

Article

A Framework for Data-Driven Agent-Based Modelling of Agricultural Land Use

Giacomo Ravaioli, Tiago Domingos  and Ricardo F. M. Teixeira * 

MARETEC-Marine, Environment and Technology Centre, LARSyS, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1, 1049-001 Lisbon, Portugal; giacomo.ravaioli@tecnico.ulisboa.pt (G.R.); tdomingos@tecnico.ulisboa.pt (T.D.)

* Correspondence: ricardo.teixeira@tecnico.ulisboa.pt

Abstract: Agent-based models (ABMs) are particularly suited for simulating the behaviour of agricultural agents in response to land use (LU) policy. However, there is no evidence of their widespread use by policymakers. Here, we carry out a review of LU ABMs to understand how farmers' decision-making has been modelled. We found that LU ABMs mainly rely on pre-defined behavioural rules at the individual farmers' level. They prioritise explanatory over predictive purposes, thus limiting the use of ABM for policy assessment. We explore the use of machine learning (ML) as a data-driven alternative for modelling decisions. Integration of ML with ABMs has never been properly applied to LU modelling, despite the increased availability of remote sensing products and agricultural micro-data. Therefore, we also propose a framework to develop data-driven ABMs for agricultural LU. This framework avoids pre-defined theoretical or heuristic rules and instead resorts to ML algorithms to learn agents' behavioural rules from data. ML models are not directly interpretable, but their analysis can provide novel insights regarding the response of farmers to policy changes. The integration of ML models can also improve the validation of individual behaviours, which increases the ability of ABMs to predict policy outcomes at the micro-level.

Keywords: policy assessment; machine learning; behavioural modelling; decision-making



Citation: Ravaioli, G.; Domingos, T.; Teixeira, R.F.M. A Framework for Data-Driven Agent-Based Modelling of Agricultural Land Use. *Land* **2023**, *12*, 756. <https://doi.org/10.3390/land12040756>

Academic Editors: Yimin Chen, Guohua Hu, Yujia Zhang, Xuecao Li and Brian Alan Johnson

Received: 25 February 2023

Revised: 22 March 2023

Accepted: 24 March 2023

Published: 27 March 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Human activity is hampering our possibility of living within planetary boundaries [1,2]. Food production is the most important driver of environmental change at a global scale [3]; it accounts for 19–29% of total anthropogenic greenhouse gas emissions [4]; it is responsible for 70% of freshwater use for irrigation purposes and 40% of land occupation and is a major driver of biodiversity loss [5]; moreover, it is the main driver of eutrophication, through the massive use of nitrogen and phosphorus fertilizers [3].

Agricultural land use (LU)—the focus of this paper, which we will refer to as LU for simplification—behaves as a complex adaptive system. Farmers are individual agents acting and interacting under the influence of past actions, peer pressure, economic market forces, governmental decisions and environmental constraints [6–8]. This causes feedback loops and non-linear responses that can lead to unexpected outcomes [7–9]. Due to path dependency, the effects of ill-designed policies might be persistent and difficult to reverse [9,10]. For these reasons, when designing policies, the literature highlights the importance of treating the complexity of LU and of socio-ecological systems more generally in order to prevent unintended consequences [7–9,11–13].

The use of computational simulations in LU studies has attracted wide scientific interest, especially since 2008 [14,15]. In particular, the interest in applying agent-based models (ABMs) to model LU experienced a steady increase [8,16,17]. The use of ABMs for agricultural policy assessment applications increased as well [15,18]. ABMs are generically identified as simulations composed of individual agents and characterized by the

importance given to their behaviours, interactions and heterogeneity [19]. Thanks to their bottom-up approach, with no centralised control and each agent acting according to its own rules, ABMs are suited for the study of complex adaptive systems [20]. Typically, LU ABMs consist of individual farmers, the main agents, deciding at each step of the simulation to maintain or change the use of the land they own, experiencing environmental changes and interacting with the government, markets and other farmers. ABMs can, therefore, overcome some important limitations of other more consolidated approaches to model LU, such as the aggregated level of analysis required by equations-based and system models (such as partial equilibrium models) or the simplification of farmers' decision-making and reciprocal influence in cellular models [21,22]. The literature stresses, in particular, some advantages of applying ABMs to LU and agricultural policies assessment [6,14,18,21,23–27], which can be summarized as follows:

- ABMs can simulate the decentralised and heterogeneous decision-making of farmers with a high level of detail and consider uncertainty regarding their behaviour. This allows for the evaluation of policy effects at the individual level.
- ABMs can explicitly model social interactions, which have an important influence on farmers' behaviour, and therefore allow the study of the diffusion of technologies and practices.
- ABMs can explicitly include a spatial dimension and the biophysical properties of land, linking it with farmers' decision-making and thus addressing the feedback between the socio-economic and biophysical spheres.
- ABMs provide a natural framework to consider out-of-equilibrium dynamics.
- ABMs can consider the complex and distributed effects of climate change on agriculture, which are likely to gain increasing relevance.

Despite the number of ABMs focusing on agricultural policies, their use for actual policy assessment is still limited [15,26]. The lack of transparency and accessibility of the models is a commonly reported reason for this [15], although transparency initiatives by journals and funders are making this problem less prevalent. The ODD protocol (which stands for overview, design concepts, details) [28–31] provides a standardised way to document ABMs and is already used for 20% of ABMs in ecology [29]. Source code and data are increasingly made available, improving the replicability and openness of ABMs [15,18,32].

An even more relevant issue is the predictive performance of ABMs. Prediction can be defined as “the ability to reliably anticipate well-defined aspects of data that is not currently known to a useful degree of accuracy via computations using the model” [33]. Reliable prediction is paramount for ex post and especially ex ante policy analysis [32,34]. Two important achievements in this regard have been the harnessing of ABMs' ability to incorporate deep uncertainty and run multi-scenario simulations, as opposed to their previous use for point predictions [26,35], and the passage from conceptual and abstract models to empirical ones, which allowed targeting specific case studies [34]. However, as Section 3 will show, ABMs have traditionally focused on equipping agents with clearly interpretable pre-set behavioural rules, striving to understand the mechanisms and processes of the system and the relation between micro-level behaviours and macro-level emergent properties. The relationship between the explanatory and predictive power of ABMs is, however, complex. There is no evidence that the imposition of theoretical rules increases the potential of an ABM to predict outcomes.

To improve their predictive ability, we argue that ABMs should integrate data-science approaches, which have an important role in the increasing efforts for evidence-based policy assessment [36–38]. Recent advances in big data, machine learning (ML) predictive algorithms and computational power allow modellers to go one step further than using empirical data only for the parametrization of models. ML enables efficient handling of large amounts of micro-level data to learn real-world patterns of behaviour [39]. ML models can, therefore, be integrated into data-driven ABMs to avoid relying on theoretical or heuristic behavioural rules. Such an approach has shown potential for outperforming

more traditional ABMs in other areas of application [39–41] but is basically absent from LU applications. LU ABMs should prioritize research along these lines as there are large amounts of LU data from field surveys and remote sensing, as well as micro-data at the individual farmer level from national surveys and research projects and spatialized biophysical data (for soils, climate, etc.) [42].

Therefore, in this paper, we started by establishing that data-driven approaches to behaviour modelling of agents are still underexplored in land use modelling. Section 2 explains how the empirical grounding of ABMs is particularly suited for LU problems, including for policy assessments. Section 3 explains how the agents' decision-making is modelled in ABMs, highlighting the potential use of micro-level data. We then propose a framework for integrating data-science approaches and ABMs to address LU problems through a completely data-driven definition of agents' behaviours and interactions. Section 4 introduces ML as a data-driven approach and shows how it has been used in combination with ABMs, focusing on its use to learn agents' behaviours directly from the available data and increase the predictive robustness of the model. It also stresses the lack of a proper framework for the integration of ML and LU ABMs. Section 5 presents our proposal for a data-driven LU agent-based modelling framework. We hope that, by providing this framework, more data-driven LU ABMs will be developed with the ability to provide robust and useful insights for policy assessment. Finally, Section 6 discusses the limitations of this framework in terms of increased complexity and data requirements and then describes how theoretical and heuristic rules can be integrated to address these issues and which other benefits they could bring.

2. Empirically Grounded Land Use ABMs

The first ABMs developed were mainly abstract and conceptual, i.e., representing a fictitious environment and agents without any use of empirical data as they aimed at demonstrating the applicability and suitability of this new modelling approach [17,21]. After successful proofs of concept of their utility and facing increased acceptance, ABMs are now increasingly complex and system-specific [17,43,44]. This caused a surge in the last two decades in the efforts to employ data in empirical LU ABMs [22,24,30,45]. This was exemplified by the proposal of modelling approaches focused on capturing observed facts such as KIDS ("Keep it Descriptive Stupid"), in direct contrast with previous ones aimed at the simplicity of the model, such as KISS ("Keep it Simple Stupid") [46]. The application of ABMs that particularly drove this shift was policy design [34]. To obtain stakeholders' and policymakers' trust in ABMs results, proper calibration and validation using empirical data [30], as well as a representation of micro-processes based on real-world observations [22,40], are, in fact, fundamental. At the same time, this empirical grounding was enhanced by the exponential increase in the availability of relevant data [34,40,43,44]. High-resolution remotely sensed data and Geographical Information Systems (GIS) products for soil, vegetation and climate provide the possibility to build spatially explicit models, directly linking agents to their biophysical environment [20,22,47]. Sample surveys and interviews [48–50], census data [47], participant observation insights and expert opinions [22,50] can be used in combination and provide micro-level data on individual farmers' socio-economic characteristics. Micro-data have been used in ABMs to instantiate agents' attributes and parametrise their rules [30,45,51], but cutting-edge applications harness micro-data to elicit individual agents' behavioural rules and interactions directly.

3. Modelling Agents' Behaviours in ABMs

Individual decision-making modelling in ABMs is critical for their outcomes' usefulness and credibility [19,21,32,52]. Modelling agents' behaviour is a delicate issue, especially in LU. Farmers' decision-making seldom follows strict economic optimization and is also influenced by culture, traditions and peer influence and subject to limited knowledge of innovations and markets [47,50,53,54]. Therefore, the use of the increas-

ingly available micro-data to empirically ground agents' behaviours in LU ABMs has received particular attention [24,52,55,56].

In this section, we survey the literature on empirical LU ABMs to characterize how agents' behaviour has been designed. We mainly based our analysis on reviews already available on the topic, such as [16,34,53,55]. We define three different approaches, increasingly reliant on data: theory-based, heuristic and data-driven ABMs (Figure 1).

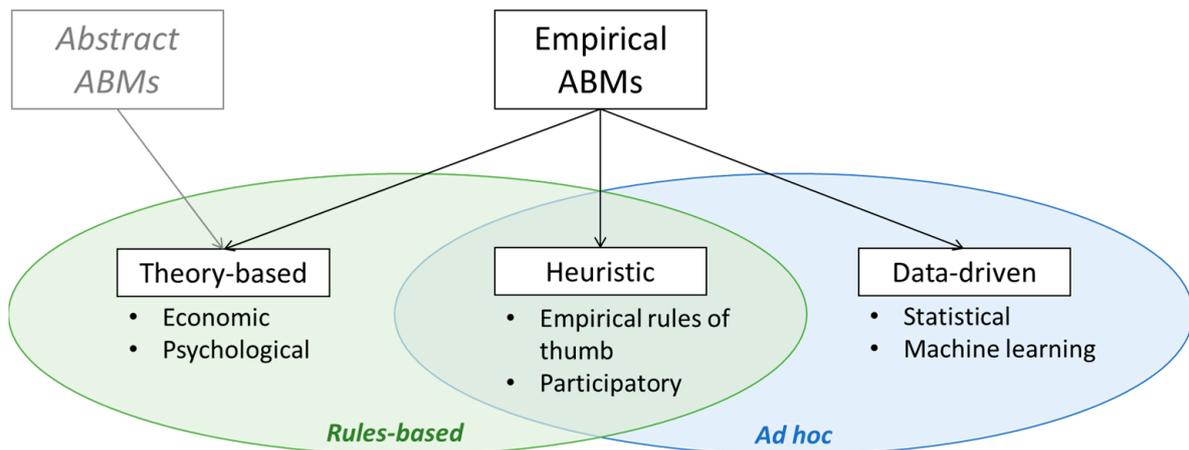


Figure 1. Approaches to design agents' behaviours in agent-based models (ABMs).

3.1. Theory-Based ABMs

Theory-based ABMs derive agents' decision-making rules from theoretical behavioural models and are the most frequent approach for abstract ABMs. The first frameworks developed for empirically grounded ABMs, such as [45], were therefore focused on the integration of empirical data to initialise agents' attributes and calibrate the parameters of their behavioural rules, already pre-defined through theories.

The most common theory-based ABMs are economic models [16]. This category encompasses various approaches, from simple evaluations of the costs and benefits of each LU choice to proper mathematical optimization decision rules, all based on the idea that agents are maximizing economic return [34,55]. A well-known implementation of this approach is the expected utility theory, in which agents choose the option that maximises their utility under risk [16]. In neoclassical economics, this translates into the Homo economicus agent, one which has complete knowledge, can perform perfect calculations and acts purely in its self-interest [16]. This paradigm has, however, been widely challenged, stressing the possibility of harnessing ABMs' individual representation to consider agents with rationality bounded by biases and a lack of information and knowledge, relaxing neoclassical assumptions [19–21,23,57]. A prominent example in this regard is the theory of Satisficing when agents review the options and stop their research as soon as they find one that matches their expectations [16,32,55]. Bounded rationality is now common in LU ABMs, even though many models continue using complete rationality [16].

The other main theoretical framework comes from psychological and cognitive models, where cognitive maps and abilities, social norms and biases are the main decision-making drivers [55]. A prominent example is the theory of planned behaviour, where perceived social pressure ("subjective norms") and internal and external barriers ("perceived behavioural control") are considered to have a fundamental role in agents' behaviour [58]. Another approach is the Consumat model, where agents engage in and switch among different cognitive strategies, such as repetition or imitation, depending on their needs and uncertainty [59]. The application of cognitive models is largely lagging behind economic models in LU ABMs. In particular, the role of emotions, values and norms, important for environmental management, are often overlooked [16,54].

3.2. Heuristic ABMs

The majority of LU ABM studies from 2000 onward adopted ad hoc implementations of agents' behaviours based on the knowledge and data collected for the specific case study and without theoretical justifications [16]. Among these, we define heuristic ABMs as ad hoc implementations that are still rules-based, i.e., where agents' behaviour consists of clearly pre-set rules and processes, such as if-then statements [32,60]. This category, therefore, encompasses various approaches sharing an easy and intuitive interpretation [34].

Some heuristic ABMs rely on empirical rules of thumb, backed by empirical insights but extracted without computationally intensive methods. These rules can be inductively derived from qualitative and quantitative data, reproducing observed patterns and behaviours, obtained from direct observations and experts' opinions or based on stylised facts abstracted from real-world studies [32,55,61]. Participatory ABMs are another heuristic approach. They directly involve stakeholders and decision-makers during each stage of the modelling process, enabling an easy design of behavioural rules and benefiting from real-time feedback [26,62]. They harness the bottom-up approach of ABMs, asking real-world entities that will be represented in the model how they would behave under different conditions.

3.3. Data-Driven ABMs

Agents' decision-making can also be modelled without expressing pre-set rules and processes but strictly relying on observed data. We define these models as data-driven ABMs. This is not because empirical theory-based and heuristic models work without data. All types of models use data to varying degrees. However, in data-driven ABMs, the desired behavioural patterns are inferred by fitting statistical models to the available data or learned by training ML models on these [34,40]. These models define agents' behaviour depending on internal parameters concerning the agents' state variables and attributes and external ones related to other agents' states and environmental conditions. With this approach, the modeller is not required to specify any pre-defined rule to express agents' behaviour or interactions between agents and reciprocal influence. This can be particularly useful when these rules cannot be based on known theories and are not intuitive [63]. This can potentially reduce—but does not cancel, as will be highlighted in the framework description—the risk for modellers to project their subjectivity and worldviews in the ABM they are developing, which is a common pitfall [64,65]. The data-driven perspective is an important mark in the ongoing shift from KISS to KIDS modelling frameworks. In ABMs, the approach has historically been to design relatively simple agents' and let complexity arise mainly from the hard-coded rules governing the interactions among them and with the environment [39]. Simple behavioural rules are attractive due to their easy interpretation but can result in simplistic representations of the agents and unrealistic results. Data-driven ABMs enable a departure from this paradigm to build complex agents, opening the possibility of considering a large number of variables within their behavioural model [39]. Moreover, some ML models can capture highly non-linear dynamics and behavioural rules, which are common in complex adaptive systems such as LU change.

Despite being hazy in many situations, a distinction between statistics and ML can be drawn in terms of purpose: while statistics focuses on inference, on understanding the data generation process, ML aims at generalisation, at prediction on out-of-sample data [66]. ML is, therefore, better suited than statistics for policy assessment, where predictive performance is fundamental [34]. ML can also handle datasets with a comparable number of variables and data points better than statistics [52,66], as can be the case in the agricultural census available at a certain aggregated level and surveys. This is also thanks to the ability of ML to build high-performance models without prior knowledge and without making explicit assumptions about the system [66]. Features (i.e., input variables) selection methods and algorithms training provide reliable methods to support the selection of the most important drivers of each behaviour. ML can, therefore, help reduce the impact of two common causes of low prediction accuracies of ABMs: the inclusion of many unreliable

assumptions and the need to decide a priori which mechanisms to include and exclude [64]. For all these reasons, we argue that ML has the potential to support the development of robust data-driven ABMs for agricultural policy assessment.

4. Machine Learning and ABMs

ML generally refers to algorithms able to learn from data automatically. ML algorithms discover patterns in the datasets they are trained on and generalize the knowledge acquired [52,67]. Over the last two decades, ML experienced dramatic development and is nowadays one of the most rapidly growing technical fields [68]. ML algorithms have been applied in many different contexts, and their integration with computational simulations has been receiving increasing attention [69,70]. ABMs have taken part in this trend as well, and there has been a growing interest in the use of ML in and for them, especially from 2013 onwards [52,71]. These uses can be categorized into three main threads.

The first thread, by far the most common, consists of using reinforcement learning to equip goal-oriented agents with a dynamic learning process [52]. An example of this approach applied to residential land growth is given by [72], where household agents learn both from their past actions and interpersonal exchange. A reinforcement learning approach modified to consider agents' bounded rationality served as inspiration for the framework proposed by [63]. This consists of generating agents training neural networks on data obtained during an "Experience" modelling stage, where the goal-oriented agents act and obtain rewards. Despite its interesting applications, such an approach is not purely data-driven, as defined here. In fact, the application of reinforcement learning does not require a dataset from which to elicit behaviours but the definition of goals and rewards, which has to rely on theoretical or heuristic rules.

The second application concerns the use of ML to support ABMs' simulation experiments and results analysis. The empirical trend of ABMs and their increase in complexity, which allowed for a growing explanatory power, came at the cost of making the process of obtaining useful insights harder [65,73]. Clustering algorithms can group similar scenarios, and classification and regression algorithms can relate input values with the produced scenarios [73,74], for instance, used support-vector machine classification to study the equilibria between intensive and extensive agriculture based on external drivers, despite misclassifying 18% of the cases. More complex models can, moreover, create issues in terms of computational efficiency and tractability, which is also true for calibration procedures, and ML algorithms can help to cope with this issue despite the risk of sacrificing accuracy [69,75]. Finally, ML has also been used to improve the validation and verification of ABMs [52].

The third thread consists of a substantial integration of ML in ABMs to model agents' behaviour by relying only on data. This corresponds to the data-driven approach suggested in this paper and will, therefore, be the focus of the following discussion. No general and accepted framework for the integration of ML in data-driven ABMs exists, despite some noteworthy attempts. In the field of innovation diffusion, mentioned [39] as the first to suggest a generic framework for data-driven ABMs. They studied the diffusion of solar panels adoption and showed the ability of data-driven ABMs to outperform a theory-based, utility-based approach. Additionally, [40] identified a handful of attempts to draft a data-driven ABM method from 2007, stressing that none presented a way to generate a model structure and individual agents from data. Therefore, they presented only "initial steps towards an agent-based modelling approach that focuses on individual-level data to generate agent behavioural rules and initialize agent attribute values" and applied it to a case study without comparing it to a more traditional approach. The absence of general guidelines is also, in part, justified by the very different context in which this modelling approach was applied and by the various possibilities in terms of the granularity and size of the data available in different fields [52].

For LU, the literature on ML and ABM integration is extremely scarce. We carried out a search on Google Scholar combining "machine learning" (or the names of the most

used ML algorithms) with “agent-based model” (or terms commonly used interchangeably as “agent-based simulations” or “multi-agent simulations”) and “land-use”, and then reviewed the most cited results to see if they were actually using the ML algorithms to model agents’ behaviours, excluding papers employing reinforcement learning. Moreover, [76] used neural networks to model agents’ preferences on spatial planning, while [41] and [77] modelled the probability of LU change. We found only two studies addressing specifically agricultural LU. Furthermore, [50] used both qualitative and quantitative data collected through interviews and surveys to build a Bayesian belief network representing farmers’ decisions on whether to participate in a scheme of payments for ecosystem services. However, they still rely on a theoretical opinion dynamic model for social interaction and properly validate only the Bayesian belief network and not the entire simulation. Thus, [56] used a clustering and regression analysis to study the relation of LUs in Portugal with landscape characteristics. Additionally, these studies have limited spatial scope as they were performed at the county level. To our knowledge, the only LU ABM that takes a data-driven approach as defined here is ref [78], which relies on two ML algorithms to model the decision of Portuguese farmers regarding the installation of sustainable pastures in a payment for ecosystem services programme, without relying on any additional pre-defined behavioural rule.

5. A Framework for Data-Driven LU ABMs

This section builds on Sections 2–4 and presents our proposal for a tentative framework to develop data-driven LU ABMs. Section 5.1 presents the flowchart of a basic data-driven ABM timestep to provide a first idea of its functioning. Section 5.2 then lists and describes the stages required to construct such a model, highlighting the key issues.

5.1. Model Timestep

Figure 2 sketches the proposed data-driven ABM framework in its simplest possible form, i.e., with a population of homogenous agents taking the same single decision at each ABM timestep. LU ABMs usually run in discrete timesteps representing a month, a season or an agricultural year. At the beginning of each timestep, the agents sense all the variables required for their decision-making process regarding the state of the system and of other agents. The behavioural model of the agents, consisting of a previously trained ML model (as Section 5.2.4 explains), takes these variables as input and outputs the outcome of the decision. Then, the agents’ state variables are updated accordingly to record the changes, and, if needed, the model’s variables are also updated (for instance, to keep track of the diffusion of a certain practice). This concludes the ABM timestep, and the following one can start.

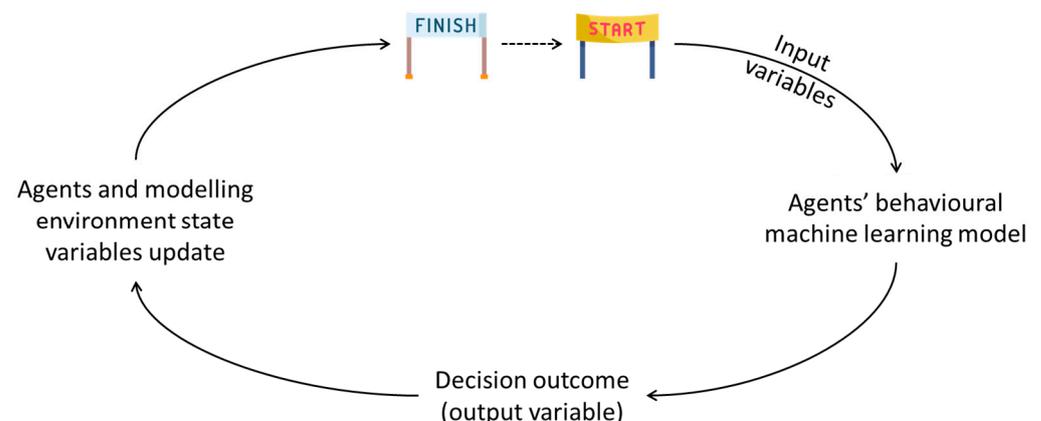


Figure 2. Flowchart of a basic timestep of the proposed data-driven agent-based models framework.

This description shows how the ABM architecture should strictly provide the required variables to the agents, trigger their behavioural model and update the state variables of

the agents and of the modelling environment. Those operations should not introduce any theoretical or heuristic rule. The key component is instead the agents' behavioural model, which completely determines the evolution of the system.

5.2. Model Implementation

Figure 3 shows the modelling stages of the proposed data-driven approach. This does not aim at providing strict guidelines but rather at sketching an implementation procedure for future case studies. It focuses on the stages concerning the ABM design and, in particular, the generation of the agents through ML algorithms. Data collection and manipulation are also extensively treated since, contrary to more traditional approaches, they highly influence the design of the ABM, and therefore, these stages have to be conducted in parallel with continuous feedback. The only analysis stage properly treated is the one concerning the ML models, characteristic of data-driven ABM. For the analysis of the ABM itself, we instead provide only some brief observations intended as initial points of discussion for future work.

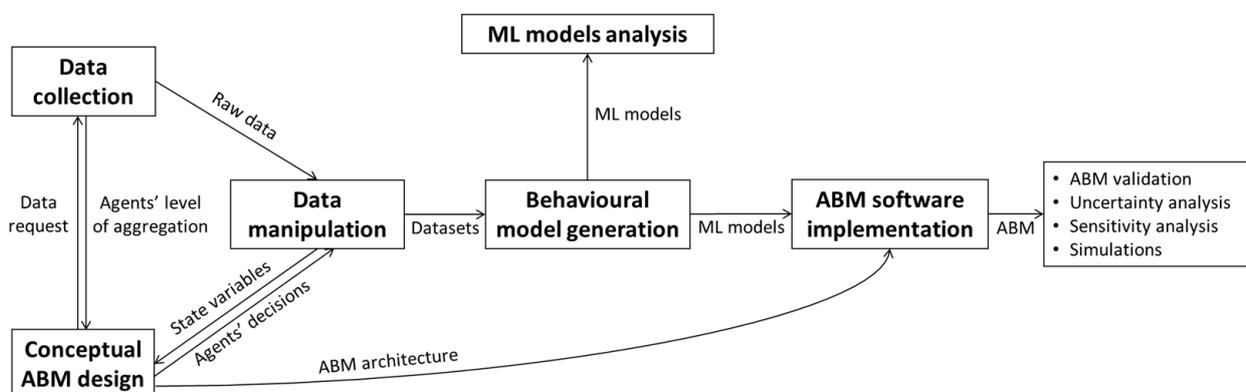


Figure 3. Modelling stages of the data-driven agent-based modelling framework and information or elements exchanged among these stages.

5.2.1. Data Collection

A data-driven approach is characterized by the need for large amounts of micro-level data to train the ML models constituting the behavioural models. These are also the same data used to instantiate the state variables of the agents and of the modelling environment in the ABM. The required data depend, therefore, on the conceptual ABM design (0) and particularly concern the characterization of the farmers, the biophysical layer and LU.

The characterization of farmers requires socio-economic and demographic data, which can be retrieved mainly from agricultural censuses and surveys/interviews. Censuses usually cover the entire farmer population. However, censuses might be available only at a certain level of aggregation, creating problems in data matching for privacy reasons. The level of aggregation also influences the decision of the main decision-makers controlling LU and, therefore, the ABM design. The default choice is farmers, but if key data are available only at a more aggregated level, agents over multiple farms or municipalities can be defined. Censuses can often lack important variables, as the questionnaire is pre-determined and not tailored to the study. Surveys and interviews collect instead the specific information needed for the study but demand time and resources to conduct them. Resources needed to perform interviews can be reduced through collaboration with farmers' organisations, which are an important proxy for social diffusion. They can, however, reach only a limited population, which may require over-sampling to represent the entire system of interest [32,44]. Acquiring data repeatedly over time for the same population is important to predict future trends (forecasting) [45]. Census data are normally available at repeated intervals, which are long and, therefore, may require interpolation for smaller timesteps.

Biophysical data holds particular importance in LU applications in order to characterise, for instance, the suitability of certain LUs. Remotely sensed data on climatic conditions and soil properties are retrievable through GIS software from geospatial datasets. The use of earth observation data enables surveying large areas systematically, which also makes it possible to create models applicable at multiple scales and for many farmers.

The accuracy of data on LU is particularly important, as it is usually the target variable of the models. If the target variables are incorrect, the ML models will output wrong predictions. LU data can also be obtained from censuses, surveys (as LUCAS) or information collected for the specific case study, such as project reports. Satellite images and maps can constitute another important source.

5.2.2. Conceptual ABM Design

The ABM design starts with the specification of who the agents are, which decisions the agents with decision-making power take, and which consequences these have. Here, there is no specification of how the agents make decisions. The framework only requires defining the sequence of decisions agents make, as each non-trivial decision requires the selection and training of one ML model. These ML models represent the single decision-making rules and compose the agent's behavioural model. The decisions depend on the state of the agents. For instance, the current LU can restrict possible future choices for farmers. These specifications may need to be hardcoded. Hardcoded rules do not hamper the data-driven character of the approach as long as these rules are not assumed or based on theoretical explanations but the result of known or trivial facts. This is often the case for the consequences of decisions taken by the agents, such as the time needed before considering resowing or a new change of LU.

Agents make decisions under the influence of other agents and the state of the system. In the data-driven approach, this influence is represented by the variables that are used to train the ML behavioural models (Section 5.2.4). For instance, hardcoded social networks, peer influence and diffusion mechanisms can be replaced by proxy variables reporting the diffusion of a certain practice in neighbourhood of each farmer. The ABM design should specify the variables stored by each entity, the rules to update these and how these are sensed by the decision-makers. The modelling environment can be used to store any global information on the state of the system that the agents need to access, such as the diffusion of a certain practice. The rules to update these variables should be trivial and not introduce any theoretical assumption. Models may include additional economic and institutional entities such as governments, providing incentives and constraints for specific LUs and markets and defining price evolution. These agents may simply store variables that may change for scenario analysis and extrapolation or, if necessary, be equipped with proper data-driven behavioural models as well.

5.2.3. Data Manipulation

The design of the ABM (Section 5.2.2) identified the agents' decisions that will be modelled through ML models. Each ML model must be selected and trained, requires a dataset composed of the target variable, i.e., the object of the agent decision, and requires the features to predict it. The various data collected (Section 5.2.1), therefore, have to be merged, linking each decision with the variables that might influence its outcome. The decision of which variables to include is key since these will be the only ones able to influence the specific process. Each data point of these datasets has to refer to a decision taken by an agent at a precise moment in time. A variable representing the time period should only be included to capture eventual changes in agents' behaviour over time that cannot be represented by other available variables.

Due to the high number of variables that can be retrieved from all the data sources mentioned, a reduction of the number of features is usually advisable, especially for predictive purposes [33]. A first screening can be performed when merging the datasets, based on prior knowledge of the system, removing from the dataset variables known to

be unrelated to a certain decision. Then, common ML feature selection methods can be employed to limit subjective decisions by the modeller. Filter selection methods can, at the same time, indicate the most important variables to keep and reveal their influence on the decisions. Some examples are the ANOVA (analysis of variance) [79] and Pearson Chi-squares test [80] or the calculation of correlation coefficients, statistics evaluating the strength of the relation between the target variable and the features. Carrying out a variance inflation factor (VIF) analysis can also help to reduce multicollinearity among features and thus improve the explanatory power of the ML models [81]. This selection process influences the ABM design since it determines the variables that each agent needs to sense and that, therefore, other agents and the modelling environment need to store.

A certain number of data points in the resulting dataset should be left aside to allow for validation of the ABM on independent data. For ABMs intended for forecasting, this split should be performed timewise. Validation data points should refer to the end of the timeframe. In this way, it will be possible to properly validate the final ABM on data for the future of those used to train the ML models.

5.2.4. Behavioural Model Generation

Each database created can, at this point, be used to select and train an ML model, which should then be saved in order to be embedded in the ABM software and compose the agents' behavioural models. This procedure should follow common ML practices: test various ML algorithms, identify the most promising ones, tweak their hyperparameters and select the best-performing model through cross-validation. For each decision, the possible choices define the algorithm to be used: classification for discrete choices, such as whether to change or not LU, and regression for continuous ones, such as the extent of adoption of a certain practice.

Stochastic elements can be included in the behavioural models to harness ABMs' capacity to account for uncertainty. Classification algorithms usually provide the possibility to output a probability distribution instead of a single class. Some decisions can be broken down into multiple stages, such as participation in a certain program: the first stage is the decision to participate or not (a binary classification), and the second is the extent of the eventual participation (a regression). This "double-hurdle" approach [82] can help to provide better estimations in datasets with a continuous response and many 0 s.

The choice of the ML algorithm has two important consequences. First, each ML algorithm can learn only a set of models. If the right model is outside this set, the algorithm will be able to reach only an approximate solution [67]. Second, some ML algorithms, such as classic tree-based ones, are unable to extrapolate. This is a problem when some features assume values outside the interval that they had in the data points used to train the model, as often happens in forecasting. It is, however, important to remember that the use of ML algorithms that are capable of extrapolating assumes that the agents would keep following the same behavioural rules under new conditions.

5.2.5. ABM Software Implementation

Due to the shift of complexity from agents' pre-defined rules and interaction to the ML models, the ABM software implementation is a relatively straightforward procedure. The design of the conceptual model (Section 5.2.2) can be directly translated into code. The state variables of the agents and of the modelling environment should be initialised with the data corresponding to the year in which the simulation starts. The individual agents' timestep can be constructed by uploading the trained ML models (Section 5.2.4) and coding their calls. However, modellers should pay attention to the choices made during the software implementation phase, which might imply assumptions about the functioning of the system. An example is the sequence of activation of the agents' behavioural models: ABM can present very different results if the agents act sequentially or in parallel.

The correspondence of the implemented software to the conceptual model should be checked for coding mistakes, to especially ensure that the ML models predict as expected.

A first check should ensure that the features are passed to the ML models in the exact order that was used to train them. A second check can consist in comparing the outcome of individual agents' decisions in the ABM with the results that are obtained running the ML models outside the ABM.

5.2.6. ML Models Analysis

An analysis of the ML model predictions should be performed at any moment after these models are generated. This analysis should aim at improving the understanding of the agents' decision-making rules represented by the ML models.

Some ML algorithms, such as those relying on polynomial relations and tree-based ones of limited size, can be directly interpreted. More complex ML models are often referred to as grey- or even black-box models for the difficulties of this process. Nevertheless, there has been significant progress in the interpretation of their results through so-called model-agnostic methods [83,84]. Simpler model-agnostic methods for specific purposes are, for instance, partial dependence plots to understand the marginal influence of one or two features on the model output [85] and permutation feature importance [86]. Arguably the most complete and theoretically robust model-agnostic method available at the moment to study ML models' predictions is the SHAP (Shapley additive explanations) framework [87], based on game theory concepts, which, however, becomes computationally expensive when large datasets are involved.

5.2.7. Following Stages

In data-driven approaches, a calibration of the ABM itself should not be necessary. Calibration is usually required to define the values of some parameters of the ABM to match empirical data and can become a cumbersome process, especially in models that strive for increased realism and, therefore, include many parameters [75]. In our data-driven framework, the calibration process is substituted by the selection and training of the ML models. None of the rules hardcoded in the ABM should require an independent calibration since these rules should only represent known or trivial facts (see Section 5.2.2).

The ABM should be validated not only at the macro-level on aggregate patterns but also at the micro-level and on data unused to train the ML models and left aside, as specified in Section 5.2.3. These data should actually remain unknown to the modeller to allow for a proper validation intended to test predictions [33]. The micro-data used to develop the model can also assess how well the model predicts individual agents' behaviour, avoiding the risk of generating the correct macro-level patterns with an incorrect micro-specification. This additional layer of micro-level validation improves the robustness of the ABM's predictions, together with the cross-validation performed for the individual ML models [39,40]. If not possible to leave aside a validation set, data collection activities with the specific aim of validating the ABM can be designed. Another approach can be to use expert opinions, despite being qualitative and less reliable.

Uncertainty and, in particular, sensitivity analysis are important for policy assessment [15,33]. Uncertainty analysis should focus on the uncertainty both embedded in the input data and caused by their manipulation, being the most important determinants of data-driven ABMs outcomes. Sensitivity analysis is critical to provide macro-level explanatory insights on the ABM dynamics complementary to the ones from the ML models analysis, which are focused on the single decision-making rules. There is no accepted standardized method to conduct such analysis for traditional ABMs [65,88], and a data-driven approach is likely to complicate the process further. The development of a proper method to conduct uncertainty and sensitivity analysis of data-driven ABMs, supported by ML as described in Section 4, would constitute an important extension of the proposed framework. ML could additionally support ABMs' output analysis, as also already described in Section 4.

6. Discussion

6.1. Challenges for Data-Driven LU ABMs

We identify two main challenges for the application of the proposed data-driven framework for agricultural policy assessment. The first is the difficulty of understanding data-driven ABMs. The literature often argues that if we are not able to understand the processes involved in the model any better than the real-world ones, the entire modelling effort is jeopardised [52,55,65,73]. A consequence is more difficult communication of the results, especially to non-scientists such as stakeholders and policymakers [30,45,54]. A better grasp of the factors determining farmers' decisions can also help identify leverage points and design more effective policies. The structure of data-driven ABMs is simpler than more traditional models. The data-driven approach uses one or few ML models instead of the many hardcoded theoretical and heuristic rules of traditional approaches. However, the data-driven approach only "hides" this complexity inside the ML models composing agents' behaviours. The introduction of the time-consuming stages of data collection and manipulation and ML model selection and training increases the difficulty of understanding the processes involved and of eliciting causal relations since these are not specified a priori and, therefore, easily recognizable [43,45,52]. The analysis of the ML models (Section 5.2.6) and a complementary sensitivity analysis on the entire ABM can help. However, these can be complex and long processes, especially when many ML models and data sources are used and are not guaranteed to provide straightforward insights. Improving these analyses is, therefore, fundamental, also because data-driven ABMs have the potential to become powerful explanatory tools. The ultimate aim of using ML algorithms is, in fact, the elicitation of patterns that are not intuitive to the modellers and that otherwise would not be included in the ABM. The interpretation of data-driven ABM can, therefore, provide even additional insights with respect to rules-based approaches, where these rules need to be pre-defined [42,89]. The modelling paradigm should, therefore, first focus on developing accurate ABMs and then use appropriate analysis to obtain reliable insights to communicate with stakeholders and decision-makers.

The second main challenge for data-driven ABMs concerns data availability and quality. Data availability restricts the applicability of our framework. For instance, it could not be applied to assess ex ante the introduction of a scheme of payments for ecosystem services, if we have no past data on the effect of the introduction of similar payments. If there are no data to train the ML models, they cannot evidently be used. Even when data are available, we should ensure their quality. The importance of the quality of the data fed to ML models' performance is often summarised with the sentence "garbage in, garbage out". Data are often more important than the ML algorithm used [43,67]. However, not all data have the same influence. The effects on the model's outcome of the assumptions made when manipulating the data should be evaluated [43], first when analysing the ML models and then through uncertainty and sensitivity analysis on the ABM [65]. This can then help guide additional data collection. However, some key variables may simply be unavailable and impossible to collect due to time and resource constraints or because the past was not recorded. If even proxies for those data cannot be found, important modifications of the ABM may be needed, such as the integration of theoretical or heuristic rules to compensate.

6.2. Integration of Rules-Based and Data-Driven Approaches

The right balance between data and theory and how to link these two apparently opposite approaches is widely debated in the ABM community, also for LU applications [30,54,90]. When some data are unavailable, or their quality is deemed insufficient, the integration of theoretical and heuristic rules is necessary. In these cases, different rules can be tested and validated, and the results may also point out the potential value of collecting missing data [16,32]. Such rules can also be used to improve data quality, for instance, supporting the disaggregation of datasets or providing important missing features that can then be used to train the ML models.

Including theoretical and heuristic rules can, however, also be useful, even if unnecessary. Rules can improve understanding and communication of the model in cases when this is particularly relevant [16,91]. Qualitative insights collected through surveys, interviews and participatory modelling can be directly included in the model as heuristic rules while providing clear and immediate insights into the drivers of agents' decision-making [60]. A specific example is the inclusion of mechanisms described by stakeholders, which, even when not improving results, can increase their engagement with the model and the odds of using results [15,43]. Heuristic insights can also help guide the screening of the features used to train the ML models and constrain them. For instance, [50] used experts' knowledge to constrain and select the best structure of a Bayesian belief network trained with quantitative data. Another issue to consider is that a data-driven approach can only learn from past patterns and is, therefore, likely to lean toward business-as-usual predictions. Participatory modelling [92] or the inclusion of adaptive expectations, such as on prices [93], can help consider different future scenarios and disruptive events that may modify the decision-making process of the agents. Finally, theoretical and heuristic rules can be the basis of using reinforcement learning to equip agents with learning, as already described in Section 4.

The possibilities for integrating theoretical and, in particular, heuristic rules in data-driven ABMs are eventually so various and diverse that a systematic treatment is probably unfeasible. For policy assessment, the validation of the model at the individual level should remain paramount in order to verify the reliability of the included rules.

7. Conclusions

In this paper, we presented and discussed a framework for the integration of ML in data-driven ABMs focused on LU applications. This framework is identified by the following characteristics, which differentiate it from other more traditional approaches:

- Use of empirical data since the very beginning of the modelling process and continuous feedback between model design and data collection and manipulation steps.
- Agents' behavioural models consisting of ML models learned from micro-data at the individual level, without relying on any pre-defined theoretical or heuristic rule.
- No assumption on agents' interaction and social networks, substituted by proxies for spatial and social influence used to train the agents' behavioural models.
- Validation performed on independent data at the macro-level and at the micro-level, improving the assessment of policy effects on the individuals.

This data-driven framework strives to harness the predictive power of ML algorithms, which is constantly improving. Data-driven ABMs could, in this way, better assess agricultural policies, evaluating their possible outcomes *ex post* and *ex ante*. The analysis of the trained ML models can shed light on decision-making rules and important drivers that would not be accessible otherwise, providing a potentially powerful but complex explanatory tool.

However, data-driven models are not completely free from assumptions. These are usually hidden, as in the choice of the features and the ML algorithm, in the tendency of ML models to reproduce past patterns and in the software implementation surrounding the ML models. There are methods to support some of these choices, e.g., for variables and algorithm selection, but modellers should keep in mind that their subjectivity can still influence the results. We identified the main challenges for the diffusion of data-driven ABMs in the increased difficulty of understanding their processes and retrieving the required data with sufficient quality. These issues could be addressed by developing proper methods for uncertainty and sensitivity analysis and treating how to integrate theoretical and heuristic rules without hampering predictive performances.

The data-driven framework we proposed constitutes a fundamental change of paradigm compared to more traditional approaches. We agree on the importance of understanding the processes within the ABM, but we argue that achieving robust predictions should be prioritized for policy assessment purposes. For complex and multidimensional human

systems, an overemphasis on easily explaining a wrong model is useless (when not damaging). Accurate models can be useful even without a complete understanding of the dynamics involved. We believe that models with better predictive performance are likely to be used more by policymakers if their potential is properly demonstrated, even without mechanistic cause–effect rules for all the processes involved. To increase policymakers’ trust, a comparison of the predictive performance of our data-driven framework with rules-based approaches will be fundamental. In fact, despite some successful examples of data-driven ABMs, this approach still has to be properly tested in LU.

Author Contributions: Conceptualization, G.R., R.F.M.T. and T.D.; formal analysis, G.R.; investigation, G.R.; writing—original draft preparation, G.R.; writing—review and editing, R.F.M.T. and T.D.; supervision, R.F.M.T. and T.D.; project administration, R.F.M.T.; funding acquisition, R.F.M.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by Fundação para a Ciência e Tecnologia through projects “LEAnMeat—Lifecycle-based Environmental Assessment and impact reduction of Meat production with a novel multi-level tool” (PTDC/EAM-AMB/30809/2017), 2021.07144.BD (to G. Ravaioli) and CEECIND/00365/2018 (R. Teixeira). The work was also funded by FCT/MCTES (PIDDAC) through project LARSyS—FCT Pluriannual funding 2020–2023 (UIDP/EEA/50009/2020).

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hardt, L.; O’Neill, D.W. Ecological Macroeconomic Models: Assessing Current Developments. *Ecol. Econ.* **2017**, *134*, 198–211. [[CrossRef](#)]
2. Raworth, K. *A Safe and Just Space for Humanity: Can We Live within the Doughnut?* Oxfam: Nairobi, Kenya, 2012.
3. Willett, W.; Rockström, J.; Loken, B.; Springmann, M.; Lang, T.; Vermeulen, S.; Garnett, T.; Tilman, D.; DeClerck, F.; Wood, A.; et al. Food in the Anthropocene: The EAT–Lancet Commission on Healthy Diets from Sustainable Food Systems. *Lancet* **2019**, *393*, 447–492. [[CrossRef](#)] [[PubMed](#)]
4. Vermeulen, S.J.; Campbell, B.M.; Ingram, J.S.I. Climate Change and Food Systems. *Annu. Rev. Environ. Resour.* **2012**, *37*, 195–222. [[CrossRef](#)]
5. Tilman, D.; Clark, M.; Williams, D.R.; Kimmel, K.; Polasky, S.; Packer, C. Future Threats to Biodiversity and Pathways to Their Prevention. *Nature* **2017**, *546*, 73–81. [[CrossRef](#)] [[PubMed](#)]
6. Gaube, V.; Haberl, H. Using Integrated Models to Analyse Socio-Ecological System Dynamics in Long-Term Socio-Ecological Research—Austrian Experiences. In *Long Term Socio-Ecological Research*; Singh, S.J., Haberl, H., Chertow, M., Mirtl, M., Schmid, M., Eds.; Springer: Dordrecht, The Netherlands, 2013; pp. 53–75; ISBN 978-94-007-1176-1.
7. Liu, J.; Dietz, T.; Carpenter, S.R.; Alberti, M.; Folke, C.; Moran, E.; Pell, A.N.; Deadman, P.; Kratz, T.; Lubchenco, J.; et al. Complexity of Coupled Human and Natural Systems. *Science* **2007**, *317*, 1513–1516. [[CrossRef](#)]
8. Rindfuss, R.R.; Entwisle, B.; Walsh, S.J.; An, L.; Badenoch, N.; Brown, D.G.; Deadman, P.; Evans, T.P.; Fox, J.; Geoghegan, J.; et al. Land Use Change: Complexity and Comparisons. *J. Land Use Sci.* **2008**, *3*, 1–10. [[CrossRef](#)]
9. Levin, S.; Xepapadeas, T.; Crépin, A.-S.; Norberg, J.; de Zeeuw, A.; Folke, C.; Hughes, T.; Arrow, K.; Barrett, S.; Daily, G.; et al. Social-Ecological Systems as Complex Adaptive Systems: Modeling and Policy Implications. *Environ. Dev. Econ.* **2013**, *18*, 111–132. [[CrossRef](#)]
10. *Agent-Based Modelling of Socio-Technical Systems*; Dam, K.H.; Nikolic, I.; Lukszo, Z. (Eds.) Springer: Dordrecht, The Netherlands, 2013; ISBN 978-94-007-4932-0.
11. Holling, C.S. Understanding the Complexity of Economic, Ecological, and Social Systems. *Ecosystems* **2001**, *4*, 390–405. [[CrossRef](#)]
12. Ostrom, E. A General Framework for Analyzing Sustainability of Social-Ecological Systems. *Science* **2009**, *325*, 419–422. [[CrossRef](#)]
13. Preiser, R.; Biggs, R.; De Vos, A.; Folke, C. Social-Ecological Systems as Complex Adaptive Systems: Organizing Principles for Advancing Research Methods and Approaches. *Ecol. Soc.* **2018**, *23*, 46. [[CrossRef](#)]
14. Berger, T.; Troost, C. Agent-Based Modelling of Climate Adaptation and Mitigation Options in Agriculture. *J. Agric. Econ.* **2014**, *65*, 323–348. [[CrossRef](#)]
15. Reidsma, P.; Janssen, S.; Jansen, J.; van Ittersum, M.K. On the Development and Use of Farm Models for Policy Impact Assessment in the European Union—A Review. *Agric. Syst.* **2018**, *159*, 111–125. [[CrossRef](#)]
16. Groeneveld, J.; Müller, B.; Buchmann, C.; Dressler, G.; Guo, C.; Hase, N.; Hoffmann, F.; John, F.; Klassert, C.; Lauf, T.; et al. Theoretical Foundations of Human Decision-Making in Agent-Based Land Use Models—A Review. *Environ. Model. Softw.* **2017**, *87*, 39–48. [[CrossRef](#)]
17. O’Sullivan, D.; Evans, T.; Manson, S.; Metcalf, S.; Ligmann-Zielinska, A.; Bone, C. Strategic Directions for Agent-Based Modeling: Avoiding the YAAWN Syndrome. *J. Land Use Sci.* **2016**, *11*, 177–187. [[CrossRef](#)]

18. Kremmydas, D.; Athanasiadis, I.N.; Rozakis, S. A Review of Agent Based Modeling for Agricultural Policy Evaluation. *Agric. Syst.* **2018**, *164*, 95–106. [[CrossRef](#)]
19. Macal, C.M. Everything You Need to Know about Agent Based Modelling and Simulation. *J. Simul.* **2016**, *10*, 144–156. [[CrossRef](#)]
20. Epstein, J.M. Agent-Based Computational Models and Generative Social Science. *Complexity* **2006**, *4*, 41–60. [[CrossRef](#)]
21. Parker, D.C.; Manson, S.M.; Janssen, M.A.; Hoffmann, M.J.; Deadman, P. Multi-Agent Systems for the Simulation of Land-Use and Land-Cover Change: A Review. *Ann. Assoc. Am. Geogr.* **2003**, *93*, 314–337. [[CrossRef](#)]
22. Robinson, D.T.; Brown, D.G.; Parker, D.C.; Schreinemachers, P.; Janssen, M.A.; Huigen, M.; Wittmer, H.; Gotts, N.; Promburom, P.; Irwin, E.; et al. Comparison of Empirical Methods for Building Agent-Based Models in Land Use Science. *J. Land Use Sci.* **2007**, *2*, 31–55. [[CrossRef](#)]
23. Dullinger, I.; Gattringer, A.; Wessely, J.; Moser, D.; Plutzer, C.; Willner, W.; Egger, C.; Gaube, V.; Haberl, H.; Mayer, A.; et al. A Socio-ecological Model for Predicting Impacts of Land-use and Climate Change on Regional Plant Diversity in the Austrian Alps. *Glob. Chang. Biol.* **2020**, *26*, 2336–2352. [[CrossRef](#)]
24. Filatova, T.; Verburg, P.H.; Parker, D.C.; Stannard, C.A. Spatial Agent-Based Models for Socio-Ecological Systems: Challenges and Prospects. *Environ. Model. Softw.* **2013**, *45*, 1–7. [[CrossRef](#)]
25. Happe, K.; Kellermann, K.; Balmann, A. Agent-Based Analysis of Agricultural Policies: An Illustration of the Agricultural Policy Simulator AgriPoliS, Its Adaptation and Behavior. *Ecol. Soc.* **2006**, *11*, 49. [[CrossRef](#)]
26. Matthews, R.B.; Gilbert, N.G.; Roach, A.; Polhill, J.G.; Gotts, N.M. Agent-Based Land-Use Models: A Review of Applications. *Landsc. Ecol.* **2007**, *22*, 1447–1459. [[CrossRef](#)]
27. Schreinemachers, P.; Berger, T. An Agent-Based Simulation Model of Human–Environment Interactions in Agricultural Systems. *Environ. Model. Softw.* **2011**, *26*, 845–859. [[CrossRef](#)]
28. Grimm, V.; Berger, U.; DeAngelis, D.L.; Polhill, J.G.; Giske, J.; Railsback, S.F. The ODD Protocol: A Review and First Update. *Ecol. Model.* **2010**, *221*, 2760–2768. [[CrossRef](#)]
29. Grimm, V.; Railsback, S.F.; Vincenot, C.E.; Berger, U.; Gallagher, C.; DeAngelis, D.L.; Edmonds, B.; Ge, J.; Giske, J.; Groeneveld, J.; et al. The ODD Protocol for Describing Agent-Based and Other Simulation Models: A Second Update to Improve Clarity, Replication, and Structural Realism. *JASSS* **2020**, *23*, 7. [[CrossRef](#)]
30. Laatabi, A.; Marilleau, N.; Nguyen-Huu, T.; Hbid, H.; Ait Babram, M. ODD+2D: An ODD Based Protocol for Mapping Data to Empirical ABMs. *JASSS* **2018**, *21*, 9. [[CrossRef](#)]
31. Müller, B.; Bohn, F.; Dreßler, G.; Groeneveld, J.; Klassert, C.; Martin, R.; Schlüter, M.; Schulze, J.; Weise, H.; Schwarz, N. Describing Human Decisions in Agent-Based Models—ODD + D, an Extension of the ODD Protocol. *Environ. Model. Softw.* **2013**, *48*, 37–48. [[CrossRef](#)]
32. Bruch, E.; Atwell, J. Agent-Based Models in Empirical Social Research. *Sociol. Methods Res.* **2015**, *44*, 186–221. [[CrossRef](#)]
33. Edmonds, B.; Grimm, V.; Meyer, R.; Montañola, C.; Ormerod, P.; Root, H.; Squazzoni, F. Different Modelling Purposes. *JASSS* **2019**, *22*, 6. [[CrossRef](#)]
34. Zhang, H.; Vorobeychik, Y. Empirically Grounded Agent-Based Models of Innovation Diffusion: A Critical Review. *Artif. Intell. Rev.* **2019**, *52*, 707–741. [[CrossRef](#)]
35. Lempert, R. Agent-Based Modeling as Organizational and Public Policy Simulators. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 7195. [[CrossRef](#)] [[PubMed](#)]
36. Data For Policy-Making in the Big Data Era: Opportunities and Challenges. In *Proceedings the Book of Data for Policy 2015 Conference, Cambridge, UK, 15–17 June 2015*; University of Cambridge: Cambridge, UK, 2015.
37. Androusoyopoulou, A.; Charalabidis, Y. A Framework for Evidence Based Policy Making Combining Big Data, Dynamic Modelling and Machine Intelligence. In *Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance, Galway, Ireland, 4–6 April 2018*.
38. Lee, J.W. Big Data Strategies for Government, Society and Policy-Making. *J. Asian Financ. Econ. Bus.* **2020**, *7*, 475–487. [[CrossRef](#)]
39. Zhang, H.; Vorobeychik, Y.; Letchford, J.; Lakkaraju, K. Data-Driven Agent-Based Modeling, with Application to Rooftop Solar Adoption. *Auton. Agent Multi-Agent Syst.* **2016**, *30*, 1023–1049. [[CrossRef](#)]
40. Kavak, H.; Padilla, J.J.; Lynch, C.J.; Diallo, S.Y. Big Data, Agents and Machine Learning: Towards a Data-Driven Agent-Based Modeling Approach. In *Proceedings of the Annual Simulation Symposium (ANSS 2018), Baltimore, MD, USA, 15–18 April 2018*; Society for Modeling and Simulation International (SCS): Baltimore, MD, USA, 2018.
41. Zhao, L.; Peng, Z.-R. LandSys II: Agent-Based Land Use–Forecast Model with Artificial Neural Networks and Multiagent Model. *J. Urban Plann. Dev.* **2015**, *141*, 04014045. [[CrossRef](#)]
42. Heppenstall, A.; Crooks, A.; Malleson, N.; Manley, E.; Ge, J.; Batty, M. Future Developments in Geographical Agent-Based Models: Challenges and Opportunities. *Geogr. Anal.* **2021**, *53*, 76–91. [[CrossRef](#)]
43. Buchmann, C.M.; Grossmann, K.; Schwarz, N. How Agent Heterogeneity, Model Structure and Input Data Determine the Performance of an Empirical ABM—A Real-World Case Study on Residential Mobility. *Environ. Model. Softw.* **2016**, *75*, 77–93. [[CrossRef](#)]
44. Janssen, M.A.; Ostrom, E. Empirically Based, Agent-Based Models. *Ecol. Soc.* **2006**, *11*, 37. [[CrossRef](#)]
45. Hassan, S.; Antunes, L.; Pavon, J.; Gilbert, N. Stepping on Earth: A Roadmap for Data-Driven Agent-Based Modelling. In *Proceedings of the 5th Conference of the European Social Simulation Association (ESSA08), Brescia, Italy, 1–5 September 2008*; p. 12.

46. Edmonds, B.; Moss, S. From KISS to KIDS—An ‘Anti-Simplistic’ Modelling Approach. In *Multi-Agent and Multi-Agent-Based Simulation*; Davidsson, P., Logan, B., Takadama, K., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2005; Volume 3415, pp. 130–144; ISBN 978-3-540-25262-7.
47. Marvuglia, A.; Navarrete Gutiérrez, T.; Baustert, P.; Benetto, E. Luxembourg Institute of Science and Technology (LIST), 5, avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg Implementation of Agent-Based Models to Support Life Cycle Assessment: A Review Focusing on Agriculture and Land Use. *AIMS Agric. Food* **2018**, *3*, 535–560. [[CrossRef](#)]
48. Acosta, A.L.; Rounsevell, D.A.M.; Bakker, M.; Van Doorn, A.; Gómez-Delgado, M.; Delgado, M. An Agent-Based Assessment of Land Use and Ecosystem Changes in Traditional Agricultural Landscape of Portugal. *Intell. Inf. Manag.* **2014**, *6*, 55–80. [[CrossRef](#)]
49. Chen, X.; Viña, A.; Shortridge, A.; An, L.; Liu, J. Assessing the Effectiveness of Payments for Ecosystem Services: An Agent-Based Modeling Approach. *Ecol. Soc.* **2014**, *19*, art7. [[CrossRef](#)]
50. Sun, Z.; Müller, D. A Framework for Modeling Payments for Ecosystem Services with Agent-Based Models, Bayesian Belief Networks and Opinion Dynamics Models. *Environ. Model. Softw.* **2013**, *45*, 15–28. [[CrossRef](#)]
51. Smajgl, A.; Brown, D.G.; Valbuena, D.; Huigen, M.G.A. Empirical Characterisation of Agent Behaviours in Socio-Ecological Systems. *Environ. Model. Softw.* **2011**, *26*, 837–844. [[CrossRef](#)]
52. Dahlke, J.; Bogner, K.; Müller, M.; Berger, T.; Pyka, A. Bernd Ebersberger Is the Juice Worth the Squeeze? Machine Learning in and for Agent-Based Modelling. *arXiv* **2020**. [[CrossRef](#)]
53. Bartkowski, B.; Bartke, S. Leverage Points for Governing Agricultural Soils: A Review of Empirical Studies of European Farmers’ Decision-Making. *Sustainability* **2018**, *10*, 3179. [[CrossRef](#)]
54. Huber, R.; Bakker, M.; Balmann, A.; Berger, T.; Bithell, M.; Brown, C.; Grêt-Regamey, A.; Xiong, H.; Le, Q.B.; Mack, G.; et al. Representation of Decision-Making in European Agricultural Agent-Based Models. *Agric. Syst.* **2018**, *167*, 143–160. [[CrossRef](#)]
55. An, L. Modeling Human Decisions in Coupled Human and Natural Systems: Review of Agent-Based Models. *Ecol. Model.* **2012**, *229*, 25–36. [[CrossRef](#)]
56. Bakker, M.M.; van Doorn, A.M. Farmer-Specific Relationships between Land Use Change and Landscape Factors: Introducing Agents in Empirical Land Use Modelling. *Land Use Policy* **2009**, *26*, 809–817. [[CrossRef](#)]
57. Farmer, J.D.; Hepburn, C.; Mealy, P.; Teytelboym, A. A Third Wave in the Economics of Climate Change. *Environ. Resour. Econ.* **2015**, *62*, 329–357. [[CrossRef](#)]
58. Ajzen, I. The Theory of Planned Behavior. *Organ. Behav. Hum. Decis. Process.* **1991**, *50*, 179–211. [[CrossRef](#)]
59. Jager, W.; Janssen, M.A.; De Vries, H.J.M.; De Greef, J.; Vlek, C.A.J. Behaviour in Commons Dilemmas: Homo Economicus and Homo Psychologicus in an Ecological-Economic Model. *Ecol. Econ.* **2000**, *35*, 357–379. [[CrossRef](#)]
60. Runck, B.C.; Manson, S.; Shook, E.; Gini, M.; Jordan, N. Using Word Embeddings to Generate Data-Driven Human Agent Decision-Making from Natural Language. *Geoinformatica* **2019**, *23*, 221–242. [[CrossRef](#)]
61. Schenk, T.A. Using Stakeholders’ Narratives to Build an Agent-Based Simulation of a Political Process. *Simulation* **2014**, *90*, 85–102. [[CrossRef](#)]
62. Gaube, V.; Kaiser, C.; Wildenberg, M.; Adensam, H.; Fleissner, P.; Kobler, J.; Lutz, J.; Schaumberger, A.; Schaumberger, J.; Smetschka, B.; et al. Combining Agent-Based and Stock-Flow Modelling Approaches in a Participative Analysis of the Integrated Land System in Reichraming, Austria. *Landsc. Ecol* **2009**, *24*, 1149–1165. [[CrossRef](#)]
63. Jäger, G. Using Neural Networks for a Universal Framework for Agent-Based Models. *Math. Comput. Model. Dyn. Syst.* **2021**, *27*, 162–178. [[CrossRef](#)]
64. Edmonds, B.; Aodha, L. ní Using Agent-Based Simulation to Inform Policy—What Could Possibly Go Wrong? In *Simulating Social Complexity—A Handbook*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 801–822.
65. Lee, J.-S.; Filatova, T.; Ligmann-Zielinska, A.; Hassani-Mahmooui, B.; Stonedahl, F.; Lorscheid, I.; Voinov, A.; Polhill, G.; Sun, Z.; Parker, D.C. The Complexities of Agent-Based Modeling Output Analysis. *JASSS* **2015**, *18*, 4. [[CrossRef](#)]
66. Bzdok, D.; Altman, N.; Krzywinski, M. Statistics versus Machine Learning. *Nat. Methods* **2018**, *15*, 233–234. [[CrossRef](#)]
67. Domingos, P. A Few Useful Things to Know about Machine Learning. *Commun. ACM* **2012**, *55*, 78–87. [[CrossRef](#)]
68. Jordan, M.I.; Mitchell, T.M. Machine Learning: Trends, Perspectives, and Prospects. *Science* **2015**, *349*, 255–260. [[CrossRef](#)]
69. Rolnick, D.; Donti, P.L.; Kaack, L.H.; Kochanski, K.; Lacoste, A.; Sankaran, K.; Ross, A.S.; Milojevic-Dupont, N.; Jaques, N.; Waldman-Brown, A.; et al. Tackling Climate Change with Machine Learning. *arXiv* **2019**. [[CrossRef](#)]
70. von Rueden, L.; Mayer, S.; Sifa, R.; Bauckhage, C.; Garcke, J. Combining Machine Learning and Simulation to a Hybrid Modelling Approach: Current and Future Directions. In *Advances in Intelligent Data Analysis XVIII*; Berthold, M.R., Feelders, A., Krempel, G., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 548–560.
71. An, L.; Grimm, V.; Sullivan, A.; Turner, B.L., II; Malleson, N.; Heppenstall, A.; Vincenot, C.; Robinson, D.; Ye, X.; Liu, J.; et al. Challenges, Tasks, and Opportunities in Modeling Agent-Based Complex Systems. *Ecol. Model.* **2021**, *457*, 109685. [[CrossRef](#)]
72. Li, F.; Li, Z.; Chen, H.; Chen, Z.; Li, M. An Agent-Based Learning-Embedded Model (ABM-Learning) for Urban Land Use Planning: A Case Study of Residential Land Growth Simulation in Shenzhen, China. *Land Use Policy* **2020**, *95*, 104620. [[CrossRef](#)]
73. Pereda, M.; Santos, J.I.; Galán, J.M. A Brief Introduction to the Use of Machine Learning Techniques in the Analysis of Agent-Based Models. In *Advances in Management Engineering*; Hernández, C., Ed.; Lecture Notes in Management and Industrial Engineering; Springer International Publishing: Cham, Switzerland, 2017; pp. 179–186; ISBN 978-3-319-55888-2.
74. van Strien, M.J.; Huber, S.H.; Anderies, J.M.; Grêt-Regamey, A. Resilience in Social-Ecological Systems: Identifying Stable and Unstable Equilibria with Agent-Based Models. *Ecol. Soc.* **2019**, *24*, art8. [[CrossRef](#)]

75. Lamperti, F.; Roventini, A.; Sani, A. Agent-Based Model Calibration Using Machine Learning Surrogates. *J. Econ. Dyn. Control.* **2018**, *90*, 366–389. [[CrossRef](#)]
76. Zhao, X.; Ma, X.; Tang, W.; Liu, D. An Adaptive Agent-Based Optimization Model for Spatial Planning: A Case Study of Anyue County, China. *Sustain. Cities Soc.* **2019**, *51*, 101733. [[CrossRef](#)]
77. Hashemi Aslani, Z.; Omidvar, B.; Karbassi, A. Integrated Model for Land-Use Transformation Analysis Based on Multi-Layer Perception Neural Network and Agent-Based Model. *Environ. Sci. Pollut. Res.* **2022**. [[CrossRef](#)]
78. Ravaioli, G.; Domingos, T.; Teixeira, R.F. Data-driven agent-based modelling of incentives for carbon sequestration: The case of sown biodiverse pastures in Portugal. *J. Environ. Manag.* **2023**, *338*, 117834.
79. Kaufmann, J.; Schering, A. Analysis of Variance ANOVA. In *Wiley StatsRef: Statistics Reference Online*; Bala-krishnan, N., Colton, T., Everitt, B., Piegorsch, W., Ruggeri, F., Teugels, J.L., Eds.; Wiley: Hoboken, NJ, USA, 2014; ISBN 978-1-118-44511-2.
80. Pearson, K.X. On the Criterion That a given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling. *Philos. Mag. Ser.* **1900**, *5*, 157–175. [[CrossRef](#)]
81. Daoud, J. Multicollinearity and Regression Analysis. In *Journal of Physics: Conference Series*; International Islamic University Malaysia: Kuala Lumpur, Malaysia, 2017; Volume 949, p. 012009.
82. Cragg, J.G. Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. *Econometrica* **1971**, *39*, 829. [[CrossRef](#)]
83. Molnar, C. Interpretable Machine Learning. Independently published. 2019; ISBN 979-8411463330.
84. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, Methods, and Applications in Interpretable Machine Learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 22071–22080. [[CrossRef](#)]
85. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Statist.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
86. Fisher, A.; Rudin, C.; Dominici, F. All Models Are Wrong, but Many Are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res.* **2019**, *20*, 177. [[PubMed](#)]
87. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
88. ten Broeke, G.; van Voorn, G.; Ligtenberg, A. Which Sensitivity Analysis Method Should I Use for My Agent-Based Model? *JASSS* **2016**, *19*, 5. [[CrossRef](#)]
89. Viana, C.M.; Santos, M.; Freire, D.; Abrantes, P.; Rocha, J. Evaluation of the Factors Explaining the Use of Agricultural Land: A Machine Learning and Model-Agnostic Approach. *Ecol. Indic.* **2021**, *131*, 108200. [[CrossRef](#)]
90. Koomen, E.; Diogo, V.; Dekkers, J.; Rietveld, P. A Utility-Based Suitability Framework for Integrated Local-Scale Land-Use Modelling. *Comput. Environ. Urban Syst.* **2015**, *50*, 1–14. [[CrossRef](#)]
91. Sun, Z.; Lorscheid, I.; Millington, J.D.; Lauf, S.; Magliocca, N.R.; Groeneveld, J.; Balbi, S.; Nolzen, H.; Müller, B.; Schulze, J.; et al. Simple or Complicated Agent-Based Models? A Complicated Issue. *Environ. Model. Softw.* **2016**, *86*, 56–67. [[CrossRef](#)]
92. Harb, M.; Garschagen, M.; Cotti, D.; Krätzschar, E.; Baccouche, H.; Ben Khaled, K.; Bellert, F.; Chebil, B.; Ben Fredj, A.; Ayed, S.; et al. Integrating Data-Driven and Participatory Modeling to Simulate Future Urban Growth Scenarios: Findings from Monastir, Tunisia. *Urban Sci.* **2020**, *4*, 10. [[CrossRef](#)]
93. Filatova, T. Empirical Agent-Based Land Market: Integrating Adaptive Economic Behavior in Urban Land-Use Models. *Comput. Environ. Urban Syst.* **2015**, *54*, 397–413. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.