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

The case of sown biodiverse pastures adoption in Portugal

Giacomo Ravaioli

giacomo.ravaioli@gmail.com

Instituto Superior Técnico, Universidade de Lisboa, Portugal

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 2009-2012. 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. 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.

1 Introduction

Food production is the most important cause of environmental change at global scale [1]. Animal production is its most impactful sub-sector, generating throughout its supply chain 14.5% of anthropogenic greenhouse gases emissions, with cattle farming for beef and milk accounting for nearly two thirds of these [2]. Since having signed the Kyoto Protocol (KP), Portugal decided to put particular attention on emission from the agro-forestry sector, with a focus on pastures management.

1.1 Sown biodiverse pastures

A critical system in this regard are sown biodiverse permanent pastures rich in legumes (SBP), referring to a rich seed mix of species originated from the Mediterranean but normally existing in little proportion in natural grasslands (as legumes). SBP exhibit, compared to semi-natural pastures (SNP), increased productivity and provide multiple ecosystem services. Increasing soil organic matter (SOM), during the first 10 years after installation SBP are a net sink of 1.55–2.13 t $CO_{2\text{-eq}}$ per hectare per year with adequate technical management [3]. Between 1990 and 2008 94,260 hectares (ha) of SBP were installed in the country and especially in Alentejo [3]. After having noticed a reduction in adoption from 2005, between 2009 and 2014 the Instituto Superior Técnico (IST) spin-off company Terraprima (TP) (http://terraprima.pt/) led a project of payments for carbon (C) sequestration through SBP adoption with the support of the Portuguese Carbon Fund (PCF), during which an additional 48,491 ha of SBP were installed [4]. This led to estimated sequestration of 1.54 million additional tons of CO_2 [4]. Figure 1 reports the area of SBP and consequent carbon sequestration observed until 2008 and forecasted by Teixeira [3] for the successive years.

To further spread SBP, a proper evaluation of its adoption dynamics is required and this can be enhanced by the use of suited tools, such as agent-based models (ABMs).

1.2 Agent-based modelling for land use

Land use/cover change (LUCC) systems are inherently complex adaptive systems (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 [5]. Therefore, ABMs appear as a suited tool for their analysis.

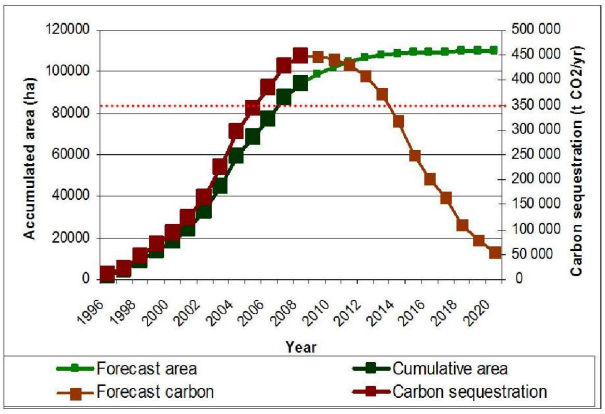

Figure 1: cumulative area of sown biodiverse pasture installed and consequent yearly carbon sequestration, observed and forecasted [3].

ABMs are generically identified as bottom-up simulations composed of individual agents and characterized by the importance given to their behaviours and heterogeneity. ABMs particularly allow to address the feedbacks and emerging properties of CAS [6]. This brough to a steady increase in the number of studies involving ABMs to study LUCC systems [7] and other topics strictly treated in them. LUCC ABMs are often used to explain land-use patters, thanks to the possibility to explicitly address space and the interaction with the environment [8]. ABMs ability to model how information diffuse in social networks drove their application to innovation diffusion [9]. This is often linked to policy design and assessment, which required a focus on the individual decision-makers which ABMs allow for [10]. The subarea of policy design involved in this work is payments for ecosystem services (PES). Finally, often the study of LUCC is finalised at the implementation of environmental assessments [11] and the combined use of life cycle assessment (LCA) with ABMs is increasingly accepted, for the possibility to include dynamic components and behavioural heterogeneity [12].

The surge of specific context applications fostered an increased use of empirical data in ABMs in the last years [13], 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. Data-driven ABMs can rely on machine learning (ML) algorithms to test various and more complex approaches to better link ABMs to the real-world [14], despite scarce literature and guidelines available in this regard. 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 [15], a key element for ABMs [16]. Theoretical behavioural models, often based on economic theories, still have an important role in ABMs for the clear interpretation they allow for [7] and data are often used in combination with them. However, in LUCC ABMs the majority of studies from 2000 on adopted ad hoc implementations without theory-based justifications [7], allowing to build more complex agents [14].

1.3 Objectives

The main goals of this analysis are: to develop and test multiple modelling frameworks in the form of ABMs 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.

This will enhance a proper assessment of the dynamics driving SBP adoption and therefore be the basis for future work on new policies aimed at its diffusion.

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.

2 Materials and methods

This analysis was divided into two main parts. A farmer-based approach (section 2.1), with a survey of 43 farmers as main data source, focused on the individual farmers as main subjects, developing both theory- and data-driven models. A municipality-based approach instead widened the scope to the entire continental Portugal through a data-driven model, thanks to the use of data from agricultural census and the PCF project (section 2.2).

The majority of the work required during this thesis was implemented through the Python coding language. All the source code is available in GitHub (https://github.com/giacrava/thesis-sbp-abm) and can be clarified on request.

2.1 Farmer-based approach

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 and the ones who did not. None of the farmer-based models included 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.

2.1.1 Data availability

The sources of data used for the farmer-based approach were:

- A survey of 43 farmers in the Alentejo region, in the scope of the Animal Future (AF) project (https://www.animalfuture.eu/), with socio-economic and agricultural management data.
- Economic data regarding the costs of installation and maintenance of SBP and SNP, their stocking rate and the feed supplementation they require [3], [17], and the payments provided during the PCF project were also retrieved.

2.1.2 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 really simplified behaviour for the farmer agents, consisting only in the calculations of the expected differential net present value (EDNPV) between adopting SBP and keeping SNP. This model is an ABM implementation of the work done in [3], [18], which designed the PCF project through supply and demand curves at the aggregated level.

Not having data on when the different farmers adopted, the ABM runs in only one step, consisting of the calculation by the farmers agents of the EDNPV over 10 years, the average lifetime of SBP (without resowing). The NPV is differential since it considers only costs that are different from adopting SBP and keeping SNP. These are related to the maintenance of SNP (harrowing) and the installation and maintenance of SBP (harrowing, liming and fertilization). Feed costs are the only one linked to livestock assumed to change when installing SBP, due to their increased productivity. The differential NPV is also expected since it includes a confidence factor whose role is to represent uncertainty in the calculations and risk propension, creating agents with bounded rationality. In the Farmer-based Toy-ABM, the confidence factor depends only on the farmers' education level: the lower the education level, the lower the confidence factor and the lower is the EDNPV calculated by the farmer, who trusts less the system and therefore expects to have to switch back to bare the maintenance costs of SNP after installing.

If the EDNPV of adopting SBP respect to keep SNP has a positive value, the farmer installs SBP, otherwise keeps SNP (assumed as initial pasture type for all).

2.1.3 Farmer-based Calibrated ABM

The *Farmer-based Calibrated ABM* was based on the architecture of the Farmer-based Toy-ABM but modified in two ways, to allow for a deeper understanding of which factors drive farmer's decision-making process and how.

The first was the inclusion of more proxies than only farmers' education level in the calculations of the confidence factor, which became a weighted sum of these. The analysis of the AF survey dataset and the application of filter-based features selection methods (Spearman's ρ correlation coefficient and Pearson's Chi-Squared test) allowed to reduce the number of variables it contained. This was necessary due to the little sample available, which had to be further reduced to 30 farmers since some important characteristics were available only for these.

The second modification was the calibration of the influence of these proxies on the available data. This was implemented through an iterative grid search over the weights of the proxies, running the model with the different combinations until maximizing the F1 score of its predictions.

2.1.4 Farmer-based Logistic Regression

The *Farmer-based Logistic Regression* consisted in a data-driven approach, which avoided any theoretical assumption and in particular the one of farmers as profit maximizing agents. This approach was implemented through the use of regularized logistic regressions tested through cross-validation (CV), using the same features used for the Farmer-level Calibrated ABM. The CV was implemented as internal loop of a nested procedure aimed at testing different splits of the dataset, which, due to the limited sample size, highly influenced the results.

2.1.5 Models validation

The validation of the farmer-based models consisted in the comparison of their output with the adoption observed in the AF survey data. At the micro-level precision, recall and F1 score were calculated. At the macro-level, the modelled and observed percentage of farmers that adopted were compared.

2.2 Municipality-based approach

The part of the analysis at the municipality level developed a data-driven ABM, the *Municipality-based Data-driven ABM*. The available data allowed in fact to widen the scope of the analysis over the years from 1996 to 2012 and over whole continental Portugal and to simulate various 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 not available in the AF survey and therefore provide different and complementary insights. 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.

The chosen framework resulted in a model combining ML, to estimate SBP adoption in each municipality in each year, and ABM, to consider time and spatial dimensions and the reciprocal influence among agents. While the ABM provides the dynamic environment in which agents are located, their internal model, constituted in a ML model, senses all these information and outputs an estimation for the area of SBP adopted in the specific municipality in the specific year.

2.2.1 Data availability

The sources of data used for the municipality-based approach were:

- Socio-economic and agricultural data at the aggregated level for the municipalities from the agricultural census of 1999, collected by The Instituto Nacional de Estatística (INE, https://www.ine.pt/) .
- Historic SBP adoption, available from 1996 to 2008 at regional level [18] and from 2009 to 2012 at the individual farmers level from the PCF project.
- Environmental data, consisting in climate and soil properties provided as geospatial maps from the version 21.0e of the E-OBS dataset from the EU-FP6 project UERRA (https://www.uerra.eu) and the Land Use/Land Cover Area Frame Survey (LUCAS) programme.
- Value of the payments per hectare offered during the PCF project.

These datasets were all reported to the level of aggregation of the municipalities. The SBP adoption prior to 2008 had to be disaggregated as if was originally reported for wider regions. The disaggregation was done proportionally to the pasture area of the municipalities included in the regions, considering only the ones that adopted also during the PCF project, and all municipalities within the region if there was no adoption during the PCF project. Climate and soil data were averaged over the municipalities through QGIS (https://www.qgis.org/en/site/) and the climate ones also over the period 1996 - 2018. All the features extracted were merged having a dataset with one row for each municipality and each agricultural year considered (which starts the 1st of September). The target variable of this dataset was the hectares od SBP adopted in the municipality in the year, divided by the pasture area of the municipality.

2.2.2 Municipality-based Data-driven ABM architecture

The architecture of the Municipality-based Data-driven ABM aimed only at being the dynamic environment providing to the main agents, the municipalities, the variables that their internal model uses to estimate the adoption in each year. The need for a proper simulation was due to the fact that the variables regarding previous SBP adoption are endogenous to the ABM, i.e. their value depend on the outcome of the previous steps.

The step of the simulation, representing one agricultural year, consists in the estimation by each municipality agent of its adoption. The agents are updated synchronously, which means that the adoption variables are updated only at the end of each model step, to simulate the time lag which information diffuses with in reality. Therefore, the order of activation of the municipalities does not influence the model's outcome.

2.2.3 Agents' internal model for the estimation of individual municipalities adoption

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. It resorted in particular to a double-hurdle model [19], dividing the estimation into a first classification stage, predicting the presence or not of adoption in the municipality, and a second regression stage, predicting for the municipalities with adoption the amount of area installed in each municipality. This allowed to study and analyse independently the two processes.

The estimation of SBP installed in the year in the ABM therefore 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 or not. Second, if there is, the regression stage estimates the hectares that are installed in the year.

The selection of both ML models was based on the dataset with all the features changing the target variable to categorical for the classification stage and removing instances with 0 adoption for the regression. The number of features was reduced to reduce overfitting risk and make results more significative, through Spearman ρ coefficients to and variance inflation factors to quantify respectively correlation with the target variables and multicollinearity.

Various classifiers and regressors, linear and non, were tested on the relative datasets through CV and multiple grid search rounds to select the best hyperparameters. The classifiers were tested on the based of their logistic loss, the regressors of their root mean squared error. The two best classifiers and regressors based on dissimilar architectures were further tested running the Municipality-based Data-driven ABM 100 times from 1996 to 2020 for each combination. The final combination was chosen on the base of the average prediction of the yearly adoption in Portugal, assessed through the adjusted R^2 compared to observed data from 1996 to 2012 and the most realistic extrapolation from 2013 to 2020.

The influence of the various features on their estimations and consequently on SBP adoption was analysed through the SHAP (SHapley Additive exPlanations) framework [20].

2.2.1 Municipality-based Data-driven ABM validation

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. 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 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.

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

The Municipality-based Data-driven ABM, allowed to quantify the amount of CO_2 the additional pastures installed thanks to the PCF project will sequester over their lifetime. The yearly adoption in Portugal after 2008 if the PCF project would not have taken place was obtained through a counterfactual simulation, consisting in running the Municipality-based Data-driven ABM over the years from 2009 to 2020 but with no payments offered to install SBP. Its estimations from 2009 to 2012 and from 2013 to 2020 were subtracted year-by-year respectively from the observed adoption during the PCF project and from the adoption obtained running the Municipality-based Data-driven ABM with the actual value of the payments (the same model validated in the previous section). Summing these two yearly difference over the years and multiplying them for the sum of the C sequestration factors for SBP for the 10 years after installations, retrieved from [21], provided the amount of C sequestered over its lifetime by the additional area installed respectively during the PCF project and after it conclusion. 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.

3 Results

3.1 Farmer-based approach

3.1.1 Farmer-based Toy-ABM

Regarding micro-validation, the Farmer-based Toy-ABM 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.

The EDNPVs depending on the highest education level of the farmers were -243.05 €/ha for primary, -76.11 €/ha for secondary and 90.84 €/ha for graduate and undergraduate.

3.1.2 Farmer-based Calibrated ABM

The farmers variables kept for this analysis were the surface of their land in hectare, the percentage of this land rented, their education level (encoded with raising value for raising level) and their legal form (encoded as 1 if individual and 0 if associated). Their weights obtained by the calibration procedure were (ordered as above) of 0.75, 0.25, 1.25 and -0.5.

Regarding micro-validation, the Farmer-based Calibrated ABM 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 2. The confidence factors present the same distribution, ranging from -0.20 to 1.97.

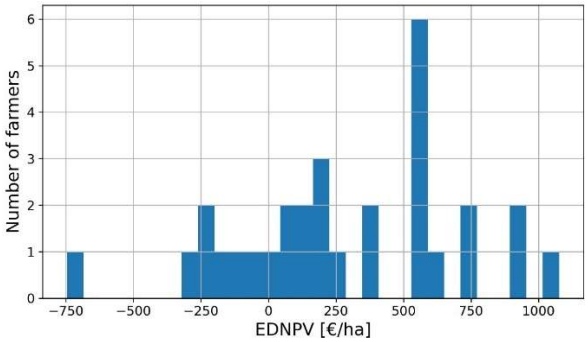

Figure 2: distribution of expected differential net present values (EDNPVs) of adopting sown biodiverse pastures respect to keep semi-natural pastures resulting from the Farmer-based Calibrated agent-based model after calibration.

3.1.3 Farmer-based Logistic Regression

Regarding micro-validation, the Farmer-based Logistic Regression reached a precision of 0.65, recall of 0.89 and F1 score of 0.75 on average over the validation sets of the CV. At the macro-level, the model predicted 23 out of 30 farmers (76.7%) switching to SBP.

Their coefficients obtained training the model for the pasture area, the percentage of pasture rented, the education level and the legal form were (in order) of 0.15, -0.24, 0.16 and -0.35.

3.2 Municipality-based approach

The double-hurdle model selected was the one adopting nonlinear support vector machine for both stages, which obtained a macro-level adjusted R^2 score of 0.794. Combinations with tree-based models obtained slightly better adjusted R^2 scores, but these ML models are not able to extrapolate properly. Figure 3 reports the yearly

adoption in Portugal observed and estimated by the Municipality-based Data-driven ABM, while Figure 4 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 5 reports the estimated and observed total area installed in each municipality until 2012 geographically.

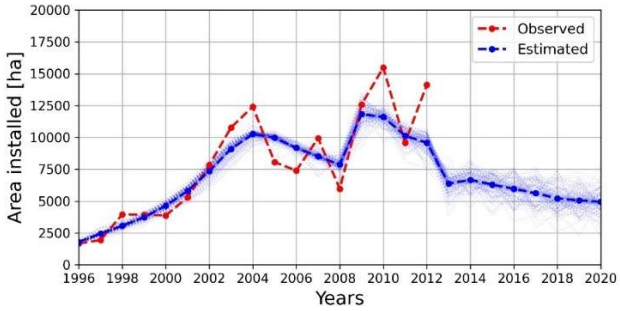

Figure 3: yearly adoption of sown biodiverse pastures in Portugal observed and estimated by the Municipality-based Data-driven agent-based model (individual runs and average are respectively in light and dark blue).

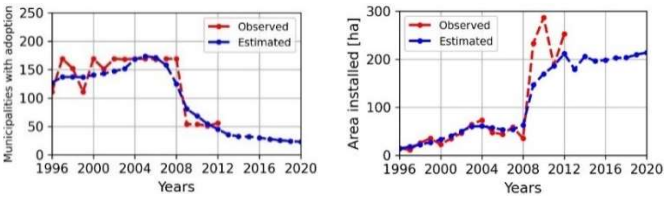

Figure 4: yearly aggregated results from 1996 to 2020 of the two stages of the internal model of the Municipality-based Data-driven agent-based model: number of municipalities with adoption estimated by the classifier (a) and average area of sown biodiverse pastures installed in the municipalities with adoption (b).

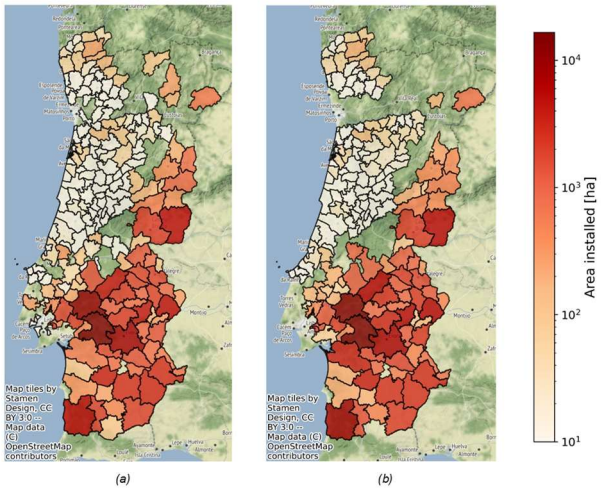

Figure 5: total area of sown biodiverse pastures installed in each municipality in Portugal until 2012, estimated by the Municipality-based Data-driven agent-based model (a) and observed (b). Municipalities with no adoption are not plotted.

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

Figure 6 reports the yearly adoption in Portugal from 1996 to 2020 if no PCF project took place, estimated by the counterfactual simulation.

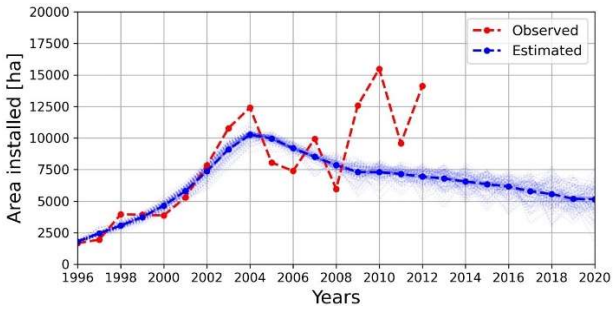

Figure 6: yearly adoption of sown biodiverse pastures in Portugal observed and estimated by the Municipality-based Data-driven agent-based model run without PCF project (individual runs and average in light and dark blue). t.

The C sequestration over its lifetime of the differential area of SBP installed thanks to the PCF project was assessed at 1.65 Mt CO_2 from 2009 to 2012 and -0.26 Mt CO_2 (indicating less area installed) from 2013 to 2020. Overall, the project will bring to the sequestration of 1.39 Mt CO_2. Figure 7 represent this graphically.

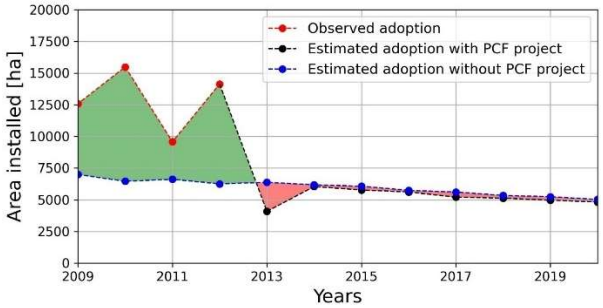

Figure 7: comparison of sown biodiverse pastures adoption with the PCF project (observed until 2012 and modelled from 2013) and without the PCF project (modelled), from 2009 to 2020[1].

4 Discussion

4.1 Farmer-based approach

4.1.1 Interpretation of results

All the farmer-based approaches overestimate adoption and do not present large performance differences in terms of F1 score. 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, this last model is less robust than the others, depending completely on the definite of the arbitrary values of the confidence factors linked to the education level. Also passing to a

[1] *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.*

completely data-driven approach with the Farmer-level Logistic Regression did not bring to better performances. The main difference between the two models based on economic calculations and the data-driven one is that while the construction of the first ones gave by itself insights on the system dynamics, the second was an approach much faster to develop and calibrate, with also the possibility to implement CV.

The EDNPVs calculated by the Farmer-based Calibrated ABM resulted in a range of values that could actually represent the expected ones by farmers, but were much higher than the ones calculated by the Farmer-based Toy-ABM. These high EDNPVs could be the consequence of other factors that the Farmer-based Calibrated ABM embedded into the economic calculations, such as peer influence, sustainability value or specific environmental conditions, which, to be properly evaluated, would need an independent consideration.

The correlation found between the features included in the Farmer-based Calibrated ABM and the Famer-based Logistic Regression and SBP adoption were not significative and the Chi-Squared test rejected every hypothesis of dependence. The weights obtained by the two models however suggest that larger farms and higher education level bring to a higher probability of adoption, as well as if the farmer is legally associated.

4.1.2 Limitations

An inevitable limitation of all the farmer-based models was the impossibility to consider temporal and spatial dimensions, which were instead introduced in the municipality-based approach. While the Farmer-based Logistic Regression was limited mainly by the limited size of the dataset, which made its results vary with the split for the CV, and its biases towards highly educated farmers, the main limitations of the Farmer-based ABMs lied in the assumptions of the economic calculations, which hampered a reliable evaluation of the extent to which economic calculations matter for SBP adoption and contributed to the overestimation of adoption.

In fact, while the analysis considered the same prices for all farmers, in reality prices can differ due to many factors. 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, which in turn depend on many variables. Together with other minor assumptions, these neglect the great variability that EDNPVs are subject to for different farmers and considering it could increase the explanatory power of economic calculation.

4.2 Municipality-based approach

4.2.1 Interpretation of results

The Municipality-based Data-driven ABM showed able to capture the underlying trend of adoption in Portugal at the macro-level and to successfully fit also at the micro-level. 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.

An important novelty exhibited by the results is the higher yearly adoption in Portugal after 2013 respect to the previous extrapolation constituting the basis of the PCF project design reported in Figure 1, which is evident in the trend estimated by the counterfactual simulation. 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. The PCF project constituted a clear discontinuity in this trend.

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 [4]. 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. The fact that the values estimated is not largely higher means that the additional area of SBP installed thanks to the PCF project estimated in this thesis is lower than previously thought. 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 1. 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 the payments offered should have been lower. However, this thesis cannot affirm that the counterfactual estimated here is more plausible than the counterfactual used to establish additionality during the PCF project, in particular for the lack of a biophysical depiction of land suitability for SBP installation which would penalize the municipalities running out of suitable area before getting to 100% of their pasture land.

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 imply that due to the PCF project less SBP were installed in this period, which probably are due to farmers that would have installed SBP in the following years but, incentivised by

the payments, decided to do it in the period 2009 – 2012. These are additional installations that should have been considered to further lower the payments. However, it should be noted that this 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.

The analysis of the SHAP values of the ML models (whose results are not reported here due to space constraints) revealed that their predictions depend mostly by the features related to SBP adoption in and until the previous year in the municipalities themselves and in Portugal. These are in fact the only features included in the datasets that vary over the years and the only ones the model can rely on to explain the temporal trend of adoption. A higher value of adoption in the previous year, especially in the municipality itself, generally brings to a higher estimated adoption in the following, confirming that more recent installations by neighbours brings more farmers to know and trust the system and therefore to install it. The models seem to rely on the total cumulative adoption in Portugal instead to decrease the estimated adoption. This does not have a realistic explanation and its likely to depend on the lack of an indicator of biophysical suitability, as explained above.

Climate features follow second in order of importance, with warm climates without temperature peaks and not too rainy being the conditions favouring the most adoption. A higher education level of the farmers in the municipality also has a strong contribution in fostering adoption. Other socio-economic factors and soil variables instead seems to have negligible influence on farmers' decisions respect to the other.

4.2.2 Limitations

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 provide the variables to the ML models, which are also the ones responsible for estimating the endogenous values related to adoption.

Figure 4 shows how the ML models incorporated biases due to the first main limitation of the analysis, the need to disaggregate of the adoption prior to the PCF. 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 explains the sudden drop in number of municipalities adopting and raise in average adoption within each municipality that the data exhibit in 2009 (Figure 4). 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 and unrealistic trends when using the model to explore scenarios after this year.

Solving the disaggregation issue would probably be also 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. However, more information which to base the disaggregation on are unlikely to be available since more precise data on adoption until 2008 do not exist. 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. This hampered the possibility to develop a predictive model, being a requirement often cited for data-driven models aiming at forecasting unknown conditions and therefore study the outcome of future policies [14], [22]. For this and for the necessity 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.

Regarding the assessment of the influence of the various drivers on adoption, the importance of the adoption in an until the previous year in the municipality resulted inflated by the disaggregation of the adoption, which favour the emergence of such trend. The disaggregation also caused the clear discontinuity before and after 2009 in the dataset labels, which the models could explain only relying on features regarding adoption. The unrealistic influences of the adoption features are however also linked to the lack of a variable linked to the biophysical capability of the land, an issue already presented for the farmer-based approach regarding economic calculations. Driving the expected productivity of SBP, 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, without the need for the model to rely on the cumulative adoption in Portugal for this. The lack of such variable is the third main limitation of the municipality-based approach. 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.

4.3 Future work

The most important single improvement for both approaches used is the consideration of the biophysical capacity of the land. For the farmer-based approach, to properly assess the expected savings of feed. For the municipality-based approach, to improve the model performance and in particular assess more reliably the influence of the features related to adoption. However, at the moment a model to estimate this variable does not exists and would therefore need to be developed. The Municipality-based Data-driven ABM would also greatly benefit from a better disaggregation of the adoption previous to the PCF project, which could be obtained collecting information through dedicated questionnaires.

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 and overcoming data protection issues to link them to the PCF database, which contains the fundamental information of which farmers adopted SBP (missing in the census).

The large number of datapoints that would be available with this procedure would allow to properly divide timewise the dataset and test the resulting model on independent data. This would make the model more reliable for policy design and future scenarios analysis, together with a proper use of uncertainty, sensibility analysis an 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, through which elicit behavioural rules or understand how information diffuse in the system for example. Most importantly these data collection methods could provide already on their own precious insights on which factors drive SBP adoption and allow to reduce the number of features included in the model – which is advised in case of developing predictive models.

A future model aimed at supporting the assessment and design of new policies for a further spreading SBP in Portugal 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 could help to design policies tailored to the farmers that will not adopt if payments are not provided.

On a broader scope, the assessment of the C sequestration thanks to SBP implemented in this analysis could be expanded to become a proper LCA of the system and evaluate its environmental effect more holistically.

5 Conclusion

The farmer-based approach showed how simplified and uniform economic calculations are not suited to represent the individual farmer's decision-making. Also the farmer-based approaches considering additional features however overestimated adoption. This suggests that socio-economic characteristics and farms management practices, the only variables available in the AF survey data, are not enough to evaluate farmers' decision-making. The municipality-based approach instead, including proxies for the interactions inside the system as previous adoption and environmental variables, could capture the underlying trend of adoption in Portugal. These results confirmed that treating the system as a CAS allowed to better represent it and that the interactions among the agents and with the environment are factors of fundamental importance.

The Municipality-based Data-driven ABM constitutes the first ABM, among the retrieved literature regarding innovation diffusion and policy design in agricultural systems, relying entirely on ML algorithms without combining them with any explicit theoretical assumption. 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.

The assessment of the PCF project suggest that the payments offered during the PCF project were higher than necessary for two reasons: more SBP area would have been adopted during the years of the project than previously estimated and a fraction of the area adopted during the project would have been adopted in the following years, if this had not taken place. This confirms that the design and IA of future effective policies aimed at expanding SBP adoption should be supported by reliable quantitative insights. These could be provided developing a farmer-level framework with wide spatial and temporal scope, unifying the two approaches here presented, surpassing data limitations and performing additional validation of model forecasts.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.