TÉCNICO LISBOA

# Data-driven agent-based modelling of incentives for carbon sequestration

The case of sown biodiverse pastures adoption in Portugal

**Giacomo Ravaioli**

Thesis to obtain the Master of Science Degree in

**Energy Engineering and Management**

Supervisors: Prof. Ricardo Filipe de Melo Teixeira

Prof. Tiago Morais Delgado Domingos

## Examination Committee

Chairperson: Prof. Susana Isabel Carvalho Relvas

Supervisor: Prof. Ricardo Filipe de Melo Teixeira

Member of the Committee: Prof. Susana Margarida da Silva Vieira

**January 2021**

# Abstract

Sown biodiverse permanent pastures rich in legumes (SBP) provide multiple ecosystem services including carbon sequestration. The goal of this thesis is to understand the factors that drive the adoption of SBP and assess the effectiveness of policies aimed at expanding their implementation using agent-based models (ABMs). ABMs are suited to study the complexity of land systems thanks to their ability to model the behaviour of landowners and their mutual interactions with their environment. The analysis involved (a) theory and data-driven ABMs developed using a survey of 43 farmers, and (b) a municipality-based ABM using data from a project funded by the Portuguese Carbon Fund (PCF) to incentivise SBP adoption between 2009-2014. This was the first country-level application of land use ABMs for innovation diffusion and policy design in the agricultural sector that relied entirely on empirical data and machine learning. Results showed that simplified economic models are insufficient to explain farmers decisions. Data-driven models involving interactions between farmers and biophysical conditions captured the underlying trend of yearly adoption in Portugal. The analysis confirmed the positive effect that the PCF project had on expanding SBP adoption but predicted a higher than expected adoption in the absence of incentives. The modelling framework here implemented constitutes the basis for future work aimed at supporting the design of new policies to further spread SBP. Conditions for future exploitation of the model involve surpassing data limitations, developing a unified farmer-level framework with wide spatial and temporal scope, and performing additional validation of model forecasts.

**Keywords:** payments for ecosystem services, policy design, grasslands, land use, innovation diffusion, machine learning.

# Resumo

As pastagens permanentes semeadas biodiversas ricas em leguminosas (PSB) providenciam múltiplos serviços de ecossistema incluindo sequestro de carbono. O objectivo desta tese é entender os factores explicativos da adopção de PSB e avaliar a eficiência de políticas para a sua expansão usando modelos baseados em agentes (MBA). MBAs são apropriados para estudar a complexidade de sistemas de uso do solo devido à sua capacidade para modelar comportamentos individuais de agricultores e suas interacções com o ambiente. A análise envolveu (a) MBAs de base teórica e empírica desenvolvidos a partir de um inquérito a 43 agricultores, e (b) MBAs à escala dos concelhos utilizando dados de um projecto financiado pelo Fundo Português de Carbono (FPC) que incentivou a intalação de PSB entre 2009-2014. Esta foi a primeira aplicação nacional de MBAs para difusão de inovação e desenho de políticas no sector agrícola baseada estritamente em dados empíricos. Os resultados demostraram a insuficiência de modelos económicos simples. Apenas modelos empíricos envolvendo interacções entre agricultores e dados biofísicos estimaram correctamente as instalações de PSB anuais em Portugal. A análise confirmou o efeito positivo do apoio do FPC na adesão mas previu uma adopção do sistema acima do esperado caso o projecto não tivesse ocorrido. O método aqui implementado poderá ser a base de futuros instrumentos de política para PSB. As condições que devem cumprir-se para tal envolvem ultrapassar limitações de disponibilidade de dados, desenvolver um modelo integrado para produtores individuais com grande resolução espacial e temporal, e realizar validação das previsões do modelo.

**Palavras-chave**: pagamentos por serviços de ecossistema, desenho de políticas, pastagens, uso do solo, difusão de inovação, aprendizagem automática.

# Acknowledgements

When I chose Select as my Master course more than two years ago, I could not have imagined what an amazing experience it would have revealed to be. All the places I have seen, all the moments I have lived and especially all the friends I have made are priceless and I could not be more aware of how lucky I am. I also could not have imagined how much I would have grown on a personal level. Despite continuously discovering new things I do not know enough about (and never having enough time for them), every week I felt a bit closer to what I aim for – which also changed a lot.

This journey was, since the very choice of undertaking it, full of decisions to take – and whoever knows me enough understands how much this can be troublesome for me. Apparently, attending a course on decision-making support models helped me just enough to take the a correct one regarding my thesis. Every moment spent working on it allowed me to learn something new, regarding topics that fascinated me but that at the beginning I barely knew. I cannot help being proud of the results achieved, but at the same time I'm aware that this, together with all the experiences of the last two years, would not have been possible without all the people that supported me and who therefore deserve the right credits and my complete gratitude.

First and foremost, I want to thank Ricardo Teixeira, the supervisor of this thesis, for the immense support it provided me – which I am aware is not common at all. Despite all the difficulties for causes out of our control – and two surely more joyful personal reasons he had – he managed to guide me throughout this year and share his knowledge with me. Even when I would have expected irritated reactions from anybody else, he always patiently solved any issue, often with a laugh. Without its supervision and dedication this thesis would not have been possible, and without his person the actual project analysed in this thesis may have never seen the light.

This thesis would not have existed without Tiago Domingos, its co-supervisor, since it was the admiration for his knowledge, clear since the first lesson of him I attended, that first convinced me to ask for possibilities to collaborate with him and MARETEC. When needed, his advice was precious, regarding this thesis and other academic matters.

Even more fundamentally (for me, at least), *I* would not have existed without my parents. This thesis is dedicated to them. If I am satisfied with where I am now and will accomplish something in this life, I owe it to them, to all the support they gave me and I am sure will keep on providing me. And to all the members of my family, especially my grandma, that helped me growing. Grazie, vi voglio bene!

Thanks to Leo and Berto, which were with me from the very first moment. Thanks to Vezzo, for being my partner in music and laughings. And thanks to Bianka, for sharing every up and down of the last year and a half. The list of all the other friends that I would like to name would not fit in these pages, but I am sure they know I am referring to them. Thanks to my friends from Ravenna, for making me willing to always come back home. Thanks to all my friends from Bologna, for making me laugh just thinking about

what we shared. Thanks to all my Select coursemates, for making me know the world and sharing this amazing experience. And thanks to my flatmates of third floor D and E, the best aspect of this last year.

I want to thank also other people who gave a direct contribution to this thesis. Tiago Morais and Manuel Paiva dos Santos for answering to all my emails, even in their busiest moments. Nuno Fachada for sharing its knowledge on agent-based modelling with me. Vasco, for his precious hints on machine learning – and on Lisbon. Chiaravo, for its help with coding – despite his hate for Python – and more in general with every problem regarding a computer.

Finally, every person from InnoEnergy who make experiences such as mine possible will always have my gratitude, in particular all the Select coordinators and Duarte.

This work was carried out in the context of projects AnimalFuture ("Steering Animal Production Systems towards Sustainable Future" – SusAn/0001/2016) and LEAnMeat ("Lifecycle-based Environmental Assessment and impact reduction of Meat production with a novel multi-level tool" – PTDC/EAM-AMB/30809/2017) funded by Fundação para a Ciência e Tecnologia.

After having enjoyed all this luck, I feel it is time for me to give something in return and I hope this thesis will only be the first glimpse of it.

Giacomo Ravaioli

December 2020

# Table of Contents

# List of Figures

# List of Tables

# List of Acronyms

| | |
|---|---|
| **ABM** | Agent-based model |
| **AF** | Animal Future |
| **C** | Carbon |
| **CAS** | Complex adaptive systems |
| **$CH_4$** | Methane |
| **$CO_2$** | Carbon dioxide |
| **$CO_{2\text{-}eq}$** | Carbon dioxide equivalent |
| **CPI** | Consumer price index |
| **CV** | Cross-validation |
| **EDNPV** | Expected differential net present value |
| **GHG** | Greenhouse gas |
| **GIS** | Geographic Information System |
| **IA** | Impact assessment |
| **IST** | Instituto Superior Técnico |
| **KP** | Kyoto Protocol |
| **LCA** | Life cycle assessment |
| **LUCAS** | Land Use/Land Cover Area Frame Survey |
| **LUCC** | Land-use/cover change |
| **$NO_2$** | Nitrous oxide |
| **NPV** | Net present value |
| **PCF** | Portuguese Carbon Fund |
| **PES** | Payments for ecosystem services |
| **RMSE** | Root men squared error |
| **ROC AUC** | Area under the receiver operating characteristics curve |
| **SBP** | Sown biodiverse permanent pastures rich in legumes |
| **SC** | Source code |
| **SDG** | Sustainabile Development Goal |
| **SES** | Social-ecological systems |

| | |
|---|---|
| **SHAP** | SHapley Additive exPlanations |
| **SNP** | Semi-natural pastures |
| **SOM** | Soil organic matter |
| **VIF** | Variance inflation factor |

# 1 Introduction

## 1.1 The environmental importance of the food system

Over the last decades, the environmental thresholds of our planet have been steadily approached – and in many cases overshot. Climate change and biodiversity loss are two of the planetary boundaries jeopardized by anthropogenic drivers [1]. The issue becomes ever more urgent if we consider the efforts that would be required to reach a safe and just life for all humanity as outlined by Raworth [2] [3]. In line also with the United Nations Sustainable Development Goals (the SDGs), environmental sustainability cannot prescind from meeting basic human needs for the entire world's population.

Among these social foundations, nutrition holds particular relevance. Despite serious imbalances in availability that cause 820 million people to be lacking sufficient food [4], worldwide caloric production is in line with global needs [5] and the second SDG, Zero Hunger, could be currently met without a significant transgression of the planetary boundaries [3]. However, world population is projected to reach 10.9 billion people at the end of this century [6], forcing to thoroughly consider the sustainability of the food system as a whole.

Food production is the most important cause of environmental change at global scale [5]. First, it is one of the main sources of greenhouse gases (GHGs) emissions worldwide, namely 9,800 to 16,900 megatons of carbon dioxide equivalent ($CO_{2\text{-eq}}$) per year, or 19%-29% of the total anthropogenic GHG emissions [7]. Second, food production is also responsible for 70% of freshwater use (for irrigation purposes) and 40% of land occupancy, thus being a major cause of biodiversity loss [8]. Moreover, agriculture is also the main driver of eutrophication especially for aquatic systems, through the massive use of nitrogen and phosphorus fertilizers [5]. Therefore, food production is a key factor regarding many planetary boundaries.

In particular, animal production is the most impactful agricultural sub-sector, generating throughout its supply chain 14.5% of anthropogenic GHG emissions (especially carbon dioxide ($CO_2$), methane ($CH_4$) and nitrous oxide ($NO_2$)), with cattle farming for beef and milk accounting for nearly two thirds of these [9]. Despite this negative impact, demand-side mitigating measures seem difficult, provided the role that meat consumption still has in the diet of the majority of people and the economic dimension of this sector of 1.4 T€ [10]. In fact, meat demand is estimated to increase at the rate of 1.3% per year until 2050 [11]. Due to these problems in acting on the demand-side in the short term, efforts should be also focused on supply-side management and optimization.

## 1.2 Land-use system in Portugal

Since having signed the Kyoto Protocol (KP), Portugal decided to put particular attention on emission from the agro-forestry sector. Portugal was one of the only two countries (together with Denmark) to choose "Grassland Management", "Cropland Management" and "Forest Management" in the framework

of the optional Agriculture, Forestry and Other Land Uses activities. The first of these, "Grassland Management", is strictly related with the topic of this thesis.

Land-use distribution in Portugal in 2012 can be seen in Figure 1: pastures account for 24% of Portuguese land, being the second most spread land use.

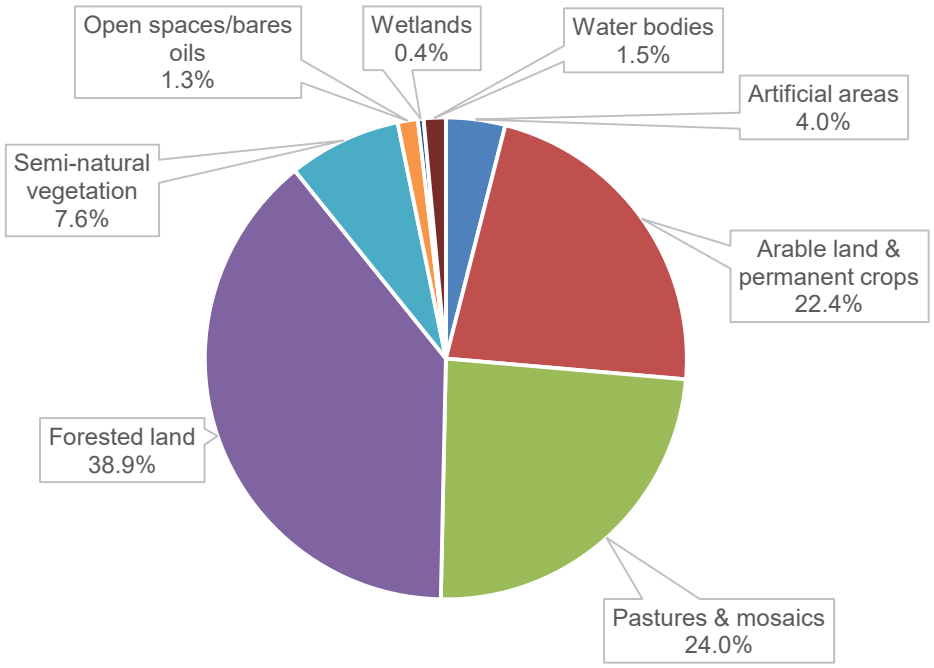



*Figure 1: land use distribution in Portugal in 2012 [12]*

The majority of these grasslands[1], especially if we consider permanent ones, are located in the Central-South of Portugal in the Alentejo region [13]. In Alentejo, pastures are typically found in an agro-silvo landscape known as *montado* in Portugal, a human-engineered savanna-like forest dominated by cork and/or holm oaks [14]. This ecosystem covers in fact 730,000 ha of Alentejo [15]. *Montado* is an important biodiversity hotspot which relies on human management for its existence [16], [17]. However, its progressive abandonment due to falling cork prices and changes in the Portuguese economy caused a strong degradation of its biodiversity and ecosystem services [17]. In particular, unmanaged *montado* is subject to shrub encroachment, which causes a shift of underground carbon (C) to aboveground ground plants and a reduction of trees growth, hampering the C storage potential of this ecosystem and inducing its desertification [14], [17].

Therefore, in Portugal, and in Alentejo in particular, grassland management is an extremely relevant issue in order to mitigate the environmental impact of the agro-forestry sector and in particular of livestock, whose grazing is the usual use of pastures. In fact, grasslands usually are, to a higher degree than croplands, net C sinks, storing more C than emitted also on a global scale [18].

[1] To simplify the language and avoid misunderstanding, the terms "pasture" and "grassland" are treated as synonyms throughout this thesis.

Three main types of permanent pastures are present in Portugal, natural, semi-natural and sown, corresponding to increasing degrees of intensification [19]. Natural pastures and semi-natural pastures (SNP) are both spontaneous grasslands in which native species grow, with the difference being the human intervention in the latter mainly for fertilization and shrub control purposes. Sown pastures instead differentiate from both due to the artificial introduction of non-spontaneous species. One important system in the country are sown biodiverse permanent pastures rich in legumes (SBP), whose adoption is the focus of this thesis.

## 1.3 Sown biodiverse pastures and their role

SBP refer to a rich seed mix of species originated from the Mediterranean but normally existing in little proportion in natural grasslands (as legumes) that can be adapted to the specific location, exhibiting, compared to SNP, increased productivity and other advantages that will be further discussed [14]. Actually, the implementation of SBP in Portugal was the main reason for the choice of the “Grassland Management” activity of the KP [19].

The first SBP areas were implemented by David Crespo in his family’s property “Herdade dos Esquerdos” in the second half of the 1960s. According to Fertiprado (http://www.fertiprado.pt), the main firm selling pasture seed mixes in Portugal, between 1990 and 2008 94,260 hectares (ha) of SBP were installed in the country and especially in Alentejo [19]. After having noticed a reduction in adoption from 2005, hypothesised to be mainly due to the saturation of the farmers share for which SBP adoption was economically viable [19], between 2009 and 2014 the Instituto Superior Técnico (IST) spin-off company Terraprima (http://terraprima.pt/) led a project of payments for C sequestration through SBP adoption with the support of the Portuguese Carbon Fund (PCF), a financial instrument created by the Portuguese government to comply with the KP objectives. Dry matter productivity is 50-100% higher in SBP than in SNP, allowing for higher stocking rates [19], but their installation can cause economic losses in the first year that could be considered unbearable or too risky by many farmers, due to a high initial investment and the limitation posed to grazing during this timeframe [19]. During the PCF project, SBP were adopted in an additional area of 48,491 ha among 1095 farmers still located for the large majority in Alentejo, reaching 4% of the national agricultural land [14]. Terraprima estimates that the additional area of SBP installed under PCF payment for the C sequestration service is sequestering 1.54 million additional tons of $CO_2$ [14]. Figure 2 reports the area of SBP and consequent carbon sequestration observed until 2008 and forecasted by Teixeira [19] for the successive years.

The link between the PCF and SBP adoption was possible thanks to the performance of this system in terms of reduction of livestock GHGs emissions. Increasing soil organic matter (SOM), during the first 10 years after installation SBP are a C sink of on average 6.5 t of $CO_2$ per hectare per year. This resulted in an estimated $CO_2$ sequestration of 3.5 million tons between 1996 and 2008 [19] and 1.54 additional million tons during the PCF project (according to Terraprima). If we consider the increase in GHGs emissions due to required limestone application, nitrification/denitrification cycle of legumes (emitting $N_2O$) and increased stocking rate (raising the $CH_4$ emissions), in the first 10 years SBP are still a net sink of 1.55–2.13 t $CO_{2\text{-eq}}$ per hectare per year with adequate technical management [19]. These data are particularly interesting considering that, to meet the Paris Agreement target of not increasing global

temperature of more than 2°C by 2100 (relative to 1861-80 temperatures), Willett et al. [5] estimate that land use, land use change, and forestry will have to pass from emitting 5 Gt of $CO_{2\text{-eq}}$ per year to storing 10 Gt of $CO_{2\text{-eq}}$ net per year. Carbon sequestration in soil is also reported by the Food and Agriculture Organization of the United Nations (FAO) among the measures to mitigate livestock impact on climate change [9]. Taking into consideration the whole life cycle, SBP can avoid 25% of the emissions from beef production, thanks mainly to the replacement of concentrated feed [20]. Concentrate has in fact a relevant impact, accounting for 41% of global livestock supply chain emissions [9].

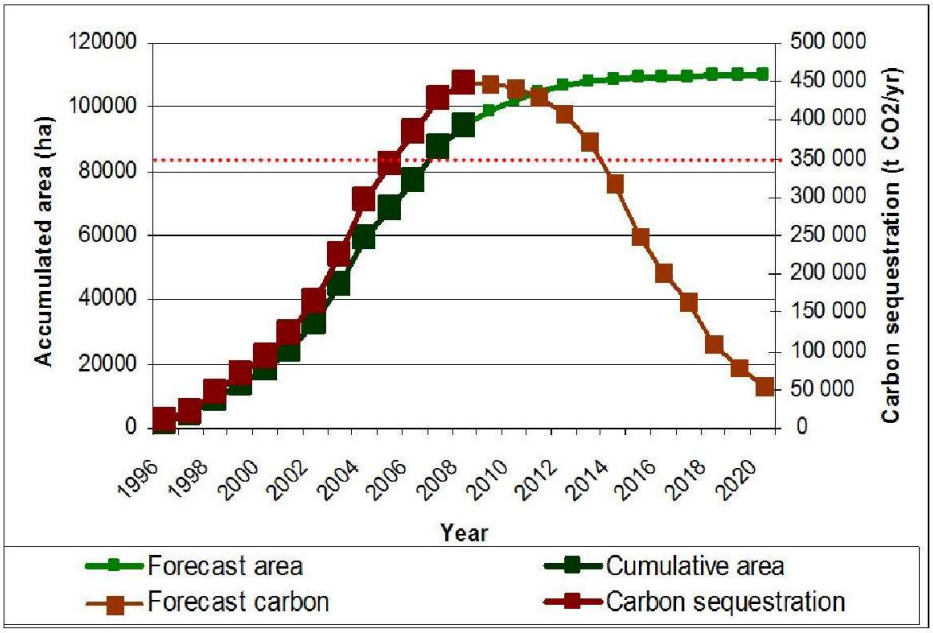



*Figure 2: cumulative area of sown biodiverse pasture installed and consequent yearly carbon sequestration, observed and forecasted [19].*

After around 10 years, SOM pools in SBP reach a steady level and the C sequestration service stops [19]. However, besides the saving in emission for feed substitution that still apply, SBP have many other environmental benefits compared to SNP that remain valuable. Increased SOM improves water retention and soil fertility [14]. Thanks to nitrogen (N) fixation by legumes, SBP do not require the use of synthetic N fertilizers [14], whose environmental impact is of ever increasing concern. The Haber-Bosch process produces ammonia through the transformation of unreactive atmospheric dinitrogen into a reactive form [21]–[23]. This industrial process accounts for 1% of global energy consumption and between 75 and 90% of this ammonia is used for N fertilizers [24]. Moreover, the extensive use of N fertilizers is associated with increased GHG emissions and contamination and eutrophication of water bodies [25]–[27]. In particular, anthropogenic emissions of nitrous oxide ($N_2O$), which are mostly related to fertilizers use in croplands, increased by 30% over the past forty years [28].

The main drawback of the system is instead the requirement of phosphate fertilizers, that also causes a risk for farmers regarding its price and availability [14]. Moreover, when installed in acidic soils SBP often require pH correction through liming [14]. Effects on biodiversity loss and land use still require further considerations. Despite the fact that for each hectare of SBP installed, 0.5 hectares of land that would have been used for feed production on average are saved [29], the impact on these two categories

at a macro level would strongly depend on the enhanced possible stocking rate's feedbacks, in particular if the consequence is an increase in livestock or a reduction of land dedicated to pastures.

Eventually, SBP can provide many environmental services while not causing significantly more damages than SNP for most environmental impact categories [19]. Therefore, SBP have the potential to positively impact many planetary boundaries and their extension could represent a step for Portugal towards the achievement of its Paris Agreement commitments.

To further spread SBP over the 4% of agricultural land achieved after the PCF project, specific new policies need to be designed. These policies will need to be new since the farmers that will evaluate to adopt SBP in the future may not be the ones that adopted in the past and may respond to different incentives. However, a clear understanding of the PCF project outcome and more in general of SBP adoption dynamics are still lacking and are a fundamental prerequisite to design effective policies in the future. This analysis can provide a proper understanding of which have been the factors determining SBP adoption in the past and provide a modelling framework that can be applied in the future to build on this improved understanding to evaluate the influence of policies and their environmental outcomes.

## 1.4 Objectives

The main goals of this thesis are:

- To develop and test multiple modelling frameworks that can be used to estimate adoption of SBP.
- To understand which are the main drivers that influenced the farmers decision-making process regarding the adoption of SBP and the consequent observed patterns of area sown.
- To assess retroactively the outcome of the PCF project, in terms of additional SBP area that was installed thanks to it and consequent C sequestered, during its duration and until today, compared to an estimation of a counterfactual scenario without the PCF project.

The main tool used to reach these objectives was agent-based modelling for its ability to build simulations focused on decision-making of individual or aggregated agents. The achievement of these goals will enhance following analysis aimed at the design of effective new policies to further expand the adoption of SBP in Portugal.

The focus on SBP, and consequently on Portugal where most of these pastures are located, is justified by their potential as a C sink. SBP can help in reducing GHG emissions and promote additional ecosystem services that can be of great benefit to landscapes facing degradation such as the *montado*.

This thesis is developed in the framework of the LEAnMeat project ("Lifecycle-based Environmental Assessment and impact reduction of Meat production with a novel multi-level tool"), led by IST of the University of Lisbon with the aim to contribute to the environmental assessment of livestock production systems and to identify and test the performance of improved farm management systems and technologies in the period 2020-2030. LEAnMeat will develop an innovative multi-level model where background and foreground processes are modelled for the Alentejo region, testing policy instruments for improvement in 2020-2030 using land use scenarios. An ABM is included in the planning of the

project to evaluate agent behaviour under global change using scenarios for global food demand, resource scarcity and climate change. The ABM developed in this thesis is conceived as a first step in the direction of this broader-scope ABM, focusing on SBP adoption due to their central relevance.

## 1.5 Structure of the thesis

The thesis starts by presenting the current state of the art of complexity theory applied to socio-ecological systems, in section 2. ABMs became a widely used tool for the study of complex systems. Their application specifically to land-use/cover change issues is presented next, with a specific focus on the integration of empirical data. Section 3 gets deeper into the specific case study of the thesis and, after the presentation of the available data in section 3.1, describes the method of the two approaches or modelling frameworks implemented to reach the defined objectives. There are two main modelling approaches:

- A *farmer-based approach* (section 3.2), with the individual farmers as main subjects. The farmer-based approach involved the development of both *theory-driven ABMs*, based on theoretical rules, and *data-driven ABMs*, completely based on the available data and avoiding any theoretical basis for agents' behaviours and interactions. All these ABMs are however empirically grounded, for they exploit empirical data. They are used for beginning to understand the key reasons for adoption of the pasture system at the individual farmer level.
- A *municipality-based approach* (section 3.3), whose main agents are the Portuguese municipalities aggregating all the farmers within them. The municipality-based approach consisted of a single data-driven ABM, employing machine learning (ML) algorithms to develop the agents' internal model whose selection is described in section 3.3.2, separately from the following presentation of the ABM architecture. The municipality-based approach enabled to study the reasons for adoption at an aggregated level and to evaluate the PCF project, comparing its outcome in terms of SBP installations with a counterfactual simulation reproducing the scenario in which payments were not provided, as described in section 3.3.4.

Section 4 presents the results obtained using both approaches, whose explanatory power is then discussed in section 5. This section also explores the limitations of the modelling approaches and how these could be addressed in future work, focusing particularly on the future combination of some characteristics of both modelling approaches into an integrated framework. Eventually, section 6 presents the conclusive remarks and how the objectives of the thesis were reached.

# 2 State of the art

## 2.1 Social-ecological systems as complex adaptive systems

Modelling the adoption of SBP means to consider the integration of ecosystems and human societies and thus to get into the field of social-ecological systems (SES). Due to the many different entities, layers and feedbacks that characterize SES, the importance of treating them as complex adaptive systems (CAS) has been largely reported in literature [30]–[33]. To understand which are the implication of this approach, a brief introduction to CAS and the field that study them, complexity science, will be useful.

Complexity science arose from the need to capture reality in a more accurate and comprehensive way than possible with Newtonian science [34]. Despite all the progresses and results that it allowed for, by modelling the world solely through a reductionist approach – i.e. explaining it only through the motion and interaction of its fundamental parts with linear equations – Newtonian science struggles to account for the properties and behaviours that emerge from these interactions, increasingly as the complexity of the systems under study and its elements grows[2]. Nowadays, the limits of a reductionist approach are evident in sciences accounting for higher complexity and decreasing generality, as psychology and sociology where human behaviour comes into play. Especially in these cases, a mechanistic representation of the world is unlikely to be a successful formalism.

Therefore, complexity science has at its core the realization of this impossibility of capturing all the properties of a system through one single formalism: complex systems cannot be treated in a unique way but need to be addressed through different approaches, all necessary since their conclusions cannot be derived from each other [34]. Moreover, in complex systems the interaction of their elements cause the emergence of macro-properties that could not be observed and understood studying these elements independently and that are key to capture fundamental aspects [34]. Thus, to the Newtonian linear reductionism complexity science counterpose non-linear holism, where “the whole is more than the sum of its parts”. Classic examples of systems showing emergent properties are languages, with words’ meaning emerging from the connection of letters, sentence meaning emerging from the connection of words and so on.

CAS are one of the objects of study of complexity science, being systems composed of individual agents (entities able of individual behaviours, often heterogeneous and that can span different scales from cells to humans to countries), which act taking into consideration other agents’ actions and their environment ([35] as cited in [36]). The multitude of individual behaviours, that adapt over time to changes in their context, is the cause of the overall behaviour of CAS that thus experience a decentralized and dispersed control. This adaptability stemming from the interaction and influence among agents and between the agents and the environment creates feedback loops that cause CAS to have highly non-linear responses

[2] If we assume a non-deterministic universe, the limits of Newtonian science are at a metaphysical level and thus complexity science is necessary. However, even though we accept that “God does not play dice” and assume a completely deterministic universe in theory explainable through a reductionist approach, in practice systems of growing complexity are, and will still be for a long time, mathematically intractable through reductionism.

to perturbations and to exhibit path-dependency, i.e. to have high sensibility to previous conditions, which state can cause the system to evolve in completely diverging ways [36]. Therefore, CAS future states result unpredictable, not because of random non-deterministic events, but due to the impossibility to characterize the system with a sufficient degree of precision and thus follow its non-linear evolution. Systems with this behaviour are subject to a property named chaos (a well-known example of it is the Lorenz's Butterfly Effect [37]). Chaotic systems allow for the prediction of attractors, final states to which the system tends to converge, but not for the forecast of which attractor will represent the final state [36].

Many of these characteristics identifying CAS are easily recognizable in SES. Feedbacks and complex interactions between human agents and the environment, especially the biophysical one, are at the very core of SES and are often explicitly addressed [32]. Heterogeneous actors compete for the limited resources provided by ecosystems, adopting complex behaviours and bringing to the emergence of macroscopic properties that in turn influence the single agents and ecosystems [31], causing non-linear responses that can result in surprising and even unintended outcomes [31], [32]. To avoid such consequences, when designing policies for SES it is of fundamental importance to consider their complexity and adaptability [31]. Past conditions need to be taken into account and errors can be difficult to fix due to path-dependency [31]. Directing SES towards the desired state is even more challenging since attractors may not be clearly defined and evolve over time, with sudden thresholds among them [38]. Not only different time scales but also spatial ones need to be addressed, to avoid unwanted feedbacks involving other part of the system [31], [39]. These different space and time scales creates a panarchy, i.e. a hierarchical structure of nested adaptive cycles of stabilization and innovation [30]. This structure allows the system to be resilient, adapting within the current basin of attraction or transforming to a new state as a response to internal and/or external perturbations while maintaining the same identity and function [38]. When designing policies addressing SES for instance, resilience is an important property to consider, that can be enhanced to reinforce beneficial behaviours and conditions or tried to be overcome to change harmful ones [31].

Due to their property as CAS, proper understanding and explanation of SES cannot be achieved focusing exclusively on individual agents' behaviour. Individuals still need to be studied, but even if we would manage to describe them perfectly, the emergent properties – that often are the ones we are interested in – could not be observed, and thus not be explained, if the system is not considered in its wholeness. What is required additionally is an environment in which the agents can interact, growing the emergent properties under our eyes. The key question that should be answered is "how could the decentralized local interactions of heterogeneous autonomous agents generate the given regularity?" [40]. This is at the base of what Epstein calls "Generativist Science", born to address this research question through generativist experiments: "situate an initial population of autonomous heterogeneous agents in a relevant spatial environment; allow them to interact according to simple local rules, and thereby generate – or "grow" – the macroscopic regularity from the bottom up" [40]. The main tool to conduct generativist experiments and study emergent properties, thanks to its intrinsic characteristics, is agent-based modelling [40].

## 2.2 Agent-based modelling

ABMs are generically identified as simulations composed of individual agents and characterized by the importance given to their behaviours and heterogeneity [41]. A unique, agreed-upon definition is however still lacking, despite the fact that the history of ABMs now spans over more than 45 years [41].

AMBs' origins are not clearly defined. Schelling's segregation model is often referred to as the first proper ABM [42]. Conway's Game of Life is also considered by some as a first simple example of ABM [43]: many early ABMs were in fact based on cellular automata, being composed of 2D grids of cells switching among a finite set of states [41]. Another source of inspiration has been object-oriented programming, which constitutes the natural computational framework for ABMs where agents are usually instantiated from classes [40]. Also, the field of CAS contributed to ABMs' development, especially for the concept of properties emerging from the bottom-up [41]. Since 1996, year that signed the beginning of a wide diffusion of ABMs thanks also to the publication of the book Growing Artificial Societies [44], the publications of articles referring to ABMs increased steadily. Nowadays, ABM is used in various fields, spanning from pure social sciences to energy, from ecology to economics, and many of the most important simulations journals regularly publish papers with ABMs [41]. It is also possible to find books introducing newcomers to the field [36], [45], many dedicated software available to facilitate ABMs construction [46] and communities built around them, as the Network for Computational Modelling in Social and Ecological Sciences (CoMSES Net; https://www.comses.net/).

With this growth in interest, over time other features have been commonly associated with ABMs. An important one is a bottom-up approach, where either the single agents' behaviours are know and the exploration of emerging properties is the goal, or these macro-scale patterns are observed and the aim is to understand which micro-specification (the initial specification of the agents and their environment) generates them[3]. Therefore, through ABMs we can understand the micro-to-macro mapping of systems that do not have a centralized control: each agent is autonomous, in the sense that react to internal and external stimuli according to its own rules [40], and can also be adaptive, learning and changing its rules during the simulation [41]. Moreover, since there is no need to necessarily model populations as homogeneous through representative agents and to resort to using variables and equations at an aggregate level, ABMs allow to have a one-to-one model of the reality [40], [41], whose complexity limits are due only to computational tractability and model usefulness.

There are two characteristics that got strongly linked to ABMs while contributing to their raise in popularity. The first is the possibility to explicitly address space, that can vary from real maps – through the use of Geographical Information Systems (GIS) – to idealized grids and even abstract spaces such as social networks [40]. The second is the possibility to explicitly model local interactions among the

[3] Epstein [40] discouraged the use of the term "emergence" due to its classic meaning linked to absolute inexplicability, which he criticized noting that, as science progresses, emergent phenomena cease to be seen as such and thus their emergence is just a temporary state relative to a certain theory. Furthermore, he advocates that this classical definition is not compatible at all with the role of agent-based models and the generativist scientist role, as the explanation of these emergent phenomena is their very purpose. Actually for Epstein agent-based modelling is in this sense even reductionist, since the micro specification of the model is enough to explain the systems' dynamics. Despite these interesting points, the term emergence is still widely used in the ABM community, with no relation to inexplicability but just to refer to properties "arising from the local interaction of agents" [44].

agents through their environment and thus address social networks and how information spreads across them [40], of fundamental importance for applications such as innovation diffusion. These characteristics will be reiterated, being important factors in the application of ABMs in land use/cover change (LUCC) systems, the one treated in this thesis.

The agreement on linking these features to ABMs and the growth in interest however do not necessarily suggest the existence of consolidated best practices for their implementation. The resources cited above present general guidelines, such as the order of the steps to be taken from the formulation of the problem to the model validation [36]. But these guidelines remain generic and many important decisions are left to the modeller. A reason for this can be adducted to the diverse fields in which ABMs has been applied, which caused the development of many different approaches in parallel often suited for the specific information available. Also due to the complexity they enable, the possibilities to specify some components of ABMs, such as agents' behaviour and reciprocal interaction, are potentially infinite. This results in challenges for acceptance and comparability of ABMs. Despite the progresses made, they remain incompletely solved and therefore specific case studies still require tailored solutions [41]. The validation of ABMs is a perfect example of a critical process often criticized for its lack of consensual methodology. Due to their micro-level specification, ABMs are unique in their need for a micro-level validation at the agents level, in addition to a more usual one on aggregated output [41]. However, this is often not conducted properly or even lacking completely [47].

This point will be increasingly clarified while treating LUCC ABMs in the following sections, where the examples presented will show a series of approaches tailored to the specific contexts rather than with a unique and common structure.

## 2.3 Land-use/cover change ABMs

LUCC systems are inherently SES, considering the influence that both natural and socio-economic factors have on them [39]. LUCC systems also behave as CAS, due to the feedbacks they present, within and between the environmental and the societal spheres, and the heterogeneity they involve, both in terms of agents (human and biophysical) and of spatial and temporal dimensions [48]. Gaube & Haberl [39] note how the awareness of the importance of using a multidisciplinary and integrated approach in addressing LUCC even brought to the new notion of "integrated land-system science".

ABMs can overcome some important limitations of other more consolidated approaches to model LUCC, such as the aggregated level of analysis required by equation-based and system models or the simplification of decision-making and reciprocal influence in cellular models [49], [50]. The literature stresses in particular the following advantages of applying ABMs to LUCC [39], [49], [51]–[53]:

- ABMs are able to represent decentralized and heterogeneous decision-making to a degree of accuracy not possible in other modelling approaches.
- ABMs can explicitly capture social interaction and its influence on decision-making.
- ABMs can report spatial diversity and its link to agents' decision-making, even through spatially-explicit biophysical models, thus being able to understand when feedbacks create non-linear responses.

These characteristics, together with the holistic possibility to consider in more detail the complexity of LUCC, brought to a steady increase the number of studies involving ABMs [48], [54], [55]. These LUCC ABMs can be further divided into more specific applications: the next section describes the most relevant ones for this thesis.

## 2.4 Applications of LUCC ABMs

### 2.4.1 Explaining land use patterns

Probably the most trivial (and most generic) application of ABMs to LUCC is the explanation of observed land use patterns. This is in fact a common application found in literature for agriculture but not only, as it has been applied also for urban and ancient societies settlement studies [49], [52].

Given the benefits of ABMs explained in the previous section, this should not come as a surprise: in land use analyses, the spatial component is obviously fundamental and the interaction between the environment and the human agents' is as well. Moreover, the explicit involvement of human agents with particular motivations besides economic reasons can provide more insights into land use change and help to get better predictions than classic economic theory [56].

Murray-Rust et al. [57] developed Aporia, an open-source framework with the goal of simplifying the construction of high-fidelity land use ABMs. The authors applied Aporia to two case-studies, to study economic and agricultural output and provision of ecosystem services depending on different land management decisions. Acosta et al. [16] developed an ABM to assess the influence of the evolution of socio-economic characteristics and biophysical constraints on land use patterns, considering climate change effects. The model was applied to an area of 44 $km^2$ in the Portuguese region of Alentejo, running scenarios for 2050 to understand the possible impacts on the *montado* ecosystem. The results highlighted significant challenges in the conservation of this traditional landscape, which however remains dominant in the region. Dullinger et al. [51] used an ABM to simulate changes in land use due to socio-economic and climate conditions as well, with the aim to study their relative impact on plant biodiversity in the Austrian Alps. To do this, the ABM was coupled with a species distribution model.

### 2.4.2 Innovation diffusion

Innovation diffusion refers to the process through which members of a social network adopt an idea, practice or object that they perceive as new [47]. Innovation diffusion is another perfect example of a field in which the advantages of ABMs come at hand, since simple economic considerations are often not sufficient to elicit an explanation [58]. Diffusion dynamics are mostly related to social networks characteristics and how information spreads in them. Contrary to models focusing on aggregate trends, ABMs give the possibility to explicitly model the interconnection between agents and their local interactions, while testing different networks [47]. Another fundamental component of innovation diffusion is its spatial dimension, which ABMs can easily integrate: spatial clustering of adopters can emerge and different nested spatial scales and their interdependencies often need to be considered [49]. Agents' heterogeneity can also be a critical driver of different diffusion patterns [47], [49]. ABMs can help, for instance, identify which characteristics differentiate the early adopters from the majority and

from the laggards. All these elements can eventually be integrated in the decision-making of the agents', for example considering how information diffusion usually decreases the uncertainty about the model's outcomes.

In a highly cited paper, Berger [59] built an ABM for farming innovations in Chile, including agents learning from past experience. The study modelled farmers' reciprocal influence through both a threshold of peer adoption required to consider the innovations and feedbacks on the hydrologic cycle. It concluded with a praise of ABM with spatial components as a useful tool to study innovation and resource use. Other studies instead stressed the importance of considering the temporal dimension, as Alexander et al. [60] with time lags arising from spatial diffusion to avoid overestimating adoption. This study also resorted to an ABM with a threshold on neighbours' positive experience to model energy crop rotation by farmers in the UK, considering also the interaction between biomass power plant investors (demand) and farmers (supply).

The topic of this thesis is intrinsically one of innovation diffusion applied to the agricultural field, since it aims at understanding the adoption of SBP by farmers in Portugal. Agricultural sociology is the field in which traditional diffusion research started [61]. Agricultural innovation and farming is also one of the most common application of ABMs for innovation diffusion based on empirical data, second only to sustainable energy and conservation technologies according to Zhang & Vorobeychik [47]. However, as of 2019 only 10 published papers belong to this category [47] while the majority of studies dealing with agricultural diffusion rely only on simple economic decision rules [58].

The study of innovation diffusion is usually aimed, as in the case of this thesis, at accelerating the spread of the innovation of interest. This is often strictly interlinked with policies design, which is the application of LUCC ABMs treated next.

### 2.4.3 Policies analysis and planning

Policy design is a task that requires careful considerations, for the very practical and complex implications it has. Increasingly, these considerations take the form of an impact assessment (IA) process, which enables a better understanding of the possible economic, social and environmental policies' outcomes and therefore to support sustainable development [62]. IA requires well-chosen tools and when dealing with LUCC agricultural system models showed to be a relevant option, which attracted large scientific interest especially from 2008 onwards [62].

To provide useful insights in this policy context, these models should include implicitly or explicitly a spatial dimension, account for the heterogeneity of farmers with sufficient detail both in terms of their characterization and behaviour, include farms interactions both among them and with the environment and comprehensively use sensitivity and uncertainty analysis [63]. The literature particularly stresses the importance of considering the influence of agricultural policies on individual's decision-making, especially considering the heterogeneous nature of farms and therefore the variety of possible outcomes [64], [65]. However, the main efforts have been so far focused on regional or crop level, not considering adequately the importance of the key decision-makers, the farmers, which should instead be a priority for a research agenda [62].

ABMs can comply with all the requirements presented [52], [63] and their bottom-up approach can help to study individuals' reactions [64]. However, until 2007 ABMs encountered difficulties in being recognized as a useful tool [52]. Lempert [66] attributes this difficulty to the fact that early ABMs were used to give point predictions, when instead they should showcase their power to deal with deep uncertainty, which they can account for through scenario assessment. Uncertainty quantification in ABMs has increasingly been acknowledged by researchers and the policy-making community. Together with the inclusion of empirical data (which will be treated in section 2.5), this sparked the use of ABMs especially from 2008 on [65]. Reidsma et al. [62] analysed the papers that used models for policy IA published between 2007 and 2015 and reported that 25 of the 31 papers (about 81% percent) modelling farmers' interactions among them and with other actors were ABM. Additionally, 22 out of the 30 (about 73%) considered structural changes of the farms due to policies. In general, about 15% of the 184 articles they analysed used ABM as the main tool.

Another relevant benefit of ABMs is that the individual representation they allow for is actually the most familiar to us: we can all be considered agents, with attributes representing our state and actions characterizing our behaviour. However, the exact responses to stimuli is often difficult to conceptualize in specific mechanistic equations. This generated increasing interests in the particular field of participatory ABMs [52]. Thanks to the direct involvement of stakeholders and decision-makers from the very construction of the model, this approach allows to directly take into consideration all the different points of view and facilitate discussion, having an important role in policy design [49], [67].

The specific case of LUCC policy relevant for this thesis is the design of payments for ecosystem services (PES). In defining ecosystem services for decision-making purposes, Fisher et al. [68] described them as "the aspects of ecosystems utilized (actively or passively) to produce human well-being", stressing the necessity of human beneficiaries. The ecosystem service treated in this thesis is the C sequestration consequent to SBP adoption, paid for during the PCF project. Money incentives to support farmers in providing ecosystem services are a market-oriented tool that gained relevance to orient LUCC towards sustainable pathways [69]. However, farmers' choices cannot be restricted to profit maximization and risk minimization, and ABMs allow to consider other decision-making factors [69]. ABMs can also consider the complexity of the system, which has often been neglected in designing PES schemes [70].

Despite these benefits of ABMs, few studies address PES design through ABM. All examples found were are applied to case-studies in China. Chen et al. [71] developed an ABM for reenrolment in a payments scheme to prevent soil erosion converting sloping croplands to forest or grasslands. The study focused on evaluating the influence of social norm, demonstrating the benefits of leveraging it. Chen et al. [72] highlighted the fact that ABMs enable to address the uncertainties in agents response to policies and the dynamic human-nature interactions, being able to assess different types of payments to reduce logging and increase afforestation. An et al. [70] went further in stressing ABMs advantages, remarking in particular the possibility to get in detail into these complex systems and to get realistic, spatially and temporally explicit insights. The ABM was applied to a case of out-migration from a natural reserve and its relations with PES.

### 2.4.4 Environmental assessment

Integration of environmental components into ABMs usually aims to quantify the impact of human activities on biophysical cycles and vice versa. The assessment of environmental consequences is often an important objective in LUCC [52]. This coupling needs to consider a multitude of complex features in order to properly inform decision-makers and enable to effectively manage these impacts, such as micro- and macro-level agent's behaviour, adapting over time to changing conditions and subject to local interactions [73].

Environmental assessment and management have been integrated with model typologies such as system dynamics and Bayesian networks. Kelly (Letcher) et al. [74] identified ABMs as particularly suited for system understanding and social learning purposes, due to their unique ability in considering interactions between individuals. Lobanova et al. [75] acknowledged the necessity of including human influence in analysing natural systems. Therefore, they coupled an ABM with a hydrological process-based model to assess the hydrological evolution of the Tagus river under climate change, with the aim of using it to facilitate policy making. Gaube et al. [67] combined a ABM and a stock-flow model of C and N cycles, interfaced thorough a spatially explicit land use model based on GIS data. The model was developed through a 2-years long participative approach in an Austrian municipality and it found evidence of strong relations between socio-economic and ecological components. The study stressed the importance of balancing the level of complexity of the socio-economic (the ABM) and the biophysical (the stock-flow model) parts, in order to focus on the interface between the two.

A modest number of studies includes additional analysis of environmental effects, through the integration of ABMs and life cycle assessment (LCA), a widely acknowledge method to quantify the environmental impact of products and processes throughout their life cycle [76]. Micolier et al. [77] found only 31 articles combining ABM and LCA addressing in total 13 case-studies. Marvuglia et al. [78] suggested that the main reason for this would be the complicated implementation that may discourage non-specialists in both fields, which are still in scarce number.

Still, ABMs are facing an increasing acceptance in the LCA community: the main reasons for this is the possibility to include dynamic components and behavioural heterogeneity in a typically static and homogeneous assessment such as LCA [77], [78]. This interest was concretized in two complementary and non-mutually exclusive integrations. The first consists in the use of ABMs' output to enhance LCA results, in order to assess the influence of individual behaviours on environmental impact. This is the approach applied in Bichraoui-Draper et al. [58], where the ABM enabled the study of how socio-economic factors and the conditions under which the adoption of switchgrass for ethanol production occurs influence its environmental performance. The second coupling is the use of LCA outputs to enhance ABM results, to account for how environmental performance affect ABM dynamics. For instance, Navarrete Gutiérrez et al. [79] used an LCA-enhanced ABM to explore how farmers' environmental awareness influences their activity and therefore their impact, with a case-study in Luxembourg.

## 2.5 Empirically grounded LUCC ABMs

### 2.5.1 Empirical data use in ABMs

ABMs have been used to both obtain general knowledge and study specific cases. The initial studies adopting ABMs for LUCC were mainly theoretical and abstract, as they aimed at demonstrating the applicability and suitability of this new modelling approach [55]. After successful proof of concept of their utility and facing increased acceptance, in the last decade and a half ABMs could be exploited for different applications, causing the main focus to shift to the study of specific contexts and the modelling of particular systems [55], [80]. This thesis, aimed at studying a specific context involving a large size of decision-makers, fully belongs, to the case-study category of ABM as defined by Janssen & Ostrom [80], similarly to many other LUCC ABMs.

The surge of specific context applications was one of the main reasons fostering the increased use of empirical data in ABMs in the last years [81], which eventually caused a shift from *theory-driven ABMs*, which exploit theoretical rules, to *data-driven ABMs*, which avoid any reference to theoretical frameworks and are completely based on rules extracted from the available data[4]. An increased used of data in LUCC ABMs should not surprise when the applications reported above are considered: the explanation of land use patterns requires an observed pattern to explain; innovation diffusion already has an extensive theoretical background and the main use of ABM is to study or support adoption in specific context; environmental assessments need defined conditions to be conducted. However, the application that most contributed to the inclusion of data in LUCC ABMs has been policy design [47]. In order to obtain stakeholders confidence, proper calibration and validation of the model using empirical data is of fundamental importance [81], as well as a representation of micro-processes reflected in real world observations [50], [82].

The last main driver of the increase in use of empirical data in ABMs was the exponential increase of available data and computational power, with the consequent explosion of fields such as data analytics and ML [47], [82], that will be further treated in section 2.5.3. Some of the commonly used data sources in LUCC ABMs include sample surveys and interviews regarding socio-economic and environmental variables [16], [69], [72], participant observations insights [50], [69], census data [78], experts opinion [69] and remote sensing and GIS maps [50], [78].

### 2.5.2 Modelling agents' decision-making

While the use of macro-level data for validation is common to many different modelling approaches, empirically grounded ABMs are particularly interesting for the possibility to exploit data at the micro-level, to represent micro-processes and in particular individual agents' behaviour [50].

The possibilities for individual decision-making modelling in ABMs are various and the choice of which one to use is key in order to obtain useful outcomes [49]. Modelling agents' behaviour is a particularly

[4] In this thesis, "empirically grounded ABM" refers to any ABM which makes use of empirical data. "Theory-driven ABM" can be empirically grounded whenever they use data. "Data-driven ABM" are always empirically grounded by the definition given.

delicate issue in LUCC. In fact, farmers' decision-making seldomly follows simple economic optimization and is instead influenced by culture, traditions and peer influence and subject to limited knowledge of innovations and market [69], [78]. Therefore, the inclusion of empirical and as much as possible unbiased data got the interest of the LUCC ABMs community [56]. However, the collection of such data has to acknowledge that these factors may also influence farmers' direct answers to in-person interviews and surveys. Farmers can be reticent to take part in these studies [56] and the ones accepting are often the most innovative ones, due to their will or the more interest in them by researchers.

However, theoretical behavioural models still have an important role in ABMs, since they allow for clear interpretation and therefore communication while enhancing generalization and reusability [54]. Most importantly, the use of theories and of data are not mutually exclusive and they often coexist: the required data are not always available, making some theoretical assumptions often necessary, and, even when they are, backing empirical results with consolidated theories can help to inform policy makers [54].

The most common theory-driven ABMs are economic models [54]. This category encompasses various approaches, from simple evaluations to proper mathematical optimization decision rules, all however based on the idea that agents are seeking high economical return [47], [83]. A well-known implementation of this approach is the Expected Utility Theory, in which agents choose the option that maximises their utility under risk [54]. In neoclassical economy this translates into the *Homo economicus* agent, one which has perfect and complete knowledge, can perform perfect calculations whenever needed and acts purely in its own self-interest [54]. This paradigm has however been widely challenged, stressing the need to consider agents with rationality bounded by lack of information and knowledge and by their biases [40], [49]. For instance, Dullinger et al. [51] used a stochastic modification of the Expected Utility Theory to model agents oriented towards a balance between workload and income. This work is also an example of theory-driven but nevertheless empirically-grounded ABMs, since the authors used census data to initialize the agents and link them to the particular land they managed. Going one step further, through the specification of the individual behaviours ABMs can relax classical microeconomic assumptions, assessing the importance of each [41]. Another prominent theory in this regard is Satisficing, when agents review the options and stop their research as soon as they find one that matches their expectation [54], [83]. Bounded rationality has been the most adopted solution in LUCC ABMs, even though still lots of models use complete rationality and the gap is strangely reducing over time, despite the recommendations [54].

The other main theoretical framework comes from psychological and cognitive models, where cognitive maps and abilities, social norm and biases are the main decision-making drivers [83]. A prominent example is the Theory of Planned Behaviour, where perceived social pressure ("subjective norms") and internal and external barriers ("perceived behavioural control") play a fundamental part in agents' behaviour [54]. An example of the application of this theory applied to the field of innovation diffusion is [84], which concluded that to explain the adoption of organic farming in Latvia and Estonia both economic and social factors need to be taken into consideration, despite the larger influence of the firsts. This work used empirical data as well, in this case to inform some parameters as the threshold for

adoption. 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 [47]. The application of cognitive models is largely lacking behind economic ones in LUCC ABMs and especially the role of emotions, knowingly important for environmental management, is often overlooked [54].

Despite the described benefits of referring to specific theories and due to the difficulties described above in eliciting the complex rules underlying farmers behaviour, in LUCC ABMs the majority of studies from 2000 on adopted ad hoc implementations without theory-based justifications [54]. These approaches go under different names in literature, as stochastic or heuristic models[5] [47], [83]. The specific subcategory of these models relevant for this thesis, that will be referred to as *data-driven ABM*, are the ones which rely on agents' behavioural rules and interactions completely based on empirical data. These rules are obtained through the use of computation and statistical analysis that can assess the relative importance of various features and parameters not defined a priori [47]. This approach marks an important shift: while in theory-driven ABMs the approach has historically been to design relatively simple agents' and let complexity arise mainly from the interaction among themselves and with the environment, data-driven ABMs enable to depart from this paradigm and build complex agents, which can already include most spatial and social influence features within their behavioural model [85].

A particularly interesting sub-category of data-driven ABMs are the ones using ML algorithms to handle the available data. The combination of ML and ABMs got a lot of interest in the last years, thanks to the ability of ML models to extract patterns in large amount of data, which in certain situations can be of great help to infer agents' behaviour [47], [86]. This is why the next sections are dedicated to a more in-depth analysis of this cutting-edge hybrid modelling approach.

### 2.5.3 Machine learning and ABMs

ML is a term that generally refers to algorithms able to automatically learn from data, being able to discover more or less complicated patterns in the datasets they are trained on and generalize the acquired knowledge [86], [87]. Over the last two decades, ML experienced a dramatic development and is nowadays one of the most rapidly growing technical fields [88].

ML algorithms have been applied in many different contexts and are often merged with other computational modelling approaches. Also ABMs have taken part in this trend and there has been a growing interest in the use of ML in and for them, especially in the last five years that saw the publications of more than half of the papers published until 2019 dealing with it [86]. In some cases, which are not the focus on this thesis, ABMs have also been used to support ML, generating data to train algorithms on [86].

[5] There is no consensus in literature on what exactly these nomenclatures mean. For instance, "heuristic" is used in Zhang & Vorobeychik [47] to indicate models still based on behavioural rules, which however are not grounded in any particular known theory and are selected ad hoc to match empirical data, while "stochastic" models adopt no rules at all but rely instead in a proper stochastic evaluation of features importance. Instead, An [83] uses the terms "empirical" and "heuristic" indifferently to encompass all models not based on theories (therefore including "stochastic" models as well).

The already cited increase of data available was one of the main reasons for this interest. ML provided the possibility, thanks to its ability in handling efficiently big amounts of data, to test various and more complex approaches to better link ABMs to the real-world [85]. For instance, ML enables modelling of really adaptive agents' with dynamic learning behaviours, especially through the use of reinforcement learning [89] for goal-oriented agents, which has been the most common application of ML in ABMs [86].

The data-driven trend of ABMs and their increase in complexity, root causes of their growing explanatory power, came at the cost of making the process of getting useful insights harder: 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 [86], [90]. Therefore, another important application of ML in ABM is to handle model experiments and results analysis, helping to understand relevant patterns, the various parameters importance and input-output relations [90]. Apart from this "simulation analytic challenge", Dahlke & Bogner [86] also stress the dangers that more complex models can pose in terms of computational efficiency and tractability, noticing how ML algorithms can in some cases help cope with this issue (despite at the risk of trading it with accuracy).

However, the opposite is also true: ML algorithms are often grey- or black-box approaches, whose interpretation can require particular effort or even not being possible at all. When the understanding of decisions' drivers is an objective, some ML algorithms could create more difficulties than solutions [83]. Furthermore, the process of deciding which algorithm to choose, calibrating its parameters and training it over the dataset is often difficult and time consuming [86]. Lastly, even though explicit theoretical assumptions are avoided, the use of a particular ML algorithm and of the features to consider in the first place imply some assumptions: the representation of the algorithm, i.e. the way it is rendered in computing language, defines the space of the models it can learn. If the right model is outside this space, the algorithm will not be able to reach the right solution[6] [87]. Therefore, the application of ML in ABMs should be carefully evaluated and limited to the cases in which the amount and quality of the data available make it meaningful.

All these challenges, combined with the relatively young age of the field and the lack of methodological clarity in ABMs developing process itself already highlighted in section 2.2, are the cause for the lack of common methods and frameworks for the combination of ABMs and ML (and for data-driven ABMs more in general). The first effort in this sense reported in literature – remarkably focused on LUCC applications – was a workshop held in 2005 on micro-level empirical approaches for ABMs, of which Robinson et al. [50] reported the resulting classification. Kavak et al. [82] identify the first attempt to draft a data-driven ABM method in 2007, stressing the need of general-purpose models due to the increased availability of data. In the field of innovation diffusion, Zhang & Vorobeychik [47] identify the work done in Zhang et al. [85] as the first one reporting a generic framework for data-driven ABMs. The lack of clear guidance 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 granularity and size of the available data. This implies that the architecture of the model and the method to develop it needs to be tailored to the specific situation, even

[6] An example of this issue is when an algorithm that can only learn linear relations between the target variable and the independent ones is used to learn from data whose underlying dependence is non-linear.

though the possibility of extracting general knowledge should be harnessed whenever possible [47]. However, Zhang et al. [85] report a framework that have resulted particularly useful for the method used in this thesis, reporting the following features and methodological notes as characterizing fully data-driven ABMs:

- No assumption on the social network of the model, which is substituted by spatial and social influence features included in the agents' behavioural model.
- Model parametrized through statistical learning method at the individual agent level, without any calibration of pre-defined dynamics on empirical data.
- Validation performed on independent data, not used for training the model.
- Validation performed not only at the macro-level over aggregated model outcomes, but also at the micro-level of agents' behaviour.

The lack of a stable methodological reference for combining ABMs and ML is particularly felt in applications to LUCC. In fact, literature is extremely scarce in this regard: from the really limited number of studies found (such as Zhao et al. [91], dealing with spatial planning, and Li et al. [92], on urban planning), only Sun & Müller [69] treated agricultural land use. All these studies were located in China and with a spatial scope limited to the county level. Sun & Müller [69] used both qualitative and quantitative data collected through interviews and surveys designed specifically for the study to build a Bayesian belief network representing agents' decision-making, applying it to a PES scheme

## 2.6 Positioning of this thesis

This thesis encompasses all the fields described in sections 2.4.1 to 2.4.4: it aims at explaining the spatial diffusion of SBP in Portugal through socio-economic and biophysical indicators, evaluating past and possible future PES schemes and the potential C captured as a result. Therefore, ABMs are a suited tool to address the complexity of the system under study and reach the objectives posed in section 1.4.

In terms of theory/data-driven trade-off, the thesis tried both approaches to compare their results and the robustness of the assumptions embedded in them, when applied to this specific case study. The model that assumes profit maximizing agents is the starting point of the analysis, which is then expanded with the consideration of other variables as proxies for bounded rationality. However, the aim of studying the specific context suggested fully data-driven approaches as a suited way to increase the robustness of the analysis. The many and various sources of data available allowed for it and made an approach based on the integration of ML in the ABMs an appealing option.

The scarce literature available regarding the combination of ML and ABMs for agricultural systems, the study of innovation diffusion in LUCC, the design of PES and the integration of environmental assessment in ABMs are all factors contributing to the novelty of this thesis. Moreover, this work constitutes the first application of ABMs to whole continental Portugal in the context of LUCC and develops the first ABM in the context of innovation diffusion and policy design in agricultural systems which relies entirely on ML algorithms, without combining them with any explicit theoretical assumption.

# 3 Materials and methods

The novelty of this analysis and the lack of clear guidelines for data-driven ABMs highlighted in the literature review meant that no ready-made process could be followed, requiring the development of an original method to achieve the objectives specified in section 1.4. The choice of the various modelling approaches was therefore strongly tailored to the available data, in order to link them to and validate their results with empirical observations. The study of the data showed the necessity of dividing the analysis in two separate parts.

The first part of the work used detailed information at the single farmer level from a survey of 43 farmers (described in section 3.1.1), to focus on their individual decision-making process regarding the choice of adopting SBP or not. To get the most insightful results and avoid unnecessary overcomplication, this thesis applied different approaches involving different levels of complexity and constructed three different models. The first consisted in a *Farmer-based Toy-ABM* (section 3.2.1) based on the calculation of a differential net present value (NPV) between the two options a farmer faces: maintaining SNP, the most frequent pasture type in Portugal, or adopting SBP, with improved agronomic performance and incentivised through a PES scheme. To address uncertainty both in terms of prices and of farmers decision-making, this model included farmers' education as a proxy for their risk propension, in line with the only analysis carried out for the same issue [19], [93]. The second approach modified the first model trying to get deeper into the factors that can influence the NPV expected by farmers, developing the *Farmer-based Calibrated ABM* (section 3.2.2). This model expanded the proxies for risk perception to other characteristics of the farmers, selected through analysis of the data and system knowledge. Most importantly, to lose the assumptions and further tie the results to the empirical data, the importance of these proxies was calibrated on the adoption observed in the survey. The third approach aimed at avoiding any theoretical assumption, to check if a purely data-driven model was suited to estimate adoption decisions. Due to the lack of time and space dimensions, this did not require the construction of a proper simulation and was simply implemented through a ML model, specifically a logistic regression presented as *Farmer-based Logistic Regression* in section 3.2.3.

The second part of the work relied on data for the whole continental Portugal from an agricultural census (section 3.1.5). The SBP area adopted was also used, as it is available until 2008 as aggregated values for agronomic regions (section 3.1.3) and from 2009 to 2012, during the PCF project, at the individual farmers level (section 3.1.4). However, the impossibility to match the information on individual adoption to the demographic and socio-economic characteristics of each farmer required shifting the level of the analysis to the municipalities, aggregating the farmers into single agents based on geographical location. The aim of this part was to link the yearly adoption in each municipality to its demographic, socio-economic, agricultural and environmental characteristics, by predicting both the presence or not of adoption in the municipality, a 1/0 problem, and the hectares of SBP installed in each municipality, a continuous variable problem. The great amount of data available and the lack of correspondence of agents with real world decision-makers suggested to directly use a fully data-driven approach, combining ML to predict adoption in each municipality and ABM to model the spatial and time dimension and the interconnection among municipalities, in the *Municipalities Data-driven ABM* (section 3.3.3).

Table 1 summarizes the organization of the work done just described and Figure 3: graphical representation of the steps of the analysis.Figure 3 the steps of the analysis.

*Table 1: organization of the work done, with the specification of the dataset and models included in each part.*

| Level of the analysis | Scope | Datasets considered | Model(s) | Models' output |
|---|---|---|---|---|
| **Farmers** | 43 farmers | • Animal Future survey<br>• Economic | Farmer-based Toy-ABM | Decision to adopt SBP or not (classification problem) |
| | | | Farmer-based Calibrated ABM | |
| | | | Farmer-based Logistic Regression | |
| **Municipalities** | Continental Portugal (except Algarve) | • SBP adoption previous to the PCF project<br>• PCF project<br>• Census<br>• Climate and soil<br>• Portuguese municipalities shapefile | Municipality-based Data-driven ABM | Area of SBP adopted (regression problem) |

*SBP – sown biodiverse pastures; PCF – Portuguese Carbon Fund; ABM – agent-based model*

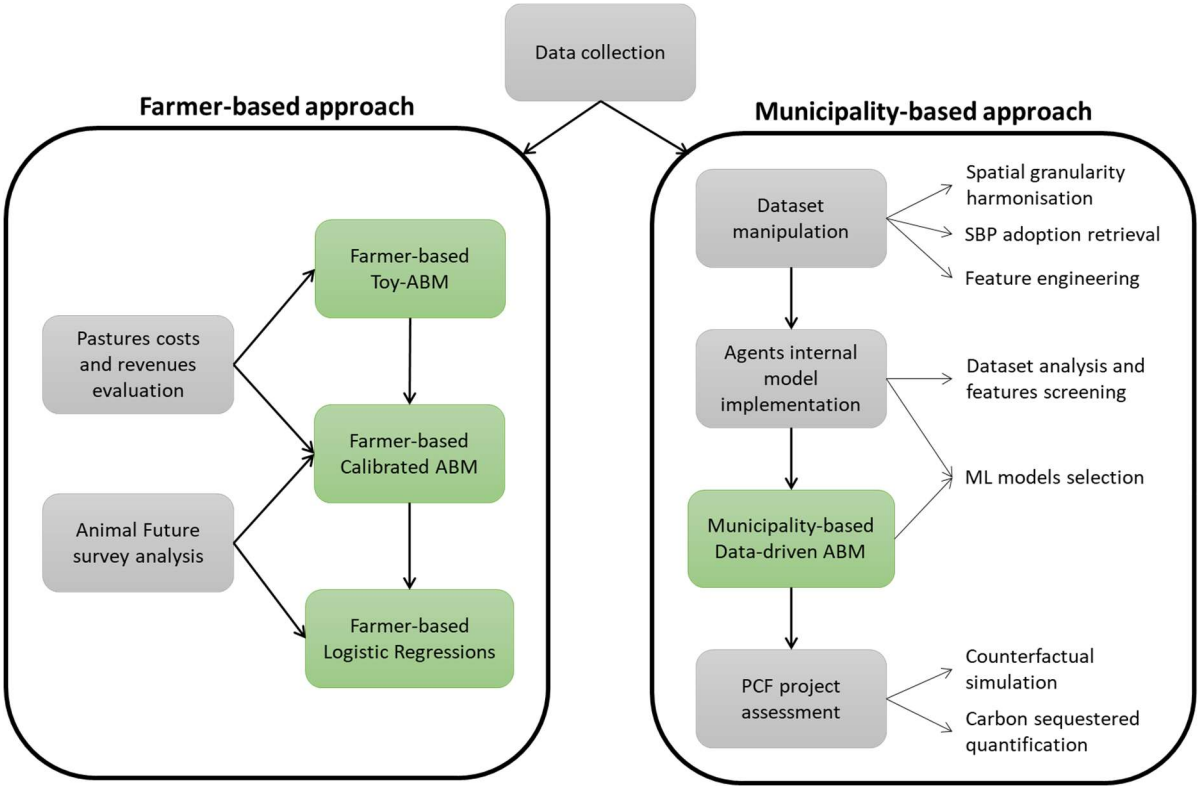



*Figure 3: graphical representation of the steps of the analysis.*

*The model names refer to their implementation and validation.*
*ABM – agent-based model; SBP – sown biodiverse pastures; PCF – Portuguese Carbon Fund.*

The majority of the work required during this thesis was implemented through the Python coding language and written on the Jupyter Notebook web application (https://jupyter.org). The proper ABM simulations of the Farmer-based Toy-ABM, the Farmer-based Calibrated ABM and the Municipality-based Data-driven ABM were instead written on the Spyder integrated development environment (IDE) (https://www.spyder-ide.org) using the mesa package (https://mesa.readthedocs.io/en/master/), an open-source ABM framework developed in Python. When not specified otherwise, the ML models where instantiated from the classes provided by the scikit learn library (https://scikit-learn.org), which also provided the functions used to tune and analyse them. All the source code (SC) is available as a GitHub repository (https://github.com/giacrava/thesis-sbp-abm) and can be clarified on request. The path to the source code for each section will be specified in the footnotes, starting from the GitHub repository home folder and introduced by the text "SC: ". Codes referring to the municipality-based approach cannot be run, since the relative databases could not be provided due to data protection. Simple calculations such as the ones presented in section 3.2.1.1 were performed in Excel.

# 3.1 Data availability

## 3.1.1 Animal Future survey data

The first dataset available is a survey conducted with farmers predominantly in the Alentejo region, in the scope of the project "Animal Future" (AF) (https://www.animalfuture.eu/). This is a declarative source of information, collected through in-person interviews conducted between May and October 2019[7] by the researchers involved in the project.

This survey is the most detailed dataset on farmers used during this work in the sense that it was the only source with data for individual farmers that includes agronomic, economic, social and environmental data. Each farm is characterized by a code composed of the letters *PT* to indicate Portugal and a number of two figures to differentiate them. The farm *PT00* has been excluded from the analysis since it's located in the North of Portugal and therefore presented very distinct characteristics from the other farms. The AF survey data is divided into various sections:

- General data: general information regarding land use. Only 30 farmers answered to this section of the survey completely.
- Social data: social characterization of the farmers. Only 30 farmers answered to this section of the survey completely.
- Economic data: economic characterization of each farmer, consisting of the breakdown of their revenues and costs and the subsidies they are benefitting from. All 43 farmers answered to this section of the survey.
- Environmental data: description of the management practices used by farmers that can influence the environmental impact of their farms. All 43 farmers answered to this section of the survey.

[7] These data are unpublished but are available in the GitHub repository linked to this thesis at the path farmer_level_analysis\survey_data\all_data as multiple spreadsheets.

The lists of both features included and excluded from the rest of work are reported in Table A.1 and Table A.2, with the explanation of their meaning and the rationale for the decision whether to consider them or not in the analysis here, as some of the features had to be discarded due to the limited sample available. The choices were guided by system understanding, literature review and specific data characteristics.

### 3.1.2 Economic data

The second dataset used regards the economic characterization of SNP and SBP. Teixeira [19] reports prices and other relevant variables for installation and maintenance of 1 hectare of SBP in 2009, using field data from the farm Quinta da França for prices and other sources for input quantities ([94], Nuno Rodrigues, personal communication). Morais [95] reports the data required to calculate the maintenance of 1 hectare of SNP for 2019. Appendix B reports these values.

Other than costs for installation and maintenance, livestock feeds are an important fraction of pastures costs. The costs of feed per hectare of each pasture type depends on the supplementation of feed required per hectare of pasture, which is on average 672 kg/ha.y for SBP and 1047 kg/ha.y for SNP if well-managed [95], and the cost of the feed components, which for silage was 0.09 €/kg in 2019 [95] and for concentrate feed was 0.311 €/kg in 2009 [96].

### 3.1.3 SBP adoption previous to the PCF project

Teixeira [93] reports the area of SBP installed each year in Portugal corresponding to the seed sold from 1996 – when the first SBP were installed – to 2008 – right before the beginning of the PCF project. The first row of the dataset – "*Region"* – reports the geographical area that the adoption is referred to, which however does not correspond to any official division of the Portuguese territory, being a combination of municipalities, districts and larger agronomic regions. The other columns report the area in hectares in which SBP was adopted during the year corresponding to the columns' header.

### 3.1.4 PCF project data

The PCF project generated important data for the scope of this thesis. Access to these data is restricted and was kindly provided for this thesis by Terraprima, and can be requested to ambiental@terraprima.pt. The PCF project database consists of one spreadsheet reporting specifications about the portions of land where SBP were installed during the PCF project, from 2009 to 2012 for a total of 1113 entries. The information was collected by field technicians who performed verification and farmer advisory during the PCF project. They visited the farms 3 to 4 times per year to assess the state of the pastures and provide management recommendations. This was a qualitative collection of data and there were no quantitative criteria for the evaluation. The information in this dataset pertains only the plots where SBP were installed and not the rest of the farms. In the years 2013 and 2014 the project was still running but not accepting new farmers, only paying the ones who already joined in the previous years.

The dataset includes only the installations that were successful. It can be considered that it contains all the parcels of land where SBP were successfully installed in the period 2009-2012. Moreover, since the PCF project contracted only new installations, all the parcels refer to land where SBP was never installed

before (no resowing). The possibility that some farmers did not want to take part in the project for personal reasons or that some installed even when payments were not provided remains, but in case this adoption was surely residual compared to the total area switched to SBP under the PES scheme. From 2009 to 2012 in fact the maximum area the PCF project would have paid for was not reached and this makes it very unlikely that anyone installed without being paid.

The database included information on the year in which SBP was installed in the parcel (from 2009 to 2012), the municipality where the parcel was located and the size of the parcel in hectares. Additionally, Terraprima provided data on the PES offered to farmers based on the year they joined the programme, reported in Table 2.

*Table 2: yearly payments to farmers installing SBP during the PCF project based on the year of adoption, in €/ha.*

| Year | 1st year | 2nd year | 3rd year |
|---|---|---|---|
| **2009** | 50.72 | 50.72 | 51.82 |
| **2010** | 65.86 | 66.24 | - |
| **2011** | 38.94 (1st phase) - 43.58 (2nd phase)[8] | 43.58 | 43.58 |
| **2012** | 69.18 | 68.78 | - |

### 3.1.5 Census data

The Instituto Nacional de Estatística (INE, https://www.ine.pt/) carries out every 10 years and makes available upon request an agricultural census for Portugal. Since the census for 2019 is not available yet and the one for 2009 could not be retrieved in time for this thesis, the census used was the one for 1999. This was available as a set of spreadsheets, one for each of the historical provinces in which Portugal was divided until 1976 apart from Algarve, which was unavailable. Additionally, the census did not report data for the municipalities of Porto and Lisbon, which however include mostly urban area. Data are provided at the level of entire regions, sub-regions and single municipalities.

Each spreadsheet is composed of over 100 sheets (the same fields are sometimes reported in more than one sheet to cover all the municipalities), encompassing a wide range of information regarding land use (cultivations, pastures, other land uses), socio-economic and demographic characteristics (education, external activities, labour, age, income, housing) and farm management information (irrigation, livestock, machines, legal form). From these, the work used only some. The first block were the total area of permanent pastures in hectares (*pastures_area_munic*) and the correspondent number of data point collected (*expl_number*) for each municipality. The municipality of São João da Madeira had a null area of pastures and for this reason was excluded from the rest of the analysis[9]. The second block corresponded to the data in the sheets *Principais características do Produtor Singular,* whose details are reported in Appendix C. The third block of features referred to the economic situation of the farmers reported in the sheets *Produtor singular segundo a Dimensão Económica e as Classes de*

[8] In 2011 a second phase of the project started and farmers received different payments for the first year depending on the moment of adoption.
[9] SC: municipality_level_analysis\data_preparation\census\Pasture area for each municipality from census data.ipynb

*Idade* and are also reported in Appendix C. All the values collected from these two sheets were divided by the relative value of *individual_prod_num* (apart from this entry itself), in order to report them as fraction of total number of farmers in the municipality. The last feature used was the ratio of land rented in each municipality over the sum of owned and rented land (*land_rented*)[10].

### 3.1.6 Climate data

The data used regarding the Portuguese climate was the version 21.0e of the E-OBS dataset from the EU-FP6 project UERRA (https://www.uerra.eu) and the Copernicus Climate Change Service, and the data providers in the ECA&D project (https://www.ecad.eu) [97].

The E-OBS dataset consists of a raster geospatial dataset available at 0.1 and 0.25 degrees regular grid for daily values of various climate variables, covering Europe (the area 25N-71.5N x 25W-45E) and the years from 1950 to 2018. The data are provided in a Network Common Data Form (NetCDF-4) format (https://www.unidata.ucar.edu/software/netcdf/) and come as a 100-member ensemble dataset, i.e. a climate dataset that consists of a number of equally probable realizations generated from input stations measurements and related to data in gridded format, constructed through the procedure reported in Cornes et al. [97] Specifically, the ensemble mean values and 90% uncertainty range are available, as a measure respectively of the "best-guess" value and of its uncertainty.

Datasets with 0.1 degrees regular grid were used, since to retrieve data at the municipality level a finer division is more precise. The climate variables considered for the following work are the daily mean, minimum and maximum temperature in °C and the daily precipitation sum in mm/pixel.day[11], as temperature and precipitation are the most important weather variables determining agricultural yields.

In order to reduce the size of the databases, which would have slowed down its following usage, an initial manipulation was carried out through the netCDF4 Python library (https://pypi.org/project/netCDF4/)[12]. Data was restricted to an area framing Portugal (36N-42.5N x 10W-6W) and for the years from 1995 onwards (to include all the years for which data on SBP adoption are available). Since the time step used for temporally explicit models was one agricultural year, the period starting on 1st September and ending the 31st August of the following solar year[13] was used. The daily values were processed to obtain variables referring to this period, which were reduced to only 24 values of the time variable in the NetCDF-4 database, one for each year from 1995 to 2018. Table 3 reports the variables obtained, together with their meaning and the procedure to obtain them. The

[10] SC: municipality_level_analysis\data_preparation\census\Census features collection.ipynb
[11] In the unit of precipitation sum, pixel refer to an area value with 0.1° side, that is the definition of the used climate dataset. Therefore, mm/pixel.day does not actually correspond to a quantity over a fixed area, since each degree of latitude and longitude has a different values in meters at different latitudes. However, over the latitudes to which the data were restricted (from 36°N to 42.5°N), this difference can be neglected introducing a small approximation. In fact, from rough calculations based on a converter of degrees into meters based on the latitude (http://www.csgnetwork.com/degreelenllavcalc.html) resulted that an area of 0.1° x 0.1° at 36°N correspond to nearly 100 $km^2$, while at 42.5°N to nearly 91 $km^2$.
[12] SC: municipality_level_analysis\data_preparation\environmental\climate\Climate data first manipulation.ipynb
[13] For instance, the year 1995 refers to the period from 1st September 1995 to 31st August 1996. For the rest of the thesis, the term "year" will refer to the agricultural year as defined here.

manipulation was aimed at generating a mix of variables juxtaposing to average conditions extreme ones, which together can give a clear picture of the climate of each year.

*Table 3: variables obtained from the manipulation of the climate data and procedure followed to obtain them.*

| Variable | Unit | Procedure to obtain it |
|---|---|---|
| Average daily mean temperature over one year | °C | Averaged the daily mean temperature over each year |
| Days over one year in which the mean temperature was over 20 °C | days | Counted for each year the number of occurrences of a daily mean temperature higher than 20 °C |
| Days over one year in which the mean temperature was over 25 °C | days | Counted for each year the number of occurrences of a daily mean temperature higher than 25 °C |
| Average daily minimum temperature over one year | °C | Averaged the daily minimum temperature over each year |
| Days over one year in which the minimum temperature was below 0 °C | days | Counted for each year the number of occurrences of a daily minimum temperature lower than 0 °C |
| Average daily maximum temperature over one year | °C | Averaged the daily maximum temperature over each year |
| Days over one year in which the maximum temperature was over 30 °C | days | Counted for each year the number of occurrences of a daily maximum temperature higher than 30 °C |
| Sum of the precipitation fallen over one year | mm | Summed the daily precipitation sum over each year |
| Days over one year in which the precipitation fallen was 0 mm | days | Counted for each year the number of occurrences of a daily precipitation sum equal to 0 °C |
| Highest number of consecutive days over one year in which the precipitation fallen was 0 mm | days | Counted for each year the longest number of consecutive occurrences of a daily precipitation sum equal to 0 °C |

### 3.1.7 Soil data

The Land Use/Land Cover Area Frame Survey (LUCAS) programme is a survey on land cover and land use started in 2001 by Eurostat (the statistical office of the European Union). In 2009, the LUCAS survey was extended with a topsoil assessment component (“LUCAS – Topsoil”) to collect and analyse 20,000 soil samples and create a dataset on soil at European level [98]. For Portugal 476 samples were collected, of which 99 in grasslands, 116 in croplands, 193 in woodlands and 52 in shrubland. Azores and Madeira districts are not included in the database.

Maps of soil chemical properties in the form of raster geospatial dataset with a 500 m resolution for EU-26 (not including Croatia and Cyprus) were generated from the discrete sampling points, through Gaussian process regression based on environmental auxiliary variables [99]. The European Soil Data Centre (ESDAC), (https://esdac.jrc.ec.europa.eu/) can provide these maps in a GeoTIFF format upon request [100]. Table 4 reports the soil properties for which maps are available their unit and the rationale for considering them in the following work.

*Table 4: soil properties for which LUCAS topsoil maps are available, their units and decision for further consideration in this thesis.*

| Soil property | Unit | Rationale |
|---|---|---|
| **pH (measured in $H_2O$)** | - | There is a relatively narrow range of pH values suited for SBP, as soil cannot be too acid nor alkaline |
| **Calcium carbonates ($CaCO_3$)** | g/kg | $CaCO_3$ is added (liming) for pH correction when installing SBP |
| **C:N ratio** | - | The organic C to nitrogen ratio indicates the organic matter turnover and nitrogen availability |
| **Nitrogen (N)** | g/kg | N and P are the main limiting nutrients for SBP growth |
| **Phosphorus (P)** | mg/kg | |

### 3.1.8 Portuguese municipalities shapefile

An ESRI shapefile [101] for the Portuguese municipalities is publicly available at dados.gov (https://dados.gov.pt/), the Portuguese Public Administration's open data portal. This database is produced by the Portuguese Agência para a Modernização Administrativa [102].

From this database a new file was created, keeping only the following fields, renaming some in a more explicit way and excluding the rest since consisting only of other codes to further characterize each municipality: *NAME_1* (renamed to *District*), the name of the district of the municipality; *NAME_2* (renamed to *Municipality*), the municipality name; *Geometry*, *the* specification of the polygon representing the position and shape of the municipality; *CCA_2*, a unique code to identify each municipality.

## 3.2 Farmer-based approach

The first part of the analysis consisted of a farmer-based approach, based on the study of the individual farmers decision-making regarding SBP adoption and therefore capable of providing useful insights on the decision processes within individual farms. This required the use of dataset at the individual farmer level, the AF survey dataset and the economic data (section 3.1.1 and 3.1.2). Three models were developed under the farmer-based approach. The purpose of all was to reproduce the specific pattern of SBP adoption observed in the AF survey data, classifying the farmers in the ones who adopted SBP at least in one portion of land and the ones who did not adopt at all. The amount of area adopted was not a variable considered in the farmer-based approach due to insufficient size of the sample. None of the farmer-based models include a spatial or temporal dimension explicitly due to the fact that the geospatial location of the farms and years of pasture installation were not collected.

The starting point was the Farmer-based Toy-ABM (section 3.2.1), based on economic calculations and the farmers' education level as a proxy for uncertainty. To properly understand the factors driving adoption the analysis proceeded with the Farmers-based Calibrated ABM (section 3.2.2), based on the Farmers-based Toy-ABM but with two major differences: the consideration of more characteristics of the farmers as proxies for uncertainty and the use of a calibration process to understand their importance. The last farmer-based approach, the Farmer-based Logistic Regression (section 3.2.3), avoided any explicit assumption through the use of ML algorithms to model farmers' decisions. The implementation strategy for each model is described in the next sections.

### 3.2.1 Farmer-based Toy-ABM

The first farmer-based model was named “Farmer-based Toy-ABM” since it runs in just one step and assumes a simple behaviour for the farmer agents, consisting only in the calculations of the expected differential NPV (EDNPV) between adopting SBP and maintaining SNP. The differential NPV is expected in the sense that farmers the maintenance costs and the revenues are subject to uncertainty due to the possibility of pasture settlement failing due to mismanagement or climate. This specific model includes farmers’ education level as an arbitrary proxy for uncertainty in the calculations, probability of mismanagement of the pasture, and risk propension of the farmer, creating agents with bounded rationality. The lower the education level, the more sceptical the *Farmer* was modelled to be regarding SBP adoption, as in [19]. This model is an ABM implementation of prior work [19], [93] carried out when the PCF project was designed, using supply and demand curves at the aggregated level. This approach allowed to place even more focus on the explanation of the individual farmers’ behaviour. Before describing the model architecture, the following section reports the economic calculations needed to initialize it.

#### 3.2.1.1 Pastures costs and revenues

This section reports the procedure to calculate the yearly cash flows for adopting SBP and maintaining SNP for a farm where SNP are currently present. These values are required in the Toy-ABM to calculate farmers’ EDNPV. Since these NPVs are differential, the cash flows can be restricted to costs and revenues that differ between SBP and SNP.

The NPVs are calculated over 10 years, since this is the approximate average time that well managed SBP can be sustained without resowing. These calculations are referred to the year 2009, for which data were available regarding SBP maintenance and installation and in which the payments to offer SBP were the highest. Therefore, data referred to 2019 had to be adjusted for inflation multiplying them for the Consumer Price Index (CPI) for Portugal relative to these years ($CPI_{2019\rightarrow2009}$). This was calculated as the ratio of $CPI_{2019\rightarrow2010}$ and $CPI_{2009\rightarrow2010}$, the CPI for Portugal respectively for 2009 and 2010 referred to 2010, 98.617 and 110.624 [103].

To calculate costs for installation and maintenance for SBP in 2009, the total cost (CT) of each individual operation of Table A.3 was calculated as (using the same notation introduced in the table)

$$CT = h_{lab} * C_{lab} + h_{lab} * CV + CF + C_{amort} + Q * P. \quad (1)$$

Then, the individual operations were summed by their type to obtain the total cost aggregated by type of operation. Lastly, the costs of all installation operations were summed to obtain the total cost for the installation of 1 hectare of SBP. The costs for maintenance of SNP, corresponding to just harrowing, were also calculated through Equation (1) and then adjusted for inflation reporting it to the corresponding value in 2009.

To evaluate costs and revenues depending on the livestock units in the farm, some additional considerations were necessary. It has been observed that farmers do not adopt SBP in their entire farms. This is probably due to the higher amount of feed produced by SBP than SNP: in fact, SNP have a maximum possible stocking rate of 0.44 CU/ha (where CU stands for Cattle Units, equivalent to one

adult cow), while SBP on average can support almost three times as much, at 1.18 CU/ha [94]. The analysis assumed that the higher stocking rate of SBP does not translate into a change in livestock units in the farm and therefore revenues such as steer selling and support for breeding cows related only to the number of animals were not included in the calculations, being the same regardless of the pasture installed. This is justified by the empirical observation that farmers sow SBP to provide extra feed and not to expand herd size (Terraprima, personal communication).

Feed costs are an important livestock-related cost that, being linked to structural differences in the two pastures and not only to livestock units, needed to be considered. The amount of feed supplementation needed per hectare of SBP and SNP reported in section 3.1.2 considers their different stocking rates. As the stocking rate for SBP is higher than for SNP, more livestock units can be hosted and fed on one hectare of the former with the reported supplementation. However, the amount of feed saved when converting from SNP to SBP has only been obtained in the literature for average cases per hectare and not per farm [104]. Those estimations cannot be directly used in the calculations in this thesis, which require the consideration of the effects on the entire pasture area. To consider feed savings, the calculation of the amount of feed saved per hectare of SNP switched to SBP (FS) was assumed, for simplification, to be obtained as

$$FS = FR_{SNP} - FR_{SBP} * \frac{SR_{SNP}}{SR_{SBP}}, \quad (2)$$

where $FR_{SNP}$ and $FR_{SBP}$ are respectively the feed supplementation required per hectare of SNP and SBP and $SR_{SNP}$ and $SR_{SBP}$ are respectively the stocking rates of SNP and SBP. The ratio of $SR_{SNP}$ and $SR_{SBP}$ gives the hectares of SBP that need to be installed in order to host the maximum livestock units of one hectare of SNP. Put simply, for each hectare of SNP switched to SBP the feed saved equals the feed supplementation required in the hectare of SNP before switching minus the feed supplementation required in the hectares of SBP installed to substitute that hectare and host all its livestock units at maximum capacity. To get the savings in euros per hectare per year, FS and the cost of the feed per kilogram were multiplied (assuming that farmers adopt until the point of substituting all purchased feed in the periods of production also implies that every kilogram produced by the pastures saves the purchase of one kilogram of feed). Being a mix of silage and concentrate, the formula to calculate the price per kilogram of feed is (FP) is

$$FP = FS * PS + FC * PC, \quad (3)$$

where FS and FC are the fraction respectively of silage and concentrate in the feed. Assuming a low silage feed formulation, these fractions are respectively 0.3 and 0.7 [105]. PC is the price of concentrate reported in section 3.1.2, already referred to 2009. PS is the price of silage reported in section 3.1.2 adjusted for inflation and thus reported to 2009. Feed savings were considered as additional costs to bare in maintaining SNP and therefore entered in the calculation as a differential, which means that no extra costs for feed were considered in SBP (consistent with the differential nature of the NPV calculations).

To all costs, general costs of 5% and an interest on circulating capital of 1.5% were added, as done in [93]. The last step required was to sum all the costs and expected revenues for each pasture system in

order to get the yearly cash flows per hectare. Regarding SBP, the cash flow for the year 0, i.e. the year in which the decision regarding adoption of SBP is taken, corresponds to the installation costs. Maintenance operations are supposed to be conducted with the frequency reported in Appendix B starting from the $X^{th}$ year after the installation, where X is their frequency. Regarding SNP, the maintenance was assumed to start in the first year after installation (year 1). The cash flow for the year 0 is therefore null, since in the first year SBP cannot be grazed during the spring, which is the most productive season, to ensure good settlement and establish the seed bank (of self-reseeding hard seeds), and therefore farmers still incur costs with feed supplementation. In every other year instead the feed saved adopting SBP was considered as a cost of maintaining SNP, assuming average productivity of SBP throughout their lifespan.

It should be noticed that the cash flow calculations were made under the assumption that prices are constant for 10 years. All prices considered here have a very high variance, but since farmers have no information on how they will evolve at the moment of the decision, it seemed plausible to use present values and assume them constant throughout the time frame considered.

#### 3.2.1.2 Model description[14]

The model description follows the ODD + D protocol, an extension of the ODD (Overview, Design concepts, Details) protocol for describing individual- and agent-based models with an additional focus on human decision-making [106], [107]. The last update to the ODD protocol described in [108] is also taken into consideration. The choice of this protocol is due to its current diffusion, especially in the field of ecology where it was used in 20% of the published papers adopting ABMs in 2018 [108]. Moreover, the use of a standardised and less technical protocol as the ODD facilitates communication across disciplines and with relevant stakeholders, while allowing for a comprehensive description of the model. However, in the interest of saving space and avoid repetitions parts previously described will be omitted, as the relative sections of the protocol. The “Design concepts”, “Initialization” and “Submodels” sections of the protocol are reported in Appendix D.

*Entities, state variables and scales*

The *Model*[15] represents the global environment and is an abstract entity controlling the submodels. *Farmers* are the main agent in the model and represents the farms’ owners. *Farms* represent the farms where the *Farmer* who owns them can decide to install SBP.

*Pastures* are the objects representing the type of pasture each farm can have. For simplicity and since the aim is to predict only adoption or not, each *Farm* can have only one type of pasture. *Pasture* entities do not hold any attribute representative of farms, but only generic characteristics of the pasture type. This allowed to have one pasture of each type, avoiding to instantiate a different *Pasture* object for each farm. *Pasture is* however a base entity, serving only as an abstract class for the more specific pasture types. If a pasture is not considered for adoption, it is directly a subclass of *Pasture*: this is the case of

[14] SC: farmer_level_analysis\toy_abm
[15] In the following description, names in italic and capitalized will refer to model entities, to differentiate from the real-world counterparts.

*NaturalPasture.* Pastures types that are considered for adoption are subclasses of *AdoptablePastures* (another base entity subclass of *Pasture*). The only adoptable pastures in this model are SBP, whose relative entity is *SownPermanentPasture*, subject to governmental payments.

*Market* is an abstract economic entity that reports all the economic values necessary for NPV calculations to the *Farmers.* It represent all the retailers and service providers which the farmers interact with to buy the necessary products and equipment to install and maintain the pastures in their farms. Each pasture type has its own *Market,* thus in the model are present a *NaturalPasturesMarket* and a *SownPermanentPasturesMarket*. *Governments* represent whoever is responsible for setting and offering incentives to farmers to adopt a new pasture. A different one is present for each pasture subject to payments and/or subsidies. In this model, the only one is *SownPermanentPasturesGovernment,* representing the PCF project and the payments offered within it. The attributes of each entity are reported in Table A.6.

The model runs in only one step, considering the most beneficial economic situation of 2009 with the highest payments offered. This step can be though as spanning from 2009 to 2019, since for the architecture of this model there is no difference on whether farmers adopt at the beginning or at the end of this period.

*Process overview and scheduling*

The schedule of the models' time step is the following (the submodels not completely specified are explained in detail in the section "Submodels" of Appendix D):

1. Each *Farmer* agent calls its *Farm* submodel to evaluate the adoption of alternative pastures, which implies:
    I. The *Farm* gets the pastures that can consider to adopt, checking which can be adopted (the ones included in the *Model*'s *adoptable_pastures* attribute) that are not already implemented in its farm.
    II. If any pasture can be adopted, the *Farm* executes its submodel to calculate the differential NPVs per hectare between switching to the adoptable pastures and maintaining the actual one. In order to do this, the following actions are executed:
        i. The *Farm* calls its *Pasture's* "NPV keeping" submodel to calculate the NPV of maintaining the actual pasture.
        ii. The *Farm* calls, for each adoptable pasture, its "NPV adoption" submodel to calculate the expected NPV of adopting the relative pasture, passing the level of education of the *Farmer* who owns it.
        iii. The *Farm* subtract to each ENPV of adopting a new pasture to the NPV of maintaining the actual one, getting the EDNPVs.
    III. If any EDNPV is positive, i.e. if switching to a different pasture is considered economically convenient, the *Farm* changes its pasture to the *Pasture* entity which presented the highest EDNPV (modifying its *pasture_type* attribute).

The entities *Farmer* are the ones called to execute the action, to make the model more realistic. Their activation sequence does not have any influence on the model's outcome, since farmers do no influence each other's decision. However, the actual calculations happen inside the *Farm* and *Pasture* entities, in order to facilitate the call of the required attributes. In particular, the NPVs per hectare are retrieved from the *Pastures* since they do not depend from the specific *Farm*, but only from the pasture type. In this way, they need to be calculated only once for each *Pasture* and not for each *Farm.*

#### 3.2.1.3 Model validation and output analysis

The validation of the Farmer-based Toy-ABM consisted in the comparison of its output with the adoption observed in the AF survey data[16]. In order to conduct a proper micro-validation, the analysis extracted from the AF survey the real adoption of each farmer, creating a dataset reporting for each farmer if they adopted SBP even just in a portion of their field or not[17]. This dataset was compared with the models' prediction regarding adoption for each farmer, which enabled the calculation of precision, recall and F1 score[18]. The model was validated also at the macro-level, comparing the modelled and observed percentage of farmers that adopted. Finally, to understand the micro-predictions of the model, from the output at the agents level the work extracted the EDNPV to adopt SBP calculated by *Farmer* agents of different education level. Since this is the only variable influencing the EDNPV values, all *Farmers* with same education have the same expected value of differential NPV.

### 3.2.2 Farmer-based Calibrated ABM

The Farmer-based Calibrated ABM was based on the Farmer-based Toy-ABM with two important modifications. The first is the inclusion of more proxies than only farmers' education level to quantify uncertainty in their economic calculations and risk propension. The second is the calibration of the influence of these proxies on the available data. This allowed for a deeper understanding of which factors drive farmer's decision-making process and how. Since the AF survey presented far more variables than observations, a first analysis of these data to select the features to include in the ABM was required and is reported in the following section.

#### 3.2.2.1 AF survey data analysis and features screening

In order to consider all the relevant features presented in Table A.1, only the observations corresponding to the 30 farmers that answered to the whole survey could be included in the analysis, neglecting the 12 for which the general and social part was missing. Table 5 reports the manipulation of the relevant fields of the AF Survey data to create the features used during the analysis[19].

[16] SC: farmer_level_analysis\toy_abm/toy_abm analysis.ipynb
[17] A farm adopted SBP in a portion of the field if at least one of the values corresponding to the farm of the *AREA_ID* column of the environmental section of the survey contains the string "Sown permanent pasture".
[18] Precision reports how many of the farmers classified as adopters actually adopted, recall is the ratio of positive instances correctly classified (how many of the farmers that have adopted were classified as adopters) and the F1_score is the harmonic mean of precision and recall [109].
[19] SC: farmer_level_analysis\survey_data\Survey data manipulation.ipynb

*Table 5: final set of manipulated features from the survey data, their type (if categorical or numerical) and procedure to generate them.*

| Feature name | Meaning | Unit / Values | Type | Manipulation procedure |
|---|---|---|---|---|
| *Adopted SBP* | Label, reporting if the farmer adopted or not SBP before 2019 | 1 if the farmer adopted SBP, 0 otherwise | Categorical (binary) | Encoded as 1 if an *AREA_ID* belonging to the farm contains the string "Sown Permanent Pasture", 0 otherwise |
| *Percent RentedLand* | Percentage of land rented over total land owned | [0, 1] | Numerical | $\frac{RentedLand}{OwnLand}$ |
| *LegalForm* | Legal form of the farm | "Individual" if individual, "Associated" if part of a cooperative | Categorical (binary) | "Individual" values left as such, while grouped the other values into "Associated" |
| *Farmer Since* | Years of experience as farmer | Years | Numerical | Equal to the *FarmerSince* feature |
| *Pasture Surface* | Farm's pasture area | Hectares | Numerical | Sum of all the *Surface* entries referring to the farm |
| *Distrito* | District where the farm is located | Any district in Portugal | Categorical | Equal to the *Distrito* feature |
| *Concelho* | Municipality where the farm is located | Any municipality in Portugal | Categorical | Equal to the *Concelho* feature |
| *Cattle Percentage* | Percentage of cattle over total livestock | [0, 1] | Numerical | $\frac{\sum_{LivestockType=cattle} AverageNumber}{\sum_{LivestockType} AverageNumber}$ |
| *Highest Educational Degree* | Highest education the farmer completed | Primary, Secondary, Undergraduate, Graduate | Categorical (ordinal) | Equal to the *HighestEducational Degree* feature |
| *Highest Agricultural Educational Degree* | Highest degree related to agriculture the farmer obtained | None, Undergraduate, Graduate | Categorical (ordinal) | Equal to the *HighestEducationalAgriculturalDegree* feature, coding NaN values as "*None*" |
| *Expectation Family Succession* | Reports if the farmer is expecting a family member to inherit the farm | "Yes" if the farmer expects family succession within the family, "No" otherwise | Categorical (binary) | Encoded as "Yes" if *ExpectationFamilySuccession* is equal to 2, "No" otherwise |

The following step consisted of a brief exploration of the dataset, to get the distribution of the labels and plotting the distribution of both numerical and categorical features[20]. For the categorical features each

[20] SC (until the end of the section): farmer_level_analysis\survey_data_analysis\Survey data analysis and features screening.ipynb

plot reported also the distribution of the labels within each category, to have an immediate visual idea of the relevance of each (Figure A.2).

The work proceeded with the study of the importance of each feature on predicting the decision to adopt or not SBP. I applied two filter based feature selection methods, one for numerical and one for categorical features. These are methods based on the specification of a metric used to evaluate the dependence of the target variable from each feature. For numerical features, I calculated the Spearman ρ correlation coefficient with the label *Adopted SBP*, which quantifies the existence of monotonic relations. For categorical features, I performed a Pearson's Chi-Squared test [110]. A Pearson's Chi-Squared test calculates the Chi-Squared score ($X^2$), a metric that quantifies the extent to which the observed frequencies for a categorical feature match the expected ones. The null hypothesis is that they are equal, i.e. that the categorical variable and the target one (the label) are independent. The p-value that is calculated in the test is the probability to obtain the observed frequencies under this null hypothesis of independence. Therefore, a lower p-value makes it easier to accept the alternative hypothesis of dependence and that the categorical feature has an influence on the label. The results of this first analysis allowed for a first screening of the features, aimed at reducing the risk of overfitting due to the few data points available in the survey.

#### 3.2.2.2 Model description[21]

Since the architecture of the Farmer-based Calibrated ABM derives from the one of the Farmer-based Toy-ABM, this section describes briefly only the few modifications implemented without reporting the entire ODD description. The main difference between the Toy ABM and the Farmer-based Calibrated ABM is the definition of the confidence factors. While the former expressed it only as a function of the education level of the farmers, the latter modelled it as a weighted sum of the proxies corresponding to the features selected as described in the previous section which, as reported in section 4.1.2.1, were *HighestEducationalDegree, PastureSurface, PercenRentedLand* and *LegalForm*. The first of these was already included in the Farmer-based Toy-ABM. The values for the other three proxies instead were retrieved from the AF survey data and added to the spreadsheet with farm data inputted to the model and are stored as attributes of the *Farm* entities during their initialization. In order to allow for its calibration, during the initialization the *model* receives the proxy weights as input and passes them to the *Farmers*, which directly calculate the confidence factor multiplying them for the relative proxies retrieved from its and its *Farm's* attributes and store it as an attribute, which is directly retrieved by the *AdoptablePasture* entities at the beginning of the "NPV adopting" submodel.

#### 3.2.2.3 Model calibration[22]

While in the Farmer-based Toy-ABM the values of the confidence factors for each level of education are arbitrarily predefined, the objective of the Farmer-based Calibrated ABM calibration was to determine the weights of the proxies included in the EDNPV calculations. The procedure consisted in an iterative grid search over the weights values, starting with a coarse discretization and halving it until no further

[21] SC: farmer_level_analysis\calibrated_abm
[22] SC: farmer_level_analysis\calibrated_abm\calibrated_abm calibration.ipynb

increase in the performance of the model could be observed. The performance metric chosen was the F1 score of the models predictions over the entire dataset. At each step, the analysis checked the distribution of F1 score and the trend with the runs, to avoid the risk of calibrating for a local minimum.

The grid search was initialized with values spanning from -1 to 1 for each weight and a step of 0.5. These boundaries were the ones considered significant, since the proxies are normalized and the values of the confidence factor that would make sense are between 0 and 1 (even though values larger than 1 can indicate an overestimation of SBP benefits). However, this bounding of the weight did not pose a hard constrain on the confidence factor values and moreover if the best value of any weights was found in the boundaries of the interval, values outside this interval were tested as well. This was a precise choice, stemmed from the consideration of the uncertainty in the NPV calculations that suggested to avoid using hard constrains.

#### 3.2.2.4 Model validation and output analysis[23]

The best model found, i.e. the one with the weights corresponding to the best ones found trough the calibration process, was validated over the dataset with real adoption with the same procedure described in section 3.2.1.1 for the Farmer-based Toy-ABM. In addition, the analysis of the model's output included the plot of the distribution of EDNPV and of confidence factors calculated, which in this model, differently than in the Farmer-based Toy-ABM, are different for each farmer depending on 4 variables.

### 3.2.3 Farmer-based Logistic Regression

The Farmer-based Logistic Regression was the last farmer-based approach implemented and consisted in the use of ML algorithms to model farmer's decision-making[24]. This is a data-driven approach, which avoided any theoretical assumption and in particular the one of farmers as profit maximizing agents, at the base of the previous two farmer-based models. The chosen model was logistic regression with penalty elastic net, which allowed to include both Lasso (l1) and l2 regularizations and vary their relative contribution. The inclusion of regularization parameters is used to reduce overfitting, since they add to the optimization function terms aiming at limiting the values of the attributes' weights. Lasso regularization is considered an embedded features selection method, as it tends to bring the weights of features bringing little information to 0. This allowed to tune the logistic regression with all the features of the survey reported in Table 5, excluding only not usable or obviously redundant ones. However, the work tested the logistic regression also with the reduced set of features used in the Farmer-based Calibrated ABM, to be able to compare the two and further reduce overfitting.

After standardizing the numerical features and encoding the categorical ones (one-hot encoding all the categorical apart from *HighestEducationalDegree*, which was ordinally encoded), the work implemented for both set of features a randomized grid search with 1000 iterations to find the best performing combination of its hyperparameters. Specifically, the values of *l1_ratio*, the mix of the two regularization

[23] SC: farmer_level_analysis\calibrated_abm\calibrated_abm best model analysis.ipynb
[24] SC: farmer_level_analysis\Logistic Regression.ipynb

parameters[25], were sampled from a uniform distribution from 0 to 1, and the ones of *C*, the inverse of the regularization strength, from a uniform distribution from 0.001 to 1. The grid searches used the average F1 score on the validation sets of 3-fold cross-validation (CV) as evaluation metric to select the hyperparameters of the best models.

The split of the data for CV was of particular concern due to the limited size of the dataset. Therefore, the randomized grid search was actually implemented within a nested procedure, running it for 30 different splits of the data stratified to maintain the same ratio of adopters and non-adopters in each fold as in the whole dataset. The final best model selected was the best performing one across the ones resulting from all the grid searches with different splits. This was tested for overfitting, comparing the average score on the CV validation sets with the score of the same model trained and tested on the entire dataset evaluating precision, recall and F1 score. At the macro-level, the validation considered the number of adopters predicted by the model. To study how the best models obtain their predictions, the analysis retrieved the features' weights after re-training them on the entire dataset. Lastly, the thesis used the same randomized grid search procedure to select the best hyperparameters without CV but through an overfitting procedure training and testing the models on the entire dataset, using only the features included in the Farmer-based Calibrated ABM. This additional analysis aimed at getting the same result of the Farmer-based Calibrated ABM through the logistic regression, in order to compare the set of weights obtained with and without the economic calculations.

## 3.3 Municipality-based approach

The second part of the analysis consisted of a municipality-based approach, where the municipalities, and not the farmers, are the agents whose behaviour was modelled. With the exclusion of the AF survey and economic data already used in the first part, the available data include information on adoption from 1996 to 2012 and regarding all farmers in Portugal. These time and spatial scopes provided the possibility to build a model able to simulate scenarios for SBP adoption at the country level. Moreover, the variety of data available in these sources had the potential to consider drivers of adoption unavailable in the AF survey and therefore provide different and complementary insights. However, because data for the period 1996-2008 was only available at an aggregated and inconsistent level, it was impossible to implement an individual agent model. The target variable of this approach was not categorical, as in the farmer-based approach, but continuous: the area of SBP adopted in each municipality in each year. Specifically, this area concerns only new installations, without considering resowing, for two reasons. First, since the lifespan of SBP is 10 years, if some of the area installed in the 12 years previous to the PCF project was resown it was probably a minimal fraction. Second, the PCF project only supported new installations and therefore did not include resown plots.

One ABM was developed under this approach, the Municipality-based Data-driven ABM. The decision to develop a data-driven ABM directly stemmed from the fact that municipalities, depicted as aggregated, representative agents in this approach, do not have a proper decision-making process that can be simulated. Their "behaviour" actually emerges from the combination of the behaviours of all the farmers

[25] A value of *l1_ratio* of 1 corresponds to using l1 regularization, a value of 0 to l2.

that are located within them and base the model on theoretical assumptions without a real-world counterpart could be problematic. Moreover, the lack of information on which factors guide farmers decisions required a thorough analysis of the data sources available, whose variety meant that any theory-driven approach trying to consider them all would need to be complicated and based on many assumptions. These considerations resulted in a model combining ML, to estimate SBP adoption in each municipality in each year (section 3.3.2), and ABM, to consider temporal and spatial dimensions and the reciprocal influence among agents. While the ABM provides the dynamic environment in which agents are located, their internal model, which is a ML model, senses all the information and outputs an estimation for the area of SBP adopted in the specific municipality in the specific year. The resulting Municipality-based Data-driven ABM is described in section 3.3.3.

## 3.3.1 Municipality-level data manipulation

Due to the variety of data sources available, the municipality-based approach started from their preparation. This consisted in the necessity of reporting them to the same level of geographical aggregation and then manipulating them to extract features considered significative. The data used in the municipality-based approach were:

- Geospatial data from the Portuguese municipalities shapefile introduced in section 3.1.1
- Socio-economic and agricultural data available in the census described in section 3.1.5
- Historic SBP adoption, obtained through the combination of the adoption previous to the PCF project (section 3.1.3) and the data collected during the PCF project (section 3.1.4)
- Environmental data, corresponding to climate and soil properties presented in sections 3.1.6 and 3.1.7

### 3.3.1.1 Spatial granularity harmonisation

The spatial scope of the analysis encompasses continental Portugal, therefore excluding the Autonomous Regions of Madeira and Azores, and also excluding the Algarve region (corresponding to the current Faro district)[26]. The exclusion of Algarve was necessary for the lack of census data for it, while the one of Madeira and the Azores for the lack of environmental data. However, this did not have a large influence on the analysis results, since these regions accounted respectively just for the 0.3% and 3.7% of area sown before 2009 and they did not present any adoption during the PCF project.

To merge the datasets for the remaining regions, the first issue to address was to handle the different spatial granularity that each presented[27]. Since the municipality level was the most common over the databases available and it corresponds to an official division of the Portuguese territory, easing the interpretation of the results, this was the spatial granularity adopted and thus the one which all datasets were reported to. This choice avoided the need to further aggregate the PCF project and census data, rich in information, while requiring the disaggregation of only the adoption previous to the PCF project.

[26] In the continuation of the thesis, Portugal will indicate continental Portugal excluding Algarve.
[27] SC: municipality_level_analysis\data_preparation\Spatial granularity harmonization.ipynb

From this, it followed the necessity of having the municipalities as main agents of the Municipality-based Data-driven ABM.

The first step to combine the datasets over the municipalities was to compare the municipalities of the three datasets referring to this geographical level and ensure their correspondence. The shapefile's column *Municipality* was taken as benchmark, since it was already known that the number of its entries corresponded to the one of the municipalities in Portugal. It was compared to the census data *pastures_area_munic* entry and the municipalities included in the PCF project data, solving any incongruence modifying the entries spelled in the less common way in the relative datasets as specified in Appendix F.

The mixture of spatial granularity of the data for SBP adoption previous to the PCF project required a case-by-case methodology to be disaggregated. The procedure consisted of three steps. The first was to link each entry of the Region column to a set of municipalities assumed it referred to. To do this, the analysis checked each *Region* value for correspondence with a municipality, district or bigger region. The original data obtained did not use a single, consistent classification, and the registered locations varied in size and administrative type. Values corresponding only to a municipality were directly linked to it. Values corresponding to both a Portuguese district and municipality were considered as districts, thus linking them to all the municipalities of the district apart from the ones already reported as separate entry. Values corresponding to bigger regions were linked to all the municipalities included in those, except for the ones already referred to previous entries. The second step was to choose the municipalities among which disaggregate the area adopted in each entry, among the ones already linked to it. The criterion chosen was to prioritize the municipalities that adopted SBP during the PCF project: if among the linked municipalities some adopted during the PCF project, the area was disaggregated among them. If none did, the area was disaggregated among all the linked municipalities. The third and last step consisted in actually disaggregating the area of SBP installed in each year among the chosen municipalities, based on their permanent pasture area (which was retrieved from the census data as described in section 3.1.5) using

$$SBP\ area_{municipality,\ y} = \frac{pasture\ area_{municipality}}{\sum_{\substack{i\ =\ municipalities\\ corresponding\ to\\ the\ pre-PCF\ region}} pasture\ area_{\ i}} * SBP\ area_{pre-PCF\ region,\ y}, \quad (4)$$

where *SBP area* indicates the adopted SBP area in the adoption previous to the PCF project dataset, *y* is any year for which it reports data and *pre-PCF region* is the value of the column *Region* to which the municipality corresponds. The remaining municipalities were assigned an area adopted of 0 hectares in each year.

To extract values at the municipality level from the raster databases for climate and soil data I used QGIS (https://www.qgis.org/en/site/), an open-source desktop GIS application, manipulating the datasets in different files. Another shapefile was created excluding the municipalities of the districts of Madeira, Algarve and Faro. This shapefile required to fix its geometries in order to correct polygons self-intersections, through the QGIS' "Fix geometries" tool. This file constituted the vectorial layer of the geographical manipulation, to which I clipped the selected climate variables (Table 3) and soil property

maps (Table 4) using the QGIS' "Clip Raster by Mask Layer" tool with the additional command-line parameter "-wo CUTLINE_ALL_TOUCHED=TRUE"[28], restricting the data to the pixels touching the area covered by the Portuguese municipalities. Previous to this passage, the climate variables data had to be reprojected using the QGIS' "Warp" tool to the coordinate reference system of the project, since it was not specified, and saved to a GeoTiff file to make the clipping process work. To calculate values referred to the municipalities' area the QGIS' "Zonal statistics" tool[29] was used. QGIS then allowed to extract the mean of each climate and soil variable over each municipality, for the climate ones for each year. The calculation of the mean of the yearly precipitation meant its value to represent the average quantity of precipitation per pixel area fell in one year in the municipality and not the total quantity in the municipality, generating a variable comparable over the different municipalities. All the calculated values were added to the relative municipalities shapefiles and exported from QGIS in ESRI shapefile format.

#### 3.3.1.2 Target variable retrieval: SBP adoption

The area of SBP sown over the period 1996 – 2012[30] was obtained merging thee SBP adoption previous to the PCF project, as modified in section 3.3.1.1 to refer to the single municipalities, and the PCF project database. While the former was already in the required format, the latter required some manipulation[31]. After dropping 13 entries which presented a null value for the size of the parcel, the dataset was grouped over the years of installation and the municipalities, summing the size of the parcels. This created a database reporting for each year and for each municipality the area of SBP sown. The municipalities that adopted during the PCF project resulted 73, while the area adopted in the remaining ones in the years 2009 – 2012 was set to 0.

Merging the two manipulated datasets resulted in a new dataset reporting the area of SBP installed for each municipality in Portugal and for each year from 1995 to 2012, which constituted the target variable of the Municipality-based Data-driven ABM. Since the adoption in the previous year was considered as a feature, a column of 0 values for 1995 was added as it was necessary to use for 1996. In fact, the adoption previous to 1996 can be assumed 0 (Terraprima, personal communication).

#### 3.3.1.3 Feature engineering and datasets aggregation

After the retrieval of the target variable, the analysis proceeded with the collection and manipulation, from the other databases, of various features considered significant to estimate the target variable retrieved in the previous section. Features regarding SBP adoption over time were directly obtained from the target variable values[32]. The first was simply the adoption in each municipality in the previous year. The second was the cumulative SBP adoption in each municipality, calculated in two different ways: the total cumulative adoption, adding all the SBP area installed since 1996 in the municipality,

[28] This parameter set to True allows to keep all the pixels overlapping to the shapefile, instead of only the ones whose centre falls in the shapefile.
[29] For how the QGIS' "Zonal statistics" tool calculates statistics and handles pixels only partially overlapping the polygons of the shapefile, see the following link: https://gis.stackexchange.com/questions/276794/how-does-qgis-zonal-statistics-handle-partially-overlapping-pixels.
[30] Due to the data available, the area from 1996 to 2008 is all the area that was sown regardless of the outcome, while the area from 2009 to 2012 the area that was successfully sown.
[31] SC: municipality_level_analysis\data_preparation\adoption\SBP adoption dataset creation.ipynb
[32] SC: municipality_level_analysis\data_preparation\adoption\SBP adoption dataset creation.ipynb

and the cumulative adoption over 10 years (the lifespan of well-managed SBP), adding for each year the area of pastures installed during the previous 10 years[33]. To check whether this disaggregation of the adoption prior to the PCF project originated unrealistically high and thus unreliable values of adoption, the code retrieved the municipalities which presented a total cumulative adoption larger than 100% and 50% of their pasture land and retroactively modified the disaggregation if necessary. Larger areas were also considered, with the adoption in the previous year, total cumulative adoption and cumulative adoption over 10 years in Portugal, obtained summing the respective databases over all the municipalities. The same values were calculated, for each municipality, over its neighbouring ones, and retrieved by initializing a simple ABM through the library mesa-geo (https://github.com/Corvince/mesa-geo)[34], an extension of mesa to develop ABM with GIS data. In order to avoid the size of the municipality to influence the results, the manipulation of the features regarding SBP adoption and of the target variable was completed by dividing them for the permanent pasture area in the region considered, retrieved for each municipality from the census data in section 3.1.5.

The shapefiles with climate and soil features generated in section 3.3.1.1 were checked for errors[35]. Features referring to precipitation and temperatures had to be divided respectively by 10 and 100, since their reprojection in QGIS caused them to be multiplied by these factors. Apart from this, all their variables were already prepared to be used as features. Additionally, the analysis also calculated the average values over the period 1995 – 2018 for each municipality.

Another important variable was the PES offered depending on the year of adoption. For the years previous to the PCF project, payments were not provided and therefore their value was set to 0 €/ha. For the years of the PCF project, the values were calculated as the sum of the payments offered for all the years following the installation, reported in Table 2. For the payments of the first year for farmers who adopted in 2011, as this analysis was unable to distinguish between who adopted during the first phase of the project and who did so in the second, the average of the two values was used.

The last set of features manipulated were the ones extracted from the census (reported in Table A.4), with the aim of understanding if their number could be reduced without losing too much information. The results of this manipulation and the specifications of which of the original variables were kept are reported in Appendix C.

The target variable and all the features extracted (summarized in Table A.10) from the datasets were aggregated in a unique merged dataset[36]. The structure of this dataset had to correspond to the desired output of the Municipality-based Data-driven ABM and therefore each of its rows was referred to one municipality in a certain year, for a total of 4,403 instances. The merging process therefore reported all the features to this structure, considering their dependence on the spatial dimension (the municipalities), the temporal dimension (the years) or both and reporting values multiple times where necessary.

[33] For years up to 2006, since no data were available for the previous 10 years I considered all the previous years from 1996.
[34] SC: municipality_level_analysis\data_preparation\adoption\neighbouring_municipalities
[35] SC: municipality_level_analysis\data_preparation\environmental\Shapefiles with soil and climate data manipulation after QGIS.ipynb
[36] SC: municipality_level_analysis\data_preparation\Municipalities adoption analysis - Final merging.ipynb

### 3.3.2 Agents' internal model for the estimation of individual municipalities adoption

This section describes the procedure to build the internal model of the main ABM agents, the municipalities, i.e. the model which, getting as inputs all the required variables for a specific municipality and a specific year, outputs the area of SBP installed. This model was required before the developing of the ABM, since from the architecture of the internal model depended the one of the ABM, especially in regard to the variables that each agent had to be able to access. The choice of using a data-driven approach meant that the agents' internal model had to be learnt entirely from the available data and for this the work relied on ML algorithms. The work resorted in particular to a double-hurdle model, firstly introduced by Cragg [110]. The idea behind this type of model is to split the extent of participation of an individual in an activity into two stages, with the first regarding the decision to take part or not in it and the second concerning the extent of participation of the individuals who decided to get involved. In the Municipality-based Data-driven ABM, and therefore in the internal model of its agents, the first stage regards if SBP are adopted in the municipality or not, while the second refers to the hectares of area adopted in the municipalities that adopted. These two stages can be guided by different processes and the use of a double-hurdle model allowed to study independently the influence of the various factors on both. Moreover, a double-hurdle model can facilitate estimations for datasets where the target variable is continuous but with many null values: the adoption in the municipalities in each year, the target variable of the model, presents around half null values.

The use of a double-hurdle approach implied the need of selecting and training two ML models: a classifier, to predict if SBP is adopted or not in a certain municipality and in a certain year, and a regressor, to estimate how much SBP area farmers sown in the municipalities where they do. This section describes the method followed, with a first analysis of the aggregated dataset (section 3.3.2.1) and the deployment of ML techniques to develop the final models (section 3.3.2.2).

#### 3.3.2.1 Dataset analysis, features screening and outliers detection

Both stages of the double-hurdle model were based on the merged dataset presented in section 3.3.1.3, which was however modified to suit the objective of each. The first stage required a categorical target variable, therefore its values were left as 0 if no SBP was adopted in the municipality in the year and encoded as 1 otherwise. The target variable was instead already in the right format for the second stage, which however regarded only municipalities and years that saw adoption. Therefore, the entries with a null value of the target variable were removed.

A first analysis of the datasets was then necessary in order to both get an understanding of their characteristics and perform a first screening of the large number of features they included. Reducing it could help at the same time to decrease overfitting and increase the significance of the results, due to the many variables included in the datasets and the multicollinearity and redundancy they presented, due to the similarity of the included ones. The same analysis was conducted for both datasets[37].

[37] SC: municipality_level_analysis\ml_models\Classification - First exploration and features analysis.ipynb and SC: municipality_level_analysis\ml_models\Regression - First exploration and features analysis.ipynb.

The first step consisted of a correlation analysis, using the Spearman ρ coefficient to quantify monotonic relations between the features and the label. This analysis also included the calculation of the same coefficients based only on the instances referring to the years of the PCF project, from 2009 to 2012, the ones for which the dataset includes values of the target variable which are sure and not disaggregated. Comparing the correlations over all years with the ones only during the PCF project provided more insights and helped avoiding the exclusion of important features.

The work continued with the implementation of a multicollinearity analysis to reduce existing correlations among the features, which could influence their significance in the ML models and therefore hamper the interpretation of their importance. A common tool to quantify multicollinearity is the calculation of variance inflation factors (VIFs) for the various predictors [111]. Correlation matrices helped to identify high correlations among features. When these occurred, only the most correlated with the target variable were kept except when other particular reasons specified in the results emerged, until all VIFs presented a value below 10. If the orderings of the correlation scores calculated on all instances and only on the instances referring to the PCF project years were discordant, the average of these two values was calculated and the feature with the highest absolute value was kept. The average of the scores was used also to keep the same features for both the classification and regression and ease the comparison of the results of the two stages, when this did not mean to keep a feature with a significantly low score (in absolute value) in one of the two cases. The last criterion used for selection was to keep only one feature for each category of census data (which is identified by the prefixes of their names).

Additionally, the code checked the dataset for outliers in respect to the target variable and the other features concerning adoption[38]. Since their values were obtained from the disaggregation of the adoption prior to the PCF project and this prioritized the municipalities that also adopted during the PCF project, some municipalities could have been assigned an unrealistic estimated adoption that is too high in some years, resulting in an unrealistically high cumulative adoption as well. Boxplots[39] were used to identify a threshold value for each feature visually separating larger values than the majority and the relative instances were removed. At the same time the procedure also aimed at checking the existence of a second degree dependence of the target variable from the cumulative adoption in the municipality. In fact, a firstly growing and then decreasing adoption of SBP with the cumulative adoption in the municipality would reflect an initial increase of interest due to spatial influence followed by a saturation of the pastures area suited for SBP, which could be the reason for the logistic trend in Figure 2. This was also the expected logistic trend due to innovation diffusion theory. This was implemented fitting a second degree polynomial for the dataset created for the regression stage of *tot_cumul_adoption_pr_y_munic* with *adoption_pr_y_munic* as dependent variable. The result was assessed calculating the $R^2$ score of the resulting polynomial function and plotting the polynomial curve obtained over the data points.

[38] The source code for the analysis in this paragraph can be found in the Notebook for the regression stage, which path was specified in footnote 37.
[39] Not reported in the results but available in the relative notebook.

#### 3.3.2.2 ML models first selection and training

Having created the datasets with the final set of features, the work proceeded evaluating various classifiers and regressors to choose the ones to further test in combination with the ABM[40]. The models tested, both for classification and regression, were linear and polynomial regression (logistic for classification) with elastic net regularization, linear and nonlinear support vector machines, decision tree, random forest, extremely randomized trees and gradient boosting with decision tree as base classifier[41]. In addition, the linear model was also tested with the addition of the square of the feature *tot_cumul_adoption_pr_y_munic*, to evaluate a quadratic relation of the adoption in the year with it.

After a first quick exploration to get descriptive statistics and study the distribution of the features and of the target variable, the analysis standardized the features of the dataset and then tested the models through one or more rounds of grid and randomized search on the most important hyperparameters[42], to get a first idea of the potential of the different architectures and thus perform a first selection of the most promising ones. The searches evaluated the different hyperparameters combination through the average validation score of 3-fold CV on the entire datasets, stratifying the split of the dataset to maintain the ratio of instances referring to each year in all the folds, important due to the PES being non-zero only in the last 4 years. The performance metrics used were logistic loss for classification, since the ABM used probabilities of adoption, and root mean squared error (RMSE) for regression. The code also calculated the adjusted $R^2$ for regression and the area under the curve of the receiver operating characteristics (ROC AUC) curve [112] for classification[43]. The same scores were compared to the one obtained training and evaluating the models on the entire dataset, to quantify their generalization error. Stochastic models, as the ones based on decision trees, where only evaluated once, since the aim in this stage was only to select some and their stochasticity was considered when combined with the ABM. The performance metrics of the regression models were also compared with the ones of a “dumb” estimator, predicting for all data points an adoption equal to the average value of the labels.

The hyperparameters of the models with the best performances were further tuned through Bayesian search, which contrary to grid and randomized search takes into consideration past evaluations to choose the next hyperparameters to test [113], using the optuna library (https://optuna.org) but the same performance metrics and CV procedure. The scores obtained allowed to select the best classifier and regressor to test with the ABM. If models based on different architecture were performing similarly, all were kept for the following analysis.

[40] SCs: municipality_level_analysis\ml_models\ Classification - Models evaluation.ipynb and municipality_level_analysis\ ml_models\Regression - Models evaluation.ipynb.
[41] Grandient boosting was implemented with the xgboost library (https://xgboost.readthedocs.io/en/latest/) using early stopping combined with plots of the validation error to estimate a proper number of estimators.
[42] The exact hyperparameters and relative values tested for each model can be found in the notebooks.
[43] The notebook presenting the selection of the classification model also reports the plot of the precision-recall and ROC curves for all models tested.

### 3.3.3 Municipality-based Data-driven ABM

#### 3.3.3.1 ABM description

A proper description of the Municipality-based Data-driven ABM through the ODD protocol was not considered necessary and is therefore not reported[44]. Contrary to the Farmer-based Toy-ABM in fact, the development of this data-driven model did not aim at obtaining more insights on the system but only at being the dynamic environment providing to the main agents, the municipalities, the variables required to estimate their adoption in each year. These variables correspond to the features which the classification and regression stage of the double-hurdle ML model were trained with, since these constitute the internal model of the municipality agents. The need for a proper simulation was therefore due to the fact that the variables regarding SBP adoption are endogenous to the ABM, i.e. their value depend on the outcome of the previous steps.

The ABM encompasses all the municipalities in Portugal included in the analysis and run in discrete time, with each step corresponding to one agricultural year. The initial year for the simulation, which has to be between 1996 and 2012, has to be specified when initializing the ABM. At the beginning of each step, the municipalities retrieve all the variables needed by their internal model. The ones regarding their own historical SBP adoption are directly stored as their attributes, the ones regarding adoption in Portugal as attributes of the model and the ones regarding adoption in the neighbouring municipalities are calculated at each step. All the other variables used to trained the ML models, whose values do not change with the year and therefore with the simulation step, are instead stored as attributes of various entities of the model (similar to the *Government* and *Market* of the Farmer-based Toy-ABM) and can be directly accessed by the municipalities. All these data, included in the merged dataset described in section 3.3.1.3, are provided to the model during its initialization. The order of activation of the municipalities does not influence the results, since their adoption variables and the ones relative to Portugal are updated only at the end of each model step, when all already estimated the SBP adoption in the current year. This synchronous updating of the agents is done to represent the time lag which information on pastures adoption within and across municipalities spreads with and the fact that the success or not of the adoption of SBP is visible only the following year. In practice, this implies that farmers get to know about the decisions taken by their peers during the year and that in autumn, at the moment of deciding if to adopt SBP or not, they only know and consider the diffusion of SBP in the year before, while ignoring the decisions taken in the current year by others.

The estimation of SBP area installed in the year happens in two steps, which correspond to the two stages of the double-hurdle model. First, the probability calculated by the classification stage is used for a probabilistic decision on if there is adoption in the municipality in the year. This is implemented through the generation of a random number between 0 and 1: if the number generated is smaller that the probability estimated, there is adoption in the municipality; otherwise no. Second, if there is adoption in the municipality the regression stage estimates the hectares that are installed in the year. If this estimation is negative, the value is set to 0. The probabilistic evaluation of the presence of adoption with

[44] The detailed implementation of the model can be found in the relative folder of the repository, which contains all the SC: municipality_level_analysis\municipalities_abm.

the classification stage causes the outcome of the ABM to be stochastic. Another stochastic element is present in the ABM if the ML models training process is stochastic, since these are trained at the beginning on the simulation on the relative datasets.

At the end of the simulation, three different datasets are assembled, with values reported for each year for which it ran: the yearly adoption in each municipality, the yearly adoption in Portugal and the yearly cumulative adoption in Portugal from 1996.

#### 3.3.3.2 ML models definitive selection and analysis

The first step after having developed the Municipality-based Data-driven ABM was to definitively choose the classifier and the regressor constituting the internal model of the main agents, the municipalities. Due to the stochasticity of the ABM, the simulation was run 100 times from 1996 to 2020 for each combination of classifiers and regressors previously selected (section 3.3.2.2). The selection of the final combination of ML models was based on two different evaluations at the aggregated level, based on the yearly area of SBP installed in Portugal. The ABMs estimations from 1996 to 2012 could be evaluated over observed data, through the adjusted $R^2$ score calculated comparing the ABM estimation with the SBP adoption extracted from the available datasets. The estimations regarding the years from 2013 to 2020 instead could not be compared with empirical data. Therefore, the trend of yearly SBP area sown in Portugal estimated by the ABMs (averaged over the 100 runs) was plotted to qualitatively assess its meaningfulness and the ability of the ABM to extrapolate outside years it was trained on[45].

The chosen classifier and regressor were firstly analysed in terms of underfitting or overfitting, plotting their learning curves, i.e. the training and validation errors obtained splitting the dataset and increasing the size of the training set. Then, the influence of the various features on their estimations and consequently on SBP adoption was analysed[46]. Specifically, the SHAP (SHapley Additive exPlanations) framework [114] was implemented through the Python SHAP library (https://shap.readthedocs.io/en/latest/). After splitting the dataset into a train and validation set, the Shap package allowed to plot a summary graph representing the features in order of importance and the SHAP values of each feature for each instance of the validation set, which quantifies how much the feature value influenced the prediction of the label for the specific instance. The SHAP library also included the possibility of plotting the SHAP values against the features values to better assess their individual effect[47].

[45] The 95% confidence interval for the mean of the estimation using the Student's t-distribution was also calculated but in the end omitted from the graphs, since not visible, while the lines corresponding to the various runs were preferred.
[46] SC: municipality_level_analysis\ml_models\Classification - Model analysis.ipynb and municipality_level_analysis\ml_models\Regression - Model analysis.ipynb
[47] In the results, only the graphs for the total cumulative adoption in the municipality are reported, to evaluate if the ML models learnt the second degree relation with it of the adoption as explained in section 3.3.2.1. Other graphs can be found in the notebooks.

#### 3.3.3.3 ABM validation and output analysis

The Municipality-based Data-driven ABM, with the ML models selected in the previous section, was validated against the observed[48] adoption of SBP in the years 1996 – 2012, both quantitatively and qualitatively and at the macro and micro level. The validation was based on the average outcome of 100 iterations of the ABM, to address its stochasticity[49].

At the macro-level, the adjusted $R^2$ score and the yearly cumulative adoption in Portugal had already been examined as specified in the previous section. In addition to the yearly adoption, the cumulative one in Portugal for the period 2009 – 2020 was plotted to be compared with the observed one. To assess the predictions of the classification and the regression stages of the agents' internal model separately, also the estimated yearly number of municipalities with adoption and their average adoption in each year were plotted, with also the observed relative values. The micro-level validation compared the estimated and observed adoption in each year in each municipality. Quantitatively, the adjusted $R^2$ and RMSE were calculated. Finally, to assess if the model could reproduce the spatial pattern of adoption, two maps were plotted, reporting the estimated and observed total area installed in each municipality until 2012.

### 3.3.4 Quantification of additional carbon sequestered thanks to the Portuguese Carbon Fund project

The Municipality-based Data-driven ABM, encompassing all the regions of Portugal where SBP were adopted during the PCF project, enabled the quantification of the amount of $CO_2$ that was and will be sequestered over their lifetime by the additional pastures installed thanks to the payments provided. This required an estimation of how much area of SBP would have been installed in the years after 2008 if no PCF project had taken place. The yearly adoption in Portugal was obtained through a counterfactual simulation, consisting in running the Municipality-based Data-driven ABM over the years from 2009 to 2020[50] with a value of the payments offered to install SBP of 0 in every year[51]. Its estimation from 2009 to 2012 was subtracted year-by-year from the observed adoption during the PCF project and summed over the years, giving the estimation of the additional area installed in Portugal in the years 2009 – 2012 thanks to the PCF project. Multiplying this area for the sum of the C sequestration factors for SBP for the 10 years after installations, retrieved from [115], gave the differential amount of C sequestered thanks to this installed area. To evaluate the residual effect that the PCF project had on adoption after its conclusion, the yearly area installed in Portugal from 2013 to 2020 by the Municipality-based Data-driven ABM with the ML models run with no payments provided was subtracted year-by-year from the adoption obtained running over the same years the Municipality-based Data-driven ABM with the actual value of the payments (the same model validated in section 3.3.3.3). Summing the differential yearly

[48] The term *observed* will be used to refer to the adoption data extracted from the datasets available, despite the need to disaggregate the data previous to the PCF project implies that this may not correspond to the real adoption.
[49] SC: municipality_level_analysis\municipalities_abm\model_validation\municipalities_abm validation - multiple runs.ipynb
[50] The simulation was started in 2009 to avoid its results being influenced by the error on the model estimations in the years 1996 – 2008. All the estimations in this paragraph refer to the average result of 100 runs of the ABMs.
[51] SC: municipality_level_analysis\municipalities_abm\carbon_sequestration\carbon_sequestration_PCF_project.ipynb

adoptions and multiplying them for the sum of the C sequestration factors provided the differential area installed and the consequent C sequestered after the conclusion of the PCF project that can attributed to the fact that the project took place. Finally, the sum of the differential areas installed and relative C sequestered from 2009 to 2012 and from 2013 to 2020 provided the total effect of the PCF project until 2020. The real, estimated with payments and estimated without payments yearly adoption in Portugal were plotted on the same graph to present visually these calculations.

# 4 Results

## 4.1 Farmer-based approach

### 4.1.1 Farmer-based Toy-ABM

#### 4.1.1.1 Pastures costs and revenues

This section reports the calculated costs and yearly cash flows over 10 years for adopting SBP and maintaining SNP as of 2009. $CPI_{2019 \rightarrow 2009}$ has a value of 1.12. Table 6 reports the costs for the installation and maintenance of 1 hectare of SBP in 2009. The total cost for the installation of SBP in 2009 was 678.85 €/ha. Maintenance cost for SNP resulted in 21.11 €/ha, corresponding to 23.68 €/ha in 2009 after being adjusted for inflation. Regarding feed costs, FS resulted equal to 796 kg/ha.y. The price for low silage feed formulation is of 0.261 €/kg and the yearly savings in feed per hectare of SNP switched to SBP amounted to 208.10 €/ha.y. Table 7 reports the final differential yearly cash flows for installation and maintenance of the pastures and feed purchase for one hectare of SBP and NP.

*Table 6: total costs for individual operations (CT) and aggregated by type of operation (CAT) required for installation and maintenance of 1 hectare of SBP in 2009.*

| | Operation | CT [€] | CAT [€/ha] | Operation | CT [€] | COT [€/ha] |
|---|---|---|---|---|---|---|
| **Maintenance** | **Soil correction** | | **171.19** | **Fertilization** | | **127.66** |
| | Lime | 131.00 | | Fertilizers transport | 10.86 | |
| | Lime transport | 37.39 | | Superphosphate | 82.00 | |
| | Lime application | 2.80 | | Operation | 34.80 | |
| **Installation** | **Soil preparation** | | **195.62** | **Sowing** | | **168.14** |
| | Harrowing | 124.81 | | Seeds Fertiprado AC70 | 107.50 | |
| | Scrolling | 70.81 | | Seeder | 60.64 | |
| | **Soil correction** | | **171.19** | **Fertilization** | | **143.90** |
| | Lime | 131.00 | | Fertilizers transport | 10.86 | |
| | Lime transport | 37.39 | | Operation | 82.00 | |
| | Lime application | 2.80 | | Superphosphate | 34.80 | |
| | | | | Borax | 3.50 | |
| | | | | Zinc Sulphate | 0.00 | |

*Table 7: total yearly cash flows for installation and maintenance of the pasture and feed purchase for one hectare of sown biodiverse pastures (SBP) and semin-natural pastures (SNP), in €/ha.y, for a decision made in 2009.*

| Year | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| **SBP** | -722.97 | 0.00 | -135.96 | 0.00 | -318.27 | 0.00 | -135.96 | 0.00 | -318.27 | 0.00 |
| **NP** | 0.00 | -235.40 | -210.18 | -210.18 | -210.18 | -210.18 | -235.40 | -210.18 | -210.18 | -210.18 |

#### 4.1.1.1 Model validation and output analysis

Regarding micro-validation, the model reached a precision of 0.59, recall of 0.94 and F1 score of 0.72. At the macro-level, the model predicted 27 out of 30 farmers (90%) switching to SBP while 17 (56.7%)

adopted SBP in the real data. Table 8 reports the EDNPV for adoption of SBP calculated by the *Farmer* agents in the model, depending on their education level.

*Table 8: expected differential net present value (EDNPV) in €/ha.y for sown biodiverse pastures (SBP) adoption calculated by the Farmer agents in the Farmer-based Toy-agent-based model, depending on their education level.*

| Farmer education | EDNPV SBP |
|---|---|
| Primary | -243.05 |
| Secondary | -76.11 |
| Undergraduate / Graduate | 90.84 |

## 4.1.2 Farmer-based Calibrated ABM

### 4.1.2.1 AF survey data analysis and features screening

Table 20 reports the results of the Spearman ρ coefficients between the numerical features created from the AF survey data and the target variable *Adoption SBP*. Table 10 reports the results of the Chi-Squared test for the categorical features, i.e. the $X^2$ score (with 5 degrees of freedom and a sample size of 30) and the p-values under the null hypothesis that the label *AdoptedSBP* and each feature are independent. All the p-values are over 0.25, meaning that a test with a significance level of 5% would fail on each. Actually, the null hypothesis cannot be rejected for any categorical feature for a significance level lower than 27%, a value too high to enable concluding that any feature has a role in predicting adoption.

*Table 9: Spearman ρ correlation coefficients between the numerical features obtained from the Animal Future survey data and the target variable, i.e. the decision to adopt sown biodiverse pastures or not.*

| **Feature** | *PastureSurface* | *PercentRentedLand* | *CattlePercentage* | *FarmerSince* |
|---|---|---|---|---|
| **Spearman ρ** | 0.14 | -0.21 | -0.05 | 0.02 |

*Table 10: $X^2$ and p-values for the Chi-Squared test on the categorical features obtained from the Animal Future survey data, under the null hypothesis that adoption and each feature are independent.*

| **Feature** | *Distrito* | *Concelho* | *Legal Form* | *Highest Educational Degree* | *Highest Agricultural Educational Degree* | *Expectation Family Succession* |
|---|---|---|---|---|---|---|
| **$X^2$** | 5.13 | 19.68 | 3.83 | 1.47 | 1.05 | 0.02 |
| **p-value** | 0.27 | 0.41 | 0.28 | 0.70 | 0.59 | 0.90 |

These results enabled a first screening of the features, reducing the ones that actually were used in the Farmer-based Calibrated ABM. The features that remained were:

- *PastureSurface* and *PercenRentedLand*: exhibited the highest correlation in absolute value.
- *HighestEducationalDegree*: being already used in the Farmer-based Toy-ABM as a proxy for farmers' risk perception, an additional evaluation of its importance was considered necessary.
- *LegalForm*: it exhibited a relatively low p-value in the Chi-Squared test, meaning that a relation between it and SBP adoption is more likely than for other features with higher p-value.

The following were instead the ones discarded:

- *Concelho:* looking at its distribution in Figure A.2 it is clear that including this feature would not give any insight, since almost every farm has a different value for it.

- *Distrito:* this feature had interesting results in the Chi-Squared test, with the lowest p-value. However, being a not ordinal categorical variable, it would generate 5 different dummy features in the model when encoded, one for each value it has, and for each there would be a small sample of farmers. It would be more interesting to transform this feature in a numerical one, representing for example the adoption level in the district of the farm.
- *HighestAgriculturalEducationalDegree*: in the Chi-Squared test exhibited just slightly better values that the *HighestEducationalDegree* feature. However, the latter also includes a distinction between primary and secondary education and it was the feature used in the Farmer-based Toy-ABM.
- *CattlePercentage* and *FarmerSince*: exhibited a low and insignificant correlation with adoption.
- *ExpectationFamilySuccession*: had the highest p-value in the Chi-Squared test, meaning that its independence of SBP adoption is highly likely.

#### 4.1.2.2 Model calibration

The calibration process ran for 4 iterations. Already the third one showed no improvement of the F1 score and found various weights combinations corresponding to the best score. However, due to its still relatively coarse step of 0.125, a fourth round was performed to confirm that no improvement was possible with a step of 0.05 and with boundaries defined to include all the best combinations of the previous step.

*Table 11: values of proxy weights tested and results for each iteration of the Farmer-based Calibrated agent-based model calibration.*

| Iteration | Step | Lower and upper boundaries for proxy weight | | | | Best set of weights (ordered as in the previous 4 columns) | Best F1 score |
|---|---|---|---|---|---|---|---|
| | | *Highest Educational Degree* | *Pasture Surface* | *LegalForm* | *Percent RentedLand* | | |
| **1st** | 0.5 | -1.0, 1.0 | -1.0, 1.0 | -1.0, 1.0 | -1.0, 1.0 | 1.0, 0.5, 0.0, 0.0 | 0.70 |
| **2nd** | 0.25 | 0.5, 1.5 | 0.0, 1.0 | -0.5, 0.5 | -0.5, 0.5 | 1.25, 0.75, -0.5, 0.25 | 0.80 |
| **3rd** | 0.125 | 1.0, 1.5 | 0.5, 1.0 | -0.75, -0.25 | 0., 0.5 | 1.25, 0.75, -0.5, 0.25* | 0.80 |
| **4th** | 0.05 | 1.25, 1.5 | 0.5, 0.875 | -0.75, -0.50 | 0.25, 0.375 | 1.25, 0.75, -0.5, 0.25* | 0.80 |

**These combinations are just representative, since different ones gave the same result. However, in all the cases the signs of the weights were the same as in the combinations presented.*

#### 4.1.2.3 Model validation

Since all the best sets of weights found return the same result on the dataset, any of these could be chosen for the model validation. The values selected were 1.25, 0.75, -0.5 and 0.25 respectively for *HighestEducationalDegree, PastureSurface, LegalForm* and *Percent RentedLand* features.

Regarding micro-validation, the model reached a precision of 0.70, recall of 0.94 and F1 score of 0.80. At the macro-level, the model predicted 23 out of 30 farmers (76.7%) switching to SBP. The distribution of EDNPVs resulting is plotted in Figure 4. The confidence factors present the same distribution, ranging between -0.20 and 1.97. Differently than the validation performance measures, these values depend on the particular set of weights chosen.

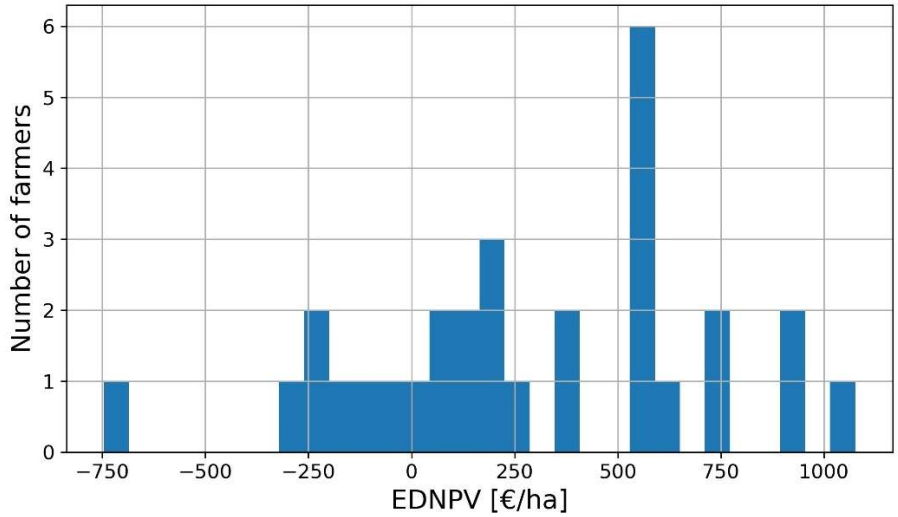



*Figure 4: distribution of expected differential net present values (EDNPV) resulting from the Farmer-based Calibrated agent-based model after calibration.*

### 4.1.3 Farmer-based Logistic Regression

The features excluded from any logistic regression were *Concelho*, for the high number of categories with just one instance (Figure A.2), and *HighestAgriculturalEducationalDegree,* for its redundancy with *HighestEducationalDegree*. Table 12 reports the performance metrics and Table 13 the hyperparameters and the intercept and coefficients after training on the whole dataset for the best logistic regressions found through the grid searches with the different procedures used.

*Table 12: evaluation metrics of the best Farmer-based Logistic Regressions found through grid search, for each procedure used.*

| Procedure | Validation scores | | | Training scores | | | % of adopters predicted |
|---|---|---|---|---|---|---|---|
| | Pr | Re | F1 | Pr | Re | F1 | |
| **All features** | 0.59 | 1.00 | **0.74** | 0.59 | 0.94 | **0.73** | 90.0 |
| **Reduced features** | 0.65 | 0.89 | **0.75** | 0.61 | 0.82 | **0.70** | 76.7 |
| **Selected through OP[52]** | *0.53* | *0.89* | ***0.65*** | 0.61 | 1.00 | **0.76** | 93.3 |

*Validation scores are the average scores on the validation sets of cross-validation, while training scores refer to the score obtained training and testing the model on the entire dataset. Validation scores are the average scores on the validation sets of cross-validation, while training scores refer to the score obtained training and testing the model on the entire dataset. Pr – precision; Re – recall; F1 – F1 score.*

[52] Since this model was not selected through CV, the CV scores were calculated only after having chosen it through an overfitting procedure to compare with the other.

*Table 13: hyperparameters, intercept and feature coefficients of the best Farmer-based Logistic Regressions found with the different procedures used, after training them on the entire dataset.*

| | All features | Reduced features | Selected through OP |
|---|---|---|---|
| **Hyperparameters** | | | |
| ***l1_ratio*** | 0.039 | 0.169 | 0.318 |
| ***C*** | 0.077 | 0.678 | 0.208 |
| **Intercept and features' coefficients** | | | |
| **Intercept** | 0.153 | 0.013 | 0.25 |
| ***HighestEducationalDegree*** | 0.059 | 0.155 | 0.068 |
| ***PastureSurface*** | 0.0814 | 0.154 | 0.064 |
| ***LegalForm*** | -0.060 | -0.348 | 0.0 |
| ***PercentRentedLand*** | -0.107 | -0.235 | -0.118 |
| ***CattlePercentage*** | -0.003 | - | - |
| ***ExpectationFamilySuccession*** | 0.005 | - | - |
| ***Beja*** | -0.071 | - | - |
| ***Portalegre*** | 0.090 | - | - |

*OP – overfitting procedure, training and testing on the entire dataset. The features FarmerSince, Santarém, Setùbal and Évora were omitted from the table since used only in the logistic regression trained with all the features and the relative coefficients obtained were 0.*

# 4.2 Municipality-based approach

## 4.2.1 Agents' internal model for the estimation of individual municipalities adoption

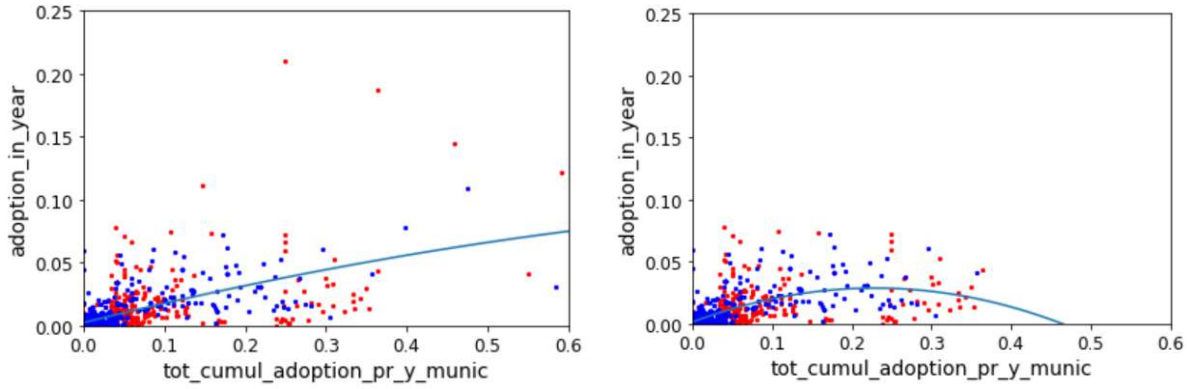
### 4.2.1.1 Dataset analysis, features screening and outliers detection

The first analysis and features screening reduced the number of variables used in the following analysis to 24 for the classification and 25 for the regression. The features kept are reported in Table A.11, together with their correlation with the target variable (considering all the years or only the PCF years) and their VIF values after having excluded the other variables, for both the classification and the regression problems[53]. While the rest of features selected were the same for both classification and regression, the only difference was the exclusion from the classification of the *sbp_payment* variable, due to its unrealistic negative correlation of -0.34 with the decision of adopting. A decision that was not based solely on correlation scores was the one to keep adoption features related to the total cumulative adoption instead of to the one only in the previous 10 years. Despite features referred to the latter showed, generally, slightly higher Spearman ρ (up to 0.02 of difference), the total cumulative adoption was considered more relevant due to the fact that the model target variable concerns new installations, for which knowing that some area after 10 years could be re-sown is irrelevant. A similar rationale is

[53] The values of these metrics for the variables excluded, together with the correlation matrices reporting the correlations among features from the same datasets, can instead be found in the section titled "Custom transformers to reduce features and multicollinearity" of the Notebooks with path municipality_level_analysis\ml_models\Classification - First exploration and features analysis.ipynb and municipality_level_analysis\ml_models\Regression - First exploration and features analysis.ipynb.

valid for aggregated census features. Being able to give an information on all the population with a single number was preferred over single categories when exhibiting a slightly lower correlation.

The outliers removed were the instances with a value of *adoption_in_year* or *adoption_pr_y_munic* over 0.1 or a value of *tot_cumul_adoption_pr_y_munic* over 0.39, for a total of 13 data points in the regression dataset and 17 in the classification one. The $R^2$ scores of the fitted polynomial before and after outliers removal were respectively of 0.40 and 0.42. Figure 5 reports the fitted polynomials.




*(a)* *(b)*

*Figure 5: data points of the dataset for regression plotted based on their value of adoption_in_year and tot_cumul_adoption_pr_y_munic with the fitted second degree polynomial degree plotted in blue, before outliers removal (a) and after (b). Blue dots are features referred to years before 2009 and red after.*

#### 4.2.1.1 ML models first selection and training

The final dataset for classification included 4,385 data points. The first analysis showed that the target variable *adoption_in_year* had 2,245 non-null values of 2,140 null ones, therefore making the dataset for classification balanced in terms of positive and negative classes. The final dataset for regression presented 2,245 instances, the ones with non-null value of the label, with a mean value of the target variable of 0.00615.

The results of the first round of models testing are reported in Table A.12. For comparison, the "dumb" regressor predicting the mean of the target variables for all instances presented an RMSE of 0.00965. The classifiers selected for further hyperparameter tuning were nonlinear support vector machines, random forest and gradient boosting, while the regressors were nonlinear support vector machines, random forest and extremely randomized trees. The results of the second round of hyperparameter tuning is available in Table A.12 as well. Based on these results, the two classifiers selected to be tested with the ABM and the relative hyperparameters were nonlinear support vector machine with *C* 8.500, *gamma* 0.053 and *kernel* 'rbf', and gradient boosting with *learning_rate* 0.115, *subsample* 0.932, *min_child_weight* 1, *max_depth* 9, *gamma* 1, *colsample_bytree* 0.251, *reg_lambda* 0.382 and *reg_alpha* 0.056. The two regressors selected were nonlinear support vector machines with *C* 0.043, *gamma* 0.058, *epsilon* 0.002 and *kernel* 'rbf' and extremely randomized trees with *n_estimators* 73, *max_depth* 19, *min_samples_leaf* 2, *min_samples_split* 5 and *max_features* 0.8289[54].

[54] Additional hyperparameter not specified here were left with the default value of the class that was used of the relative Python library used, where are reported also the meaning of each. The best hyperparameter combinations for the models that were not selected for the second round of tuning can be found in the notebooks.

### 4.2.2 Municipality-based Data-driven model

#### 4.2.2.1 ML models definitive selection and analysis

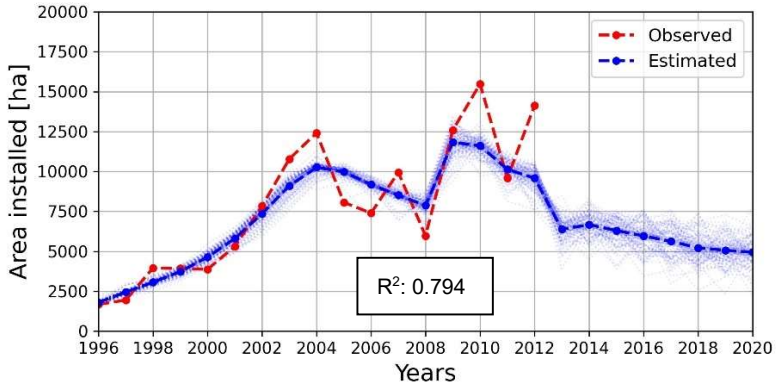
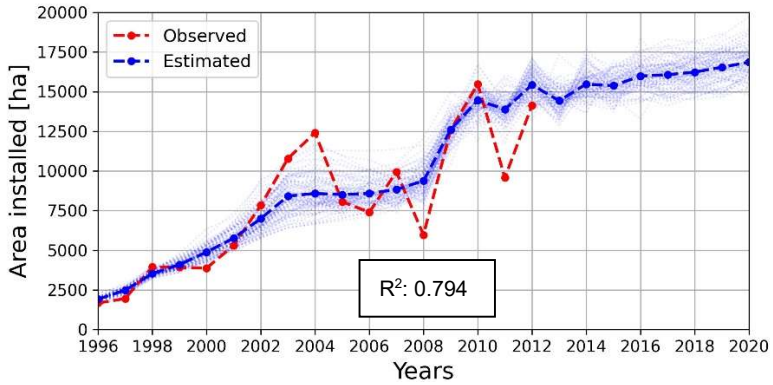
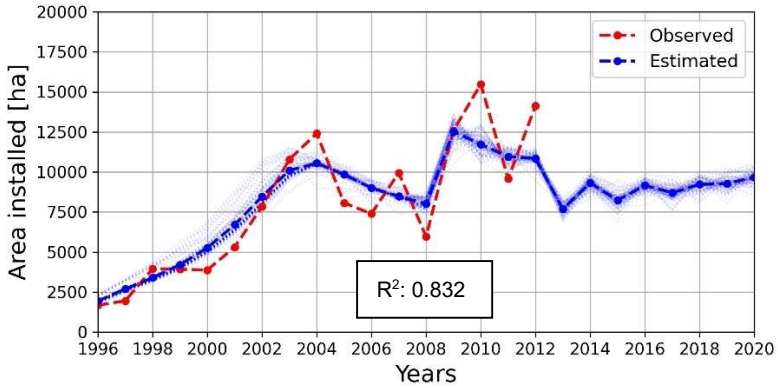
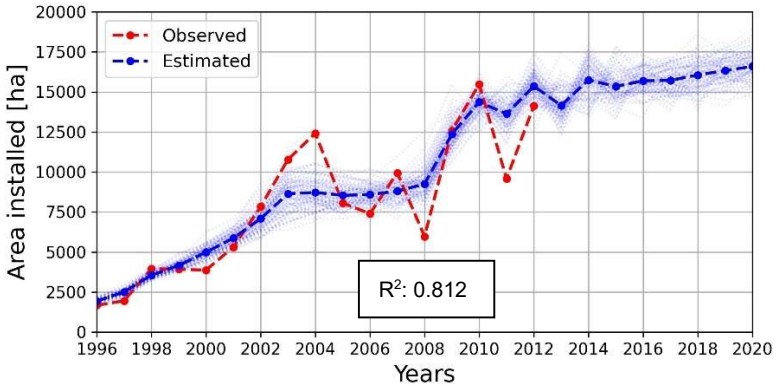
Table 14 reports the adjusted $R^2$ score and the plot of the estimated adoption of the Municipality-based Data-driven ABM with all the combinations of classifier and regressor constituting the agent's internal model tested[55].

The version of Municipality-based Data-driven ABM chosen was the one using nonlinear support vector machines both for classification and regression. Despite the fact that these versions did not present the highest $R^2$ score, support vector machines are a type of ML model able to extrapolate outside the range of values of the features on which it was trained [116]. On the contrary, tree-based models such as gradient boosting and extremely randomized trees are unable to extrapolate[56] and this can result in bad performances in running the model after 2012, when the features regarding cumulative adoption increase over the values in the dataset used for training the ML models. This is clear from the trends estimated by the models in Table 14 using a tree-based regressor, which keep growing even after the payments stop in 2012 and are therefore unrealistic as it would mean that the financial incentives had a negligible effect. This justified the decision of avoiding tree-based models completely and therefore of relying on nonlinear support vector machines for both classification and regression. Moreover, the trend obtained with these ML models is the one which best reflects the expectations of innovation diffusion theory and previous literature, as shown in Figure 2.

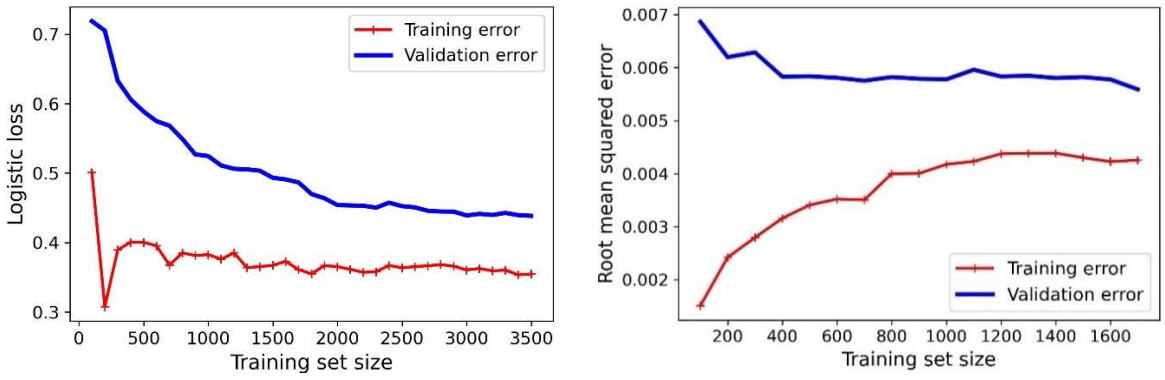
Figure 6 reports the plot of the learning curves of the chosen combination of ML models, to evaluate their overfitting. Figure 7 reports the summary of the results of the application of the SHAP framework to evaluate the importance of the various features on the ML models estimations. Red dots represent high values of the relative feature on the vertical axis, while blue represents low values. The effect of these values on the prediction of the model can be read on the horizontal axis. Therefore, clustering of red dots on the right (left) side of the vertical axis representing the 0 imply that higher-than-average values of that feature increase (decrease) the prediction of the target variable by the model. An unclear division of red and blue dots imply a more complex effect of the feature and its likely due to its interaction with others. Figure 8 focuses on the effect of the total cumulative adoption in the municipality, more clearly reporting the effect of its increase on models' prediction.

[55] Additional metrics and graphs resulting from testing the various combinations of ML models can be accessed at the path thesis-sbp-abm\municipality_level_analysis\municipalities_abm\model_validation\results.
[56] For a discussion on this, see https://www.kaggle.com/c/web-traffic-time-series-forecasting/discussion/38352.

*Table 14: yearly adoption of sown biodiverse pastures in Portugal observed and estimated by the Municipality-based Data-driven agent-based model for all the combinations of classifiers and regressors tested and relative adjusted $R^2$ score.*

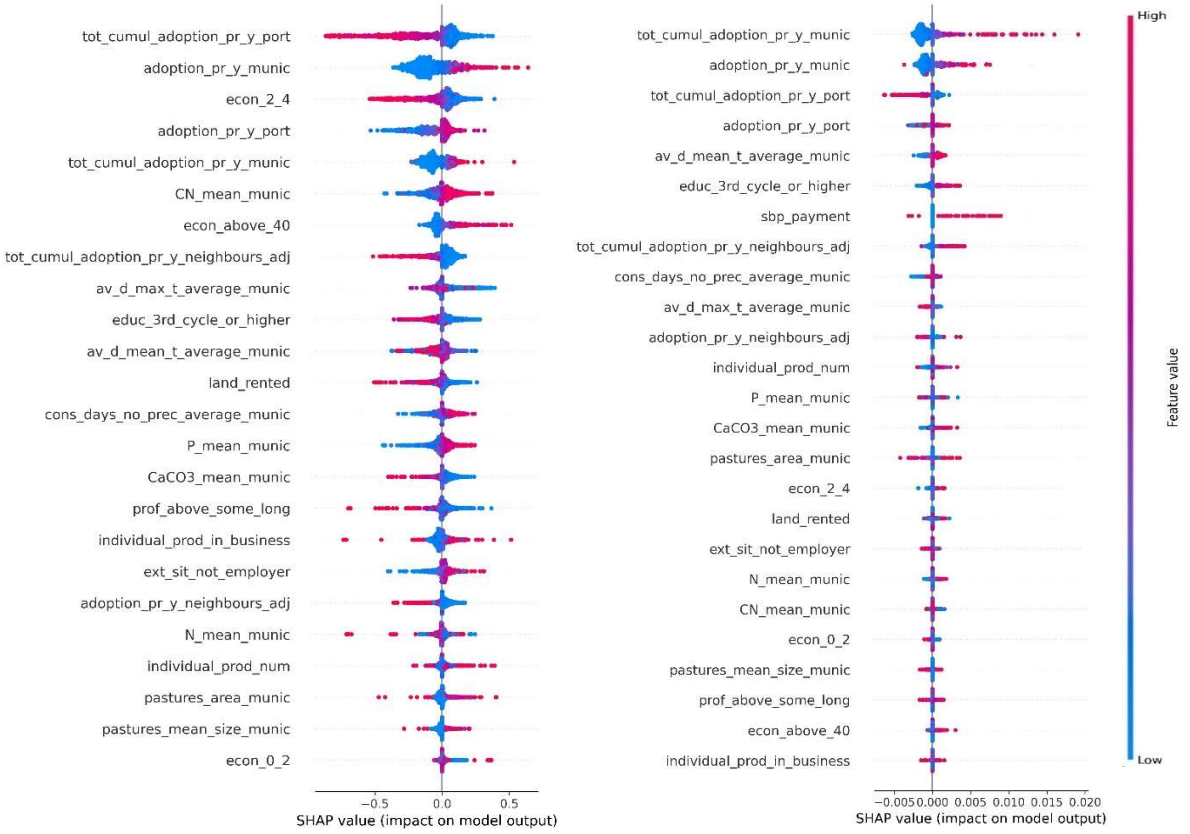
| | **NLSVM** | **XRT** |
|---|---|---|
| **NLSVM** |  |  |
| **GB** |  |  |

*Each graph was obtained with the classifier reported on the vertical axis and the regressor reported on the horizontal one. The graphs report the adoption estimated by all the 100 runs of the Municipality-based Data-driven agent-based model (light blue) and their average (dark blue). NLSVM – nonlinear support vector machines; XRT – extremely randomized trees; GB – gradient boosting with decision tree as base estimator.*




*(a)* *(b)*

*Figure 6: learning curves of the final classifier (a) and regressor (b) chosen for the agents' internal model of the Municipality-based Data-driven agent-based model.*




*(a)* *(b)*

*Figure 7: SHAP values for the classifier (a) and the regressor (b) constituting the Municipality-based Data-driven agent-based model agents' internal model.*

*The features (which are reported with the internal name given during the analysis and whose meaning is reported in Table A.10) appear on the vertical axis, ordered from the one with the most influence on the model output at the top to the one with the least at the bottom. The SHAP values quantify how much the value of a given feature (compared to a baseline value) impacted the prediction of the model for the specific instance. The horizontal axis report these values, which represent probability of having adoption in a municipality for classification and area installed for regression. Each row reports a dot for each instance of the validation set used to calculate the SHAP values. The further a dot is from the central axis, the more the value of the feature relative to the row impacted the model output. The colour of the dots shows whether the value of the feature was high or low for the relative instance. SHAP – SHapley Additive exPlanations.*

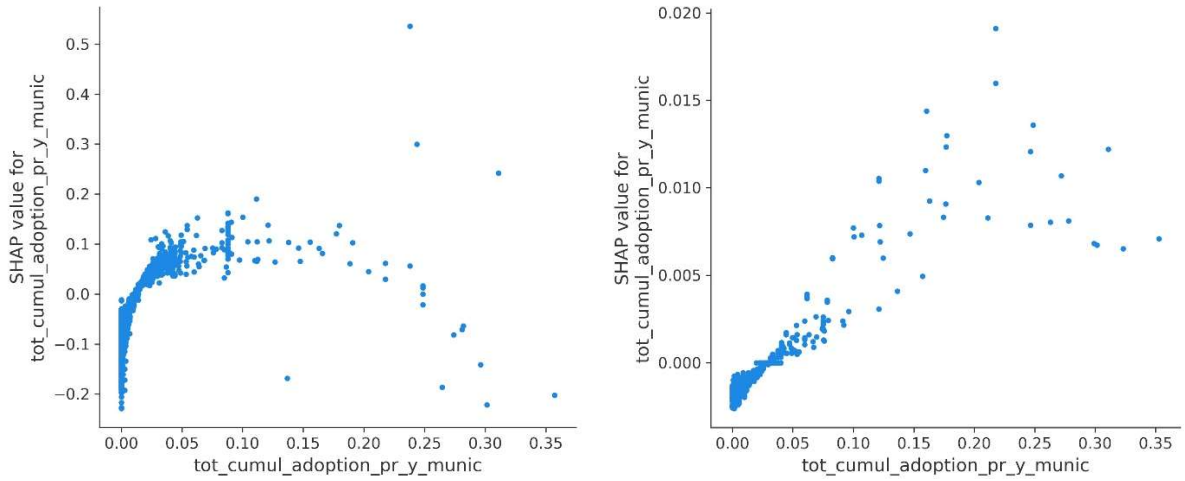



*Figure 8: SHAP values for the feature tot_cumul_adoption_pr_y_munic for the classifier (a) and regressor (b) constituting the Municipality-based Data-driven agent-based model agents' internal model.*

*For the meaning of the SHAP values refer to the subcaption of Figure 7.*

#### 4.2.2.2 ABM validation and output analysis

Figure 9 reports the cumulative adoption in Portugal observed and estimated by the Municipality-based Data-driven ABM, while Figure 10 shows the aggregated results of the classification and regression stages of the agents' internal model. The adjusted $R^2$ score and RMSE at the micro-level were respectively of 0.352 and 0.00801.

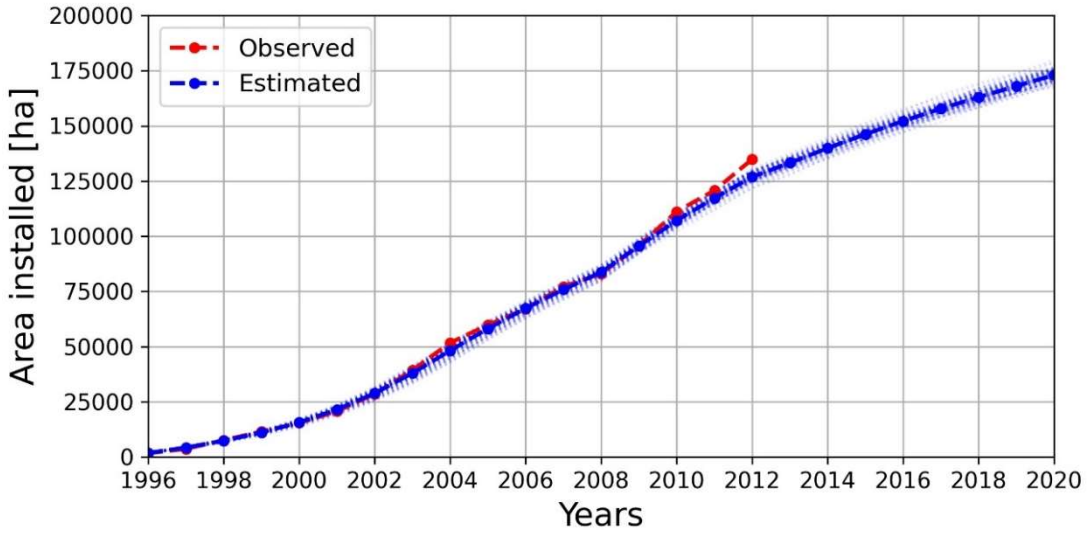
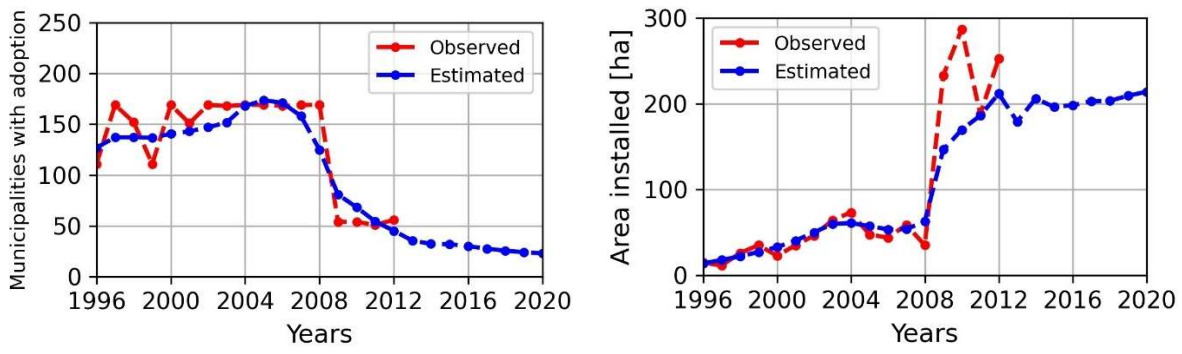


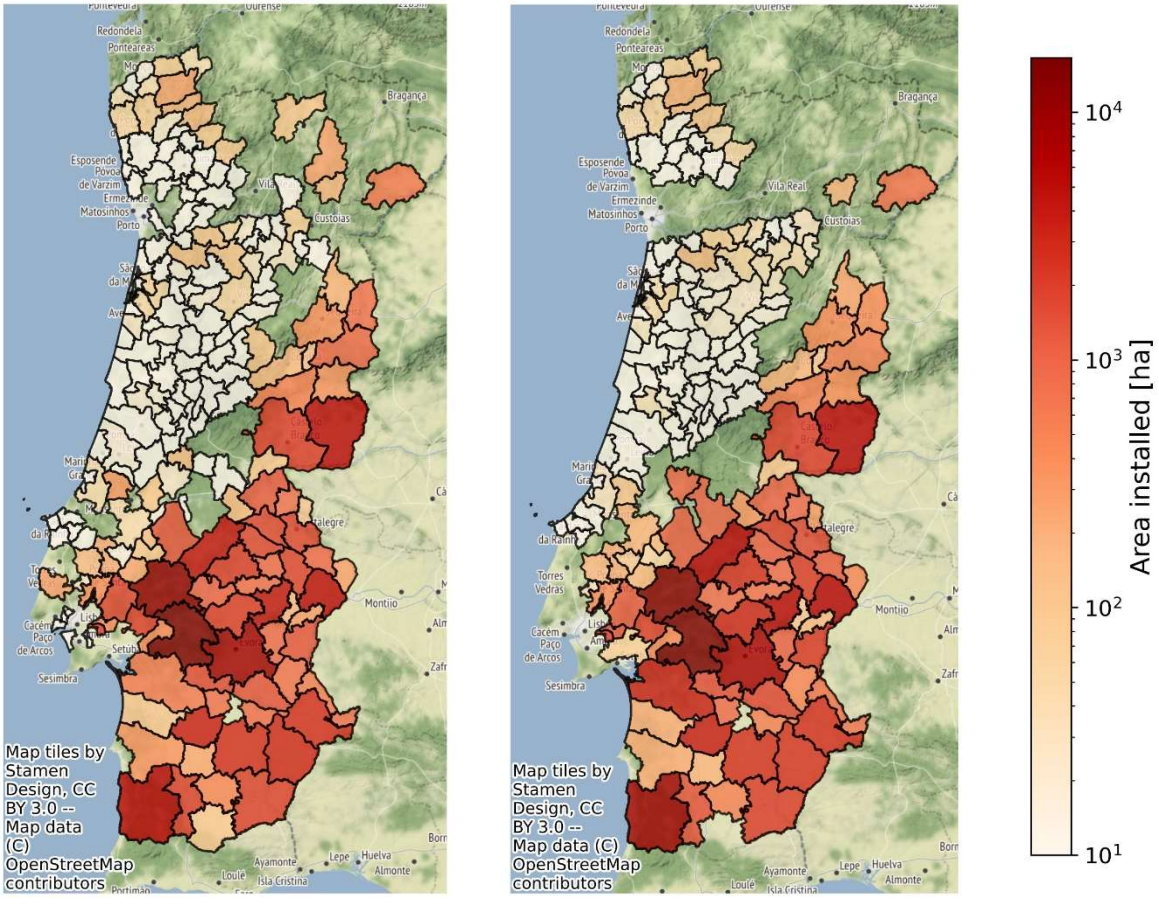

*Figure 9: total cumulative adoption of sown biodiverse pastures in Portugal observed and estimated by the Municipality-based Data-driven agent-based model from 1996 to 2020 (individual runs in light blue and average dark blue).*




*(a)* *(b)*

*Figure 10: yearly aggregated results from 1996 to 2020 of the two stages of the internal model of the Municipality-based Data-driven agent-based model: number of municipalities with adoption estimated by the classifier (a) and average area of sown biodiverse pastures installed in the municipalities with adoption (b).*

Figure 11 reports the estimated and observed total area installed in each municipality until 2012. The municipalities without adoption in the map from observed data (Figure 11.b) are likely to be less than in reality due to the disaggregation done of the adoption before the PCF project. The Municipality-based Data-driven ABM estimated that some area was installed in a larger number of municipalities.




*(a)* *(b)*

*Figure 11: total area of sown biodiverse pastures installed in each municipality in Portugal until 2012, estimated by the Municipality-based Data-driven agent-based model (a) and observed (b). Municipalities with no adoption were not plotted.*

### 4.2.3 Quantification of additional C sequestered thanks to the PCF project

Figure 12 reports the yearly and cumulative adoption in Portugal from 1996 to 2020 if no PCF project took place, estimated by the counterfactual simulation. The estimated cumulative trend of Figure 12.b is directly comparable to the expectations for the progression of adoption made in 2010 when designing the PCF project, reported in Figure 2.

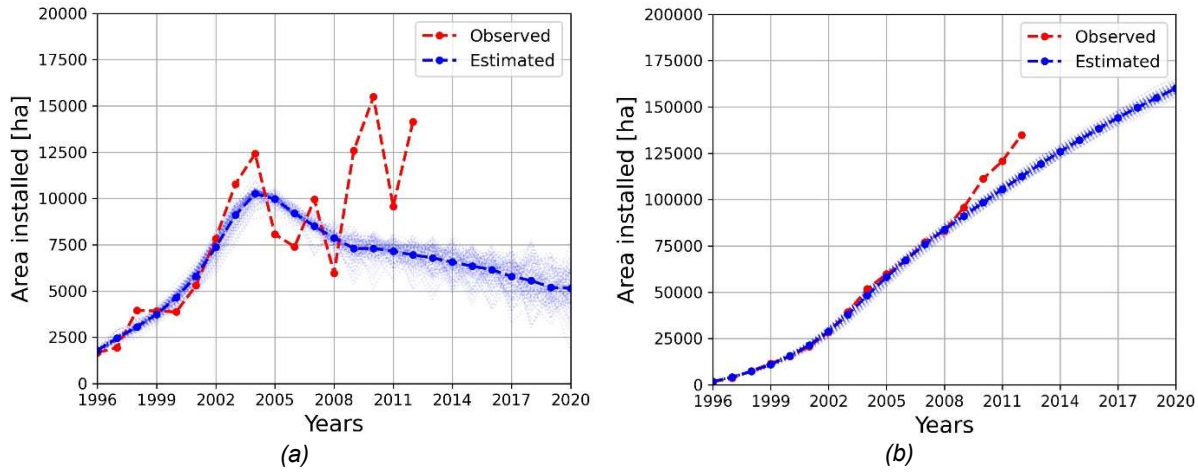



*Figure 12: yearly (a) and cumulative (b) adoption of sown biodiverse pastures in Portugal observed and estimated by the Municipality-based Data-driven agent-based model run without the PCF project, i.e. with no payments provided (individual runs in light blue and average dark blue).*

Table 15 reports the effect of the PCF project (over its duration, after its conclusion and overall) in terms of SBP area sown and the consequent C that this area already sequestered and will sequester over its lifetime (if properly managed). The negative values for the period 2013 – 2020 imply that after its end the PCF project caused a lower area to be installed, respect to the one that would have been expected without it. Figure 13 gives an approximated graphical representation of the differential areas.

*Table 15: effect of the Portuguese Carbon Fund (PCF) project on the sown biodiverse pastures (SBP) area installed during the project (2009 – 2012), after its conclusion (2013 – 2020) and total until 2020, and consequent carbon sequestered during the lifetime of the pastures.*

| | 2009 – 2012 | 2013 – 2020 | 2009 – 2020 (total) |
|---|---|---|---|
| Differential area of SBP installed thanks to the PCF project [kha] | 25.46 | -3.98 | 21.48 |
| Carbon sequestered by the differential SBP installed over their lifetime [Mt $CO_2$] | 1.65 | -0.26 | 1.39 |

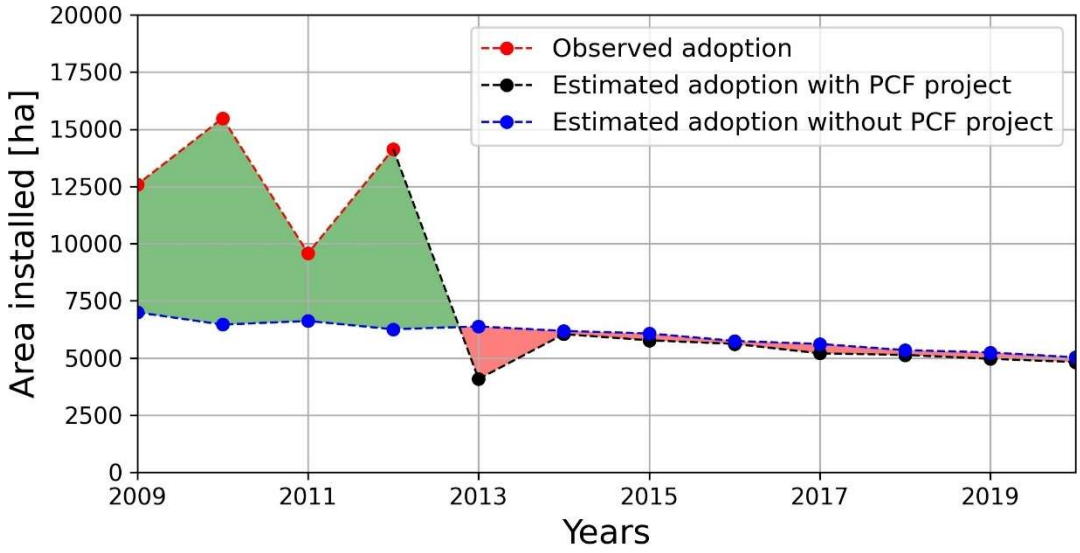



*Figure 13: comparison of sown biodiverse pastures adoption with the PCF project (observed until 2012 and modelled from 2013) and without the PCF project (modelled), from 2009 to 2020.*

*The area coloured in green in the figure gives an indication of the additional sown biodiverse pastures area installed thanks to the PCF project, the one in red of the reduced area installed due to it. However, they do not correspond to these areas and are depicted only to aid the visual identification of increased or reduced area, since the analysis was done per year. PCF – Portuguese Carbon Fund.*

# 5 Discussion

## 5.1 Interpretation of results

### 5.1.1 Farmer-based approach

#### 5.1.1.1 Comparison of the different approaches results

All the approaches used to model the adoption of the farmers in the AF survey data overestimate adoption, predicting percentages of adopters varying from 76.7% (the Farmer-based Calibrated ABM and the logistic regression trained with the reduced number of features) to 93.3% (the logistic regression selected through the overfitting procedure), as opposed to the actual one in the dataset of 56.6%. The models do not present large performance differences in terms of F1 score, ranging from of a minimum of 0.72 (the Farmer-based Toy-ABM) to a maximum of 0.80 (the Farmer-based Calibrated ABM).

It is interesting to note that the approaches based on empirical data, through calibration or logistic regression, are not significatively outperforming the Farmer-based Toy-ABM, which is based on pure economic calculations and a pre-defined confidence factor. However, the results of the Farmer-based Toy-ABM can drastically change with the definition of different values of confidence factors for the education levels, which makes it a non-robust method. In fact, once the economic calculations are defined, the model's predictions completely depend on the definition of the confidence factors and on the education level of the farmers. This means that the reflection of the model's performance can be predicted simply looking at the number of adopters with different education level. While the model's results in Table 8 imply that the Farmer-based Toy-ABM predicts that all and only the *Farmers* with education level Graduate or Undergraduate adopt, Figure A.2 shows how this is not the case in the available data.

The overestimation of number of adopters by the Farmer-based Toy-ABM reflects the overestimation of SBP area sown in the years 2009 and 2010 by Teixeira [19], whose approach to model individual decision-making was the basis for the Farmer-based Toy-ABM. Teixeira [19] estimated the adoption of 42,000 ha of SBP in the two years, while the area successfully sown was of 28,000 ha and the difference is too large to be accounted for by failed installations. However, some distinctions are necessary between the two models. First, the Farmer-based Toy-ABM estimates the number of adopters while Teixeira [19] the hectares adopted, which depend also on the area sown by each adopter. Second, the scope of the Farmer-based Toy-ABM is much more limited than the one in Teixeira [19], which considers the entire area sown in Portugal. Despite these differences, the results of both works suggest that approaches based only on economic considerations and farmers education level tend to overestimate SBP adoption. However, this problem is not solved by any of the other farmer-based approaches undertook. The Farmer-based Calibrated ABM overestimates adoption as well, despite considering additional drivers than solely education. The same result is observed in the logistic regressions tested, which considered even more drivers and did not include any economic calculation. This overestimation is attributable to different factors concerning data limitations, discussed in section 5.2.1.1.

At the individual farmers level, the fact that the Farmer-based Calibrated ABM performs better than the logistic regression models selected through CV was expected, since the first was calibrated specifically to overfit the dataset. Even though the logistic regression was trained with more features as well, the additional ones were expected to be less relevant, being excluded from the Farmer-based Calibrated ABM exactly for this expectation.

Two not easily explainable and unexpected results also emerged from the analysis. The first is the fact that the logistic regressions selected through CV performed better in CV than when retrained and evaluated on the entire dataset, as Table 12 shows. This strange outcome is probably due to the limited sample available, which causes the way the data are split in the different folds to have large influence over the results. Since the procedure used to select the best model iterated over different splits, it is possible that the best selected model corresponds to a split of the data that makes the predictions on the test sets easier than when the model is trained and predicts on the entire dataset.

The second unexpected result is the lower F1 score of the logistic regression selected through the overfitting procedure compared to the Farmer-based Calibrated ABM. Since the cash flows regarding SBP installation are equal for all the farmers, the Farmer-based Calibrated ABM can be reduced to the same decision-making structure of the logistic regression: it calculates a weighted sum of the different characteristics of each farmers and if the result is higher than a certain threshold the farmer adopts. While in the logistic regression this threshold is 0.5 (the default value[57]), in the Farmer-based Calibrated ABM it is the value of the confidence factor that makes the EDNPV equal to 0. Another difference is that the logistic regression also fits an intercept, i.e. a parameter that does not multiply any feature. However, this additional parameter should even give the logistic regression more freedom to find the best solution given the same set of features. The reason for this result is unclear but could be due to the different algorithms used to find the best solution. While the iteratively refined grid search for the Farmer-based Calibrated ABM is ensured to find the best solution if enough iterations are done, the *LogisticRegression* class implemented in scikit-learn uses a particular kind of stochastic gradient descent called “SAGA” [117], which may not converge to the global optimum.

Despite this practical similarity, both Farmer-based Calibrated ABM and Farmer-based Logistic Regression were used for their different approach to the same problem. The Farmer-based Calibrated ABM (as the Farmer-based Toy-ABM) does not present any spatial and temporal dimension and runs in just one step. It could have been replaced by simply an equation representing the EDNPV calculation for each farmer. However, the process of building a proper simulation on the one hand helped to understand the system’s dynamics and on the other hand allows to easily increase the complexity of the analysis if needed. In fact, this model was already configured to accommodate the possibility to include some improvements in the future considering not only SBP but land use change more generically. For instance, *Farmers* and *Farms* were set as different entities to allow for property change dynamics, as *Farmers* buying and selling F*arms,* F*armers* owning more than one *Farm* and *Farms* inheritance. Moreover, the nested structure of *Pasture*, *Market* and *Government* entities and the process of the

[57] The analysis also tried to change the threshold of the LOGRs, however in all cases the best F1 score corresponded to the default threshold and therefore the results of this tests are not presented (SC: farmer_level_analysis\Logistic Regression.ipynb).

*Farms* to check all the pastures that can be adopted and not only SBP were designed to simplify the inclusion of additional pasture types.The use of a logistic regression was much faster to develop and calibrate, with also the possibility to implement CV. The choice of logistic regression as a ML algorithm was due to the simplicity of interpreting its results though the features weights of the fitted model. Also, logistic regression was found in the literature to be a good model for binary decisions on adoption [71], [85].

#### 5.1.1.2 Adoption drivers at the farmer level

Both the Farmer-based Toy-ABM and the Farmer-based Calibrated ABM are based on the assumption of farmers acting on the basis of economic calculations, with the inclusion of proxies for their uncertainty. Two facts reiterate the importance of such economic considerations. The first is that the F1 score of the Farmer-based Toy-ABM is actually similar to the one obtained by the other models. The second is that the values of EDNPV resulting from the Farmer-based Calibrated ABM are within the range of values that could actually represent the expected ones by farmers, varying from around -750 €/ha to 1000 €/ha over 10 years with an initial investment of 723 €/ha to install SBP. In fact, only one confidence factor calculated by the Farmer-based Calibrated ABM has a negative value, which does not make sense since a value of 0 already expresses certainty that SBP installation would fail. The other values range between 0 and 2. If a value of 1 expresses full confidence in a positive outcome of the installation, values larger than 1 can be interpreted as the farmers overestimating the economic benefit of SBP. However, an overestimation of twice the benefit is unlikely to derive only from economic consideration. These high EDNPVs could be the consequence of other elements that the Farmer-based Calibrated ABM embeds into the economic calculations, being the only discriminant for decision considered. Examples of these factors are SBP diffusion and peer influence, sustainability value or specific environmental conditions, and they should be considered separately in order to assess their influence, as for example carried out with the available factors in the municipality-based analysis. However, to properly understand the role of economic calculations and the extent to which they can predict farmers' decision-making, these would need to differentiate farmers since the assumptions of all calculating the same costs per hectare is not realistic (issue that will be treated further in section 5.2.1.2). In particular, the highest EDNPV obtained in the Farmer-based Toy-ABM was 98 €/ha, a number too low to convince many farmers to adopt facing the initial investment of 720 €/ha (and a total that would be negative without the payments), suggests that the calculations cannot reflect the calculations done by the majority of farmers.

For what concerns the other farmers' characteristics studied, the results of the first data analysis in section 3.2.2.1 provides some first insights. In Table 9, the highest positive correlation is exhibited by *PastureSurface,* meaning that farms with a larger surface dedicated to pastures tends to adopt more. This can be easily explained by some economy of scale factors not considered in the model, which could lower costs per hectare in case more surface is sown. Larger farms typically have a higher turnover and more free area to experiment with innovative systems. On the contrary, *PercentRentedLand* has the most negative coefficient, suggesting that the more area of its farm is rented, the less likely is a farmer to adopt SBP. This is an expected result as well, since the more land a farmer rented, the less willing they could be to undertake long-term investments. However, the absolute values of the correlation

coefficients do not go over 0.21 and in general no strong correlation was found between adoption and the numerical attributes. Regarding the categorical features, already from Figure A.2 no relation between any of these and the adoption clearly appears. Almost all the categories of each feature present both adopters and not adopters in a balanced way. This was confirmed by the Pearson's Chi-Square test results in Table 10, which was inconclusive in rejecting the null hypothesis of independence for all the features with any significance level lower than 25%, with only the district of the municipality and its legal form being close to this boundary.

The logistic regressions confirm the insights obtained from this first analysis. In the logistic regression trained with all the features, the 4 selected from the data analysis present the highest coefficients in absolute values (Table 13), with the exclusion of two categories generated from the *Distrito* feature (Beja and Portalegre) which was excluded only for its large number of categories. This result suggests a spatial dimension in the model would be an important addition. Instead, the almost null coefficients of the percentage of cattle, for how long a farmer has been in the profession and its expectation over family succession confirm their lack of correlation with SBP adoption. In all logistic regressions, the signs of *PastureSurface* and *PercentRentedLand* agrees with the ones of the first data analysis done. The importance of the legal form of the farm varies greatly among the models, from having a coefficient of 0 in the logistic regression selected through the overfitting procedure to being the most relevant in the logistic regression selected through CV with the reduced number of features. However, when it takes a non-0 value, its coefficient is negative, meaning that farmers legally associated in any form are more likely to adopt, which can be explained by the diffusion of information within farmers associations which can improve SBP credibility. The level of education always has a positive impact on adoption, even though its effect is never the strongest. A note regarding the results of the logistic regression is that the stochasticity of the optimization algorithm used to train it, together with the limited sample available to train the model on, can cause the weights of the fitted model to change in different runs and therefore the values reported should not be taken as definitive.

The results of the Farmer-based Toy-ABM and Figure A.2 show how the level of education of the farmers cannot predict adoption by itself. However, in the Farmer-based Calibrated ABM results the education level has the largest coefficient, meaning it is the most important driver of adoption according to this model. The coefficients in the Farmer-based Calibrated ABM (and also in the Farmer-based Toy-ABM), contrary to the ones of the logistic regression, have a specific meaning, being directly related to the uncertainty in the outcome of the adoption and in economic calculations. This, together with the inclusion of the additional term of degree 0 in the logistic regression (the intercept), can explain why some features have different relevance in the Farmer-based Calibrated ABM and in the logistic regression, but not the different sign of their coefficients. Another important difference in the weights obtained from the Farmer-based Calibrated ABM from the coefficients of the logistic regressions is the positive one for the percentage of rented land, despite having the smallest absolute value among the features included. The positive contribution of the percentage of rented land in the Farmer-based Calibrated ABM can instead be explained by checking its Spearman ρ coefficient with the pasture surface of the farm. This revealed a value of -0.25, meaning that largest farms in the dataset tend to be more privately owned. Therefore,

the positive coefficient of *PercentRentedLand* is likely due to the need to compensate the high positive one of *PastureSurface*.

In general, the lack of large variations in the performance of the models, together with the results of the data analysis, suggests that the considered characteristics are insufficient to reliably predict the decision of the individual farmers in the available dataset.

## 5.1.2 Municipality-based approach

### 5.1.2.1 Municipality-based Data-driven ABM

The validation of the Municipality-based Data-driven ABM highlighted its ability to forecast at the macro-level the system dynamics and in particular the yearly area installed of SBP in Portugal with an $R^2$ score of nearly 0.80. The plots of yearly and cumulative adoption in Portugal follow the underlying trend of the observed data, despite being smoother and not reproducing all its interannual variability. This was however expected, since the only features changing values over the years in the simulation (excluding the one reporting the payment value) are the ones related to adoption, which are likely to be related with the underlaying trend and not with irregular yearly variations. If the model had been able to capture all the variability of the observed adoption, it would have been an indicator that it was highly overfitting the data. The yearly variability is instead likely to be related to factors contingent to the specific years, like patterns emerging from series of casual events or external perturbations of the model, that are impossible and often undesirable to fully capture in a model. This constitutes a known issue of validation through comparison with historic data [36], which does not hamper the validity of the model. However, some limitations regarding the available data can also contribute to the lack of variability and will addressed in section 5.2.2.1.

The uncertainty in the models' macro-level estimations due to its stochasticity appears in the trend of the individual runs, which presents oscillations around the average value after 2013. The span covered by the different runs grows larger with the years, implying that as expected the uncertainty grows as the model tries to extrapolate more in the future. This resemble the behaviour of dynamic systems oscillating around a point of equilibrium and is due to the interaction of the features regarding previous adoption.

Even though the ML models constituting the agent's internal model of the ABM are trained on the entire dataset available at the beginning of the simulation, the micro-level RMSE and adjusted $R^2$ are worse than the validation scores obtained by the chosen regressor alone, passing respectively from 0.00643 to 0.00801 and from 0.543 to 0.352. This difference is due to the increased complexity of the ABM in respect to the regressor alone. Features regarding SBP adoption are endogenous to the ABM and are therefore calculated at each time step, propagating the errors, and the ABM's performance also depends on the classification stage of the agents' internal model.

The map of adoption in the different municipalities in 2012 (Figure 11) shows that the ABM is able to reproduce the emerging spatial pattern of adoption in Portugal, concentrating the majority in Alentejo as in reality. However, it predicts some more municipalities adopting. The ones with a significative SBP adoption are mostly located in the north and could actually represent municipalities where SBP were

adopted before 2008 but not during the PCF project, which do not present adoption in the observed data due to the disaggregation rule used of prioritizing municipalities with adoption in the PCF.

An important novelty exhibited by the results is the higher yearly adoption in Portugal after 2013 in comparison to the previous extrapolation constituting the basis of the PCF project design reported in Figure 2, which is evident in the trend estimated by counterfactual simulation. While the yearly adoption in Portugal without PCF project was previously forecasted to decisively decrease after peaking in 2004 and reach 0 already in 2020, with a consequent plateau of the cumulative adoption, the simulations showed a slower decrease, with around 5,000 ha of SBP still installed in 2020. The consequence is that the estimated yearly adoption, which until 2004 raises similarly to the probability density function of a logistic distribution, after 2004 decreases only linearly when no payments are provided (Figure 12.a). The PCF project constituted a discontinuity in this trend of adoption, which after 2013 returns to a similar trend than if there would not have been the project.

#### 5.1.2.2 Assessment of the PCF project outcome

This thesis estimated 1.65 Mt of $CO_2$ sequestered by the additional SBP area installed thanks to the PCF project over the years 2009 – 2012, 0.11 Mt more than the previous assessment done by Terraprima [14]. However, while the previous estimations considered only the C sequestered over the duration of the project, the one in thesis considered the entire lifetime of these pastures of 10 years, which will be completely stored in 2022. However, the additional area of SBP installed thanks to the PCF project estimated in this thesis is lower than the previous estimation. This lower value directly follows from the slower decrease in adoption after 2004 when no payments are provided estimated by the counterfactual simulation respect to the extrapolation in Figure 2 discussed in the previous section. This extrapolation constituted the counterfactual used to design the PCF project, which means that the payments per hectare of SBP installed offered to the farmers would have been lower with the counterfactual case used in this thesis. This is due to the fact that the PCF only paid for additional C, i.e. C sequestered in the area estimated to be sown regardless of the existence of the project was removed from the amount of C paid. The result of this thesis therefore imply that, with the same payments per tonne of $CO_2$ sequestered, the payments offered to the farmers per hectare of SBP installed would have been lower, since more area would have been installed without them. However, this thesis cannot affirm that the counterfactual estimated here is more plausible than the counterfactual used to establish additionality during the PCF project. The main reason for this is the absence of a biophysical depiction of land suitability for SBP installation, as debated in section 5.2.2.3. Here, all pasture land was implicitly considered to be equally suited for installing SBP, and therefore no penalty was assigned to any municipality for each hectare installed up to 100% of the pasture area in the municipality. In reality, municipalities would run out of suitable area before getting to 100% - due to naturally low-quality areas of low yields, technical difficulties with some of the terrain, etc. A biophysical model of pasture quality, which does not exist at the moment, is therefore a critical future step as a follow-up to this analysis.

In addition, this thesis estimated for the first time also the residual effect of the PCF project after its conclusion and until 2020. The negative value obtained implies that due to the PCF project less SBP

were installed in this period. However, the reduction in area installed was small compared to the increase until 2013 and the net effect of the project was to increase adoption and therefore C sequestration. The likely main explanation for this is that the PCF caused two categories of farmers to install, additionally to the ones that would have installed in any case during the PCF project years. The first consisted of farmers that would have installed SBP in the following years but, economically incentivised by the payments or because the increased spread of the system made them know and trust it before, decided to do it in the period 2009 – 2012. The second category are the farmers who would not have adopted SBP if the project had not taken place, either because the adoption would not have been economically feasible for them of because they would not have known it before 2020. During the PCF project, the first category installed an area equal to the negative difference of area installed after 2013 and the second to the net difference of area over the entire period. Payments to the farmers in the second category, who would not have adopted otherwise, were effective and had a net positive impact. Instead, payments to farmers that would have installed in any case but just later did not contribute to increase the additional adoption of SBP and should not have been taken into account when defining the payments offered. The values calculated imply that almost 88% of the payments corresponded to effectively additional C sequestration, a high value considering that the payments were provided for all new installations, without any further condition differentiating among farmers. This can be explained by the fact that the majority of farmers interested in the system and for whom its adoption was already economically beneficial adopted it before the PCF project and therefore could not benefit from the incentives. It should also be noted that the category of farmers who would have ended up installing pastures was unobservable and impossible to estimate when the project started, and that their decision to anticipate the installation helped the Portuguese State to comply with the Kyoto Protocol.

#### 5.1.2.3 Adoption drivers at the municipality level

The SHAP analysis of the classifier and regressor constituting the agent's internal model reported in Figure 7 aimed at studying the factors influencing respectively the presence or not of SBP adoption in a municipality in a certain year and, where and when there is adoption, how much area is installed. While in a farmer-level approach the classification stage is more relevant (and was in fact the only one conducted), since the area of SBP installed relatively to the pasture area of the farm usually does not vary much and in any case corresponds to a small value compared to the aggregated adoption, in a municipality-level approach the opposite is true. Municipalities in different years saw very different fractions of pastures area being installed and these variations corresponded to hundreds of hectares of difference, giving a great relevance to the regression stage. Instead, the classification stage treats indifferently municipalities with SBP adoption regardless of the amount of adoption, while the difference between municipalities with 1 and all farmers adopting is much more important than the one between municipalities with 0 and 1 farmer adopting. Moreover, as better explained in section 5.2.2.3, the need for disaggregating the adoption previous to the PCF project exacerbated this issue, causing many municipalities to be labelled as adopting than what is expected in reality. For these reasons, the following analysis focuses more on the regression stage.

While the outcome of the regressor is determined mostly by a small number of important features, the one of the classifier depends more evenly on all the features used. For both models however the features reporting SBP adoption in and until the previous year, especially regarding the municipality itself and Portugal, are the ones having the highest influence. This was expected, as these are the only features that change between the years that were included in the datasets (apart from the payments in the last 4 years for the regressor) and therefore the only ones that allow the models to capture the interannual variability and to explain the aggregated trend of yearly adoption in Portugal. The results show that a higher value of adoption in the previous year, especially in the municipality itself, generally brings to a higher estimated adoption in the following. This confirms that more installations by neighbours in the previous year bring more farmers to know and trust the system and therefore to install it. Figure 8 shows that while the probability of adoption peaks with a value of total cumulative adoption in the municipality (divided by pasture area) of around 0.10 and then starts decreasing, the amount of area adopted keeps increasing, and only seems to start plateauing for values of total cumulative adoption above 0.3. This means that the regressor did not learn the relation that was shown in Figure 5. Instead, the models seems to rely on the total cumulative adoption in Portugal to decrease the estimated adoption, since this is the only adoption feature among the 4 most important which has a negative contribution. The fact that the total cumulative adoption in Portugal has an opposite effect to the one in the municipalities does not have a realistic explanation. The same applies to the features regarding the adoption in the neighbouring municipalities, which, despite being less influent, have opposite contributions for the classifier and the regressor. In the former their contribution is also opposite to the contribution of the adoption in the municipality. The explanation for these unrealistic trends can be related to two major limitations of the municipality-based approach that will be further treated in section 5.2.2.3, the biases introduced with the disaggregation of the adoption previous to the PCF project and the lack of a variable representing the biophysical capacity of the soil.

Climate features are the second group with the most influence on the regressor estimations, especially the average mean daily temperature in the municipality, whose higher values bring to more adoption. The average maximum daily temperature instead has the opposite contribution, suggesting that more farmers adopt SBP in municipalities with a warm climate where, however, the temperature does not have the highest peaks during the day. Regarding precipitations, the best conditions for adoption are an average of 60 to 70 consecutive days without rain. Reduced periods without rain discourage adoption of SBP, probably because rainy locations will have naturally high yields even for other grassland systems. None of the soil features have negligible contributions, implying that simple soil chemical properties are not considered by farmers when deciding.

The value of the payments was included as a feature only for the regressor. As expected, their presence generally leads to higher adoption, apart from a few cases which could be explained with the interaction with other features that would need to be further investigated and could be simply due to outliers.

Regarding census features, particularly relevant is the fact that the more farmers of a municipality completed at least the third cycle of basic education the more SBP area is installed. This confirms the hypothesis of the economic calculations by Teixeira [19] were based on and the results of the farmer-

based approach. The other census features were much less important in driving the regressor predictions.

#### 5.1.2.4 Additional experiments

Some additional procedures were implemented during the analysis of the ML models and in particular of the regressors, which were not reported in the main text due to lack of space and to the fact that they were not strictly necessary for understanding the main method[58].

The first experiments aimed at confirming some decision regarding the first screening of the features. In particular, random forest regressors were trained and evaluated on three different datasets: one keeping the most correlated census features for each category instead of giving priority to the combined ones, one substituting the total cumulative adoption with the one considering only the 10 years before and one including all the features before the first screening, i.e. all the ones reported in Table A.10. The evaluation of the models was performed through the same search method used for the first round of hyperparameters tuning of the random forest, so that the results could be directly compared. The analysis confirmed that the decisions made regarding the census variables and adoption features and the first screening did not influence negatively the results, since for all the random forests the RMSE was similar to the one actually trained.

Then, to see if a further reduction of the features could decrease the generalization gap of the models, a recursive features elimination technique was tested for a random forest regressor, based on the coefficients obtained from a LR. However, the gap reduced only slightly while the validation error increased.

A final batch of experiments aimed at checking if a nonlinear support vector machines could learn a second degree relation of the adoption in the year with the *tot_cumul_adoption_pr_y_munic* feature, if the other features concerning adoption were reduced. The first test removed *tot_cumul_adoption_pr_y_port* and *adoption_pr_y_neighbours_adj* and the second also the other two features relative to the adoption in Portugal and in the adjacent municipalities, but none learnt the desired relation and in all an increase in the total cumulative adoption of the municipality caused an increase in the adoption in the year. The same experiment was also implemented while training the models only on the data points referring to the years before 2009, to understand if the models did not learn this relation because of the increase in adoption due to the PCF project. However, also in this case the results did not exhibit such trend.

Apart from the experiments regarding the ML models, a different model to run the counterfactual simulation required to evaluate the outcome of the PCF project (section 3.3.4) was tested. This consisted in a version of the Municipality-based Data-driven ABM with, as internal model of the agents, the same classifier and regressor previously selected but with hyperparameters tuned and parameters fitted only

[58] The implementation and the results of all these experiments can be found in the notebook located at municipality_level_analysis\ml_models\ Regression - Models evaluation.ipynb, in the section named “Additional experiments”.

on instances of the relative databases regarding the years previous to 2009[59]. Since all these instances presented the feature *sbp_payment* equal to 0, this procedure ensured that the value of the payments had no influence on the models estimations. This ABM and the one trained on all instances were both run from 1996 to 2008 and their macro-level adjusted $R^2$ scores, i.e. calculated on the yearly cumulative adoption in Portugal, were compared. Since the ABM trained on all instances obtained a better $R^2$ score, it was chosen for the counterfactual simulation and C sequestration calculations.

# 5.2 Limitations of the work

## 5.2.1 Farmer-based approach

### 5.2.1.1 AF survey data limitations

A first intrinsic and important limitation of the AF survey data for the purposes of this thesis is the lack of information on when each farmer adopted. This forced the work to refer all the calculations to the same year, considering the same payments and the prices for all the farmers. The selected year, 2009, was the one with the highest payments offered, since it is fair to assume that the majority adopted in the year when it was most beneficial to do so. However, this could be one of the reasons driving the adoption overestimation of the models based on economic considerations: some farmers could not have been aware of SBP in 2009 and considered adoption only afterwards, when the payments were lower or not provided at all, therefore not adopting. Moreover, not knowing the year of adoption prevented the retrieval, from the data sources used in the municipality-level approach, of the level of adoption in their municipality when they decided to adopt.

The first exploration of the AF survey gave insights on other limitations which hamper the possibility to generalise the insights obtained outside the sample of farmers interviewed. The data have a limitation in the spatial scope, since the farmers in the survey are predominantly located in the Alentejo region (with the exception of two, which however are in close proximity in the Lisbon district). Due to the heterogeneity in socio-economic and biophysical characteristics at the national level, conclusions for one region cannot be considered valid for others. The sample also includes two important biases. First, the majority of farmers adopted SBP and therefore adopters are overrepresented in respect to the actual situation in Portugal. Second, the farmers who answered the interview have a particularly high level of education, with 27 out of 30 farmers having a degree. As noted in section 2.5.2, a bias towards more innovative farmers is common in interviews conducted in the scope of research projects. A solution for this is to rely on compulsory data collection results, as done in the analysis at the municipality level where the census data include all farmers in Portugal and the Terraprima data all farmers who participated in the PCF project.

The overestimation of farmers adopting by all farmer-based models can be linked to the fact that some of the farmers may have tried to adopt SBP but the installation failed and therefore are not reported as

[59] The SC for the tuning and training of these ML model can be found in the section called “Final model chosen trained only on the years before 2009” of the notebooks at the paths municipality_level_analysis\ml_models\Classification – Models evaluation and municipality_level_analysis\ml_models\Regression – Models evaluation.

adopters in the AF survey data. Moreover, some farmers in the survey sew SBP many years ago and, not having any sown species left, may have reported the pastures as spontaneous or semi-natural. Another explanation could be that the characteristics of the farmers that were retrieved from the survey data lacked some important variables that make farmers estimate a lower economic value in adopting SBP. Important ones in this regard would be proxies for the expected productivity of the pastures (further discussed in the following section) which influence the amount of feed saved, a critical variable to make SBP adoption economically viable.

The inconclusive results of the correlation analysis and the Pearson's Chi-Square test and the small variability in the models results when considering only education or the entire set of attributes, show how the small sample of farmers in the AF survey data do not allow for general trends to emerge over the noise generated by individual stochastic variations. This limitation makes it impossible to establish a solid link between the features and the adoption of SBP (assuming obviously the existence of such relationships).

An important limitation of using the AF survey data for the aim of explaining SBP adoption is that the survey was designed for purposes not directly linked to it. The characteristics included in the survey force any analysis to consider only socio-economic characteristics and farms management practices. They do not include any usable data linked to biophysical characteristics and information diffusion, which are supposedly important drivers of adoption. The LUCC ABMs found in literature based on surveys, tailored the interviews specifically for the purposes of the study, to get direct insights on the decision-making process of the farmers which allowed for the construction of a behavioural model [16], [69].

The final issue of the AF survey data was its size, which did not allow to leave out a test set to assess the ability of models to generalize on independent data. The Farmer-based Calibrated ABM had therefore to be calibrated on the entire dataset. This imply that the application of the results for a sample outside the farmers included in the dataset could be misleading, even though the precautions taken to limit overfitting allow for the cautious conclusions reported above. The logistic regression was trained through CV and therefore its results, despite not being tested on a separated dataset as well after hyperparameter tuning, was tested on data not used for training. However, the biases and limitations of the dataset highlighted in the previous section suggest care in extracting general knowledge from all these models, limiting their explanatory power to the farmers involved.

#### 5.2.1.2 Method limitations

Apart from the limitations concerning the AF survey data, all the approaches modelled variables influence on the decision of adopting as linear. This was however a precise choice taken with the aim of reducing the risk of overfitting for the Farmer-based Calibrated ABM and the logistic regressions, which was significant due to the limited sample available. To understand if this choice caused an important limitation, the analysis also tested a random forest, a model able to learn non-linear relations. This model (whose testing was not reported in the text) provided similar results, confirming that the use of a linear model did not hamper the analysis. Moreover, none of the models considered temporal and spatial dimensions. While this is an intrinsic limitation of logistic regressions (unless features related to these are provided), the ABMs could have the possibility to include it abstractly, as a social network for diffusion

of information for instance. This limitation was addressed with the municipality-based approach, where space is explicitly addressed.

Being an approach based on ML techniques, the logistic regression included less assumptions than the Farmer-based Toy- and Calibrated ABM, which instead aimed at reproducing the real interactions among the different entities influencing farmers decisions. However, the a-temporality and a-spatiality of the ABMs reduced these interactions only to the farmers retrieval of information, which is not dynamic. The main difference between the logistic regressions and the ABMs lie instead in the economic considerations that are at the core of the latter.

The EDNPV calculations required in fact the inclusion of various assumptions, which hampered a proper evaluation of the extent to which economic calculations matter for SBP adoption and contributed to the overestimation of adoption, especially by the Farmer-based Calibrated ABM. While the analysis considered unique prices for all farmers, in reality prices can differ due to many factors, such as the specific retailer, the year and the size of the farm and therefore the quantities purchased (and the location, but as specified the farmers interviewed are located in the same region). This causes the EDNPVs to vary largely among farmers and considering this could increase the explanatory power of economic calculation. Other neglected factors that if considered would contribute to increase this variability are the use of micronutrients such as zinc for SBP maintenance fertilization, that would be required but many farmers do not actually use (but that in any case has a small impact on the total costs) and the subsidy of 25% for the investment offered by the Rural Development Programme for Continental Portugal 2007-2013 (PRODER) during the years of the project, which could not be considered due to the impossibility of knowing which farmers benefitted from it and which did not. Additionally, Equation (2) assumes that farmers own the livestock that saturates the maximum capacity of their SNP before adopting SBP and that they adopt the area of SBP which, considering its higher stocking rate supported, can host all the livestock already owned. It also neglects the feed produced by the remaining area not converted to SBP, if this is kept as SNP. This area could also be converted to other land uses, but no data were available on the decisions that farmers take in this regard. In any case, the use of the additional land translates in additional revenues that could make the adoption of SBP more convenient and explain why the EDNPV calculated by the Farmer-based Toy-ABM are so low.

Another main assumption was to consider an amount of supplementation required for SBP based on data collected from pastures optimally managed. In practice, even though the seed mix of SBP is usually tailored to the soil conditions, the productivity of SBP varies largely depending on the biophysical capability of the land in a given year, which in turn depend on many variables such as soil type and texture, initial SOM, climate and weather, legacy effects from prior uses, coverage with trees and stocking rate. As a consequence, the costs of feed evaluated in reality by farmers when adopting SBP could be higher than the ones calculated in the analysis done here, causing the actual costs for maintaining SNP, which include the differential costs of feed in respect to adopting SBP, to be lower. This issue could be solved with a model evaluating the productivity of the pastures.

An additional economic assumption was to consider that as soon as EDNPV is positive a farmer adopts. This is unlikely to hold in reality, where the return on the investment, i.e. the return obtained in relation

to the investment required, is also an important indicator. While being a factor that could radically change the performance of the Farmer-based Toy-ABM, due to its predefined parameters, for the Farmer-based Calibrated ABM this represents mostly a problem in terms of interpretability of the weights obtained from the calibration procedure. In fact, the addition of a threshold for the EDNPV could represent another parameter to calibrate, with a similar function as the one of the intercept in the logistic regressions, which could make the weights more significant. The model's performance would probably instead not change much and therefore the model does not include this term to reduce overfitting.

Another issue is the requirement of a base case to calculate EDNPVs. This required the inclusion of an initial land use for the area switched to SBP, set as SNP. Despite this being surely the most common case, it is not trivial that all the SBP area was SNP before. Moreover, the model neglects that farmers could have decided against installing SBP because they decided to dedicate the land to another use and therefore did not adopt even if SBP were a better option than SNP. These assumptions are not embedded in the logistic regressions, since these models predict whether it is advantageous for a farmer to sow SBP or not, regardless of its previous land use and the alternatives available to them and opportunity costs that do not need to be defined.

## 5.2.2 Municipality-based approach

### 5.2.2.1 Municipality-based Data-driven ABM

The limitations of the Municipality-based Data-driven ABM are strictly related to the ones of the ML models constituting the internal model of its agents. In fact, the ABM architecture does not introduce other assumptions but simply provides the features to the ML models, which are also the ones responsible for estimating the endogenous variables related to adoption.

The graphs in Figure 10 show how during the simulation the classifier and regressor reproduce the trends respectively of yearly number of municipalities with adoption and of their average adoption embedded in the labels of the datasets used to train them. These trends however expose the first main limitation of the municipality-based approach, which is the need to disaggregate the adoption previous to the PCF project. Its main consequence was to assign a small adoption (relative to their pastures area) to a large number of municipalities in regions where no municipalities adopted during the PCF project, which explained the sudden drop in number of municipalities adopting and raise in average adoption within each municipality that the data exhibit in 2009[60]. The fact that the ML models learnt these discontinuities caused by the disaggregation introduce an important error in the Municipality-based Data-driven ABM at the micro-level. The trends after 2013 show that the consequences of this on the extrapolated estimations of the ABM are an ever decreasing number of municipalities adopting and increasing area adopted in each of these municipalities, which combined imply an agglomeration of adoption in few municipalities. Despite not creating evident problems for the extrapolation by the ABM of the aggregated yearly trend of adoption at the macro-level until 2020, these biases could create issues

[60] The disaggregation had also consequences which however probably had lower influence on the results. In the regions where some municipalities adopted during the PCF project, their prioritization probably caused their adoption to be overestimated. At the same time, in these regions municipalities that did not present adoption during the PCF project because they already had a lot of SBP installed before 2009 were assigned no adoption (if present).

and unrealistic trends when using the model to explore scenarios after this year. This was outside the scope of the thesis but is a key point of the work that should follow it, in order to design new policies to further expand SBP adoption.

Apart from incorporating these biases in relation to the real-world situation, on the purely computational side the analysis of the learning curves in Figure 6 showed that both ML models also present some degree of overfitting, since there is a gap between the validation and training error that however reduces with the increasing size of the training set. A further reduction of the features was not beneficial (see the additional experiments reported in section 5.1.2.4), the models hyperparameters were tuned thoroughly and less complex models underperformed the chosen ones. Therefore, important improvements in the models performance can be probably reached only through data cleaning and collection, which however present important challenges. Solving the disaggregation issue would probably be the most important improvement in terms of data cleaning, since it would provide the models with more reliable labels and values of the features regarding adoption, shown in section 5.1.2.3 to have a fundamental role in the ML models predictions. However, it is unlikely that more information which to base the disaggregation on will be available since more precise data on adoption until 2008 are protected by commercial secrecy. A strategy to solve this issue could be to test different disaggregation approaches and choose the one performing better on an independent test set with reliable labels, which in this case should be composed only of instances referred to the years 2009 – 2012.

This brings to the issue of data availability and collection, since the data points during the PCF project are the only ones reporting a non-null value of the payments and therefore cannot be used only for testing but are required also for training. Their number however is intrinsically limited by the fact that the payments were provided only for 4 years. This was the reason for which leaving out a set to test the ABM on independent data was considered unfeasible and therefore not done. The learning curves of the ML models justified this choice, showing how all the data available were required for training them and even more would have been needed. In particular, the fact that the features with reliable labels and non-null payments values were the last 4 years of the 17 considered hampered the possibility to develop a predictive model, since an evaluation on independent data split time-wise is a requirement often cited for data-driven models aiming at anticipating unknown conditions and therefore study the outcome of future policies [85], [118]. For this and for its necessity in order to test a better disaggregation, the lack of evaluation of the Municipality-based Data-driven ABM on data in the future in relation to the data used for training constitutes the second main limitation of the municipality-based approach and a further problem for both the extrapolation of yearly adoption done until 2020 and future work related to its use for new policies design. Even if commercial data on adoption after 2012 could be retrieved, these would with most probability be provided only at a similar aggregation level of the ones previous to the PCF project since they were not collected in the scope of any project and therefore would be unlikely to include all the features necessary for a reliable test set. Moreover, the payments variable would be 0 and therefore could not be used to test the reliability of the ABM estimations under the effect of incentives.

These limitations in terms of data suggest that a large improvement of the Municipality-based Data-driven ABM performance is difficult. There are however other issues with the features used that could be instead addressed more easily and, despite having a lower impact than the disaggregated adoption, could positively impact the Municipality-based Data-driven ABM and in particular its lack of timewise variability highlighted in section 5.1.2.1. An example is the use of census features referred to 1999. Even though socio-economic and demographic characteristics do not usually change abruptly, especially in the agricultural context, 17 years are a timeframe long enough to introduce important changes. The impossibility to retrieve the census data for 2009 did not allow to consider these changes for instance through interpolation and forced to leave them static. Climate and soil features also introduced important simplifications. Apart from the intrinsic uncertainty that these maps presented, their calculations as averages over the entire municipalities meant to also include areas unrelated to pastures. For instance, the soil samples were collected not only from grasslands but also from croplands, woodlands and shrublands. Regarding climate features, only their averages over the entire period and average yearly values were considered. Features more focused on Autumn, the period when farmers usually decide if to adopt or not SBP, could be introduced and help in explaining some of the ups and downs of the yearly adoption in Portugal.

#### 5.2.2.2 Assessment of the PCF project outcome

The use of Municipality-based Data-driven ABM to asses the PCF project outcome was subjected to some additional limitations. First, the comparison of the observed yearly adoption in Portugal during the PCF project with the counterfactual simulation results underestimated the additional yearly adoption thanks to the payments. In fact, while the observed data report only successful installations, the counterfactual simulation was obtained through ML models trained also on adoption data previous to the PCF project, which include failed installations as well. Therefore some of the installations corresponding to the counterfactual simulation that would have failed are included in its estimation, meaning that a lower value of C would have been stored without the PCF project. However, because during the PCF project farmers received technical support, the probability of failure was much lower than it would have been in the counterfactual case of no project. Second, the lack of adoption data after 2012 forced the use of a simulation to estimate the yearly adoption after 2013 considering the PCF project as well and not only for the counterfactual. Third, as specified in the previous section, the lack of testing of the ABM on independent data and the biases introduced by the disaggregation of the adoption previous to the PCF make the extrapolations of the Municipality-based Data-driven ABM subject to errors.

#### 5.2.2.3 ABP adoption drivers at the municipality level

The reliability of the analysis on how the ML models predicted SBP adoption, aimed at understanding which factors drove SBP adoption and how, was hampered by the need to disaggregate the adoption previous to the PCF and in particular by the assignment of a small adoption to a lot of municipalities before 2009. This constituted a problem in particular for the classification, since many municipalities that

did not present adoption in reality were wrongly labelled with adoption in the dataset[61]. However, due to the focus of the approach on the municipality level, insights from the classification stage are less significative than from the regression stage, as already justified in section 5.1.2.3.

Both models however suffered issues surrounding the assessment of the influence of the features related to adoption in the previous years, which can explain the unrealistic results obtained. First, the importance of the adoption in and until the previous year in the municipality was inflated by the disaggregation of the adoption. In fact, this prioritized the municipalities which adopted during the PCF project, which can also explain why the quadratic trend with the total cumulative adoption in the municipality was not learnt by the model: the adoption assigned to these municipalities before the PCF project was probably higher than the real one and, despite this, they kept increasing their adoption even more after 2008. The disaggregation also caused that before 2009 the same municipalities adopted over all the years, with the exception of the years when there was no adoption in the corresponding region. Second, the disaggregation also caused the clear discontinuity before and after 2009 in the yearly number of municipalities with adoption and the average area of SBP installed observed in Figure 10, which the models could explain only through features regarding adoption, the only variable that changes over the years (together with the payments value for the regressor).

The unrealistic influences of the adoption features are however also explained by the lack of a variable linked to the biophysical capability of the land, an issue already presented for the farmer-based approach regarding feed savings (section 5.2.1.2). Local yields influence the expectations of farmers regarding the productivity of SBP and this is in fact a factor that could be determinant in the decision of farmers to adopt or not SBP and that could explain the reduction in adoption observed between 2005 and the beginning of the PCF project. Depending on various factors, the biophysical capability of a municipality is unlikely to depend linearly on its fraction of pastures area in which SBP was already installed, probably presenting a more complex behaviour and therefore requiring separate consideration. The lack of such variable is the third main limitation of the municipality-based approach. In its absence, the model had to rely only on the features regarding adoption to explain the reduction of adoption before the PCF project. Considering the biophysical capacity of the municipalities could free the adoption features from this role and allow them to be a proxy only for peer influence and trust in the system. Moreover, marginal and average biophysical capacity could be important predictive features that improves the performance of the ABM as a whole.

The sharp decrease of the yearly number of municipalities after 2009 due to the disaggregation of the adoption previous to the PCF project explains the unrealistic negative correlation of the payments value with the presence of adoption in the municipality. This brought to its exclusion from the dataset to train the classifier, to avoid associating the presence of incentives with a decreased number of municipalities with adoption. Due to the sharp increase in average adoption in the municipalities presenting it, the risk for the regression stage was instead to overestimate the importance of the payments. However, the fact

[61] The regressor probably was less influenced by this issue, since there is little difference in instances presenting a null or really small value, apart from the fact that the second were included in the regression dataset while the first no. This avoided this issue to hamper the significance of the Municipality-based Data-driven ABM's estimation especially at the macro-level, since the area adopted was the final variable of interest.

that the average area adopted after 2013 remains in the same order of magnitude as during the project (Figure 10.b) implies that this did not happen.

The evaluation of the influence of the features obtained from the census, climate and soil data faced also the problems already highlighted in the last paragraph of section 5.2.2.1. The difficulties related to obtaining reliable insights can more generally be linked to the choice of using a data-driven ABM for the municipality-based approach. This enabled the exploitation and study of all the different sources of data available, but at the same time complicated the analysis of the ABM estimations, particularly in combination with the issues regarding data availability. The implementation of alternative techniques to analyse the outcome of the data-driven ABM requires careful planning and a longer time execution time than the time available for completing the present thesis.

## 5.3 Future work

The interpretation of the results and the limitations identified point towards a clear and detailed plan for the work ahead. The activities of this plan configure a proposal for the next steps towards the refinement of the conclusions and achievements of the objectives of this thesis.

The most important single improvement for both approaches used is the consideration of the biophysical capacity of the land. For the farmer-based approach, it could help to properly assess the expected savings of feed, a decisive factor for the economic benefit of adopting SBP. For the municipality-based approach, it could be a key variable to improve the model performance and in particular assess more reliably the influence of the features related to adoption and therefore of peer influence. Moreover, it could also help in the estimation of the number of failed installations, which is likely to be a deterrent for further adoption and an important biophysical feedback. The estimation of such a variable, due to the many factors it depends on already presented in section 5.2.1.2, would require an independent study and the development and calibration of the biophysical model as the first step. At the moment such a model does not exist.

The Municipality-based Data-driven ABM would also greatly benefit from a better disaggregation of the adoption data for the period previous to the PCF project. As explained in section 5.2.2.1, despite the absence of further data to perform this disaggregation at the municipality-level, dedicated questionnaires could be designed to obtain more information on where the adoption before 2009 was concentrated. These questionnaires could be made to a sample of farmers or to farmer advisors who usually have a good grasp on the hotspots of adoption of SBP. Also, some municipalities where it is known that there was no adoption to a relevant extent could be directly excluded from the analysis.

The most interesting way forward consists in an approach merging the spatial and temporal scope and modelling framework of the municipality-based approach with the level of agency used in the farmer-based approach, which allows for insights on the single farmers' behaviour. This will be possible by retrieving census data at the individual farmer level for 2009, with some caveats to consider first. These census data would in fact miss the fundamental information on which farmers adopted SBP and when, information that is included in the PCF project dataset for all farmers who adopted between 2009 and 2012. However, the impossibility to pinpoint individual farmers in the census microdata due to privacy

and data protection reasons would not allow to directly match the farmers in this dataset with the ones that took part in the PCF project. To solve this issue some variables contained in both datasets, such as the municipality where farmers are located and the size of the farms, could be used.

Another issue is that the PCF project database is only referred to the years from 2009 to 2012 when payments were always provided. The large number of datapoints that would be available at the farmer-level would allow to properly divide timewise the dataset and test the resulting model on independent data. This could also provide the possibility to test different disaggregation rules of the adoption prior to the PCF project to obtain the level of adoption in the individual municipalities before 2008. The same strategy could be tried to assign this adoption to the individual farmers, since reliable individual data on which farmers adopted would most likely not be made available due to commercial secrecy. The possibility to test the model estimations could also make the model more reliable for policy design through IA and scenarios analysis, together with a proper use of uncertainty, sensibility analysis and ML techniques to analyse its outputs.

Moreover, participatory simulations involving directly the farmers and other important stakeholders and surveys tailored to the objective of the work could complement the abundance of quantitative data with more qualitative insights. These could then be used to elicit behavioural rules to integrate in the model, as in the work done by Sun & Müller [69] for instance, or understand how information diffuse through networks of farmers and explicitly model local interactions among the agents, through for instance the inclusion of farmers organizations, important sources of information for Portuguese farmers [16]. Most importantly these data collection methods could provide already on their own precious insights on which factors drive SBP adoption and reduce the number of features included in the model – which is advised by Edmonds et al. [118] in case of developing predictive models. The integration of reliable and tested theoretical rules in this future ABM should in fact be considered, since it could help to obtain clearer and more immediate insights on the drivers guiding SBP adoption, which as highlighted in section 5.2.2.3 can be complicated for data-driven ABM.

Successfully addressing the issues of the method presented here would provide a model suited to support the assessment and design of new policies aimed at further spreading SBP in Portugal. To do this properly, the model should try to limit the provision of ineffective payments, i.e. of payments to farmers that would install the system even without. Characterizing the various categories of farmers that adopted before and during the PCF project as divided in section 5.1.2.2 could help to design policies tailored to the farmers that will not adopt if payments are not provided. Moreover, the evaluation of the area in each municipality where SBP were not installed yet (and of its biophysical capability) would indicate if new policies should be aimed also at fostering resowing and not only new installations.

On a broader scope, to reach the objectives of the LEAnMeat project, from which this thesis originated, the assessment of the C sequestration thanks to SBP implemented in this thesis should be expanded to become a proper LCA of the system and evaluate its environmental effects more holistically, on the line of the raising interest of its combination with ABM highlighted in section 2.4.4. The last and most long-term aim would be to expand the scope of the study to other land uses and evaluate the adoption of SBP in terms of choice between them and not only SNP, but all other possibilities farmers have.

# 6 Conclusions

This thesis tested a variety of approaches to model SBP adoption. The farmer-based approach showed how simplified and uniform economic calculations are not suited to represent the individual farmer's decision-making. The analysis could not clarify if this is due to the large heterogeneity of farms that cannot be captured due to impossibility of collecting all the required data, or due to the fact that economic considerations are actually of secondary importance in respect to other reasons for LUCC. In both cases, the implication is that to model farmers' behaviour regarding SBP adoption other factors require consideration. Moreover, the Farmer-based Logistic Regression showed that disregarding economic considerations explicitly does not translate in a decrease in model's performance. All the farmer-based approaches however overestimated adoption. This suggests that socio-economic characteristics and farms management practices, the only variables available in the farmer-level survey data, are insufficient for evaluating farmers' decision-making.

Through the use of a greater variety of data sources, the municipality-based approach was able to include interactions among farmers, such as peer influence through the previous adoption of SBP, and between them and their environment, through climate and soil variables. These allowed to consider the main causes of LUCC systems' complexity and harness the strongest analytic capability of ABMs. The resulting Municipality-based Data-driven ABM – the first encompassing all Portugal in the context of LUCC – captured the underlying trend of adoption in the country, which, through yearly variations, showed a logistic rise in adoption until 2004, followed by a linear decrease interrupted by the PCF project which caused an important increase of installations in the period 2009 – 2012. The model exhibited also a good fit at the individual farmers level. These results confirmed that treating the system as a CAS and including proxies for the interactions within it allowed to capture feedback loops and emergent properties and therefore to better represent it.

The Municipality-based Data-driven ABM constitutes also the first ABM, among the literature assessed here, regarding innovation diffusion and policy design in agricultural systems that relies entirely on ML algorithms without combining them with any explicit theoretical assumption. Despite the limitations faced in terms of data availability, this approach was successful in estimating overall adoption while avoiding the need for formulating and testing theoretical assumptions. However, with more time and resources available, the analysis could benefit from the development and integration of sound and reliable rules on agents' behaviour and how information diffuse in the system, especially in terms of obtaining a further understanding of which factors drive adoption. These rules could be elicited through dedicated surveys and interviews.

In this regard, the importance of the level of education of the farmers clearly emerges from all the approaches. More educated farmers are likely to be more informed about innovative practices and could also be more sensible to the sustainability value of the pastures. The average weather of the location also has a role, with warm climates without temperature peaks and places that are not too rainy being the conditions favouring the most adoption – while the exclusion of the weather in the previous year due to its lower correlation means that farmers do not consider it relevant to predict the weather in the

following year. Other socio-economic factors and soil variables instead seems to have reduced influence on farmers' decisions. The driver which had by large the most evident effect is the previous diffusion of SBP and therefore peer influence. Each new farmer that installs SBP increases peer influence and trust in the system in the short-term and therefore adoption. To better evaluate the long-term trend of adoption, the development of a model evaluating the biophysical capability of the soil and able to capture the saturation of the land suitable for SBP adoption would be a key issue.

The analysis for the first time estimated the area of SBP that would have been adopted if no PCF project had taken place through a validated modelling approach. The resulting estimation was higher than the one which the project was based on, implying that a lower value of additional area of SBP was installed during the PCF project than previously thought. The comparison was expanded until 2020, forecasting that the project also caused a reduced adoption after it ended compared to the counterfactual. These results suggest that the payments per hectare of SBP installed for additional C sequestration offered during the PCF project would have been lower if the results of this thesis had been used to design the project, for two reasons: more SBP area would have been adopted during the years of the project if this had not taken place and a fraction of the area adopted during the project would have been adopted in the following years.

This confirms that the design and IA of future effective policies aimed at expanding SBP adoption should be supported by reliable quantitative insights. The combination of the two approaches used in this thesis can be a sound basis for a model suited to provide such insights. The starting point would be to apply the developed modelling framework for the Municipality-based Data-driven ABM to the farmer-level, when census data with this granularity will be available. To provide more reliable forecasts and increase trust in its robustness, this wide-scope farmer-based model should be tested on independent and future data respect to the ones used for training it and resort to uncertainty and sensibility analysis..

# References


[1] W. Steffen *et al.*, 'Planetary boundaries: Guiding human development on a changing planet', *Science*, vol. 347, no. 6223, pp. 1259855–1259855, Feb. 2015, doi: 10.1126/science.1259855.

[2] K. Raworth, 'A safe and Just space for humanity: Can we live within the doughnut?', Oxfam, Feb. 2012. doi: 10.1163/2210-7975_HRD-9824-0069.

[3] D. W. O'Neill, A. L. Fanning, W. F. Lamb, and J. K. Steinberger, 'A good life for all within planetary boundaries', *Nat Sustain*, vol. 1, no. 2, pp. 88–95, Feb. 2018, doi: 10.1038/s41893-018-0021-4.

[4] FAO, UNICEF, WFP, WHO, and IFAD, 'Building climate resilience for food security and nutrition', Food and Agriculture Organization of the United Nations (FAO), Rome, 2018.

[5] W. Willett *et al.*, 'Food in the Anthropocene: the EAT–Lancet Commission on healthy diets from sustainable food systems', *The Lancet*, vol. 393, no. 10170, pp. 447–492, Feb. 2019, doi: 10.1016/S0140-6736(18)31788-4.

[6] United Nations, 'World Population Prospects 2019', Dept of Economic and Social Affairs, 2019. Accessed: Apr. 24, 2020. [Online]. Available: https://population.un.org/wpp/Graphs/Probabilistic/POP/TOT/900.

[7] S. J. Vermeulen, B. M. Campbell, and J. S. I. Ingram, 'Climate Change and Food Systems', *Annu. Rev. Environ. Resour.*, vol. 37, no. 1, pp. 195–222, Nov. 2012, doi: 10.1146/annurev-environ-020411-130608.

[8] D. Tilman, M. Clark, D. R. Williams, K. Kimmel, S. Polasky, and C. Packer, 'Future threats to biodiversity and pathways to their prevention', *Nature*, vol. 546, no. 7656, pp. 73–81, Jun. 2017, doi: 10.1038/nature22900.

[9] P. J. Gerber *et al.*, *Tackling climate change through livestock: a global assessment of emissions and mitigation opportunities*. Rome: Food and Agriculture Organization of the United Nations (FAO), 2013.

[10] R. F. M. Teixeira and T. Domingos, 'Current Practice and Future Perspectives for Livestock Production and Industrial Ecology', *Sustainability*, vol. 11, no. 15, Art. no. 15, Jan. 2019, doi: 10.3390/su11154210.

[11] N. Alexandratos and J. Bruinsma, 'World agriculture towards 2030/2050: the 2012 revision', Food and Agriculture Organization of the United Nations (FAO), Rome, ESA Working paper No. 12-03, 2012.

[12] European Environment Agency, 'Portugal land cover country fact sheet 2012', Feb. 07, 2017. https://www.eea.europa.eu/themes/landuse/land-cover-country-fact-sheets/pt-portugal-landcover-2012.pdf/view (accessed May 29, 2020).

[13] Instituto Nacional de Estatistica, 'Superfície ($km^2$) das unidades territoriais por Localização geográfica (NUTS - 2013) e Classes de uso e ocupação do solo; Não periódica', Nov. 28, 2017. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0009776&contexto=bd&selTab=tab2 (accessed May 30, 2020).

[14] R. F. M. Teixeira, V. Proença, D. Crespo, T. Valada, and T. Domingos, 'A conceptual framework for the analysis of engineered biodiverse pastures', *Ecological Engineering*, vol. 77, pp. 85–97, Apr. 2015, doi: 10.1016/j.ecoleng.2015.01.002.

[15] T. Pinto-Correia, N. Ribeiro, and P. Sá-Sousa, 'Introducing the montado, the cork and holm oak agroforestry system of Southern Portugal', *Agroforest Syst*, vol. 82, no. 2, p. 99, Apr. 2011, doi: 10.1007/s10457-011-9388-1.

[16] A. L. Acosta, D. A. M. Rounsevell, M. Bakker, A. Van Doorn, M. Gómez-Delgado, and M. Delgado, 'An agent-based assessment of land use and ecosystem changes in traditional agricultural landscape of Portugal', *IIM*, vol. 06, no. 02, pp. 55–80, 2014, doi: 10.4236/iim.2014.62008.

[17] M. N. Bugalho, M. C. Caldeira, J. S. Pereira, J. Aronson, and J. G. Pausas, 'Mediterranean cork oak savannas require human use to sustain biodiversity and ecosystem services', *Frontiers in Ecology and the Environment*, vol. 9, no. 5, pp. 278–286, Jun. 2011, doi: 10.1890/100084.

[18] T. G. Morais, R. F. M. Teixeira, and T. Domingos, ‘Detailed global modelling of soil organic carbon in cropland, grassland and forest soils’, *PLoS One*, vol. 14, no. 9, Sep. 2019, doi: 10.1371/journal.pone.0222604.

[19] R. F. M. Teixeira, ‘Sustainable Land Uses and Carbon Sequestration : The Case of Sown Biodiverse Permanent Pastures Rich in Legumes.’, PhD Thesis, Instituto Superior Técnico - Universidade de Lisboa, Lisbon, Portugal, 2010.

[20] T. G. Morais, R. F. M. Teixeira, and T. Domingos, ‘The Effects on Greenhouse Gas Emissions of Ecological Intensification of Meat Production with Rainfed Sown Biodiverse Pastures’, *Sustainability*, vol. 10, no. 11, p. 4184, Nov. 2018, doi: 10.3390/su10114184.

[21] G. V. Duinen, ‘How a century of ammonia synthesis changed the world’, *nature geoscience*, vol. 1, p. 4, 2008.

[22] L. Lassaletta, G. Billen, B. Grizzetti, J. Anglade, and J. Garnier, ‘50 year trends in nitrogen use efficiency of world cropping systems: the relationship between yield and nitrogen input to cropland’, *Environ. Res. Lett.*, vol. 9, no. 10, p. 105011, Oct. 2014, doi: 10.1088/1748-9326/9/10/105011.

[23] X. Zhang, E. A. Davidson, D. L. Mauzerall, T. D. Searchinger, P. Dumas, and Y. Shen, ‘Managing nitrogen for sustainable development’, *Nature*, vol. 528, no. 7580, pp. 51–59, Dec. 2015, doi: 10.1038/nature15743.

[24] L. K. Boerner, ‘Industrial ammonia production emits more CO2 than any other chemical-making reaction. Chemists want to change that.’, vol. 97, no. 24, Jul. 15, 2019.

[25] B. L. Bodirsky *et al.*, ‘Reactive nitrogen requirements to feed the world in 2050 and potential to mitigate nitrogen pollution’, *Nat Commun*, vol. 5, no. 1, p. 3858, Sep. 2014, doi: 10.1038/ncomms4858.

[26] K. M. Carlson *et al.*, ‘Greenhouse gas emissions intensity of global croplands’, *Nature Clim Change*, vol. 7, no. 1, pp. 63–68, Jan. 2017, doi: 10.1038/nclimate3158.

[27] J. N. Galloway *et al.*, ‘Transformation of the Nitrogen Cycle: Recent Trends, Questions, and Potential Solutions’, *Science*, vol. 320, no. 5878, p. 889, May 2008, doi: 10.1126/science.1136674.

[28] H. Tian *et al.*, ‘A comprehensive quantification of global nitrous oxide sources and sinks’, *Nature*, vol. 586, no. 7828, pp. 248–256, Oct. 2020, doi: 10.1038/s41586-020-2780-0.

[29] R. F. M. Teixeira, T. G. Morais, and T. Domingos, ‘A Practical Comparison of Regionalized Land Use and Biodiversity Life Cycle Impact Assessment Models Using Livestock Production as a Case Study’, *Sustainability*, vol. 10, no. 11, p. 4089, Nov. 2018, doi: 10.3390/su10114089.

[30] C. S. Holling, ‘Understanding the Complexity of Economic, Ecological, and Social Systems’, *Ecosystems*, vol. 4, no. 5, pp. 390–405, Aug. 2001, doi: 10.1007/s10021-001-0101-5.

[31] S. Levin *et al.*, ‘Social-ecological systems as complex adaptive systems: modeling and policy implications’, *Envir. Dev. Econ.*, vol. 18, no. 2, pp. 111–132, Apr. 2013, doi: 10.1017/S1355770X12000460.

[32] J. Liu *et al.*, ‘Complexity of Coupled Human and Natural Systems’, *Science*, vol. 317, no. 5844, pp. 1513–1516, Sep. 2007, doi: 10.1126/science.1144004.

[33] R. Preiser, R. Biggs, A. De Vos, and C. Folke, ‘Social-ecological systems as complex adaptive systems: organizing principles for advancing research methods and approaches’, *E&S*, vol. 23, no. 4, p. art46, 2018, doi: 10.5751/ES-10558-230446.

[34] D. C. Mikulecky, ‘The emergence of complexity: science coming of age or science growing old?’, *Computers & Chemistry*, vol. 25, no. 4, pp. 341–348, Jul. 2001, doi: 10.1016/S0097-8485(01)00070-5.

[35] M. M. Waldorp, *Complexity: The emerging science at the edge of order and chaos*. New York: Simon and Schuster, 1992.

[36] K. H. Dam, I. Nikolic, and Z. Lukszo, Eds., *Agent-based modelling of socio-technical systems*. Dordrecht: Springer Netherlands, 2013.

[37] E. Lorenz, ‘Predictability: Does the Flap of a Butterfly’s Wings in Brazil Set Off a Tornado in Texas?’”, presented at the AAAS Section on Environmental Sciences, New Approaches to Global Weather: GARP, Sheraton Park Plaza Hotel, Boston, 1972.

[38] C. Folke, S. R. Carpenter, B. Walker, M. Scheffer, T. Chapin, and J. Rockström, ‘Resilience Thinking: Integrating Resilience, Adaptability and Transformability’, *Ecology and Society*, vol. 15, no. 4, 2010, Accessed: May 01, 2020. [Online]. Available: https://www.jstor.org/stable/26268226.

[39] V. Gaube and H. Haberl, ‘Using Integrated Models to Analyse Socio-ecological System Dynamics in Long-Term Socio-ecological Research – Austrian Experiences’, in *Long Term Socio-Ecological Research*, S. J. Singh, H. Haberl, M. Chertow, M. Mirtl, and M. Schmid, Eds. Dordrecht: Springer Netherlands, 2013, pp. 53–75.

[40] J. M. Epstein, ‘Agent-based computational models and generative social science’, *Complexity*, vol. 4, no. 5, pp. 41–60, 2006.

[41] C. M. Macal, ‘Everything you need to know about agent based modelling and simulation’, *Journal of Simulation*, vol. 10, no. 2, pp. 144–156, 2016, doi: 10.1057/jos.2016.7.

[42] T. C. Schelling, ‘Dynamic models of segregation’, *The Journal of Mathematical Sociology*, vol. 1, no. 2, pp. 143–186, Jul. 1971, doi: 10.1080/0022250X.1971.9989794.

[43] M. Gardner, ‘The fantastic combinations of John Conway’s new solitaire game “life”’, vol. Scientific American, no. 223, pp. 120–123, 1970.

[44] J. M. Epstein and R. Axtell, *Growing artificial societies: Social science from the bottom up*. Cambridge, MA: MIT Press, 1996.

[45] U. Wilensky and W. Rand, *An introduction to agent-based modeling: Modeling natural, social, and engineered complex systems with NetLogo*. The MIT Press, 2015.

[46] S. Abar, G. K. Theodoropoulos, P. Lemarinier, and G. M. P. O’Hare, ‘Agent based modelling and simulation tools: A review of the state-of-art software.’, *Computer Science Review*, vol. 24, pp. 13–33, May 2017, doi: 10.1016/j.cosrev.2017.03.001.

[47] H. Zhang and Y. Vorobeychik, ‘Empirically grounded agent-based models of innovation diffusion: a critical review’, *Artif Intell Rev*, vol. 52, no. 1, pp. 707–741, Jun. 2019, doi: 10.1007/s10462-017-9577-z.

[48] R. R. Rindfuss *et al.*, ‘Land use change: complexity and comparisons’, *Journal of Land Use Science*, vol. 3, no. 1, pp. 1–10, Jul. 2008, doi: 10.1080/17474230802047955.

[49] D. C. Parker, S. M. Manson, M. A. Janssen, M. J. Hoffmann, and P. Deadman, ‘Multi-Agent Systems for the Simulation of Land-Use and Land-Cover Change: A Review’, *Annals of the Association of American Geographers*, vol. 93, no. 2, pp. 314–337, Jun. 2003, doi: 10.1111/1467-8306.9302004.

[50] D. T. Robinson *et al.*, ‘Comparison of empirical methods for building agent-based models in land use science’, *Journal of Land Use Science*, vol. 2, no. 1, pp. 31–55, Apr. 2007, doi: 10.1080/17474230701201349.

[51] I. Dullinger *et al.*, ‘A socio-ecological model for predicting impacts of land-use and climate change on regional plant diversity in the Austrian Alps’, *Glob Change Biol*, p. gcb.14977, Jan. 2020, doi: 10.1111/gcb.14977.

[52] R. B. Matthews, N. G. Gilbert, A. Roach, J. G. Polhill, and N. M. Gotts, ‘Agent-based land-use models: a review of applications’, *Landscape Ecol*, vol. 22, no. 10, pp. 1447–1459, Nov. 2007, doi: 10.1007/s10980-007-9135-1.

[53] P. Schreinemachers and T. Berger, ‘An agent-based simulation model of human–environment interactions in agricultural systems’, *Environmental Modelling & Software*, vol. 26, no. 7, pp. 845–859, Jul. 2011, doi: 10.1016/j.envsoft.2011.02.004.

[54] J. Groeneveld *et al.*, ‘Theoretical foundations of human decision-making in agent-based land use models – A review’, *Environmental Modelling & Software*, vol. 87, pp. 39–48, Jan. 2017, doi: 10.1016/j.envsoft.2016.10.008.

[55] D. O'Sullivan, T. Evans, S. Manson, S. Metcalf, A. Ligmann-Zielinska, and C. Bone, 'Strategic directions for agent-based modeling: avoiding the YAAWN syndrome', *Journal of Land Use Science*, vol. 11, no. 2, pp. 177–187, Mar. 2016, doi: 10.1080/1747423X.2015.1030463.

[56] M. M. Bakker and A. M. van Doorn, 'Farmer-specific relationships between land use change and landscape factors: Introducing agents in empirical land use modelling', *Land Use Policy*, vol. 26, no. 3, pp. 809–817, Jul. 2009, doi: 10.1016/j.landusepol.2008.10.010.

[57] D. Murray-Rust, D. T. Robinson, E. Guillem, E. Karali, and M. Rounsevell, 'An open framework for agent based modelling of agricultural land use change', *Environmental Modelling & Software*, vol. 61, pp. 19–38, Nov. 2014, doi: 10.1016/j.envsoft.2014.06.027.

[58] N. Bichraoui-Draper, M. Xu, S. A. Miller, and B. Guillaume, 'Agent-based life cycle assessment for switchgrass-based bioenergy systems', *Resources, Conservation and Recycling*, vol. 103, pp. 171–178, Oct. 2015, doi: 10.1016/j.resconrec.2015.08.003.

[59] T. Berger, 'Agent-based spatial models applied to agriculture: a simulation tool for technology diffusion, resource use changes and policy analysis', *Agricultural Economics*, p. 16, 2001.

[60] P. Alexander, D. Moran, M. D. A. Rounsevell, and P. Smith, 'Modelling the perennial energy crop market: the role of spatial diffusion', *J. R. Soc. Interface.*, vol. 10, no. 88, p. 20130656, Nov. 2013, doi: 10.1098/rsif.2013.0656.

[61] E. Kiesling, M. Günther, C. Stummer, and L. M. Wakolbinger, 'Agent-based simulation of innovation diffusion: a review', *Cent Eur J Oper Res*, vol. 20, no. 2, pp. 183–230, Jun. 2012, doi: 10.1007/s10100-011-0210-y.

[62] P. Reidsma, S. Janssen, J. Jansen, and M. K. van Ittersum, 'On the development and use of farm models for policy impact assessment in the European Union – A review', *Agricultural Systems*, vol. 159, pp. 111–125, Jan. 2018, doi: 10.1016/j.agsy.2017.10.012.

[63] T. Berger and C. Troost, 'Agent-based Modelling of Climate Adaptation and Mitigation Options in Agriculture', *J Agric Econ*, vol. 65, no. 2, pp. 323–348, Jun. 2014, doi: 10.1111/1477-9552.12045.

[64] K. Happe, K. Kellermann, and A. Balmann, 'Agent-based Analysis of Agricultural Policies: an Illustration of the Agricultural Policy Simulator AgriPoliS, its Adaptation and Behavior', *E&S*, vol. 11, no. 1, p. art49, 2006, doi: 10.5751/ES-01741-110149.

[65] D. Kremmydas, I. N. Athanasiadis, and S. Rozakis, 'A review of Agent Based Modeling for agricultural policy evaluation', *Agricultural Systems*, vol. 164, pp. 95–106, Jul. 2018, doi: 10.1016/j.agsy.2018.03.010.

[66] R. Lempert, 'Agent-based modeling as organizational and public policy simulators', *Proc Natl Acad Sci USA*, vol. 99, no. suppl 3, p. 7195, May 2002, doi: 10.1073/pnas.072079399.

[67] V. Gaube *et al.*, 'Combining agent-based and stock-flow modelling approaches in a participative analysis of the integrated land system in Reichraming, Austria', *Landscape Ecol*, vol. 24, no. 9, pp. 1149–1165, Nov. 2009, doi: 10.1007/s10980-009-9356-6.

[68] B. Fisher, R. K. Turner, and P. Morling, 'Defining and classifying ecosystem services for decision making', *Ecological Economics*, vol. 68, no. 3, pp. 643–653, Jan. 2009, doi: 10.1016/j.ecolecon.2008.09.014.

[69] Z. Sun and D. Müller, 'A framework for modeling payments for ecosystem services with agent-based models, Bayesian belief networks and opinion dynamics models', *Environmental Modelling & Software*, vol. 45, pp. 15–28, Jul. 2013, doi: 10.1016/j.envsoft.2012.06.007.

[70] L. An *et al.*, 'Cascading Impacts of Payments for Ecosystem Services in Complex Human-Environment Systems', *JASSS*, vol. 23, no. 1, p. 5, 2020, doi: 10.18564/jasss.4196.

[71] X. Chen, F. Lupi, L. An, R. Sheely, A. Viña, and J. Liu, 'Agent-based modeling of the effects of social norms on enrollment in payments for ecosystem services', *Ecological Modelling*, vol. 229, pp. 16–24, Mar. 2012, doi: 10.1016/j.ecolmodel.2011.06.007.

[72] X. Chen, A. Viña, A. Shortridge, L. An, and J. Liu, 'Assessing the Effectiveness of Payments for Ecosystem Services: an Agent-Based Modeling Approach', *E&S*, vol. 19, no. 1, p. art7, 2014, doi: 10.5751/ES-05578-190107.

[73] M. Hare and P. Deadman, ‘Further towards a taxonomy of agent-based simulation models in environmental management’, *Mathematics and Computers in Simulation*, vol. 64, no. 1, pp. 25–40, Jan. 2004, doi: 10.1016/S0378-4754(03)00118-6.

[74] R. A. Kelly (Letcher) *et al.*, ‘Selecting among five common modelling approaches for integrated environmental assessment and management’, *Environmental Modelling & Software*, vol. 47, pp. 159–181, Sep. 2013, doi: 10.1016/j.envsoft.2013.05.005.

[75] A. Lobanova, F. Lamperti, A. Roventini, D. Tabara, S. Liersch, and V. Krysanova, ‘Exploring socio-hydrological dynamics with a hybrid hydrological agent-based model’, Tolouse, France, Jul. 2016, p. 3.

[76] ISO 14040, ‘Environmental management — Life cycle assessment — Principles and framework’, International Organization for Standardization (ISO), Jun. 1997.

[77] A. Micolier, P. Loubet, F. Taillandier, and G. Sonnemann, ‘To what extent can agent-based modelling enhance a life cycle assessment? Answers based on a literature review’, *Journal of Cleaner Production*, vol. 239, p. 118123, Dec. 2019, doi: 10.1016/j.jclepro.2019.118123.

[78] A. Marvuglia, T. Navarrete Gutiérrez, P. Baustert, E. Benetto, and 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 Agriculture and Food*, vol. 3, no. 4, pp. 535–560, 2018, doi: 10.3934/agrfood.2018.4.535.

[79] T. Navarrete Gutiérrez, S. Rege, A. Marvuglia, and E. Benetto, ‘Sustainable Farming Behaviours: An Agent Based Modelling and LCA Perspective’, in *Agent-Based Modeling of Sustainable Behaviors*, A. Alonso-Betanzos, N. Sánchez-Maroño, O. Fontenla-Romero, J. G. Polhill, T. Craig, J. Bajo, and J. M. Corchado, Eds. Cham: Springer International Publishing, 2017, pp. 187–206.

[80] M. A. Janssen and E. Ostrom, ‘Empirically Based, Agent-based models’, *E&S*, vol. 11, no. 2, p. art37, 2006, doi: 10.5751/ES-01861-110237.

[81] A. Laatabi, N. Marilleau, T. Nguyen-Huu, H. Hbid, and M. Ait Babram, ‘ODD+2D: An ODD Based Protocol for Mapping Data to Empirical ABMs’, *JASSS*, vol. 21, no. 2, p. 9, 2018, doi: 10.18564/jasss.3646.

[82] H. Kavak, J. J. Padilla, C. J. Lynch, and S. Y. Diallo, ‘Big Data, Agents and Machine Learning: Towards a Data-Driven Agent-Based Modeling Approach’, presented at the 2018 Spring Simulation Multi-Conference, Baltimore, MD, USA, 2018, doi: 10.22360/SpringSim.2018.ANSS.021.

[83] L. An, ‘Modeling human decisions in coupled human and natural systems: Review of agent-based models’, *Ecological Modelling*, vol. 229, pp. 25–36, Mar. 2012, doi: 10.1016/j.ecolmodel.2011.07.010.

[84] P. Kaufmann, S. Stagl, and D. W. Franks, ‘Simulating the diffusion of organic farming practices in two New EU Member States’, *Ecological Economics*, vol. 68, no. 10, pp. 2580–2593, Aug. 2009, doi: 10.1016/j.ecolecon.2009.04.001.

[85] H. Zhang, Y. Vorobeychik, J. Letchford, and K. Lakkaraju, ‘Data-driven agent-based modeling, with application to rooftop solar adoption’, *Auton Agent Multi-Agent Syst*, vol. 30, no. 6, pp. 1023–1049, Nov. 2016, doi: 10.1007/s10458-016-9326-8.

[86] J. Dahlke and K. Bogner, ‘Is the juice worth the squeeze? Machine Learning in and for Agent-Based Modelling’, p. 25, 2020.

[87] P. Domingos, ‘A few useful things to know about machine learning’, *Commun. ACM*, vol. 55, no. 10, pp. 78–87, Oct. 2012, doi: 10.1145/2347736.2347755.

[88] M. I. Jordan and T. M. Mitchell, ‘Machine learning: Trends, perspectives, and prospects’, *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/science.aaa8415.

[89] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction.* The MIT Press, 1988.

[90] M. Pereda, J. I. Santos, and J. M. Galán, ‘A Brief Introduction to the Use of Machine Learning Techniques in the Analysis of Agent-Based Models’, in *Advances in Management Engineering*, C. Hernández, Ed. Cham: Springer International Publishing, 2017, pp. 179–186.

[91] X. Zhao, X. Ma, W. Tang, and D. Liu, ‘An adaptive agent-based optimization model for spatial planning: A case study of Anyue County, China’, *Sustainable Cities and Society*, vol. 51, p. 101733, Nov. 2019, doi: 10.1016/j.scs.2019.101733.

[92] F. Li, Z. Li, H. Chen, Z. Chen, and M. Li, ‘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*, vol. 95, p. 104620, Jun. 2020, doi: 10.1016/j.landusepol.2020.104620.

[93] R. F. M. Teixeira, ‘Economic incentives for carbon sequestration in grassland soils: An offer you cannot refuse’, MSc Thesis, Instituto Superior Técnico - Universidade de Lisboa, Lisbon, Portugal, 2008.

[94] J. P. Carneiro, R. C. Freixial, J. S. Pereira, A. C. Campos, J. P. Crespo, and R. Carneiro, ‘Relatório Final do Projecto AGRO 87 (“Final Report of the Agro 87 Project”, in Portuguese).’, Estação Nacional de Melhoramento de Plantas, Universidade de Évora, Instituto Superior de Agronomia, Direcção Regional de Agricultura do Alentejo, Fertiprado, Laboratório Químico Agrícola Rebelo da Silva., 2005.

[95] T. G. Morais, ‘Studies in quantitative environmental assessment of land use systems.’, PhD Thesis, Instituto Superior Técnico - Universidade de Lisboa, Lisbon, Portugal, 2021.

[96] IACA, ‘Anuário da IACA de 2011’, IACA (Associação Portuguesa dos Industriais de Alimentos Compostos para Animais), 2011.

[97] R. C. Cornes, G. van der Schrier, E. J. M. van den Besselaar, and P. D. Jones, ‘An Ensemble Version of the E-OBS Temperature and Precipitation Data Sets’, *Journal of Geophysical Research: Atmospheres*, vol. 123, no. 17, pp. 9391–9409, 2018, doi: 10.1029/2017JD028200.

[98] G. Tóth, A. Jones, L. Montanarella, European Commission, Joint Research Centre, and Institute for Environment and Sustainability, *LUCAS topsoil survey: methodology, data and results.* Luxembourg: Publications Office, 2013.

[99] C. Ballabio *et al.*, ‘Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression’, *Geoderma*, vol. 355, p. 113912, Dec. 2019, doi: 10.1016/j.geoderma.2019.113912.

[100] European Commission, Joint Research Centre (JRC), ‘Maps of soil chemical properties at European scale based on LUCAS 2009/2012 topsoil data [geo Tiff]’, 2019. Retrieved from https://esdac.jrc.ec.europa.eu/content/chemical-properties-european-scale-based-lucas-topsoil-data (accessed Jul. 26, 2020).

[101] Environmental System Research Institute, Inc, ‘ESRI Shapefile Technical Description’, Jul. 2008.

[102] Agência para a Modernização Administrativa (ama), ‘Concelhos de Portugal [shapefile]’, Nov. 12, 2018. Retrieved from https://dados.gov.pt/en/datasets/concelhos-de-portugal/ (accessed Jul. 12, 2020).

[103] International Monetary Fund, ‘International Financial Statistics and data files.’, *Consumer price index (2010 = 100) - Portugal*, 2020. https://data.worldbank.org/indicator/FP.CPI.TOTL?locations=PT (accessed Oct. 30, 2020).

[104] R. F. M. Teixeira, L. Barão, T. G. Morais, and T. Domingos, ‘The carbon and nitrogen ecological model that we mention in the project description, that we wish to couple with an ABM’, *Sustainability*, vol. 11, no. 1, p. 53, Dec. 2018, doi: 10.3390/su11010053.

[105] C. P. *et al.*, ‘Effect of low- and high-forage diets on meat quality and fatty acid composition of Alentejana and Barrosã beef breeds.’, *Animal*, vol. 6, pp. 1187–1197, 2012, doi: 10.1017/S1751731111002722.

[106] V. Grimm *et al.*, ‘A standard protocol for describing individual-based and agent-based models’, *Ecological Modelling*, vol. 198, no. 1–2, pp. 115–126, Sep. 2006, doi: 10.1016/j.ecolmodel.2006.04.023.

[107] B. Müller *et al.*, ‘Describing human decisions in agent-based models – ODD + D, an extension of the ODD protocol’, *Environmental Modelling & Software*, vol. 48, pp. 37–48, Oct. 2013, doi: 10.1016/j.envsoft.2013.06.003.

[108] V. Grimm *et al.*, ‘The ODD Protocol for Describing Agent-Based and Other Simulation Models: A Second Update to Improve Clarity, Replication, and Structural Realism’, *JASSS*, vol. 23, no. 2, p. 7, 2020, doi: 10.18564/jasss.4259.

[109] A. Géron, *Hands-on machine learning, with Scikit-Learn, Keras & TensorFlow*, 2nd ed. O’Reilly, 2019.

[110] J. G. Cragg, ‘Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods’, *Econometrica*, vol. 39, no. 5, p. 829, Sep. 1971, doi: 10.2307/1909582.

[111] J. Daoud, ‘Multicollinearity and Regression Analysis’, *Journal of Physics: Conference Series*, vol. 949, p. 012009, Dec. 2017, doi: 10.1088/1742-6596/949/1/012009.

[112] L. B. Lusted, ‘Signal detectability and medical decision-making’, *Science*, vol. 171, no. 3977, pp. 1217–1219, Mar. 1971.

[113] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, ‘Algorithms for Hyper-Parameter Optimization’, p. 9.

[114] S. M. Lundberg and S.-I. Lee, ‘A Unified Approach to Interpreting Model Predictions’, presented at the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017.

[115] T. C. Pereira, A. Amaro, M. Borges, R. Silva, A. Pina, and P. Canaveira, ‘Portuguese national inventory report on greenhouse gases, 1990 - 2018’, Portuguese Environmental Agency, Amadora, Apr. 2020.

[116] S. F. Crone, J. Guajardo, and R. Weber, ‘A study on the ability of Support Vector Regression and Neural Networks to Forecast Basic Time Series Patterns’, in *Artificial Intelligence in Theory and Practice*, vol. 217, M. Bramer, Ed. Springer US, 2006, pp. 149–158.

[117] A. Defazio, F. Bach, and S. Lacoste-Julien, ‘SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives’, *arXiv:1407.0202 [cs, math, stat]*, Dec. 2014, Accessed: Dec. 01, 2020. [Online]. Available: http://arxiv.org/abs/1407.0202.

[118] B. Edmonds *et al.*, ‘Different Modelling Purposes’, *Journal of Artificial Societies and Social Simulation*, vol. 22, no. 3, Jun. 2019, doi: 10.18564/jasss.3993.

# Appendix

## A. AF survey data

Table A.1 and Table A.2 in this section report respectively the entries of the AF survey data that were included and excluded from the following analysis, together with the rationale which guided the decision. Some farmers' attributes do not change with land use and therefore can be considered not to be influenced by a switch in pasture, such as farm area. Others instead can influence the decision since adopting SBP could affect them, as for the amount of labour or the revenues. The attributes belonging to this second category cannot be considered while studying adoption decisions in the past, since the values reported for 2019 were collected after the adoption and thus either they are different from the ones in the past or they do not have an influence on adoption. The rationale for excluding these attributes will be reported in the following paragraphs simply as "Data after adoption". The economic data were not used for two reasons. The first is that these are data collected after adoption. The second is that, since the models that make use of the survey data were all based on theoretical economic calculations of differential NPVs, empirical data were not needed.

*Table A.1: features considered from survey data and rationale.*

| Feature name | Meaning | Unit / Values | Rationale |
|---|---|---|---|
| **General data** | | | |
| *OwnLand* | Total land area of the farm | Hectares | To consider a possible economy of scale effect |
| *RentedLand* | Land area rented by the farmer | Hectares | To consider if owning the land or renting it has an effect on adoption |
| *LegalForm* | Legal form of the farmer | Individual, Sociedade Agrícola | Being associated with other farmers could affect peer influence |
| *FarmerSince* | Years of experience as farmer | Years | Experience may influence perception of innovation |
| *Distrito* | District where the farm is located | Any district in Portugal | Farmers can be influenced by the decision of their neighbours |
| *Concelho* | Municipality[62] where the farm is located | Any municipality in Portugal | Farmers can be influenced by the decision of their neighbours |
| **Social data** | | | |
| *Highest Educational Degree* | Highest education the farmer completed | Primary, Secondary, Undergraduate, Graduate | The education level of the farmer was considered a proxy for risk aversion of farmers in [93] |
| *Highest Agricultural Educational Degree* | Highest degree related to agriculture the farmer obtained | NaN, Undergraduate, Graduate | The education level of the farmer was considered a proxy for risk aversion of farmers in [93] |

[62] In the following work, municipality is used to indicate the Portuguese "município" or "concelho", the second-level division

| Feature name | Meaning | Unit / Values | Rationale |
|---|---|---|---|
| *Expectation Family Succession* | Expectation of the farmer on the future of the family. | 1: no expectation<br>2: inheritance within the family<br>3: sell the property<br>4: give up the tenancy<br>5: other | The possibility of a succession could increase the willingness for long-term investments |
| **Environmental data** | | | |
| *AREA_ID* | Land-use of the specific portion of land | Any possible land use | Required to understand the area dedicated to pasture and to SBP in particular |
| *Surface* | Area dedicated to the specific land use | Hectares | Required to understand the area dedicated to pasture and to SBP in particular |
| *Livestock Type* | Type of livestock | Cattle, sheep, goats | The fraction of cattle over total livestock could influence the decision on adopting SBP |
| *Average Number* | Average number of each livestock type kept in the farm | Number of animals | The fraction of cattle over total livestock could influence the decision on adopting SBP |

*Table A.2: features removed from survey data and rationale.*

| Features | Rationale |
|---|---|
| **General data** | |
| Quantification of work in the farm | Data after adoption |
| Overall satisfaction being a farmer | Data after adoption |
| Area under other agro-environmental measures | SBP adoption does not have an influence on the participation or not in these or if it has it would be impossible to know it |
| **Social data** | |
| Training days for employees | The farmer's level of education is considered more relevant than this indicator |
| Quality of life and labour | Data after adoption |
| Other activities besides farming | The adoption of SBP does not hamper nor foster any additional activity within or outside the farm |
| Participation in a labelling scheme | The adoption of SBP does not hamper nor foster the participation in labelling schemes |
| Expected economic viability of the farm in 10 years | Data after adoption |
| Data on animal welfare | Data after adoption |
| **Environmental data** | |
| Fraction of legumes for each land use | Depends on land use |
| Dry matter content in land | Data after adoption |
| Yield for each land use | Depends on land use |
| Crop protection agents used | Data after adoption |
| Fertilizers used | Data after adoption |
| Dead and sold animals characteristics | Data after adoption |
| Animal housing | The adoption of SBP does not have influence on animal housing |
| Water and energy consumption | Data after adoption |
| Animals diet composition | Data after adoption |

*"Data after adoption" as rationale refers to the fact that the variable is affected by the adoption of SBP and, being the data collected after the decision was made, their value in the database is probably different than the one they had before adoption.*

# B. Pastures costs

Table A.3 reports the variables required to calculated the costs for the installation and maintenance of one hectare of SBP. The cost of labour ($C_{lab}$) is 4.78 €/h [19] and soil correction is needed every 4 years and fertilization every 2 [19].

*Table A.3: variables required to calculated the costs for the installation and maintenance of one hectare of sown biodiverse pastures for 2019 [19].*

| | **Labour** | **Machinery** | | | **Materials** | |
|---|---|---|---|---|---|---|
| | **$h_{lab}$ [h]** | **CV [€/h]** | **CF [€]** | **$C_{amort}$ [€]** | **Q [kg]** | **P [€/kg]** |
| **Maintenance** | | | | | | |
| **Soil correction** | | | | | | |
| Lime | | | | | 2000 | 0.0655 |
| Lime transport | 1.7 | 8.88 | 7.37 | 6.8 | | |
| Lime application | 0.2 | 0.9 | 0.86 | 0.8 | | |
| **Fertilization** | | | | | | |
| Fertilizers transport | 1 | 2.34 | 1.95 | 1.79 | | |
| Superphosphate | | | | | 200 | 0.41 |
| Operation | 2 | 7.33 | 6.16 | 4.42 | | |
| **Installation** | | | | | | |
| **Soil correction** | | | | | | |
| Lime | | | | | 2000 | 0.0655 |
| Lime transport | 1.7 | 8.88 | 7.37 | 6.8 | | |
| Lime application | 0.2 | 0.9 | 0.86 | 0.8 | | |
| **Soil preparation** | | | | | | |
| Harrowing | 6 | 13.22 | 10.01 | 6.8 | | |
| Scrolling | 3 | 13.22 | 10.01 | 6.8 | | |
| **Sowing** | | | | | | |
| Seeds Fertiprado AC70 | | | | | 25 | 4.3 |
| Seeder | 3.3 | 8.08 | 10.45 | 7.75 | | |
| **Fertilization** | | | | | | |
| Fertilizers transport | 1 | 2.34 | 1.95 | 1.79 | | |
| Operation | 2 | 7.33 | 6.16 | 4.42 | | |
| Superphosphate | | | | | 200 | 0.41 |
| Borax | | | | | 10 | 0.35 |
| Zinc Sulphate | | | | | 7 | 1.82 |

*The entries in bold represent the type of operation which the individual operations below are part of. $h_{lab}$ – number of labour hours required by the operation for 1 hectare of pasture; CV – variable costs for 1 hour of work on 1 hectare of pasture; CF – fixed costs for 1 hectare of pasture; $C_{amort}$ – amortization costs for the machineries required for 1 hectare of pasture; Q – quantity of material in kilograms required for 1 hectare of pasture; P – price per kilogram of material.*

The costs for maintenance of SNP consist only of harrowing, which is required every 5 years and per hectare requires 0.9 h of labour ($h_{lab}$), with variable costs of 6.39 €/ha (CV), fixed of 6.65 €/ha (CF) and amortization of 5.21 € ($C_{amort}$) [95].

# C. Census data manipulation

Table A.4 reports the variables extracted from the sheets *Principais características do Produtor Singular* of the census data, with the category they belong to and the name given in the rest of the analysis. All the variables are reported in the census in number of farmers. The sheets presented a third level of specification for some variables, which was deemed not necessary and therefore neglected.

*Table A.4: variables extracted from the sheets Principais características do Produtor Singular of the census data, with the category they belong to and the name given in the rest of the analysis. All values in number of farmers.*

| Category of the variables | Variable name in the census | Variable name given | Variable name in the census | Variable name given |
|---|---|---|---|---|
| **-** | *Produtor singular* | *individual_prod_num* | *Empresário* | *individual_prod_in_business* |
| | *Autónomo* | *individual_prod_autonomous* | | |
| ***Nível de instrução*** | *Não sabe ler nem escrever* | *educ_cannot_read_write* | *Secúndário agrícola* | *educ_secondary_agr* |
| | *Sabe ler e escrever* | *educ_cannot_read_write* | *Secúndário não agrícola* | *educ_secondary_not_agr* |
| | *Basico - 1º ciclo* | *educ_basic_1st_cycle* | *Politécnico superior agrícola* | *educ_polyt_or_superior_agr* |
| | *Basico - 2º ciclo* | *educ_basic_2nd_cycle* | *Politécnico superior não agrícola* | *educ_polyt_or_superior_not_agr* |
| | *Basico - 3º ciclo* | *educ_basic_3rd_cycle* | | |
| ***Formação profissional agrícola*** | *Exclusivamente prática* | *prof_only_practical* | *Longa e curta duração* | *prof_short_and_long* |
| | *Curta duração* | *prof_short* | *Completa* | *prof_complete* |
| | *Longa duração* | *prof_long* | | |
| ***Tempo de actividade agrícola*** | *Tempo parcial* | *agr_time_partial* | *Tempo completo* | *agr_time_full* |
| ***Actividades remuneradas exterior à exploração*** | *Principal* | *ext_imp_principal* | *Secúndária* | *ext_imp_secondary* |
| ***Situação na profissão exterior à exploração*** | *Patrão/ empregador* | *ext_sit_employer* | *Trabalhador por conta de outrem* | *ext_sit_employed_by_others* |
| | *Trabalhador por conta própria* | *ext_sit_self_employed* | *Trabalhador familiar remunerado* | *ext_sit_in_family* |

From the sheets *Produtor singular segundo a Dimensão Económica e as Classes de Idade* the analysis extracted the number of farms belonging to various ranges of size in economic terms from the columns *Classes de dimensão económica (UDE),* naming them *econ_0_2, econ_2_4, econ_4_8, econ_8_16, econ_16_40, econ_40_100, econ_above_100* (where the first and second number at the end of the features name represent the interval considered with the lower boundary included and the upper excluded, except for *econ_above_100*). Table A.5 reports the features obtained from these and the ones in Table A.4 and the procedure to obtain them. Of the features above, *individual_prod_autonomous* and *agr_time_partial* were discarded, being complementary respectively to *individual_prod_in_business* and

*agr_time_full* (they sum to the total number of farmers and the features were divided for this value, therefore summing to 1).

*Table A.5: features obtained from the combination of the census variables and procedure to obtain them from the variables included in the census data.*

| Feature | Procedure to obtain it |
|---|---|
| *educ_above_basic* | Sum of *educ_secondary _agr*, ed*uc_secondary_not_agr, educ_polyt_or_superior_agr* and e*duc_polyt_or_superior_not_agr* |
| *educ_3rd_cycle_or_higher* | Same as above plus *educ_basic_3rd_cycle* |
| *prof_short_and_long_and_more* | Sum of *prof_long, prof_short_and_long* and *prof_complete* |
| *ext_act_num* | Sum of *ext_imp_principal* and *ext_imp_secondary* |
| *ext_sit_not_employer* | Sum of *ext_sit_self_employed*, *ext_sit_employed_by_others* and *ext_sit_in_family* |
| *econ_above_40* | Sum of *econ_40_100* and *econ_above_100* |
| *econ_below_4* | Sum of *econ_0_2* and *econ_2_4* |

# D. Farmer-based Toy-ABM ODD additional sections

*Table A.6: attributes of each entity in the Farmer-based Toy-ABM, their type, possible values and meaning.*

| Variable | Type | Values | Meaning |
|---|---|---|---|
| | | ***Model*** | |
| ***adoptable_ pastures*** | List of *Pasture* objects | An *AdoptablePasture* object | Contains the pasture that each *Farmer* can consider to adopt |
| | | ***Farmer*** | |
| ***farm*** | *Farm* object | A *Farm* object | *Farm* owned by the *farmer* |
| ***education*** | String; discrete | Primary, secondary, undergraduate, graduate | Level of education of the farmers |
| | | ***Farm*** | |
| ***owner*** | *Farmer* object | A *Farmer* object | *Farmer* that owns the *farm* |
| ***pasture_type*** | *Pasture* object | A *Pasture* object | *Pasture* type that the farm has |
| | | ***Pasture* (and *NaturalPasture*)** | |
| ***market*** | *Market* object | Any *Market* object in the model | *Market* responsible for reporting the economic data necessary for net present values calculations |
| ***type*** | String | Any string corresponding to the name of a pasture included in the model | The name of the pasture |
| | | ***SownPermanentPasture*** | |
| ***government*** | *Government* object | Any *Government* object in the model | *Government* responsible for setting the payments for the specific pasture type, if any |
| ***education_ confidence*** | *Dictionary* | Keys: education level. Values: between 0 and 1 | Maps each education level to the relative confidence factor that switching from semi-natural pastures to sown biodiverse corresponds to the calculated income |
| | | ***Market*** | |
| ***discount_ rate*** | Float; continuous | (0; 1) | Assumed discount rate for net present values calculations |
| ***installation*** | Float; continuous | Real | Cost in €/ha for the installation of the relative pasture (cash flow for the year 0) |

| Variable | Type | Values | Meaning |
|---|---|---|---|
| ***maintenance*** | List of 9 floats | Real (each element) | Yearly cost in €/ha for the maintenance of the relative pasture (cash flows for the years 1 to 9) |
| ***SownPermanentPasturesGovernment*** | | | |
| ***pasture_type*** | String | Any value of *Pastures' pasture_type* | Name of the pasture type they are dealing with |
| ***payments*** | List of floats | Real (each element) | Payment to farmer who install sown biodiverse pastures for each year, starting from the year of installation, in €/ha |

*Design concepts*

**Individual decision-making:** the only decision in the model regards the *Farmers'* one to switch or not their *Farms' Pasture* to an alternative pasture type, in this case SBP. This is a direct objective seeking decision, where the *Farmer* aims at maximising the additional economic benefit expected from its decision to substitute or not its current pasture with a different one, quantified through the EDNPV. The choice of the Farmer is driven by all the economic data regarding both the actual pasture and the ones to adopt retrieved from the *Market*, the payments the relative *Government* can offer to adopt a certain pasture and the education level of the *Farmer*.

**Individual sensing:** there are two main "sensing" of variables happening in the model. The first regards the *Pastures* accessing the list of adoptable pastures from the *Model*: this implies that every farmer knows all the pastures that can be considered for adoption. The second regards the economic variables required to calculate the NPV per hectare of maintaining the current pasture or adopting a different one. Also these are known with certainty by the *Farmers* (through the calculations done in their *Farm*), despite being evaluated in light of their risk attitude through their education level.

**Individual prediction:** prediction is not modelled explicitly. However, *Farmers* need to have a process of implicit prediction in order to calculate NPV. In this model, this consists of the assumption that all the market prices and costs will remain constant over the 10 years for which the NPVs are calculated.

**Interaction:** the only interaction of *Farmers* consists in calling the submodel of their *Farm* to evaluate the adoption of different pastures. *Farms* interact with their current *Pasture* to get the NPV of maintaining it and with the adoptable *Pastures* to get the NPV of adopting them, passing to the latter the education level of the *Farmer* who owns them. To perform the NPV calculations, each *Pasture* needs to interact with its relative *Market* and *Government* (the latter only if existing) to retrieve the necessary economic data. No interaction among *Farmers* nor *Farms* is included, therefore the individual decision-making is not influenced by the collective trend.

**Heterogeneity:** *Farmers* are heterogeneous due to the different education level they have. However, they do not differ in their decision-making. *Pastures* are heterogeneous since they can be adoptable or not and, among the adoptable ones, there can be some incentivized with payments and some not.

**Stochasticity:** the only process in the model using pseudorandom numbers is the activation of farmers, which however has no effect on the outcome as explained above. The decision whether to adopt or not is instead deterministic, depending on the EDNPVs.

**Observation:** at the model level, the percentage of adoption at the end of the simulation (after its first and only step) is collected as a measure of the aggregated adoption. At the agent level, the individual *Farmer*'s *Farm pasture_type* attribute is collected at the end of the simulation in the column *Pasture* of a dataset associating it to the farm's *FARM_ID*, to retrieve if the *Farmer* decided to switch or not its pasture. To this dataset, also a column with the education of the farmer (*Farmer education*) and one for each adoptable pasture with the calculated EDNPV to adopt it (in this case, only *EDNPV SBP*).

No emergent outcome is produced by the model. In fact, the results depend on a single decision per agent, based on a deterministic calculation not influenced by other agents' behaviours or any other unpredictable factor.

*Initialization*

Four datasets are needed upon initialization of the model:

- Farmers data: excel spreadsheet needed to initialize the *Farmers* in the model. It has to contain two columns:
    - *ID*: identification code unique for each farmer
    - *HighestEducationalDegree*: it represents the education level of the farmer (primary, secondary, undergraduate, graduate)

  The file cannot have two lines referring to the same farmer (i.e. with the same *ID*), otherwise the model will report an error and stop the execution. The AF survey provided these data: *ID* correspond to the code with which the farmers are identified in the survey, while *HighestEducationalDegree* to the homonym column of the social data section.
- Farms data: excel spreadsheet needed to initialize the *Farms* in the model. It has to contain two columns:
    - *FARM_ID*: identification code unique for each farm and corresponding to the one of the *Farmer* who owns it
    - *Pasture*: pasture type that the farm has installed

  *FARM_ID* corresponds to the code with which the farmers are identified in the AF survey. The file has to include exactly one *Farm* for each *Farmer* (i.e. one and only one *FARM_ID* has to be equal to each *ID* in the farmers data), otherwise the model will report an error and stop the execution. *Pasture* are all set to "Natural Pasture" and thus it is assumed that all farms have SNP installed before the beginning of the PCF project, since no data were available on the actual pasture of each farm in 2008, before the beginning of the programme. The assumption is that all the farmers in the survey did not adopt before the PCF project, justified by the fact that the average lifetime of SBP is 10 years and the data have been collected in 2019.
- Pasture costs: installation and maintenance costs over 10 years for each pasture, referred to 2009. Required in the form of a dictionary mapping each pasture type to another dictionary with two entries, "installation" and "maintenance", reporting respectively installation and maintenance costs for the pasture in €/ha.y.
- Payments: payments for the incentivised adoptable pastures. Given in the form of a dictionary mapping each pasture type to a list of the payment per hectare per year (thus, the number of

element in the list represent the number of years for which the payment is given). If the payments have different values depending on the year of installation, this dataset has to report the values corresponding to the year with the highest cumulative payment. In fact, since costs for pastures adoption are available only for 2009, it was not possible to consider their change. Therefore, if farmers do not adopt with the highest value of payments they would not adopt with lower ones and we can consider only the highest one. For SBP, the highest payments offered are the ones for 2009, as reported in section 3.1.4.This means that there is also no need to adjust for inflation, since costs reported in the Pasture costs database are also referred to 2009.

Pastures costs and payments are also parameters of the model, that can be changed to assess different scenarios. In the same way, the model is initialized with a default discount rate of 5%, i.e. with the value of the parameter *discount_rate* equal to 0.05.

The first entities to be created during the initialization of the model are *Markets*, one for each pasture type in the model (*NaturalPasturesMarket and SownPermanentPasturesMarket* in this case) and the *SownPermamentPasturesGovernment* (one only since only SBP are incentivised). *Markets* are initialized with the discount_rate parameter and the Pasture costs dataset. Upon initialization, *Markets* set their attribute *pasture_type* as the name of the pasture they refer to and *discount_rate* as the value of the discount_rate parameter. Then, they set their *installation* and *maintenance* variables for each market retrieving the installation and maintenance costs for the relative pasture from the Pasture costs dataset. *Governments* are initialised with the Payments dataset. Upon initialization, *Governments* set their *pasture_type* attribute as the name of the pasture they refer to and set their *payments* attribute retrieving the payments relative to their *pasture_type* from the Payments dataset. Then, the *Pastures* are initialized, *NaturalPasture* and *SownPermanentPasture*. These are created after the previous entities since they need to be linked to the relative *Market* and *Government* (if incentivised), setting their relative *market* and *government* attributes to point at them. The *pasture_type* attribute is set with their name. The *adoptable_pastures* attribute of the model is also set including the *Pastures* subclassed from *AdoptablePasture*. Lastly, the *education_confidence* dictionary is set with the values in Table A.7, defined in a way to correspond to the values used in Teixeira [93] that are also reported.

Lastly, *Farmers* are instantiated with the Farmers and Farm databases, retrieving their *education_level* from the *HighestEducationalDegree* column of the farmers database. Each *Farmer* initialise the *Farm* it owns (setting its *farm* attribute and the *Farm*'s *owner* attribute) with the Farms database, looking for the entry with the *FARMER_ID* corresponding to its own *ID*. Each *Farm* retrieve its *pasture_type* attribute from the *Pasture* column of the Farms database, after each string has been replaced with the relative *Pasture* entity. Note that, as specified above, in the Farms database the attribute *Pasture* is set as "Natural Pasture" for each entry.

*Table A.7: confidence factors (CoFa) that switching from SNP to SBP corresponds to the calculated income, mapped to the corresponding education level of the farmers (and relative values in [93]).*

| Education level in Teixeira [93] | Corresponding opinion on SBP | Correspoding values from the Farmers dataset *HighestEducationalDegree* column | CoFa |
|---|---|---|---|
| Cannot read or write | "I will stick to what I know" | No farmers with this education level in the dataset | 0.0 |
| Can read and write | "I would only switch to SBPPRL if I was sure they worked" | No farmers with this education level in the dataset | 0.2 |
| Basic education | "They may or may not work, I do not know" | **Primary** | 0.4 |
| Secondary education | "They could probably work" | **Secondary** | 0.6 |
| Higher education | SBPPRL are better, but sometimes their installation fails" | **Undergraduate, graduate** | 0.8 |

*Submodels*

**NPV keeping:** submodel through which a *Pasture* already installed in a *Farm* calculates the NPV for the *Farmer* owning the *Farm* of not switching to another pasture type. The process is the following:

- The *Pasture* retrieves from its relative *Market* the attributes *installation*, *maintenance* and *discount_rate.*
- Since *installation* and *maintenance* represent, respectively, the cash flows for the year of the installation (year 0) and for all the following year (years 1 to 9), the *Pasture* can recreate the cash flows for its NPV simply by joining them.
- The *Pasture* calculates the NPV of not switching as:

$$NPV = \sum_{t=0}^{9} CaFl_t * \frac{1}{(1+i)^t} \quad (5)$$

where *i* is the discount rate (equal to *discount_rate*), $CaFl_t$ is the cash flow at the $t^{th}$ year.

**NPV adopting:** submodel through which an *AdoptablePasture* calculates the NPV for a farmer of switching to it from the current one. The process is the following:

- The *AdoptablePasture* retrieves the confidence factor that switching from SNP to SBP correspond to the calculated income relative to the education level of the *Farmer* that owns the *Farm* from its attribute *education_confidence.*
- The *AdoptablePasture* retrieves from its relative *Market* the attributes *installation*, *maintenance* and *discount_rate.*
- The *AdoptablePasture* retrieves from the *Market* relative to the *Pasture* the *Farm* currently has installed the attribute *maintenance.*
- The *AdoptablePasture* calculates the expected maintenance costs for each year ($EMC_y$) as

$$EMC_y = CoFa * MC_{AP,y} + (1 - CoFa) * MC_{CP,y} \quad (6)$$

Where $MC_{AP,i}$ is the maintenance cost of the *AdoptablePasture* which is performing the calculation in the year *y* (i.e. the $y^{th}$ element of the *maintenance* attribute of its relative *Market*)

and $MC_{AP,i}$ is the maintenance cost of the *Pasture* the *Farm* currently has installed in the year *y* (i.e. the $y^{th}$ element of the *maintenance* attribute of its relative *Market*). The rationale for this calculation is that in case the installation of SBP fails this will happen during the first year and farmers will switch back to SNP, therefore having to bare the maintenance costs for SNP (harrowing and feed) and not anymore the ones for SBP.

- Denoting *EMC* the vector containing all the calculated $EMC_y$ ordered as in the *maintenance* attributes of the *Market* entities, *installation* and *EMC* represent, respectively, the cash flows for the year of the installation (year 0) and the expected ones for all the following year (years 1 to 9). The *AdoptablePasture* can recreate the expected costs cash flows of its ENPV simply by joining them.
- To calculate the total expected cash flows, the *AdoptablePasture* sums the payments provided by its relative *Government* to the costs cash flows of the corresponding years, with the first payment given in the year of the installation. Note that the payments are not part of the expected value but are taken as a sure cash flows. This is because, after adopting SBP, even if the pasture is not producing as expected farmer would still get the payments as long as they followed the correct maintenance.
- The *AdoptablePasture* calculates the ENPV of switching as

$$ENPV = \sum_{t=0}^{9} ECF_t * \frac{1}{(1+i)^t}, \quad (7)$$

where *i* is the discount rate (equal to *discount_rate*) and *ECF* is the expected cash flow at the $t^{th}$ year.

# E. AF survey data analysis plots

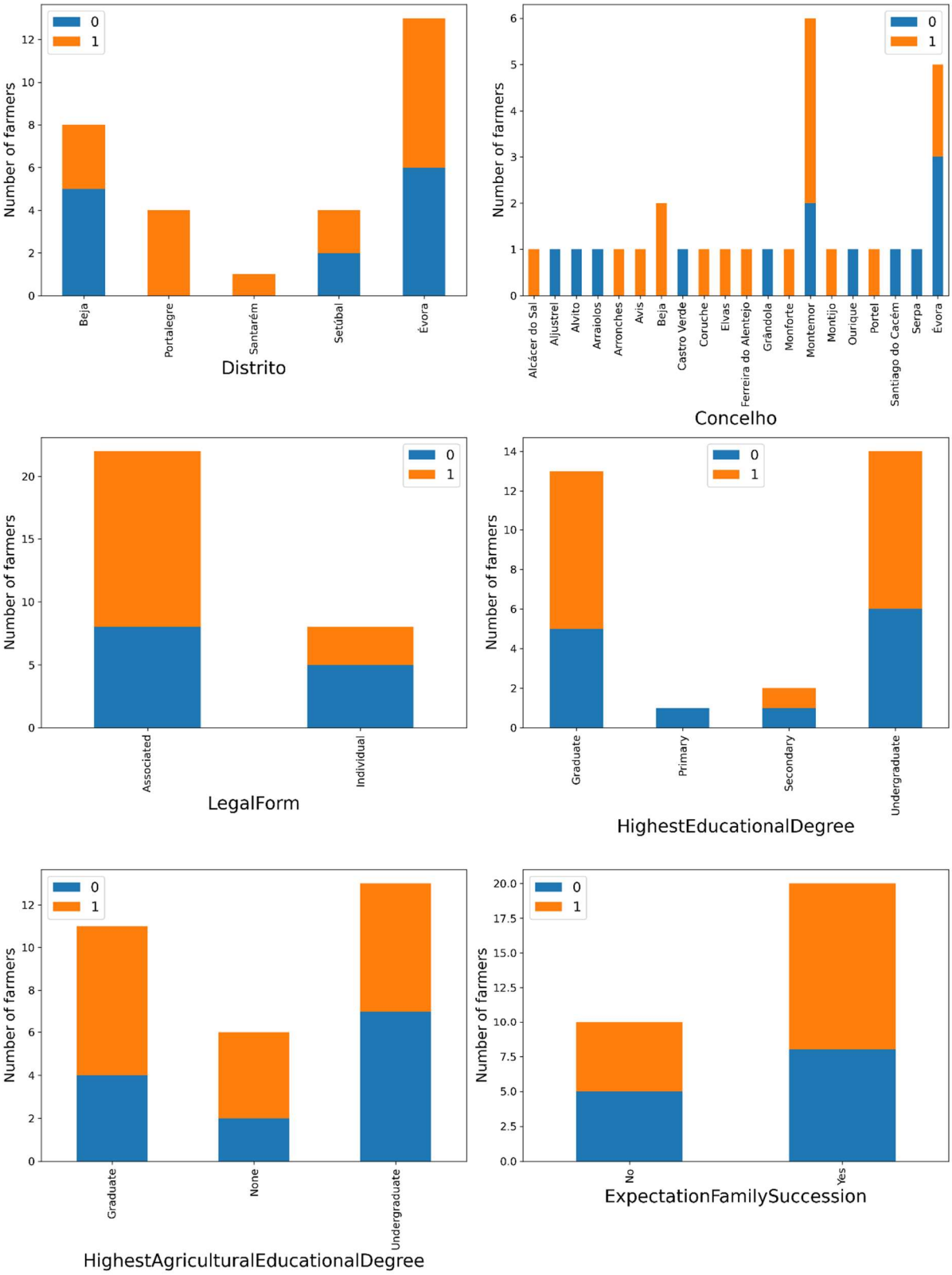



*Figure A.2: distribution of categorical attributes in the survey data, reporting also the number of adopters (in orange, labelled as 1) and not adopters (in blue, labelled as 0).*

# F. Spatial granularity harmonization

The only differences found between the census data and the shapefile with geographic data were the spelling of two municipalities, that appeared as “Ponte de Sôr” and “Mêda” in the shapefile and as “Ponte de Sor” and “Meda” in the census. To solve the conflict, the entries in the shapefile were changed to match the ones of the census since the second denomination was the most commonly found on the internet and Ponte de Sor is reported with this spelling also in the PCF project database. Comparing the PCF project dataset with the modified shapefile, several conflicts of denomination were found and the consequent modifications to the PCF project database are reported in Table A.8. When an entry was referring to two municipalities of the shapefile, due to the lack of further information it was referred to the municipality with the largest pastures area. The Region column of the adoption previous to the PCF project had 24 entries. With the initial disaggregation, the municipalities corresponding to the Oeste region presented a value over 100%. To solve this issue, the Oeste entry was considered as referring to the old agricultural region of Ribatejo and Oeste. The only entries with values over 50% were Abrantes and Tomar, which have specific entries in the and can be considered reliable, and Alcochete but only in 2012. Therefore, these were not modified. Table A.9 reports the municipalities mapped to each of the regions. Of the 262 municipalities in the area of Portugal considered, 170 were assigned a non-null SBP area installed in the years when the corresponding region saw adoption.

*Table A.8: County values from the PCF project data that do not match any value of the Municipality feature of the shapefile with geographic data, analyses of the conflict and corresponding Municipality value assigned.*

| ***County* from the PCF project data** | **Analysis** | **Corresponding *Municipality* value assigned** |
|---|---|---|
| Albernoa | Albernoa is a civil parish in the Beja municipality. | Beja |
| Alcácer do Sal - Santa Susana, Alcácer do Sal - Torrão, Alcácer do Sal - Torrão/Alvito-V.N. Baronia | None of the second names of these entries has a corresponding one in the shapefile *Municipality* column, while Alcácer do Sal has it. | Alcácer do Sal |
| Benavente/Porto Alto | Benavente has a correspondent entry in the shapefile *Municipality* column, Porto Alto has not. | Benavente |
| Elvas e Campo Maior | Both Elvas and Campo Maior are also in the shapefile. | Elvas |
| Ferreira do Alentejo /Figueira dos Cavaleiros | Ferreira do Alentejo has a correspondent entry in the shapefile *Municipality* column, Figueira dos Cavaleiros has not. | Ferreira do Alentejo |
| Lisboa - Serpa | The municipality of Lisboa includes mainly urban area and does not have any other entry in the PCF project database. | Serpa |
| Moura e Serpa | Both Moura and Serpa are also in the shapefile. | Serpa |
| Ourique | Some entries presens a typing error with a blank space following the letters. | Ourique |
| Ponte de Sor / Montargil | Ponte de Sor has a correspondent entry in the shapefile *Municipality* column, Montargil has not. | Ponte de Sor |
| Santiago do cacém | Some entries present Cacém not capitalised. | Santiago do Cacém |
| Vila Velha de Rodão | Some entries lack the accent on the first o in Ródão. | Vila Velha de Ródão |
| Évora / Montemor-o-Novo | Both Évora and Montemor-o-Novo exist in the shapefile. | Évora |

*Table A.9: specification of the corresponding municipalities assigned to each* Region *value of the adoption previous to the PCF project data.*

| ***Region* values already corresponding to a municipality** | |
|---|---|
| **Odemira, Estremoz, Elvas, Ponte de Sor, Coruche, Tomar, Abrantes** | |
| ***Region* value** | **Corresponding municipalities assigned** |
| Ferrereira[63] | Ferreira do Alentejo |
| Montemor[64] | Montemor-o-Novo |
| Aveiro | Águeda, Ílhavo, Albergaria-a-Velha, Anadia, Arouca, Aveiro, Castelo de Paiva, Espinho, Estarreja, Mealhada, Murtosa, Oliveira de Azeméis, Oliveira do Bairro, Ovar, São João da Madeira, Santa Maria da Feira, Sever do Vouga, Vagos, Vale de Cambra |
| Coimbra | Arganil, Cantanhede, Coimbra, Condeixa-a-Nova, Figueira da Foz, Góis, Lousã, Mira, Miranda do Corvo, Montemor-o-Velho, Oliveira do Hospital, Pampilhosa da Serra, Penacova, Penela, Soure, Tábua, Vila Nova de Poiares |
| Guarda | Almeida, Guarda, Pinhel, Sabugal |
| Viseu | Armamar, Carregal do Sal, Castro Daire, Cinfães, Lamego, Mangualde, Moimenta da Beira, Mortágua, Nelas, Oliveira de Frades, Penalva do Castelo, Penedono, Resende, Sátão, São João da Pesqueira, São Pedro do Sul, Santa Comba Dão, Sernancelhe, Tabuaço, Tarouca, Tondela, Vila Nova de Paiva, Viseu, Vouzela |
| Portalegre | Alter do Chão, Arronches, Avis, Campo Maior, Castelo de Vide, Crato, Fronteira, Gavião, Marvão, Monforte, Nisa, Portalegre, Sousel |
| Beja | Aljustrel, Alvito, Barrancos, Beja, Castro Verde, Mértola, Moura, Ourique, Serpa, Vidigueira |
| Castelo Branco | Belmonte, Castelo Branco, Covilhã, Fundão, Idanha-a-Nova, Penamacor, Vila Velha de Ródão |
| Évora | Évora, Alandroal, Arraiolos, Borba, Mora, Mourão, Portel, Redondo, Reguengos de Monsaraz, Vendas Novas, Viana do Alentejo, Vila Viçosa |
| Santarém | Almeirim, Alpiarça, Benavente, Cartaxo, Chamusca, Salvaterra de Magos, Santarém |
| Leiria | Óbidos, Alcobaça, Alvaiázere, Ansião, Batalha, Bombarral, Caldas da Rainha, Castanheira de Pêra, Figueiró dos Vinhos, Leiria, Marinha Grande, Nazaré, Pedrógão Grande, Peniche, Pombal, Porto de Mós |
| Oeste | Alenquer, Arruda dos Vinhos, Azambuja, Vila Franca de Xira, Alcochete, Montijo |
| Trás-os-Montes | Mogadouro, Vila Flor |
| Minho | Amares, Barcelos, Braga, Guimarães, Cabeceiras de Basto, Celorico de Basto, Esposende, Fafe, Póvoa de Lanhoso, Terras de Bouro, Vieira do Minho, Vila Nova de Famalicão, Vila Verde, Vizela, Arcos de Valdevez, Caminha, Melgaço, Monção, Paredes de Coura, Ponte da Barca, Ponte de Lima, Valença, Viana do Castelo, Vila Nova de Cerveira |

[63] The entry “Ferreira” could have corresponded to Ferreira do Alentejo (in the Beja district) or Ferreira do Zêzere (in the Santarém district). I chose the first since also present in the PCF project database and for the bigger area.
[64] The entry “Montemor” could have corresponded to Montemor-o-Novo (in the Évora district) or Montemor-o-Velho (in the Coimbra district). I chose the first since also present in the PCF project database and for the bigger area.

# G. Municipality-based Data-driven ABM features

*Table A.10: features included in the municipality-level dataset, with their values and their meaning.*

| Feature | Unit/values | Data meaning |
|---|---|---|
| **SBP adoption** | | |
| *adoption_pr_y_munic* | [0, 1] | SBP area installed in the municipality in the previous year |
| *tot_cumul_adoption_pr_y_munic* | [0, 1] | Total cumulative SBP area installed in the municipality until the previous year |
| *cumul_adoption_10_y_pr_y_munic* | [0, 1] | SBP area installed in the municipality in the previous 10 year |
| All the features regarding SBP adoption as a fraction of permanent pastures area in the concerned area. The features above were also included for Portugal, substituting in the names "*munic*" with "*port*" and for adjacent municipalities, substituting in the names "*munic*" with "*neighbours_adj".* | | |
| **Census data** | | |
| *pastures_area_munic* | ha | Total permanent pastures area in the municipality, corresponding to the *Pastagens permanentes* variable. |
| For the features included from the census data refer to Appendix C. All are reported as a fraction of total number of farmers in the municipality, therefore having values in the interval [0, 1], apart from *individual_prod_num,* which is reported in number of farmers, *pastures_mean_size_munic*, in hectares, and *land_rented,* as a fraction of the sum of its values and the owned land in hectares. | | |
| **Climate data** | | |
| *av_d_mean/min/max_t_pr_y_munic* | °C | Mean, minimum and maximum daily temperature of the municipality in the previous year |
| *days_mean_t_over_20/25_pr_y_ munic* | days | Number of days of the previous year in which the daily mean temperature was over 20/25 |
| *days_max_t_over_30_pr_y_munic* | days | Number of days of the previous year in which the maximum temperature was over 30°C |
| *days_min_t_under_0_pr_y_munic* | days | Number of days of the previous year in which the minimum temperature was below 0°C |
| *av_prec_sum_pr_y_munic* | mm/ pixel.day | Average precipitation fell over an area of 0.1° side in the municipality in the previous year |
| *days_no_prec_pr_y_munic* | days | Number of days of the previous year without any precipitation |
| *cons_days_no_prec_pr_y_munic* | days | Maximum number of consecutive days of the previous year without any precipitation |
| The features above regarding climate data were also included as averages over the period 1996 – 2018, substituting in the names "*pr_y*" with "*average*". | | |
| **Soil data** | | |
| *CaCO3_mean_munic* | g/kg | Municipality average calcium carbonates soil content |
| *CN_mean_munic* | - | Municipality average carbonium – nitrogen ratio in soil |
| *N_mean_munic* | g/kg | Municipality average nitrogen soil content |
| *P_mean_munic* | mg/kg | Municipality average phosphorus soil content |
| *pH_mean_munic* | - | Municipality average pH in water |
| **Economic data** | | |
| *sbp_payment* | €/ha | Total payment offered per hectare of SBP installed |

*A "/" in the features name means that the line is referring to multiple features with names differing for only the piece included between "_". SBP – sown biodiverse pastures. For the acronyms of the soil data features refer to Table 4.*

*Table A.11: features for the municipality-based analysis kept after the first screening, their Spearman ρ correlation score with the target variable considering only instances referred to the years of the Portuguese Carbon Fund (PCF) project and all and their variance inflation factors (VIFs), both for the regression and the classification stages of the double hurdle model.*

| Feature | Classification | | | Regression | | |
|---|---|---|---|---|---|---|
| | ρ PCF | ρ all years | VIF | ρ PCF | ρ all years | VIF |
| *pastures_area_munic* | 0.61 | 0.26 | 2.44 | -0.26 | 0.15 | 2.75 |
| *pastures_mean_size_munic* | 0.61 | 0.24 | 3.74 | 0.19 | 0.21 | 4.92 |
| *individual_prod_num* | -0.15 | -0.05 | 1.57 | -0.15 | -0.08 | 1.58 |
| *individual_prod_in_business* | 0.36 | 0.13 | 2.20 | 0.27 | 0.22 | 3.42 |
| *land_rented* | 0.36 | 0.07 | 1.96 | -0.06 | 0.20 | 2.66 |
| *educ_3rd_cycle_or_higher* | 0.46 | 0.05 | 3.17 | 0.45 | 0.28 | 4.40 |
| *prof_above_some_long* | 0.34 | 0.09 | 3.43 | 0.31 | 0.15 | 4.90 |
| *ext_sit_not_employer* | 0.15 | 0.10 | 1.93 | 0.12 | 0.24 | 2.16 |
| *econ_above_40* | 0.45 | 0.14 | 4.50 | 0.26 | 0.24 | 5.10 |
| *econ_0_2* | 0.03 | 0.13 | 1.84 | -0.23 | 0.00 | 2.89 |
| *econ_2_4* | -0.44 | -0.16 | 2.23 | -0.11 | -0.24 | 3.65 |
| *adoption_pr_y_munic* | 0.62 | 0.71 | 1.82 | 0.38 | 0.68 | 2.23 |
| *tot_cumul_adoption_pr_y_munic* | 0.48 | 0.52 | 2.02 | 0.36 | 0.70 | 2.77 |
| *adoption_pr_y_neighbours_adj* | 0.54 | 0.37 | 2.03 | 0.15 | 0.55 | 2.75 |
| *tot_cumul_adoption_pr_y_neighbours_adj* | 0.45 | 0.14 | 2.45 | 0.25 | 0.58 | 4.32 |
| *adoption_pr_y_port* | -0.01 | -0.11 | 3.21 | 0.03 | 0.43 | 3.70 |
| *tot_cumul_adoption_pr_y_port* | 0.00 | -0.19 | 3.72 | -0.02 | 0.47 | 5.31 |
| *av_d_mean_t_average_munic* | 0.47 | 0.18 | 4.31 | 0.02 | 0.34 | 4.76 |
| *av_d_max_t_average_munic* | 0.60 | 0.24 | 8.83 | 0.22 | 0.27 | 12.70 |
| *cons_days_no_prec_average_munic* | 0.60 | 0.27 | 7.93 | -0.07 | 0.32 | 13.18 |
| *CaCO3_mean_munic* | 0.06 | -0.11 | 3.66 | 0.17 | 0.25 | 3.08 |
| *CN_mean_munic* | -0.53 | -0.09 | 5.81 | -0.23 | -0.23 | 8.26 |
| *N_mean_munic* | -0.22 | -0.02 | 2.11 | -0.18 | -0.31 | 2.38 |
| *P_mean_munic* | -0.36 | -0.14 | 3.37 | 0.19 | -0.27 | 3.66 |
| *sbp_payment* | Not included | | | -0.03 | 0.29 | 2.03 |

*The colours of the cells reporting the Spearman ρ scores vary from red, for the most negative values, to green, for the most positive, therefore highlighting the sign and strength of the corelation. Yellow values are for weakly correlated variables.*

# H. ML models selection

*Table A.12: performance metrics of the models with the tuned hyperparameters after the first and the second round of tuning.*

| Model | Classification | | | | Regression | | | |
|---|---|---|---|---|---|---|---|---|
| | Validation scores | | Training scores | | Validation scores | | Training scores | |
| | LL | ROC AUC | LL | ROC AUC | RMSE | $AR^2$ | RMSE | $AR^2$ |
| **First tuning** | | | | | | | | |
| **LR/LLOGR** | 0.467 | 0.882 | 0.456 | 0.887 | 0.00726 | 0.417 | 0.00714 | 0.449 |
| **LR/LLOGR + SQ** | 0.467 | 0.882 | 0.456 | 0.888 | 0.00710 | 0.441 | 0.00695 | 0.478 |
| **PR/PLOGR** | 0.294 | 0.952 | 0.239 | 0.967 | 0.00967 | -0.883 | 0.00968 | -0.185 |
| **LSVM** | 0.497 | 0.877 | 0.490 | 0.884 | 0.00726 | 0.417 | 0.00712 | 0.452 |
| **LSVM +SQ** | 0.500 | 0.882 | 0.491 | 0.889 | 0.00710 | 0.441 | 0.00695 | 0.479 |
| **NLSVM** | 0.214 | 0.966 | 0.135 | 0.983 | 0.00646 | 0.538 | 0.00462 | 0.769 |
| **DT** | 0.260 | 0.953 | 0.160 | 0.982 | 0.00670 | 0.503 | 0.00612 | 0.596 |
| **RF** | 0.114 | 0.993 | 0.064 | 0.998 | 0.00599 | 0.603 | 0.00443 | 0.788 |
| **ERT** | 0.219 | 0.987 | 0.190 | 0.992 | 0.00587 | 0.618 | 0.00405 | 0.823 |
| **GBT** | 0.089 | 0.995 | 0.024 | 1.000 | 0.00969 | -0.038 | 0.00968 | -0.013 |
| **Second tuning** | | | | | | | | |
| **NLSVM** | 0.211 | 0.968 | 0.130 | 0.985 | 0.00643 | 0.543 | 0.00419 | 0.810 |
| **RF** | 0.101 | 0.994 | 0.051 | 1.000 | 0.00580 | 0.628 | 0.00287 | 0.911 |
| **ERT** | Not included | | | | 0.00579 | 0.629 | 0.00248 | 0.933 |
| **GBT** | 0.076 | 0.996 | 0.017 | 1.000 | Not included | | | |

*Validation scores are the average scores on the validation sets of cross-validation, while training scores refer to the score obtained training and testing the model on the entire dataset. The models tested were linear (LR) and polynomial (PR) regression – for classification logistic (LLOGR and PLOGR) – with elastic net regularization, linear (LSVM) and nonlinear (NLSVM) Support Vector Machine, Decision Tree (DT), Random Forest (RF), Extremely Randomized Trees ensemble (ERT) and Gradient Boosting with DT as base classifier (GBT). "+ SQ" after the name of the model implies that it was tested adding as feature the square of the feature tot_cumul_adoption_pr_y_munic. The performance metrics for classification are logistic loss (LL) and area under the curve of the receiver operating characteristic (ROC AUC) and the ones for regression are root mean squared error (RMSE) and adjusted $R^2$ ($AR^2$).*