Predicting agri-food quality across space: a Machine Learning model for the acknowledgement of Geographical Indications

Giuliano Resce*, Cristina Vaquero-Piñeiro†

Abstract

Geographical Indications (GIs), as Protected Designation of Origin (PDO) and Protected Geographical Indication (PGI), offer a unique protection scheme to preserve high-quality agri-food productions and support rural development, and they have been recognised as a powerful tool to enhance sustainable development and ecological economic transactions at the territorial level. However, not all the areas with traditional agri-food products are acknowledge with a GI. Examining the Italian wine sector by a geo-referenced and a machine learning framework, we show that municipalities which obtain a GI within the following 10 years (2002-2011) can be predicted using a large set of (lagged) municipality-level data (1981-2001). We find that the Random Forest algorithm is the best model to make out-of-sample predictions of municipalities which obtain GIs. Among the features used, the local wine growing tradition, proximity to capital cities, local employment and education rates emerge as crucial in the prediction of GI certifications. This evidence can support policy makers and stakeholders to target rural development policies and investment allocation, and it offers strong policy implications for the future reforms of this quality scheme.

Keywords: Geographical Indications; Rural Development; Agri-Food Production; Machine Learning; Geo-Referenced Data.

JEL Classification: C53; 013; Q18.

*Department of Economics, University of Molise, Via F. de Sanctis - 86100 Campobasso, Italy. giuliano.resce@unimol.it

†Department of Economics, Roma Tre University, Via Silvio D'Amico 77 - 00145 Rome, Italy. cristina.vaqueropineiro@uniroma3.it

1 Introduction

Geographical Indications (GIs) are the main scheme of the European Union quality policy aiming at protecting the names of specific products to promote their uniqueness (characteristics, reputation and quality) essentially or exclusively resulting from the characteristics of their region of origin as well as their traditional expertise.[1] According to the EU regulations, agri-food products can be legitimately marked as a GI if they have a specific link to the place where they are made, and always only after the European Commission's endorsement. This sign identifies products as legally tied to a specific production area (i.e. region of origin) whose environmental conditions, contextual know-how, cultural traditions, entrepreneurial practices and local actors interactions were consolidated over time becoming the drivers of the intangible value-added of these products (Bowen, 2010). GIs are, therefore, deeply rooted in their area of production and their value-added can be considered as resulting from the interaction of a set of natural and human elements coexisting in the region of origin.

Originating in Mediterranean Europe, GIs have been experiencing a massive diffusion all over the world (Huysmans and Swinnen, 2019). In the EU there are more than 3,000 GIs to which are added 30 GIs produced in non-EU countries (Qualivita, 2021). Italy is the country with the higher number of GI (Qualivita, 2021) resulting in a turnover of around €20 m. Although the main effect of GIs can be summarised as preserving the agri-food biodiversity of local high-quality production, GIs can exert several positive economic effects at both individual and collective levels (Török et al., 2020). On the one hand, obtaining a GI provides competitive benefits for producers, not only in terms of premium price (Huysmans and Swinnen, 2019), value-added (Belletti et al., 2015; Newton et al., 2015) and market access (Altomonte et al., 2016), but also high level of protection against counterfeiting and piracy (European Union Intellectual Property Office - EUIPO, 2017). In this sense, GIs are protected as intellectual property rights, a collective patent owned by the producers. The EU has concluded several bilateral and multilateral international agreements, which allow the recognition of many EU GIs outside the EU and the recognition of non-EU Geographical Indications in the EU. The most recent has been signed with China in March, 2021. In addition, the GI recognition enables consumers to trust and distinguish

[1] Regulation available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32012R1151; https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019R0034

quality products (Costanigro et al., 2019; Moschini et al., 2008). GIs guarantee to consumers that the concerned produce is made in its specific areas of origin according to the product specification, the code of practice that all the producers must follow. On the other, at the territorial level, GIs can trigger several positive socio-economic externalities (Haeck et al., 2019; Meloni and Swinnen, 2018; Charters and Spielmann, 2014),such as supporting rural development (Cei et al., 2018; Crescenzi et al., 2021), contributing to food-safety (De Rosa, 2015; Wirth, 2016), enhancing sustainability and reducing environmental impacts (Vandecandelaere et al., 2018; Belletti et al., 2017). These socio-economic effects have been investigated by a bourgeoning group of studies, and journals themselves have dedicated special issues to provide new evidence on the virtuous role of GIs. For instance, a recent paper written by Crescenzi et al. (2021) has provided, for the first time in the literature, econometric evidence that rural municipalities with GIs experience better performance in terms of local economic development than others by supporting population growth and fostering the economic reorganization towards non-farming sectors, which frequently involve higher value-added activities. Nowadays, particular attention is paid in evaluating the potential contribution of GIs in moving the agri-food sector towards a more resilient and sustainable system Vandecandelaere (2021).

It is true, however, that the socio-economic benefits differ radically among GIs (sectors and products) and across regions of origin. Most of the economic power competitiveness tends to remain spatially and sectorially concentrated. The GIs market is indeed led by products that were well-know also before they got the designation, such as Parmigiano Reggiano DOP and Prosciutto di Parma DOP (Qualivita, 2021). Evidence has accumulated that territorial-level factors are strongly associated with the success of GIs. What factors encourage producers to obtain institutional acknowledgement has, in fact, been the focus of a significant group of studies. Favourable institutional context, local actors' engagement and co-operation have been highlighted among others. A recent study by Vaquero-Piñeiro (2021), empirically demonstrated that for food GIs ex-ante socio-economic conditions are fundamental for the success (in terms of revenues) of a GI, while in the wine sector socio-cultural elements and tacit knowledge are more relevant. Therefore, while GIs may endogenously stimulate local economy and generate a potential virtuous circle between product and territory, they may also contribute to create a sort of path-dependence in market inefficiencies and rent-seeking. In this scenario, a risk of exclusionary effects exist: the largest

agribusiness capture GIs rents without any benefits flowing to smaller (Bramley et al., 2009).

Certainly, in practice, local stakeholders (e.g., institutions, farmers, administrative authorities) do not have all the information useful for concluding that a specific, maybe new, GI will be successfully for the region of origin. And, what is more, they do not know if that territory can ever be acknowledged with a GI at a certain point in time. What is certain is, in fact, that not all the traditional and quality agri-food productions existing in the world will become a GI. Selecting wines and food "of merit" on the basis of their human and natural values, as well as on the historical linkages with territories, are the criteria for rating quality within this scheme. Examining the Italian wine sector and focusing on wines which are designated with the highest level of GI (Protected Designation of Origin - PDO), this paper proposes a specific model to predict the territorial PDO acknowledgement, using a large set of lagged geo-referenced municipality-level indicators and Machine Learning (ML) algorithms. The scope is test whether ex-ante spatial features contain enough information to predict the future acknowledgment of a GI. To this purpose, we select the best algorithm from a battery of four ML models (LASSO, Random Forest, Gradient Boosting Machines, and Neural Network), which has the potential to become a valuable tool for the targeting of the quality scheme and rural development policies.

In order to design appropriate policy intervention in support of the agri-food sector and rural areas, it would be extremely useful for policy makers to have geo-located information of future GI, on the basis of territorial indicators easily available to them (e.g., agricultural census). In particular, given the on-going debate about the effectiveness of place-based policies (Barca and Rodríguez-Pose, 2012) and the evidence provided by the literature on the GIs effects on local development (Crescenzi et al., 2021; Cei et al., 2018), knowing if the territory has the potential to see their local products acknowledged as a GI, or not, would be particularly relevant for next rural development policy strategies. The advent of GIs require a full-ranging adaptation of the local economy: producers must follow product specifications, new administrative offices (such as Consortia) must be established and services mechanism activated to collectively manage and promote the GI. At the same time, knowing the potential future trajectory of a territory would be important also for investments, at both government and individual investment decisions. In rural areas, and in particular within

the EU Rural Development Policy Framework, the allocation of public resources are often devolved to governments to support strategic territorial assets. Farmers located in a specific area would also be interested in investing to set up a specific farming activity, in buying a new piece of land or in diversifying their income towards new activities, such as tourism. In Italy, the expansion of GI wines occurred in tandem with changes in consumption behavior: from the daily consumption of lower quality wines to the higher but more occasional consumption of high-quality wines (Pomarici et al., 2021). Also, in this context, farmers and winemakers are more and more attracted by areas recognised as capable of producing certified high-quality wines, and citizens, more generally, could be interested in deciding whether to live there or not.

In the recent economic literature, it has been argued that these policy problems do not require ex-post correlation or causal inference solutions but, instead, predictive inference would be of greater assistance (Kleinberg et al., 2015). From the empirical perspective, it has been highlighted that, for solving these prediction policy problems, the standard econometric models are not adequate, since these are tuned to generating unbiased estimates of coefficients rather than minimizing prediction error (Kleinberg et al., 2015; Einav and Levin, 2014; McAfee et al., 2012). In this regard, new developments in the field of machine learning have shown a great potential for addressing prediction problems. ML techniques are gaining momentum for solving problems connected to the evaluation of poverty and food security target (Jean et al., 2016; McBride and Nichols, 2018; Lentz et al., 2019; Hossain et al., 2019), in particular to identify predictors of access to healthful food (Amin et al., 2021) and in predicting calorie-based food security among poor households and communities (Hossain et al., 2019). ML models are now used to evaluate the effectiveness of public programs and spending (Andini et al., 2018; Hoffman and Mast, 2019), to improve the evaluation of public policies (Ballestar et al., 2019), to exploit the potential of historical documents and to identify areas with similar location premiums in urban economics (Combes et al., 2021; Sommervoll and Sommervoll, 2019). Focusing on the Italian context, recent works have leveraged the potential of ML to predict bankruptcy of local governments (Antulov-Fantulin et al., 2021), vaccine hesitancy in Italian municipalities (Carrieri et al., 2021), and to estimate local mortality and local inequality during the COVID-19 pandemic (Cerqua et al., 2021; Cerqua and Letta, 2021). This is the first application of a machine learning model for predicting the territorial acknowledgement of GI.

The analysis is developed at municipality level, which is not only the most disaggregated level available, but also the only appropriate level of the analysis (Crescenzi et al., 2021). Indeed, according to the rules of the assignment of GIs, the region of origin refers to an area of specific municipalities that can be much smaller than other administrative units (e.g., regions). We use census data matched with data collected directly from the Product Specifications of PDO wines. Results show that the Random Forest algorithm is the best model to make out-of-sample predictions of municipalities with PDO with an accuracy of 84 per cent. Among the area-level indicators, the wine growing tradition of municipalities and regions, local employment and education rates emerge as crucial in the prediction of the PDO acknowledgement. Given that we use census data, this evidence can support policy makers and stakeholders to target rural development policies in advance, and in advising on investment allocation. Furthermore the model also offers strong policy implications for the future reforms of this quality scheme.

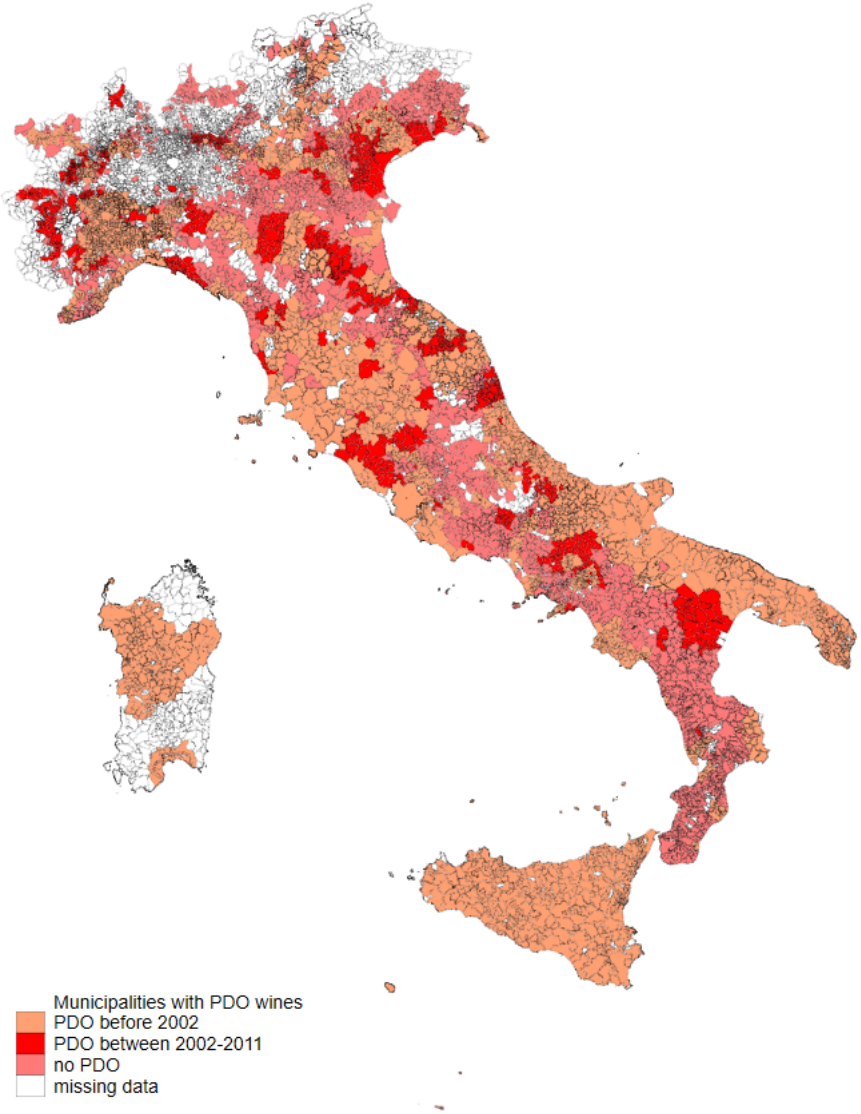
The remainder of the paper is organized as follows: Section 2 presents the institutional setting, Section 3 introduces the data and the methods, Section 4 presents the results, and Section 5 concludes the study.

2 Institutional Setting

The Regulation (EU) No 1151/2012 on quality schemes for agricultural products and foodstuffs an the Commission Implementing Regulation (EU) 2019/34 for Geographical Indications in the wine sector were intended to support high-quality agricultural and processing activities by protecting the names of products that originate from specific regions and have specific qualities or enjoy a reputation linked to the production territory.[2] According to these regulations, food and wine GIs consist of Protected Designation of Origin (PDO) and Protected Geographical Indications (PGI). Conversely, for spirit drinks only the general term of Geographical Indication can be used. The main difference between PDO and PGI is related to how much of the raw

[2]Legal documents available at: https://ec.europa.eu/info/food-farming-fisheries/food-safety-and-quality/certi.cation/quality-labels/quality-schemes-explained/regulations-food-and-agricultural-productsen;https://ec.europa.eu/info/food-farming-.sheries/food-safety-and-quality/certification/quality-labels/quality-schemes-explained/regulations-wineen

Figure 1: Spatial distribution of Municipalities with PDOs in 2011



materials can come from the region of origin or if parts of the production process can take place elsewhere. For PDOs, every part of the production, processing and preparation process must take place in the specific region. In the case of wines, grapes have to come exclusively from the geographical area where the wine is made. Conversely, PGIs requires that at least one of the stages of production, processing or preparation takes place in the region. This means that for PGI wines at least 85% of the grapes used have to come exclusively from the geographical area in which the wine is actually made. Obtaining a PDO is therefore even harder than obtaining a PGI. A PDO label is granted only to products with the strongest link with the region where they are made, not only conceptually, but also in practice given that every part of production has to be located within the region of origin, whose specific characteristics constitute the defining factor for achieving the properties of the product.

Even if currently GIs are used to preserve a wide set of different agri-food products, GIs originated in the wine sector, with France and Italy as pioneers. France laid out the rules for appellation d'origine contrôlée (AOC) wines as early as 1935, while in Italy, the national regulation goes back to the 1960s. The wine sector in Italy represents therefore an interesting setting to study GI because of the historical value of these high-quality productions in this country. 35% of the total GI wines in the EU come from Italy. In Italy, more than 520 wines are a GI (77% POD and 23% PGI) accounting for 63% of GIs. According to the Italian regulation, PDOs include both *Denominazione di Origine Controllata (DOC)* and *Denominazione di Origine Controllata e Garantita (DOCG)* wines. The wine sector is therefore particularly evocative of the GI phenomenon in Italy.

Among GI wines, we illustrate our approach focusing on the PDOs because we are interested in predicting territorial features that support the official acknowledgment of high-quality productions embedded in their area of origin. PDOs represent not only the GIs with the highest quality, but also the only one for which the entire production process must be located in the region of origin, i.e., 100% of the grapes must come from that area and producers have to follow stricter rules (Corsi et al., 2019). In 2011, around 62% of Italian municipalities were acknowledged as having produced at least one PDO wine. Among them, 24% of municipalities have obtained the first wines certified between 2002 and 2011 (Figure 1).

3 Data and Methods

To conduct the analysis we rely on a municipality-level geo-referenced database arranged by matching census data collected by the Italian National Institute of Statistics (ISTAT), remote sensing data, and data obtained directly by digitalizing product specifications. Starting from the entire list of Italian municipalities, we restrict the sample to municipalities with a positive level of viticulture (more than 0 ha), which are the only ones that can qualify for obtaining a PDO certification. After excluding municipalities which already have PDO in 2001 and municipalities with missing data (see Figure 1), we end up with 1508 municipalities, of which 999 did not receive PDO in the 2002-2011 time interval, and the remaining 509 who did receive PDO in the 2002-2011 time interval. Our task is to correctly predict the municipalities which receive PDO using predictors of three (past) time points: 2001, 1991, and 1981[3].

Formally, every municipality x^t has an associated target binary variable PDO_x^t (Protected Designation of Origin) that takes values 1 (positive sample) if the municipality is PDO, and value 0 (negative sample) otherwise. Based on the set of (lagged) Geographical-Demographic ($GD_x^{T<t}$), Socioeconomic ($S_x^{T<t}$), and Agricultural ($A_x^{T<t}$) features (Table 1) for municipality x^t, the prediction task is to find the algorithm $f(.)$ (Machine Learning model) that predicts PDO_x^t:

$$\left\{GD_x^{T<t}, S_x^{T<t}, A_x^{T<t}\right\} \xrightarrow{f(.)} PDO_x^t. \tag{1}$$

The full list of features used to predict PDO_x^t are reported in Table 1 and described in Table A1 in the appendix; they are all connected to local agriculture sector, socioeconomic and environmental conditions, prosperity and well-being conditions as well as the role of the agriculture sector in the region where the municipality is located. We also include the spatial lags of territorial features in order to account for potential spatial externalities.

The standard routine in the ML literature is to randomly divide the data in a training set, in which the model is built and tuned, and a testing set, in which its

[3]Starting from the 80s is relevant due to the fact that throughout the whole 1970–1985 period, wines with certified origin became more and more important in Italy: the number of certified wines reached 225 in 1985, covering 10% of Italian wine production (Pomarici et al., 2021)

Table 1: Features used to predict PDO_x^t

Geographical-Demographic	Socioeconomic	Agricultural
Variable	Foreign-born employment rate	Total Agricultural Area
Population	Foreign-born unemployment rate	Utilised Agricultural Area
Urban area	Share of foreign-born workers	Farms
Remoteness	Residential mobility - Foreign-born workers	Small farms
Population density	Italian/foreign school attendance ratio	Medium farms
Gender gap	Italian/foreign school indipendents ratio	Big famrs
Young population	Education gender gap	Family farms
Elderly population	Education rate	Farms' phisical size
Elderly population	High education rate (educated people)	Agricultural land intensity
Young population	Illiterate rate	Agricultural land diffusion
Erderly rate	Middle education rate (young people)	Family workers
Divorce rate	High education rate	Non-family workers
Immigration rate	High education rate (young people)	Number of employees
Young immigrants	Education rate (15-19 years)	Employment intensity
Marriage rate	Middle education rate (adults)	Vineyards
Avarage size of families	Incidence of graduates	Vineyards (dummy)
Families without childern	Male participation rate	Winegrowing farms
Families with children	Female participation rate	Winegrowing farm density
Young single-person families	Participation rate	Winegrowing specilisation
Young single parent families	Incidence of inactive young population	Vineyard diffusion
Youg families without children	Inactivity rate (young people)	Winegrowing farms' phisical size
Youg families with children	Male unemployment rate	Single-grape wines
Single-person fmilies	Female unemployment rate	Sparkling wines
Single parent families	Umemployemnt rate	Food and spirit GI
Families without childre	Youth unemployment rate	Unesco area
Families with children	Male employment rate	Main economic relavant DOP (1)
Homeownership rate	Female employment rate	Main economic relavant DOP (2)
average size of a single-family home	Employment rate	Main economic relavant DOP (3)
Potential for use of buildings	Employment turnover	Main economic relavant DOP (4)
Potential for residential use in built-up areas	Youth employment rate	Main economic relavant DOP (5)
Residential use in nuclei and scattered houses	Industrial employment	Main economic relavant DOP
Avarage age of buildings	Non-tredable sectors employmnet	Regional agricultural aoutput
Index of availability of services in the home	Tradable sectors employment	Regional winegrowing output
Buildings - good conditions	Female non-tredable sectors employmnet	Regional area
Buildings - bad conditions	High-specialisation employment	Regional vineyard diffusion
Occupied historical buildings	Specific-specialisation employment	Regional Total Agricultural Area
Occupied buildings	Unskilled sectors employment	Regional Utilized Agricultural Area
Square meters per occupant	Self-employment gender gap	Regional vineyard s for quality wines
Housing underutilisation index	Commuting rate	Regional number of farms
Housing crowding index	Extra-municipality commuting rate	Agriculture employment growth rate
Residential mobility	Municipality study commuting rate	Agriculture employment
Residential housing	Municipality work commuting rate	
Hilly municipalities	Private transport - commuting rate	
Mountain municipalities	Public transport - commuting rate	
Land municipalities	Other means of trasport - commuting rate	
Sismic municipalities	Under 30 minutes commuting rate	
Railroad	Over 30 minutes commuting rate	
Airports	Incidence of unsuitable housing	
Clay	Incidence of large families	
Core area	Families with potential economic hardship	
Connectivity	Incidence of crowded population	
Area	Young outside the labour market and training	
Altitude	Incidence of families in care distress	
Altitude classification	Population growth rate	
	Non agricultural employment growth rate	
	Tradable sectors employmnet growth rate	
	Non-tradable sectors employment growth rate	
	Employment growth rate	
	Accomodation facilities - bed	
	Accomodation facilities	
	Hotel	
	Density of accomodation facilities	
	Criminal organizations	

Sources: National Agricultural Census, ISTAT; National Census, ISTAT; Authors' elaboration from product specification and Qualivita (2021); Geographical Information System and Agenzia Nazionale per l'amministrazione e la destinazione dei beni sequestrati e confiscati alla criminalità organizzata; EU soil database

predictive power is tested (Cerqua et al., 2021; Antulov-Fantulin et al., 2021). The size of these two sets must be chosen taking in to account the trade-off between the benefits of a large training set (i.e., it is the only part of the database on which the algorithm builds the mapping) and the benefits of a quite large testing set (in order to reduce the testing error). Spending too much on training (e.g.,$> 80\%$) will not allow getting a good assessment of predictive performance because it may find a model that fits the training data very well but is not generalizable (overfitting). On the contrary, too much spent in testing ($> 40\%$) will not allow getting a good assessment of model parameters (Boehmke and Greenwell, 2019). To account for this trade-off, we follow one of the most used procedures in the literature which is to randomly divide the database as 70 percent for training and 30 percent for out-of-sample test set (Friedman et al., 2001). The hyper-parameter optimization is only done on the training set. Four different models have been analyzed:

- the Least Absolute Shrinkage and Selection Operator (LASSO): a regression statistical method that performs features selection and regularization with L1 norm to reduce over-fitting and increase prediction accuracy and interpretability (Tibshirani, 1996);

- the Random Forest (RF): a family of randomized tree-based classifier decision trees which uses different random subsets of the features at each split in the tree (Breiman, 2001);

- the Gradient Boosting Machines (GBM): the ensemble method which works in an iterative way where at each stage new learner tries to correct the pseudo-residual of its predecessors (Friedman, 2001);

- the Neural Network (NN): the model that uses an associated set of input/output units in which each connection has a weight associated, and learns by adjusting the weights to predict the correct class label of the given inputs (Ripley et al., 2016).

The performance of PDO classification prediction is assessed by the Receiver Operating Characteristics (ROC) curve (Fawcett, 2006). In our binary classification problem, the positive class is defined as the municipality with high PDO probability and the negative class is the municipality with low PDO probability. The ROC curve shows the classifier diagnostic ability by plotting the true positive rate (TPR) on

the y-axis against the false positive rate (FPR) on the x-axis since its discrimination threshold is varied (Antulov-Fantulin et al., 2021).

Machine learning models also give information on how useful each feature is in explaining PDO. Each model has a different algorithm to estimate importance (Friedman et al., 2001). In LASSO, feature importance is estimated as the absolute value of the coefficients corresponding to the tuned model. For RF, feature importance is the mean gain produced by the feature over all the trees where the gain is measured by the Gini index. The feature importance in GBM is the average improvement of the splitting on the features across all the trees generated by the boosting algorithm. The feature importance in NN is determined by identifying all weighted connections between the layers in the network.

4 Results

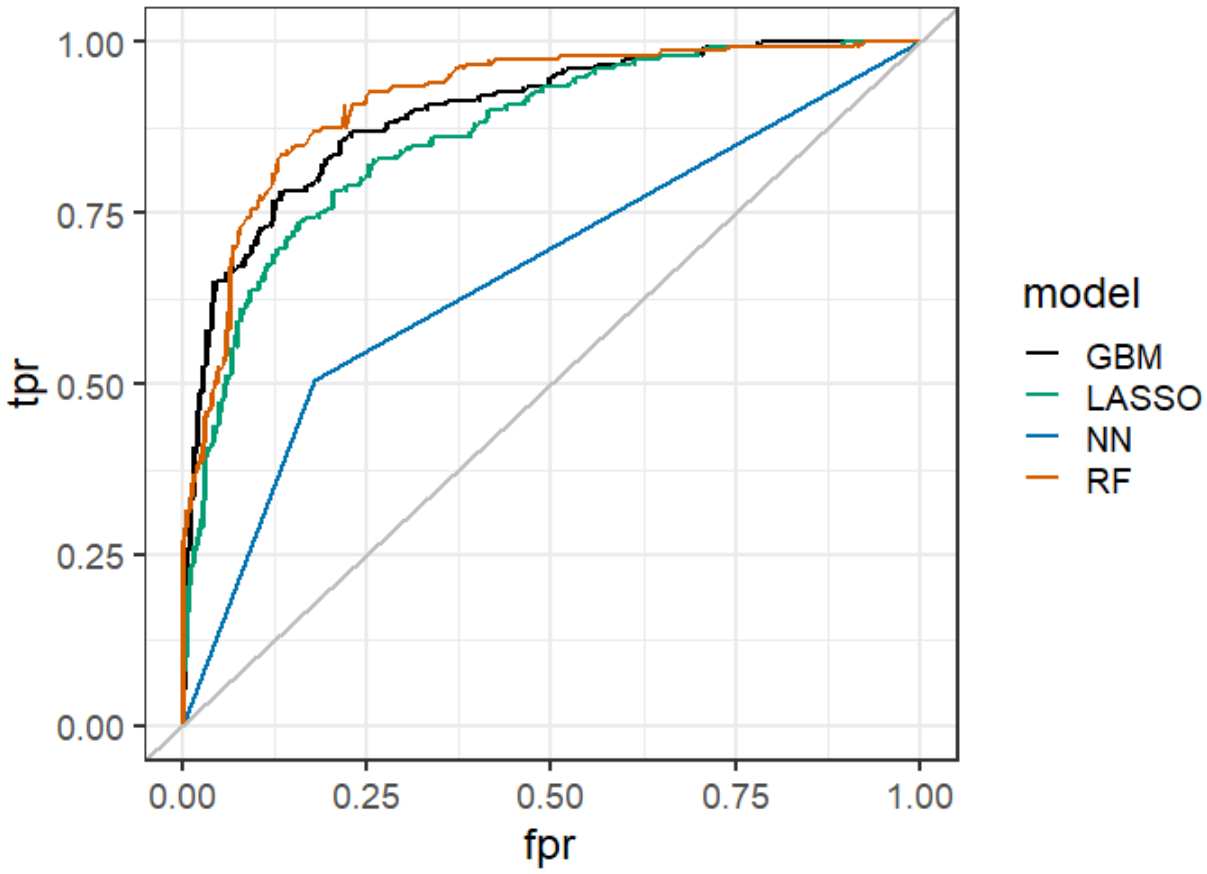
In this section we present the results of the model predicting municipalities with PDO wines. The focus will be on two main aspects: the predictability of our dependent variable (Section 4.1) and the features' importance of independent variables used for predictions (Section 4.2). In these regards, in interpreting features, it is important to clarify that we are looking at which are the potential territorial specific features that in the medium-long run (1 to 10 years) can determine the acknowledgement of a GI in that municipality.

4.1 Predictability of PDO

Figure 2 shows the ROC curves for the four models (GBM, LASSO, NN, and RF) trained on 70% of observations (1056) and tested on the remaining 30% (452). The estimates are based on a repeated cross validation algorithm, which trains and tests the model tuning the hyper-parameters with the aim of maximising the area under the ROC curve. The best model in terms of area under the curve (AUC), is RF (0.9154), the second is GBM (0.8974), followed by LASSO (0.8639), and NN (0.6633) which show lower performances.

Table 2 shows the four models' performances according to the standard measures used in the machine learning literature. The accuracy is statistically higher than the no information rate for all the four models used here (RF, GBM, LASSO, and NN).

Figure 2: ROC curve for four ML models



Models trained on 70% of observations and tested on the remaining 30%

RF and GBM, the best models, show exactly the same accuracy (0.843), while LASSO (0.805) and NN (0.715) have lower performance rates. Overall, Table 2 shows that the RF and GBM models surpass the other models in any of the metrics used: Accuracy, Sensitivity, Specificity, Detection Rate, and Balanced Accuracy. These results, in line with previous empirical applications, confirm that the tree-based models are the more competitive methods for structured binary tasks (Carmona et al., 2019; Climent et al., 2019; Antulov-Fantulin et al., 2021).

Table 2: Models' performances

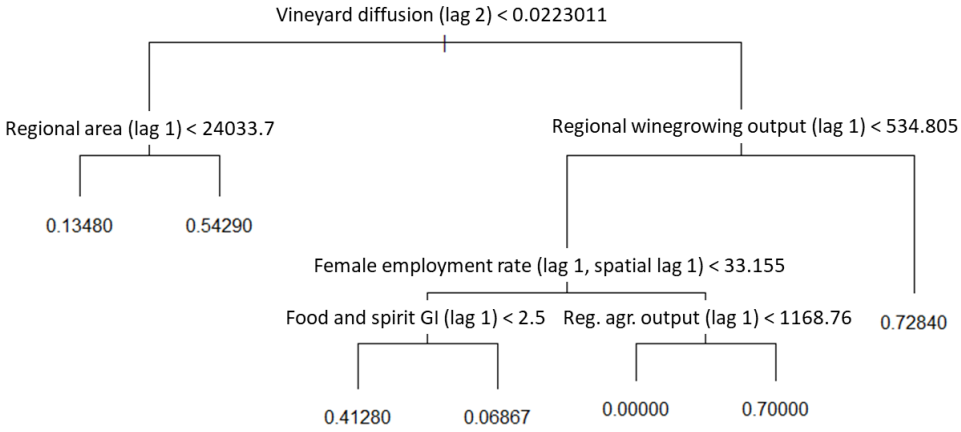
	RF	GBM	NN	LASSO
Accuracy	0.843	0.843	0.715	0.805
95% CI	(0.806, 0.8752)	(0.806, 0.8752)	(0.6706, 0.7558)	(0.7657, 0.8408)
No Information Rate	0.664	0.664	0.664	0.664
P-Value [Acc > NIR]	0.000	0.000	0.012	0.000
Sensitivity	0.937	0.940	0.820	0.850
Specificity	0.658	0.651	0.507	0.717
Pos Pred Value	0.844	0.842	0.766	0.856
Neg Pred Value	0.840	0.846	0.588	0.708
Prevalence	0.664	0.664	0.664	0.664
Detection Rate	0.622	0.624	0.544	0.564
Detection Prevalence	0.737	0.741	0.710	0.659
Balanced Accuracy	0.797	0.796	0.663	0.784

4.2 Features Importance

In this section we show the importance of each feature in the prediction task. As the two best performing models (RF and GBM) are based on combinations of different regression trees (Hastie et al., 2009), a first description of the data mapping can be represented by a decision tree (Figure 3). Figure 4 reports the first 10 most important features to predict PDO acknowledgement in Random Forest, while Figure 5 shows the first 10 important features to predict PDO acknowledgement in the Gradient Boosted Machine.

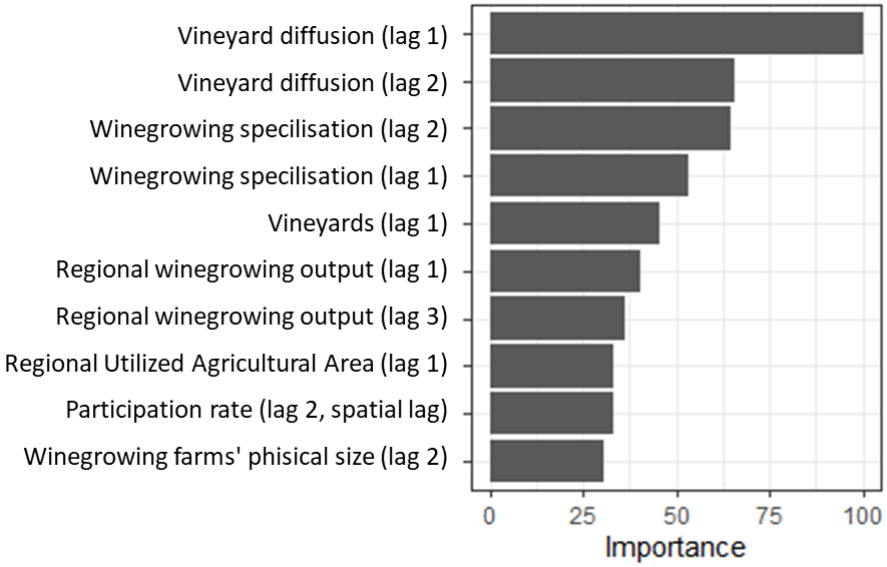
Overall, findings reveal that the main important features to predict the establishment of a PDO are spatial and territorial: the viticulture and winegrowing tradition

Figure 3: Regression Tree over the whole sample



of municipalities emerges as crucial. This is firstly evident in the decision tree, for which the historical share of agricultural area cultivated with vine in the municipality (with two time lags, which means 20 years before) is the top feature predicting the inclusion of a municipality within a PDO area. Even using the Random Forest, the historical dedication of the area to winegrowing emerges due to the fact that it is not only relevant to the share of agricultural area cultivated with vines along years in the municipality, but also to the physical dimension results and as a stand-alone important feature. However, the geographical concentration of farmers working in the wine sector also appears among the top 10 important features with the Random Forest. This indicator captures the importance of the presence of small and medium sized local producers, with relatively small holdings, rather than that of a few big farms. This finding confirms that one way to improve market access for origin-linked products produced and processed by family farmers and small enterprises is to develop GIs, as stated by Vandecandelaere et al. (2018). This is particularly relevant due to the fact that the Italian wine supply is mainly characterised by an organisation that hinders the exploitation of economies of scale and a "district" nature of a large part of the sector (Pomarici et al., 2021). Variables capturing (i) the diffusion of the vine cultivation and (ii) the structural characteristics of the local wine sector are present also by using the Gradient Boosted Machine (Figure 5). As highlighted by

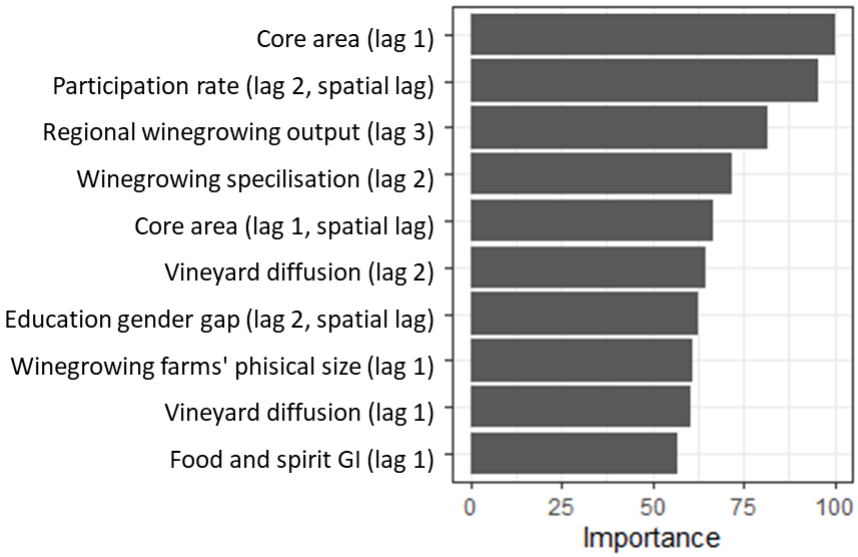
Figure 4: Feature Importance to predict PDO: the first 10 important features in Random Forest



the literature, the economic specialisation (i.e. output production in terms of million of euro) of the entire region toward agriculture and wine production can play a crucial role (Ferretti and Gandino, 2018).

Compared to previous literature, our results show some elements of novelty on the success of wine PDOs (Vaquero-Piñeiro, 2021) by providing a clear evidence of the relevance of the territorial characteristics and historical cultural tradition for wine-growing in the regions of origin. This result complements the evidence provided in Vaquero-Piñeiro (2021), who found that socio-economic predictors were not statistically significant, meaning that some territorial indicators had a stronger role and could help more than others, but not including them in the analysis. Furthermore, this paper has focused on all Italian PDO wines, rather than only on the ten most economically relevant ones as done in Vaquero-Piñeiro (2021). Our findings also show that the total number of food and spirit GIs producing in the same municipality and already registered appears as one of the most important features. The geographical concentration of GIs, especially economic relevant GIs, is still a debated question. The literature has pointed out that the primary users of this quality scheme were the Southern EU Member States, which registered seven times more food GIs per capita

Figure 5: Feature Importance to predict PDO: the first 10 important features in Gradient Boosted Machine



than in other EU countries (Huysmans and Swinnen, 2019). The leaders were Italy and France, both in terms of numbers and revenues (EC, 2020). Regarding Italy, it should also be considered that PDOs are spatially concentrated in the North-Central Italy (Vaquero-Piñeiro, 2021).

Additional important features include the share of female employment in neighbouring municipalities and the total area of the region. This is particularly relevant since we are looking at Italy, where regions with larger areas tend to coincide with regions which are more developed (except for Sicily and Sardinia islands) and characterized by an outstanding agri-food sector (e.g., Emilia Romagna). This is also relevant in an historical dual country, in which the socio-economic gap between Northern and Southern regions goes beyond the simplest economic indicators, e.g., GDP (Svimez, 2020; Greco et al., 2018). According to the Random-Forest output (Figure 3), the employment participation rate of neighbouring municipalities is one of the most important features suggesting the importance of being in a vital economic system to obtain the GI. This evidence is confirmed by applying the Gradient Boosted Machine (Figure 4). The employment participation rate of neighbouring municipalities, the gender gap between educated inhabitants, and the distance from the capital city of

the region also have roles showing-up among the most import features in several models. Even controlling for the whole set of alternative socio-economic related elements (see Table A1), such as the presence of criminal organizations, proxied by the confiscation of properties belonging to individuals convicted for mafia-related crimes (Boeri et al., 2019), being a remote municipality, far from the main economic and infrastructural centres of the areas, matters. In this regard it has been largely shown that, due to a combination of globalisation and technological change, rural regions have been characterised by lower labour-force participation and income, while many large metropolitan areas have been more prosperous in terms of income and employment (Iammarino et al., 2019), and this has increased the historical gap between the regions at the core and the regions at the periphery in many countries (Krugman, 1991). Our results show that remoteness from the core also influences the potential acknowledgement of GI, which is a typical outcome of rural areas. As the GIs have the potential to foster local economic development (Crescenzi et al., 2021), the role played by distance from the center can give rise to a duality within rural areas in addition to the well known rural–urban differential (Bourguignon and Morrisson, 1998). With reference to the core-periphery dynamics, in Italy territories characterized by a significant distance from the main centers for providing citizenship services, i.e., those services connected to the quality of life (health, education, mobility), are now targeted by specific cohesion policies: *Strategia Nazionale per le Aree Interne - SNAI*[4]. Hopefully future policies will also help such rural areas in the agri-food quality certification, fostering the endogenous rural development, i.e., strategies focused on enhancing local resources specific to a sector, such as cultural values (Mikulcak et al., 2015). Overall, in Italy vineyards are spread over the whole country, and they have contributed to maintaining adequate socio-economic conditions in some lagging regions for decades, especially after the Second World War (Corsi et al., 2019), of course, among the variables not included in this model, there may be a set of intangible factors that can explain a significant part of the acknowledgement process.

5 Conclusions

This paper aims at investigating whether the establishment of a PDO is predictable by the territorial features of the region of origin in previous years. The model is developed based on 1508 municipalities (999 municipalities which do not receive PDO

[4]https://www.agenziacoesione.gov.it/strategia-nazionale-aree-interne/

in 2002-2011 and 509 who did receive PDO in 2002-2011). Our results suggest that it is possible to make out-of-sample predictions of municipalities that have obtained the PDO status for wine productions in the period 2002-2011, with socioeconomic, agri-food sector related, and territorial characteristics of municipalities with reference to a period in the past (1981-2001).

Features' importance suggests that territorial factors play a significant role, and they are more important than socio-economic conditions to predict the inclusion of municipalities within a PDO wine area. In particular, variables capturing the historical traditions, the specialisation and the presence of local networks (local actors involved in winemaking) are among the top ten important features with all the different algorithms considered. This provided evidence of the importance not only of tangible capital (e.g., Utilised Agricultural Area - UAA), but also of intangible capital for the certification of PDO wines. As stated by Pomarici et al. (2021), the presence of local networks and linkages, some of which are formal and others informal, gives most Italian local production systems specialising in grapes and wine the characteristics of industrial districts (Sforzi, 2008), due to the local social, environmental and cultural (in a single word territorial) capital that is stratified there (Muringani J and A., 2021).

The nexus between GI and territories is one of the pillars of the EU quality scheme; however, so far the literature has not attempted to provide econometric evidence of the conceptual rationale behind that scheme. For the first time in the literature, by using a predictive model, rather than ex-post evaluation techniques, this analysis has uncovered the role of territorial factors. From the policy perspective, this paper offers a valuable tool for predicting areas that have the potential for being acknowledge with PDO wines. This is important not only in order to implement rural development policies, but also to target investment (both private and public) allocation. At the same time, more linked to the GI quality scheme, this approach could also be relevant to investigate (i) if the GI scheme has been working in line with the regulatory framework, (ii) which are the areas for which the request of an upgrade from PGI to PDO will be presumably accepted and (iii) which are the neighboring municipalities that have a greater probability of being included within the production area if there is a request for extending the demarcated area to satisfy the demand.

Our analysis does not investigate what happens in the case of PDO food and

weather territorial features also having a role also in predicting PDOs in other EU countries, but both of these issues are in our agenda for future research.

References

Altomonte, C., Colantone, I., and Pennings, E. (2016). Heterogeneous firms and asymmetric product differentiation. *The Journal of Industrial Economics*, 64(4):835–874.

Amin, M. D., Badruddoza, S., and McCluskey, J. J. (2021). Predicting access to healthful food retailers with machine learning. *Food Policy*, 99:101985.

Andini, M., Ciani, E., de Blasio, G., D'Ignazio, A., and Salvestrini, V. (2018). Targeting with machine learning: An application to a tax rebate program in italy. *Journal of Economic Behavior & Organization*, 156:86–102.

Antulov-Fantulin, N., Lagravinese, R., and Resce, G. (2021). Predicting bankruptcy of local government: A machine learning approach. *Journal of Economic Behavior & Organization*, 183:681–699.

Ballestar, M. T., Doncel, L. M., Sainz, J., and Ortigosa-Blanch, A. (2019). A novel machine learning approach for evaluation of public policies: An application in relation to the performance of university researchers. *Technological Forecasting and Social Change*, 149:119756.

Barca, F., M. P. and Rodríguez-Pose, A. (2012). The case for regional development intervention: place-based versus place-nautral approches. *Journal of Regional Science*, 52:134–152.

Belletti, G., Marescotti, A., Sanz-Cañada, J., and Vakoufaris, H. (2015). Linking protection of geographical indications to the environment: Evidence from the european union olive-oil sector. *Land Use Policy*, 48:94–106.

Belletti, G., Marescotti, A., and Touzard, J.-M. (2017). Geographical indications, public goods, and sustainable development: The roles of actors' strategies and public policies. *World Development*, 98:45–57.

Boehmke, B. and Greenwell, B. M. (2019). *Hands-on machine learning with R*. CRC Press.

Boeri, F., Di Cataldo, M., and Pietrostefani, E. (2019). Out of the darkness: Reallocation of confiscated real estate mafia assets. *Available at SSRN 3488626*.

Bourguignon, F. and Morrisson, C. (1998). Inequality and development: the role of dualism. *Journal of development economics*, 57(2):233–257.

Bowen, S. (2010). Embedding local places in global spaces: Geographical indications as a territorial development strategy. *Rural Sociology*, 75(2):209–243.

Bramley, C., Biénabe, E., and Kirsten, J. (2009). The economics of geographical indications: towards a conceptual framework for geographical indication research in developing countries. *The economics of intellectual property*, 1:109–141.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Carmona, P., Climent, F., and Momparler, A. (2019). Predicting failure in the us banking sector: An extreme gradient boosting approach. *International Review of Economics & Finance*, 61:304–323.

Carrieri, V., Lagravinese, R., and Resce, G. (2021). Predicting vaccine hesitancy from area-level indicators: A machine learning approach. *Health Economics*.

Cei, L., Stefani, G., Defrancesco, E., and Lombardi, G. V. (2018). Geographical indications: A first assessment of the impact on rural development in italian nuts3 regions. *Land Use Policy*, 75:620–630.

Cerqua, A., Di Stefano, R., Letta, M., and Miccoli, S. (2021). Local mortality estimates during the covid-19 pandemic in italy. *Journal of Population Economics*, pages 1–29.

Cerqua, A. and Letta, M. (2021). Local inequalities of the covid-19 crisis. *Regional Science and Urban Economics*, page 103752.

Charters, S. and Spielmann, N. (2014). Characteristics of strong territorial brands: The case of champagne. *Journal of Business Research*, 67(7):1461–1467.

Climent, F., Momparler, A., and Carmona, P. (2019). Anticipating bank distress in the eurozone: An extreme gradient boosting approach. *Journal of Business Research*, 101:885–896.

Combes, P.-P., Gobillon, L., and Zylberberg, Y. (2021). Urban economics in a historical perspective: Recovering data with machine learning. *Regional Science and Urban Economics*, page 103711.

Corsi, A., Mazzarino, S., and Pomarici, E. (2019). The italian wine industry. In *The Palgrave Handbook of Wine Industry Economics*, pages 47–76. Springer.

Costanigro, M., Scozzafava, G., and Casini, L. (2019). Vertical differentiation via multi-tier geographical indications and the consumer perception of quality: The case of chianti wines. *Food Policy*, 83:246–259.

Crescenzi, R., De Filippis, F., Giua, M., and Vaquero-Piñeiro, C. (2021). Geographical indications and local development: the strength of territorial embeddedness. *Regional Studies*, pages 1–13.

De Rosa, M. (2015). The role of geographical indication in supporting food safety: a not taken for granted nexus. *Italian journal of food safety*, 4(4).

Einav, L. and Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy*, 14(1):1–24.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Ferretti, V. and Gandino, E. (2018). Co-designing the solution space for rural regeneration in a new world heritage site: A choice experiments approach. *European Journal of Operational Research*, 268(3):1077–1091.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Greco, S., Ishizaka, A., Matarazzo, B., and Torrisi, G. (2018). Stochastic multi-attribute acceptability analysis (smaa): an application to the ranking of italian regions. *Regional studies*, 52(4):585–600.

Haeck, C., Meloni, G., and Swinnen, J. (2019). The value of terroir: A historical analysis of the bordeaux and champagne geographical indications. *Applied Economic Perspectives and Policy*, 41(4):598–619.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learnin. *Cited on*, page 33.

Hoffman, I. and Mast, E. (2019). Heterogeneity in the effect of federal spending on local crime: Evidence from causal forests. *Regional Science and Urban Economics*, 78:103463.

Hossain, M., Mullally, C., and Asadullah, M. N. (2019). Alternatives to calorie-based indicators of food security: An application of machine learning methods. *Food policy*, 84:77–91.

Huysmans, M. and Swinnen, J. (2019). No terroir in the cold? a note on the geography of geographical indications. *Journal of agricultural economics*, 70(2):550–559.

Iammarino, S., Rodríguez-Pose, A., and Storper, M. (2019). Regional inequality in europe: evidence, theory and policy implications. *Journal of economic geography*, 19(2):273–298.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5):491–95.

Krugman, P. (1991). Increasing returns and economic geography. *Journal of political economy*, 99(3):483–499.

Lentz, E., Michelson, H., Baylis, K., and Zhou, Y. (2019). A data-driven approach improves food insecurity crisis prediction. *World Development*, 122:399–409.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., and Barton, D. (2012). Big data: the management revolution. *Harvard business review*, 90(10):60–68.

McBride, L. and Nichols, A. (2018). Retooling poverty targeting using out-of-sample validation and machine learning. *The World Bank Economic Review*, 32(3):531–550.

Meloni, G. and Swinnen, J. (2018). Trade and terroir. the political economy of the world's first geographical indications. *Food Policy*, 81:1–20.

Mikulcak, F., Haider, J. L., Abson, D. J., Newig, J., and Fischer, J. (2015). Applying a capitals approach to understand rural development traps: A case study from post-socialist romania. *Land Use Policy*, 43:248–258.

Moschini, G., Menapace, L., and Pick, D. (2008). Geographical indications and the competitive provision of quality in agricultural markets. *American Journal of Agricultural Economics*, 90(3):794–812.

Muringani J, F. R. and A., R.-P. (2021). Social capital and economic growth in the regions of europe. environment and planning a: Economy and space. *Environment and Planning A: Economy and Space*, 53(6):1412–1434.

Newton, S. K., Gilinsky Jr, A., and Jordan, D. (2015). Differentiation strategies and winery financial performance: An empirical investigation. *Wine Economics and Policy*, 4(2):88–97.

Pomarici, E., Corsi, A., Mazzarino, S., and Sardone, R. (2021). The italian wine sector: evolution, structure, competitiveness and future challenges of an enduring leader. *Italian Economic Journal*, pages 1–37.

Ripley, B., Venables, W., and Ripley, M. B. (2016). Package 'nnet'. *R package version*, 7(3-12):700.

Sforzi, F. (2008). Il distretto industriale: da marshall a becattini. *Il distretto industriale*, pages 1000–1010.

Sommervoll, Å. and Sommervoll, D. E. (2019). Learning from man or machine: Spatial fixed effects in urban econometrics. *Regional Science and Urban Economics*, 77:239–252.

Svimez, R. (2020). *Rapporto SVIMEZ. L'economia e la società del Mezzogiorno*. Il mulino Bologna.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Török, Á., Jantyik, L., Maró, Z. M., and Moir, H. V. (2020). Understanding the real-world impact of geographical indications: A critical review of the empirical economic literature. *Sustainability*, 12(22):9434.

Vandecandelaere, E., Teyssier, C., Barjolle, D., Jeanneaux, P., Fournier, S., and Beucherie, O. (2018). Strengthening sustainable food systems through geographical indications: an analysis of economic impacts. Technical Report 13, European Bank for Reconstruction and Development (EBRD).

Vandecandelaere, E.; Samper, L. R. A. D. A. M. P. T. F. V. M. (2021). he geographical indication pathway to sustainability: A framework to assess and monitor the contributions of geographical indications to sustainability through a participatory process. *Sustainability*, 13(7535).

Vaquero-Piñeiro, C. (2021). The long-term fortunes of territories as a route for agri-food policies: Evidence from geographical indications. *Bio-Based and Applied Economics*, 10(2):89–108.

Wirth, D. A. (2016). Geographical indications, food safety, and sustainability: conflicts and synergies. *Bio-based and Applied Economics*, 5(2).

A Appendix

Table A1: Data Description

Variable	Definition
Population	Number of residents
Urban area	Share of land classified as cities or functional urban areas
Remoteness	Share of people living in remote areas
Population density	Population density
Gender gap	Ratio between male and female residents
Young population	Share of under 6 years resident population
Elderly population	Share of over 75 resident population
Elderly population	Share of over 65 years resident population
Young population	Share of under 14 years resident population
Erderly rate	Ratio between the elderly population and the working age population (15-64 years)
Divorce rate	Percentage of people legally separated or divorced
Immigration rate	Percentage of foreign population
Young immigrants	Share of under 18 years foreign population
Marriage rate	Percentage ratio of married or de facto couples with a foreign spouse to the toatle of married or de facto couples
Foreign-born employment rate	Foreign-born unemployment rate
Foreign-born unemployment rate	Share of foreign workers on national workers (persons aged 15-64)
Share of foreign-born workers	Share of foreignunemployed people on national unemployed people (persons aged 15-64)
Residential mobility - Foreign-born workers	Percentage ratio between the foreign resident population of 15-24 years enrolled in a regular course of study and / or professional and the total of the resident foreign resident population aged 15-24 years
Italian/foreign school attendance ratio	Percentage ratio between the school attendance rate of Italians and foreign resident population
Italian/foreign school indipendents ratio	Percentage ratio between the rate of Italian independents (self-employed Italians compared to italian employees) and that of foreigners (foreign self-employed compared to foreign employees).
Avarage size of families	Ratio of the resident population in the family to the number of households
Families without childern	Percentage of families without children
Families with children	Percentage of families with children (two or more)
Young single-person families	Percentage ratio people living alone (15-34 years)
Young single parent families	Percentage ratio of families with only one parents (15-34 years)
Youg families without children	Percentage ratio of young families without children (15-34 year)
Youg families with children	Percentage ratio of young families with children (15-34 year)
Single-person fmilies	Percentage ratio of older single-person (non-cohabiting) households (aged 65 and over) to the population aged 65 and over
Single parent families	Percentage ratio of families with only one parents (over 35)
Families without childre	Percentage ratio of young families without children (over 35)
Families with children	Percentage ratio of young families with children (over 35)
Homeownership rate	Percentage ratio of occupied dwellings to total occupied dwellings
average size of a single-family home	Ratio between the total area of occupied dwellings (m2) and the total number of occupied dwellings

Table A1: (Continued)

Variable	Definition
Potential for residential use in built-up areas	Percentage ratio between unoaded dwellings in built-up areas and total dwellings in built-up areas
Potential for residential use in nuclei and scattered houses	Percentage ratio between unoceded dwellings in households and scattered houses and total dwellings in households and scattered houses
Avarage age of buildings	Difference between the year of census and the year of construction of the dwelling (after 1962)
Index of availability of services in the home	Share of houses with basic services (es., drinkable water)
Buildings - good conditions	Percentage of buildings in good preservation conditions
Buildings - bad conditions	Percentage of buildings in the worst preservation conditions
Occupied historical buildings	Percentage ratio of occupied dwellings built before 1919 to total occupied dwellings
Occupied buildings	Percentage of the number of occupied dwellings built in the last decade in towns and villages and the number of those built in the previous decade
Square meters per occupant	Ratio between the total area of occupied dwellings (m2) and the total number of occupants of occupied dwellings
Housing underutilisation index	Ratio between occupied dwellings with more than 80 m2 and 1 occupant or with more than 100 m2 and less than 3 occupants or with more than 120 square meters and less than 4 occupants and the total number of occupied dwellings
Housing crowding index	Ratio between occupied homes with less than 40 m2 and over 4 components or with 40-59 m2 and over 5 components or with 60-79 m2 and over 6 components and the total number of occupied dwellings
Residential mobility	Percentage ratio between the resident population that has changed habitual residence in the last year and the total resident population
Residential housing	Ratio of occupants and rooms of dwellings occupied by resident population
Education gender gap	Difference of education between women and men (high school education; over 6 years)
Education rate	Percentage ration between the resident aged 25-64 years attending a regular course of study and / or vocational training and resident population of 25-64 years
High education rate (educated people)	Percentage ratio of resident population aged 25-64 with diploma or degree to those of the same age with a middle school license
Illiterate rate	Percentage ratio between the resident population aged 6 years and older illiterate and the resident population aged 6 and over
Middle education rate (young people)	Percentage ration between the resident aged 15-24 years with a middle school license who does not attend a regular course of study and / or vocational training and the resident population of 15-24 years
High education rate	Percentage ration between the resident of 25-64 years with a high school diploma or university degree and the resident population of 25-64 years
High education rate (young people)	Percentage ration between the resident of 25-34 years with a high school diploma or university degree and the resident population of 25-34 years
Education rate (15-19 years)	Share of resident population of 15-19 years with a lower secondary school diploma or high school diploma and resident population of 15-19 years
Middle education rate (adults)	Percentage ratio between resident population of 25-64 years with a lower average license and the resident population of 25-64 years
Incidence of graduates	Percentage ratio between resident population of 6 and more years graduated and graduated on the total population of the same age
Male participation rate	Percentage ratio between the active male resident population and the male resident population of the same age group
Female participation rate	Percentage ratio between the active female resident population and the female resident population of the same age group
Participation rate	Percentage ratio between the active resident population and the resident population of the same age group
Incidence of inactive young population	Percentage ratio between the resident population aged 6 years and older illiterate and the resident population aged 6 and over
Inactivity rate (young people)	Percentage ratio between resident population of 15-29 years old not student and not employed and resident population of 15-29 years old
Male unemployment rate	Percentage ratio between the male resident population aged 15 and over seeking employment and the male resident population aged 15 and over.
Female unemployment rate	Percentage ratio between resident population aged 15 and over seeking employment and resident population aged 15 and over.

Table A1: (Continued)

Variable	Definition
Youth unemployment rate	Percentage ratio between the resident population of 15-24 years looking for work and the resident population of 15-24 years active
Male employment rate	Ratio of the employed male resident population to total the male resident population aged 15 and over
Female employment rate	Ratio of the employed female resident population to the female resident population aged 15 and over
Employment rate	Ratio of employed resident population to total resident population aged 15 and over
employment turnover	Percentage ratio of employed people over 45 to those aged 15-29
Youth employment rate	Percentage ratio between the employed resident population of 15-24 years the resident population of 15-24 years
Agriculture employment	Share of economically active population working in agriculture, forestry and fishing sectors
Industrial employment	Share of economically active population working in industrial sectors
Non-tredable sectors employmnet	Share of economically active population in non-tradable sectors
Tradable sectors employment	Share of economically active population in tradable sectors
Female non-tredable sectors employmnet	Share of economically active famele population in non-tradable sectors
High-specialisation employment	Share of economically active population in high-specialisation sectors
Specific-specialisation employment	Share of economically active population in specific-specialisation sectors (e.g., handicraft, agriculture)
Unskilled sectors employment	Share of economically active population in unskilled sectors
Self-employment gender gap	Ratio between male and female self-employment
Commuting rate	Percentage ratio between resident population who travel daily to go to the place of work or study and resident population aged up to 64 years
Extra-municipality commuting rate	Percentage ratio between resident population who travel daily to go to the place of work or study and resident population aged up to 64 years - different municipality
Municipality study commuting rate	Percentage ratio among resident population who travel daily for work purposes outside the municipality of habitual residence and resident population who move daily for work reasons within the municipality of habitual residence
Municipality work commuting rate	Percentage ratio among resident population who travel daily for study purposes outside the municipality of habitual residence and resident population who move daily for work reasons within the municipality of habitual residence
Private transport - commuting rate	Percentage ratio among resident population who travel daily for work or study purposes outside the municipality of habitual residence and resident population who move daily for work reasons within the municipality of habitual residence - private transport
Public transport - commuting rate	Percentage ratio among resident population who travel daily for work or study purposes outside the municipality of habitual residence and resident population who move daily for work reasons within the municipality of habitual residence - public transport
Other means of trasport - commuting rate	Percentage ratio among resident population who travel daily for work or study purposes outside the municipality of habitual residence and resident population who move daily for work reasons within the municipality of habitual residence - bycicle or by foot
Under 30 minutes commuting rate	Percentage ratio among resident population who travel daily for work or study purposes outside the municipality of habitual residence and resident population who move daily for work reasons within the municipality of habitual residence - 30 minutes or less
Over 30 minutes commuting rate	Percentage ratio among resident population who travel daily for work or study purposes outside the municipality of habitual residence and resident population who move daily for work reasons within the municipality of habitual residence - more than 30 minutes
Incidence of unsuitable housing	Percentage ratio between the number of other types of housing and the total number of dwellings
Incidence of large families	Percentage ratio between the number of households with 6 or more members and the total number of households
Incidence of families with potential economic hardship	Percentage ratio between the number of families with children with the reference person aged up to 64 years in which no member is employed or withdrawn from work and the total number of families

Table A1: (Continued)

Variable	Definition
Incidence of crowded population	Percentage ratio between the population residing in dwellings with an area of less than 40 square meters and more than 4 occupants either in 40-59 sqm and more than 5 occupants or in 60-79 sqm and more than 6 occupants, and the total population residing in occupied dwellings
Incidence of young people outside the labour market and training	Percentage ratio between resident population of 15-29 years in non-professional condition other than student on resident population of the same age
Incidence of families in care distress	Percentage ratio between the number of households with at least two members, without co-inhabitants, with all members aged 65 and over and with the presence of at least one member aged 80 and over, and the total number of families
Population growth rate	Growth rate of the absolute number of inhabitants
Non agricultural employment growth rate	Growth rate of the share of economically active population in other (tradable and non-tradable) sectors
Tradable sectors employmnet growth rate	Growth rate of the share of economically active population in other tradable sectors
Non-tradable sectors employment growth rate	Growth rate of the share of economically active population in non-tradable sectors
Agriculture employment growth rate	Growth rate of the share of economically active population working in agriculture, forestry and fishing sectors
Employment growth rate	Growth rate of the share of economically active population
Accomodation facilities - bed	Number of beds
Accomodation facilities	Number of accommodation facilities (Hotels; holiday and other short-stay accommodation; camping grounds, recreational vehicle parks and trailer parks)
Hotel	Number of hotels
Density of accomodation facilities	Number of accommodation facilities per km2
Hilly municipalities	Dummy = 1 if hilly municiplaities
Mountain municipalities	Dummy = 1 if mountain municiplaities
Land municipalities	Dummy = 1 if land municiplaities
Sismic municipalities	Dummy = 1 if seismic
Criminal organizations	Dummy = 1 if municipalities with confiscated properties belonging to individuals convicted for mafia-related crimes
Railroad	Km of railroad
Airports	Dummy = 1 for municipalities with airports
Clay	Percentage of clay in the soil
Core area	Distance from the major city of the region (meters)
Connectivity	Distance from the major city of the region (minutes)
Area	Area (km2)
Altitude	Avarage of the level of altitude
Altitude classification	Categorical variable classifying municipalities according to the level of altitude: low, moderate and high altitude
Total Agricultural Area	Total Agricultural Area (km2)
Utilised Agricultural Area	Utilized Agriucultural Area - UAA (km2)
Farms	Number of farms
Small farms	Number of farms with UAA between 1-10 ha
Medium farms	Number of farms with UAA between 10-50 ha
Big famrs	Share of farms with UAA more than 100 ha
Family farms	Share of family employees
Farms' phisical size	Utilized Agriucultural Area/number of farms
Agricultural land intensity	Ratio between the Utilized Agricultural Area and the Total Agricultural Area

Table A1: (Continued)

Variable	Definition
Agricultural land diffusion	Utilized Agriucultural Area/area
Family workers	Number of family workers employed in farms
Non-family workers	Number of non-family workers employed in farms
Number of employees	Number of employees working in farms
Employment intensity	Ratio between Utilized Agriucultural Area and the number of workers
Vineyards	Utilised Agricultural Area for vines
Vineyards (dummy)	Dummy
Winegrowing farms	Number of farms specialized in winegrowing
Winegrowing farm density	Winegrowing farms per km2
Winegrowing specilisation	Ration between winegrowing farms and the total number of farms
Vineyard diffusion	Share Utilised Agricultural Area for vines (ha)
Winegrowing farms' phisical size	Utilized Agriucultural Area for vines/number of farms specilised in winegrowing
Single-grape wines	Dummy = 1 for municipalities with single-grape IG wines
Sparkling wines	Dummy = 1 for municipalities with IG sparkling wine (Spumante or Prosecco)
Food and spirit GI	Number of total food GIs acknowledged the municipality
Unesco area	Dummy = 1 if municipality is within a vineyard UNESCO area
Main economic relavant DOP (1)	Dummy = 1 for municipalities where Mozzarella di Bufala Campana DOP is produced
Main economic relavant DOP (2)	Dummy = 1 for municipalities where Mozzarella di Bufala Campana DOP is produced
Main economic relavant DOP (3)	Dummy = 1 for municipalities where Parmigiano Reggiano DOP is produced
Main economic relavant DOP (4)	Dummy = 1 for municipalities where Gorgonzola DOP is produced
Main economic relavant DOP (5)	Dummy = 1 for municipalities where Grana Padano DOP is produced
Main economic relavant DOP	Dummy = 1 for municipalities where one of the most economically relevant Italian DOP is produced
Regional agricultural aoutput	Regional output of agricultural sector - basic and producer prices
Regional winegrowing output	Regional output of winegrowing activities - basic and producer prices
Regional area	Regional area (ha)
Regional vineyard diffusion	Share of regional Utilised Agricultural Area for vines (ha)
Regional Total Agricultural Area	Regional Total Agricultural Area (ha)
Regional Utilized Agricultural Area	Regional Utilized Agricultural Area
Regional vineyard s for quality wines	Regional Utilized Agricultural Area for DOC and DOCG wines(ha)
Regional number of farms	Number of farms located in the region