). Retrieved from https://geohub.lio.gov.on.ca/datasets/1ce266ee55c44ffca2d457bc5db13b92

Ontario Ministry of Natural Resources and Forestry. (2020). Ontario Hydro Network (OHN) - Waterbody. Retrieved from https://geohub.lio.gov.on.ca/datasets/22bab3c9f37a4dd0845eb89e7b247a9f_25?geometry=-151.490%2C38.917%2C-17.984%2C58.786

Pannell, D. J., Llewellyn, R. S., & Corbeels, M. (2014). The farm-level economics of conservation agriculture for resource-poor farmers. Agriculture, ecosystems & environment, 187, 52-64.

Parker, D. C., & Robinson, D. T. (2017). Agent-Based Modeling. International Encyclopedia of Geography: People, the Earth, Environment and Technology, 1–9.

Railsback, S., & Grimm, V. (2019). *Agent-based and individual-based modeling : A practical introduction*. Princeton: Princeton University Press.

Raman, N., & Leidner, J. L. (2019, June). Financial Market Data Simulation Using Deep Intelligence Agents. In International Conference on Practical Applications of Agents and Multi-Agent Systems (pp. 200-211). Springer, Cham.

Rand, W. (2006). Machine learning meets agent-based modeling: when not to go to a bar. In *Conference on Social Agents: Results and Prospects*.

Rand, W., & Stonedahl, F. (2007). The El Farol bar problem and computational effort: Why people fail to use bars efficiently. Northwestern University, Evanston, IL.

Rind, D. (1999). Complexity and climate. science, 284(5411), 105-107.

Robinson, D. T., & Brown, D. G. (2016). Representation: Dynamic Complex Systems. International Encyclopedia of Geography: People, the Earth, Environment and Technology: People, the Earth, Environment and Technology, 1-11.

Robinson, D., Brown, D., Parker, D., Schreinemachers, P., Janssen, M., Huigen, M., Wittmer, H., Gotts, N., Promburom, P., Irwin, E., Berger, T., Gatzweiler, F., & Barnaud, C. (2007). Comparison of empirical methods for building agent-based models in land use science. Journal of Land Use Science, 2(1), 31–55. https://doi.org/10.1080/17474230701201349

Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., & Santos, J. (2018). Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier]. ieee ComputatioNal iNtelligeNCe magaziNe, 13(4), 59-76.

Schreinemachers, P., & Berger, T. (2006). Land use decisions in developing countries and their representation in multi-agent systems. Journal of land use science, 1(1), 29-44.

Schroter, J. B., & Kansas, H. (n.d.). Grassed Waterways versus Underground Outlet Tile. Retrieved from https://www.nrcs.usda.gov/wps/portal/nrcs/detail/ks/newsroom/features/?cid=stelprdb1242792

Shafie-Khah, M., & Catalão, J. P. (2014). A stochastic multi-layer agent-based model to study electricity market participants behavior. *IEEE Transactions on Power Systems*, *30*(2), 867-881.

Shalizi, C. R. (2006). Methods and techniques of complex systems science: An overview. In Complex systems science in biomedicine (pp. 33-114). Springer, Boston, MA.

Sheikhha, F., Malazi, H., & Amjadifard, R. (2009). Adaptive Parasitized El Farol Bar Problem. 2009 WRI World Congress on Computer Science and Information Engineering, 7, 422–426. https://doi.org/10.1109/CSIE.2009.1107

Simon, H. A. (1991). The architecture of complexity. In Facets of systems science (pp. 457-476). Springer, Boston, MA.

Simon, H. A. (1997). *Models of bounded rationality: Empirically grounded economic reason* (Vol. 3). MIT press.

Starfield, A., Smith, K., & Bleloch, A. (1990). How to model it : problem solving for the computer age . McGraw-Hill.

Teranet Incorporated (2018). Regional Municipality of Waterloo Property Parcels [computer file].

Torrens, P., Li, X., & Griffin, W. (2011). Building Agent-Based Walking Models by Machine-Learning on Diverse Databases of Space-Time Trajectory Samples. Transactions in GIS, 15(s1), 67–94. https://doi.org/10.1111/j.1467-9671.2011.01261.x

Uchmański, J. (2000). Individual variability and population regulation: an individual ‐ based model. Oikos, 90(3), 539-548.

Ul Hassan, C., Khan, M., & Shah, M. (2018). Comparison of Machine Learning Algorithms in Data classification. 1–6. https://doi.org/10.23919/IConAC.2018.8748995

United States Department of Agriculture. (2007, December). Engineering Field Handbook -Chapter 7.

University of Waterloo. (n.d.). Welcome to Agricultural Water Futures (AWF). Retrieved from https://uwaterloo.ca/agricultural-water-futures/

Upper Thames River Conservation Authority. (n.d. a). Farmland BMPs: UTRCA: Inspiring A Healthy Environment. Retrieved from http://thamesriver.on.ca/landowner-grants-stewardship/farmland-bmps/

Upper Thames River Conservation Authority. (n.d. b). Riparian Buffer Strips. Retrieved from http://thamesriver.on.ca/wp-content/uploads//farmlandbmps/Buffer-factsheet.pdf

Upper Thames River Conservation Authority. (n.d. c). Windbreaks. Retrieved from http://thamesriver.on.ca/wp-content/uploads//farmlandbmps/Windbreak-factsheet.pdf

Upper Thames River Conservation Authority. (n.d. d). Medway Creek. Retrieved from http://thamesriver.on.ca/education-community/watershed-friends-of-projects/medway/

Upper Thames River Conservation Authority. (n.d. e). MEDWAY CREEK COMMUNITY BASED ENHANCEMENT STRATEGY. Retrieved from http://thamesriver.on.ca/wpcontent/uploads/MedwayCreek/MedwayCBESreport.pdf

Utah State University. (n.d.). Best Management Practices. Retrieved August 18, 2020, from https://extension.usu.edu/waterquality/protectyourwater/howtoprotectwaterquality/bmps/index

Van Dam, K. H., Nikolic, I., & Lukszo, Z. (Eds.). (2012). Agent-based modelling of socio-technical systems (Vol. 9). Springer Science & Business Media.

Wiegand, T., Jeltsch, F., Hanski, I., & Grimm, V. (2003). Using pattern‐oriented modeling for revealing hidden information: a key for reconciling ecological theory and application. Oikos, 100(2), 209-222.

Wilensky, U., & Rand, W. (2015). An introduction to agent-based modeling : modeling natural, social, and engineered complex systems with NetLogo . Cambridge, Massachusetts: The MIT Press.

Wojtusiak, J., Warden, T., & Herzog, O. (2012). Machine learning in agent-based stochastic simulation: Inferential theory and evaluation in transportation logistics. Computers and Mathematics with Applications, 64(12), 3658–3665. https://doi.org/10.1016/j.camwa.2012.01.079

Zeman, K. R. (2019). Modeling farmer decision-making using an agent-based model for studying best management practice adoption rates: A typological approach (Doctoral dissertation).

Zhou, Q. (2018). Integrating a hierarchical SOM-based intervention method in an agent-based pertussis model.