Bio-based and Applied Economics 9(1): 85-107, 2020
DOI: 10.13128/bae-8464

Full Research Article

The role of group-time treatment effect heterogeneity in long standing European agricultural policies. An application to the European geographical indication policy

Leonardo Cei[1,*], Gianluca Stefani[2], Edi Defrancesco[1]

[1] University of Padova (Italy)
[2] University of Florence (Italy)

Abstract. In recent years, the European Union is stressing the importance of monitoring and evaluating its policies, among which the common agricultural policy plays an important role. Policy evaluation, in order to provide reliable results on which to take important legislative decisions, should rely on robust methodological tools. A recent strand of literature casts some doubts about the reliability of the two-way fixed effect estimator when the effect of a treatment is heterogeneous across groups of units or over time. This estimator is widely used in agricultural economics to estimate the effect of policies where effect heterogeneity may be at stake. Using the European geographical indication (GI) policy, we compared the two-way fixed effects estimator with a novel non-parametric estimator that accounts for the issues created by effect heterogeneity. The results show that the two estimators, consistently with the concerns expressed by the technical literature, may lead to different estimates of the policy effect. This suggests that treatment effect heterogeneity is likely a concern when assessing the impact of GI-type policies. Therefore, the use of the standard estimator may lead to misleading conclusions and, as a result, to inappropriate policy actions.

Keywords. Treatment heterogeneity; geographical indications; impact assessment; two-way fixed effects; policy evaluation

JEL codes. Q18, Q56.

1. Introduction

In recent years, the European Union (EU) is stressing the need to move toward an ever more evidenced-based policy making. Despite the renewed attention it is attracting nowadays, evidence-based policy making is not a new concept. The discussion about the need to use empirical evidence to understand how policies work and to identify their results was already in place in the 1990s (e.g., OECD, 1994; Pawson and Tilley, 1997).

*Corresponding author. E-mail: leonardo.cei@phd.unipd.it

Editor: Fabio Gaetano Santeramo.

ISSN 2280-6180 (print)
ISSN 2280-6172 (online)

Sanderson (2002) claims that two kinds of evidences are required to improve the governmental action. On the one hand, it is necessary to understand whether the policy action is effective. On the other hand, acquiring knowledge about how a certain policy works is of fundamental importance. In the language of Yin (2013), this corresponds to answer, respectively, a "what" and a "why" question.

Especially the former aspect plays an important role in the current EU Common Agricultural Policy (CAP), where the legislator stresses the importance of a constant monitoring and evaluation of its measures, also providing indicators and methodological guidelines, as well as some ex-ante evaluations on quantitative goals. On the verge of the new CAP programming period (2021-2027), the policy course that aims at providing evidences about the effectiveness of the policies and measures of the CAP is confirmed and stressed. The new CAP Regulation proposal states that "the current Common Monitoring and Evaluation Framework (CMEF) and the current monitoring system of Direct Payments and Rural Development would be used as a basis for monitoring and assessing policy performance, but they will have to be streamlined and further developed" (European Commission, 2018: pg. 9).

The rising interest in evidence-based policy making, however, requires proper tools to collect evidences, analyze them and interpret the results. In this respect, a useful reservoir of approaches, methods and techniques to be used in the evaluation process is represented by quasi-experimental approaches. Adopting an *ex-post* perspective (i.e., after the policy has been implemented), the main goal of quasi-experiments is to identify the effect that a certain policy, program or treatment produces on some indicator that measures the policy objectives. Basically, this requires to clearly identify the causal relationship between the treatment and the outcome, in order to isolate the effect of the policy from the role played by other confounding factors (Khandker et al., 2009). The identification of this causal link, however, constitutes the major effort in real socio-economic contexts. Different policy settings have different pitfalls that hinder the correct identification of the causal effect. To overcome these issues, researchers came up, over the years, with strategies and techniques tailored to specific policy settings. To cite some examples, regression adjustment and matching are ways to account for the effect of observable covariates; instrumental variables and difference-in-differences (DID) can get rid of the influence of unobservable factors (Cerulli, 2015); the regression discontinuity design is well suited in contexts where the administration of the treatment is based upon certain thresholds. As a result, before starting an impact analysis, the researcher should pay attention to the policy he/she aims at evaluating and to the setting where the policy is implemented.

The ideal policy setting for impact analysis involves a binary treatment that is administered to one group of individuals, while another group can be used as a control. The two groups can be observed at a single point in time or over a couple of periods. However, some policies are characterized by more complex settings, as is the case of several EU agricultural policies. This is especially the case of long-standing policies, where the participation is voluntary, the enrollment in the treatment not simultaneous, and individuals can be observed for multiple time periods. The policy we refer to in our article, the geographical indication (GI) policy, is an example of this situation. Provided that their farm is located in the area of origin of a GI product, farmers do not have any obligation about whether or when to join the specific GI system.

Policy settings where the treatment administration is based on voluntariness and is not simultaneous can be included in the category that is referred to as event study designs (Borusyak and Jaravel, 2017) or staggered adoption designs (Athey and Imbens, 2018). An important aspect in event study designs is that the effect of the treatment might not be constant across groups of individuals or time periods, a condition that is referred to as group-time treatment effect heterogeneity. The standard econometric model that has been used so far to deal with this kind of policy frameworks is the two-way fixed effects (TWFE), a panel fixed effect estimator with group (or individual) and time effects. De Chaisemartin and D'Haultfoeuille (2019) noted that the TWFE was used in 20% of the empirical articles published on the American Economic Review between 2010 and 2012. This tool is used in agricultural economics as well, where is exploited to study a variety of topics. Dawson (2005), for example, used a TWFE regression to measure the contribution of agricultural exports in less developed countries, finding a positive effect of agricultural exports on economic growth. Lien and Hardaker (2001), in a study on Norwegian farmers, showed that, in the choice of the optimal farm plans, subsidy schemes, market conditions and available labor have more importance than the farmer's risk attitude. In the context of the GI policy, Raimondi et al. (2019) investigated the effect of these quality labels on trade, highlighting that the GI policy promotes the export of agri-food products and has positive effects on export prices, while it has weak negative effects on imports. Despite the wide use of the TWFE, however, a recent bunch of literature questions the validity of this estimator when estimating the impact of the treatment in presence of group-time effect heterogeneity, claiming that it does not provide easy-to-interpret estimates (Goodman-Bacon, 2018; Athey and Imbens, 2018; Imai and Kim, 2019) and, more important, that this estimator can produce, in some cases, biased results (de Chaisemartin and D'Haultfoeuille, 2019; Borusyak and Jaravel, 2017; Abraham and Sun, 2018).

Given the practical relevance of impact analysis, biased results are a serious concern, especially when institutions stress the link between policy making and empirical evidence, as in the European case. Moreover, the European agricultural context is quite rich in policies that have an event study structure, such as the GI policy, the organic certification, or the rural development programs. Some studies tried to investigate the effects of these policies. Torres et al. (2016) compared, over a 25 years period, the performance of organic and conventional citrus farms in Spain using profitability indicators to evaluate farms investments. Nordin (2014) and Nordin and Manevska-Tasevska (2013) assessed the impact of the grassland support on agricultural employment in Sweden at the municipality and farm level, respectively. Within the GI context, Cei et al. (2018a) and Raimondi et al. (2018) estimated the impact of GIs at the regional level, respectively, on agricultural value added in Italy and on agricultural value added and employment in Italy, France and Spain. To our knowledge, however, so far, no study explored the relevance of group-time treatment effect heterogeneity in measuring the impact of this kind of policies and measures in Europe. In light of this, the objective of this paper is to understand whether group-time treatment effect heterogeneity is a concern when estimating the effects on the agricultural value added of the GI policy, an EU agricultural policy characterized by both voluntariness, not-simultaneity of the treatment and persistence of the treatment over time. This is done comparing the results of the standard TWFE estimator with the results obtained using a novel estimator proposed by Callaway and Sant'Anna (2018) that spe-

cifically accounts for the presence of group-time treatment effect heterogeneity. Ideally, if group-time treatment effect heterogeneity is not an issue in the studied context, the results of the two estimators should coincide. Understanding the relevance of group-time treatment effect heterogeneity would help in identifying the best strategies to correctly assess the impact of this kind of EU policies.

In the next section, we provide a brief overview of the GI policy in the EU and of the economic effects of this policy on the rural economy, and we review the technical literature addressing the issue of group-time effect heterogeneity in impact analysis. Here, we also present the novel non-parametric estimator that we will use as a comparison for the TWFE estimator. The third section describes the data and methods we used in the analysis, while in the fourth section we present our results, that will be discussed in a critical way in the fifth section. We end the article drawing some conclusion and highlighting the relevant research and policy implications of our work.

2. Policy and technical background

2.1 Geographical indications in Europe and their economic impact

Geographical Indications (GIs) are defined as "indications which identify a good as originating in the territory of a Member, or a region or locality in that territory, where a given quality, reputation or other characteristic of the good is essentially attributable to its geographical origin" (WTO, 1994, article 22). In Europe, geographical indications were given a common legal framework in 1992, but some countries (especially Mediterranean ones) already had in place, by that time, national provisions regulating GIs. According to the European definition of GIs, the quality of a GI product directly stems from specific and unique characteristics of the area where the product is produced, i.e. from the *terroir*. The GI policy regulates two types of GIs, the protected designation of origin (PDO) and the protected geographical indication (PGI), but the link between the product quality and the *terroir* is stronger for the PDO, whose entire production process must take place in the delimited area of origin, while the PGI just requires that at least one of the production steps takes place in the area of origin. The distinctive sign of the EU GI policy is that, in contrast to what happens in other countries, where the protection of GIs is mainly based on trademarks, PDO and PGI are public-owned signs. Farmers are thus free to join GI schemes, provided they are located within the area of origin and they comply with the rules contained in the product specification.

The strong link between GI products and the territories from which they originate is reflected in the objectives of the policy. Reg.(EU) No 1151/2012, that currently regulates the European GI system, places a considerable importance on the value adding function of the GI certification, claiming that this legislative tool is able to improve the income of local farmers. In turn, this would reflect in positive effects for the local economy and rural development.

The idea that GIs can positively affect the economy of the area where their production takes place relies on several economic foundations. First, GIs are widely recognized to be market instruments that reduce the information gap between producers and consumers (Marette et al., 1999; Josling, 2006; Anania and Nisticò, 2004). Providing additional infor-

mation to consumers is expected to raise their willingness to pay for the product. If this added value manages to be transferred up the supply chain, it will turn into an economic benefit for producers. Another function fulfilled by the GI certification is to act as a substitute for producer's reputation (Menapace and Moschini, 2012), which, according to Shapiro (1983), needs time to be built, but eventually grants a price premium on the market. Finally, GI-type certifications are able to create a rent for a limited number of producers because of the excluding mechanisms that operate in this kind of systems (Moran, 1993; Perrier-Cornet, 1990; Josling, 2006; Thiedig and Sylvander, 2000) as a consequence of area restrictions, yield limits, or both (Landi and Stefani, 2015; Hayes et al., 2004).

The value-creation function of GIs and quality schemes in general is supported by several studies that approach the problem from a theoretical and modeling perspective (Anania and Nisticò, 2004; Menapace and Moschini, 2014; Moschini et al., 2008; Zago and Pick, 2004). On the other hand, however, empirical studies offer a more controversial scenario. Consumers usually attach a greater value to GIs, despite the occurrence of positive label effects is heterogeneous across GI products (see Deselnicu et al. (2013), Leufkens (2018) and Santeramo and Lamonaca (2020) for some meta-analysis of studies on GIs and regional products, Garavaglia and Mariani (2017), Menapace et al. (2011) and De-Magistris and Gracia (2016) for specific studies) However, some difficulties are identified for that value to be transferred to agricultural producers (Cei et al., 2018b). With respect to proper impact evaluation analysis, to our knowledge, to date, only two studies have addressed the topic from this perspective. Cei et al. (2018a), found a positive impact of the GI protection on regional agricultural value added in Italy while Raimondi et al. (2018) estimated a positive impact of GIs on regional employment in France, Italy and Spain, and a positive effect on labor productivity in Spain.

2.2 Group-time treatment effect heterogeneity in impact evaluation

The GI policy allows farmers to voluntarily start the production of a GI product provided they are located in the area of origin and they comply with the GI product specification. Moreover, the policy has been continuously in place for more than 25 years, so that its activity has been observed for several periods. These characteristics create suitable conditions for the presence of group-time treatment effect heterogeneity. While treatment effect heterogeneity is defined as "the degree to which different treatments have differential causal effects on each unit" (Imai and Ratkovic, 2013), in line with the relevant literature (see, for example, Athey and Imbens (2018), Borusyak and Jaravel (2017), Abraham and Sun (2018)), group-time treatment effect heterogeneity arises when the effect of the treatment varies across groups of individuals (group heterogeneity), over time (time heterogeneity), or both. In this respect, group-time treatment effect heterogeneity can be considered a specific case of the general treatment effect heterogeneity, where the effect varies not at an individual level, but at the level of groups of individuals (e.g., groups that receive the treatment in the same year, groups for which the effect is estimated in a certain year). Three types of group-time heterogeneity can be distinguished according to Callaway and Sant'Anna (2018). The first type, which they refer to as *Selective treatment timing*, is the pure group heterogeneity case, where the effect of the treatment depends on when an individual is treated for the first time (groups are made of individuals who receive the first

treatment at the same time). Time heterogeneity is decomposed into a *Dynamic treatment effect* and a *Calendar treatment effect*. The former considers the possibility that the effect of the treatment may depend on the amount of time an individual has been exposed to the treatment. The latter lets the treatment effect vary according to the moment (period) when the effect is measured.

In contexts where group-time treatment effect heterogeneity can show up, researchers usually exploit a standard parametric way to measure the average treatment effect on the treated (ATT), the two-way fixed effects model. The TWFE model, whose specification is reported in (1), is a modification of the classical fixed effects regression.

$$Y_{it} = \alpha_i + \delta_t + \beta D_{it} + \theta X_{it} + \varepsilon_{it} \quad (1)$$

In (1), i and t are the individual/group and year subscripts, Y is the outcome, X is a set of covariates that account for possible confounders, and ε denotes the error term. α_i and δ_i are, respectively, the unit/group and time fixed effects. β, the coefficient associated to the treatment variable D_{it}, is the estimator for the ATT.

The TWFE is a regression-based DID estimator (Abadie, 2005) and as such is able to get rid of the selection bias introduced not only by observable factors (which can be directly included in the set of covariates X), but also by unobservable factors, provided that these factors are constant over time. This characteristic makes the classical fixed effects the perfect parametric counterpart of the DID method in the basic impact analysis setting where two groups (treated and controls) are observed over two periods (before and after the treatment). Similarly, working with multiple groups and multiple periods, the TWFE is expected to provide an average estimate of the treatment effect. This average estimate is the result of the aggregation of the various group-time ATTs, i.e. the ATTs for each group of individuals measured in a specific time period[1]. The aggregation of the group-time ATTs, however, involves a not linear weights structure, which makes the interpretation of the β coefficient not straightforward (Imai and Kim, 2019; Athey and Imbens, 2018; Goodman-Bacon, 2018). More importantly, in contexts characterized by group-time treatment effect heterogeneity, some of the group-time ATTs can receive negative weights when aggregated into the TWFE estimator (Abraham and Sun, 2018; Borusyak and Jaravel, 2017; de Chaisemartin and D'Haultfoeuille, 2019). Negative weights are a potential risk not only for the interpretation, but also for the reliability of the estimator, since they alter the sign of some ATTs that form the aggregated estimate and thus introduce a bias.

To face this issue, several authors suggested some novel estimators, either parametric (Imai and Kim, 2019) or non-parametric (de Chaisemartin and D'Haultfoeuille, 2019; Callaway and Sant'Anna, 2018), that do not involve negative weights. In our study, we use the one suggested by Callaway and Sant'Anna (2018) (hereinafter referred to as the CSA estimator) and we compare its results with the estimates obtained using the TWFE. The CSA estimator computes the ATT for each group of treated units (g) in each time period (t). Treated units are those observations that receives the treatment at some point in time during the observation period and they are assumed to not withdraw from the treatment

[1] The group-time ATTs are not actually estimated by the TWFE, but some authors offer several decompositions of the TWFE estimate in terms of group-time ATTs (Imai and Kim, 2019; Athey and Imbens, 2018; Goodman-Bacon, 2018).

once they received it. Each group g of treated units is composed of individuals that are treated for the first time in period g (i.e., they are not treated at $t < g$). Controls are the units that never receive the treatment.

The authors provide two versions of the estimator, one for balanced panel data, reported in (2), and one for repeated cross sections.

$$ATT(g,t) = E\left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{\hat{p}_g(X)C}{1-\hat{p}_g(X)}}{E\left[\frac{\hat{p}_g(X)C}{1-\hat{p}_g(X)}\right]}\right)\left(Y_t - Y_{g-1}\right)\right] \tag{2}$$

In (2), G_g is a group binary indicator that identifies individuals first treated at time g, C is a binary variable identifying control units, Y is the outcome variable, and $\hat{p}_g(X)$ is the generalized propensity score[2], estimated on a set of covariates X, that estimates the probability of a certain unit to be first treated at time g. The idea behind the estimator resembles the one in Lemma 3.1 in Abadie (2005) for the classical two groups-two periods setting. Basically, control units are weighted down when they have characteristics that are uncommon in the treated group, and weighted up when their characteristics are frequent in the treated group. This mechanism guarantees the balancing of the covariates between the treated (g) and the control group (Abadie, 2005; Callaway and Sant'Anna, 2018).

Basically, the CSA strategy computes, for each (g,t) pair, a DID estimate weighting control units on the basis of a propensity score measure. The propensity score is estimated for each (g,t) sample, i.e. using all control units and those treated units that form group g. It is important to note that the ATT can be estimated even for pre-treatment periods, i.e. with $g > t$. Because the treatment is supposed to not affect the outcome before it is administered, the analysis of the pre-treatment ATTs allows to verify that the conditional parallel trends assumption (i.e., the trends of the outcome variable in the treated and control groups are parallel conditional on X for all g and t) holds. The parallel trend assumption is common in DID settings, where we assume that the change in the outcome variable would have been the same in the treated and control group had the treatment not been administered. In a setting with multiple groups and multiple periods, it is required that the parallel trends assumption holds for all $g \leq t$. This assumption is fundamentally untestable (Callaway and Sant'Anna, 2018), but once we extend it to cover also pre-treatment periods it can be tested looking at the significance of the pre-treatment ATT estimates.

The means through which the CSA strategy addresses the group-time treatment effect heterogeneity issue are the avoidance of making "functional form assumptions about the evolution of potential outcomes" (Callaway and Sant'Anna, 2018, p.9) and the devising of several summary measures that avoid the drawback of negative weights. The main summary measures suggested in Callaway and Sant'Anna (2018) are reported in Table 1. The first measure (*Simple weighted average*) is a simple average where each ATT(g,t) is weighted by the number of treated observations in the respective (g,t) subsample. The *Selective treatment timing*, the *Dynamic treatment effects* and the *Calendar treatment effects* meas-

[2] This definition is provided in Callaway and Sant'Anna (2018), despite the term "generalized propensity score" is used with different meanings in the literature. In Rosenbaum and Rubin (1984), it refers to a form of the propensity score that accounts for missing data in the covariates, while Hirano and Imbens (2004) use the same term to indicate a propensity score that also accounts for cases when the treatment is not a binary variable.

Table 1. Summary Parameters of the ATT Proposed by Callaway and Sant'Anna (2018).

Summary parameter	First level	Second level
Weighted average[1]	$\theta = \frac{1}{k}\sum_{g=2}^{T}\sum_{t=2}^{T} 1\{g \le t\}ATT(g,t)P(G=g)$	
Selective treatment timing	$\theta_s^*(g) = \frac{1}{T-g+1}\sum_{t=2}^{T} 1\{g \le t\}ATT(g,t)$	$\theta_s = \sum_{g=2}^{T} \theta_s^*(g)P(G=g)$
Dynamic treatment effects[2]	$\theta_d^*(e) = \sum_{g=2}^{T}\sum_{t=2}^{T} 1\{t-g+1 = e\}ATT(g,t)P(G = g\|t-g+1=e)$	$\theta_d = \frac{1}{T-1}\sum_{e=1}^{T-1} \theta_d^*(e)$
Calendar time effects	$\theta_c^*(t) = \sum_{g=2}^{T} 1\{g \le t\}ATT(g,t)P(G = g\|g \le t)$	$\theta_c = \frac{1}{T-1}\sum_{t=2}^{T} \theta_c^*(t)$
Selective + Dynamic[3]	$\theta_{sd}^*(e,e') = \sum_{g=2}^{T}\sum_{t=2}^{T} \delta_{gt}(e,e')ATT(g,t)P\big(G = g\|\delta_{gt}(e,e')=1\big)$	$\theta_{sd}(e') = \frac{1}{T-e'}\sum_{e=1}^{T-e'} \theta_{sd}^*(e,e')$

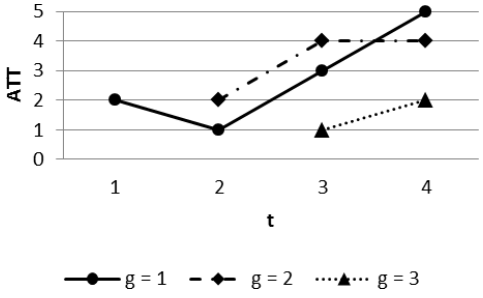
1. The weighted average parameter had a single level of aggregation. The term k assures the normalization of weights, and is equal to $\sum_{g=2}^{T}\sum_{t=2}^{T} 1\{g \le t\}P(G=g)$.
2. e represents the number of periods (years) after a group g of units receive the first treatment.
3. e' is a specific number of periods, selected by the researcher, after a group g of units receives the first treatment. $\delta_{gt}(e,e')$ abbreviates the logic function $1\{t-g+1=e\}1\{T-g+1 \ge e'\}1\{e \le e'\}$.

ure the three types of heterogeneity that we mentioned in the first part of this subsection, where the effect is thought to vary according either to the group, to the length of exposure to the treatment, or to the moment when the effect is estimated, respectively. The last summary measure, *Selective + Dynamic*, is a combination of *Selective treatment timing* and *Dynamic treatment effects*. Each of these summary measures has two levels of aggregation. The first level indicates the ATT within each group (g), number of periods after the treatment (e), or period (t). The second level measure is an average of the first level measures. As we can see from Table 1, to obtain these measures, the group-time ATTs are weighted on the basis of the size of the samples of interest (which vary according to the different summary measures). In this way, weights are assured to be always positive and meaningful, thus avoiding the occurrence of any bias or difficulty in their interpretation.

To better clarify the meaning of the summary measures, we propose a simple practical example. In Figure 1, we report hypothetical ATTs for three groups of individuals. One group (bold line) is first treated at period 1 ($g = 1$), the second group (dashed line) at period 2 ($g = 2$) and the last group (dotted line) is treated for the first time at period 3 ($g = 3$).

According to the formulas in Table 1, these ATTs are aggregated into the first level summary measures, which are reported in Figure 2. The *Selective treatment timing* ("Selective" pane of Figure 2) highlights that the average effect is larger for the group that receive the

Figure 1. Hypothetical group-time ATTs: each line identifies a group of treated units that receive the treatment in a specific time period.



first treatment in the second period. The *Dynamic treatment effect* ("Dynamic" pane) shows that, on average, the longer individuals stays in the treatment, the higher the treatment effect. Finally, the *Calendar treatment effect* ("Calendar" pane) suggests that the effect measured in the last periods ($t = 3$ and $t = 4$) is higher. For this simple example, all this information were easily retrievable from Figure 1, but the summary measures gain importance when the number of groups and periods gets large.

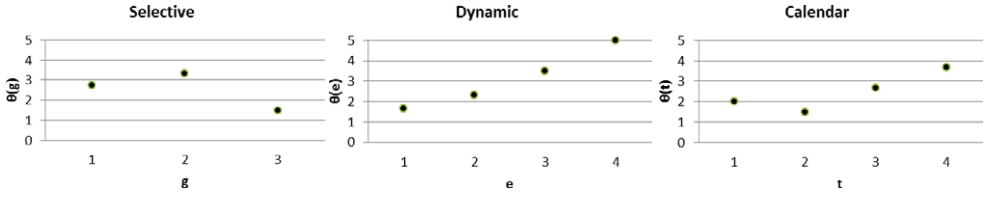
A final major contribution of Callaway and Sant'Anna (2018) is the derivation of the respective asymptotic theory for both the ATT(g,t) estimators and the summary parameters. Specifically, they derived both a consistent estimator of the variance and a specific bootstrap procedure. The suggested bootstrap procedure in particular has some advantages over traditional bootstrap. It avoids the re-estimation of the propensity score in each draw; it includes, in each iteration, observations from each group; and it allows to compute confidence bands simultaneously valid in g an t.

3. Data and methods

3.1 Data sources and samples

In our study we used two data sources: the Italian Farm Accountancy Data Network (FADN) and the EU *eAmbrosia* database[3]. The FADN data we worked with cover a nine years period, from 2008 to 2016. FADN is an unbalanced panel collecting farm-level data using a stratified sample design common to all EU Member states[4]. The database reports

Figure 2. First-level summary measures for the hypothetical group-time ATTs reported in Figure 1.



[3] The *eAmbrosia* database is accessible at: https://ec.europa.eu/info/food-farming-fisheries/food-safety-and-quality/certification/quality-labels/geographical-indications-register/

[4] The FADN field of observation consists of commercial farms, which are defined according to country-specific economic size thresholds (see Reg.(EC) 1242/2008). For Italy, the threshold is set to 4000 euros until 2014

data on farm structure, the farmer and workforce characteristics, the production process and several economic indicators. A specific section of the database reports whether a farm is involved in GI production and details which crop (or animal type, in case of livestock production) is under PDO and/or PGI certification.

eAmbrosia (formerly DOOR), is an European database where all the registered GI products are listed. For each product, several information is reported, including the product specification.

To identify the GI case studies on which to perform the analysis, we crossed the information from the two databases. Specifically, we know, from FADN, whether a farm is involved in the GI production, to which crop/animal type the certification refers, and where the farm is located. Rearranging the information from product specifications, we know which GIs can be produced in the area where the farm is located. Using this information, we selected two cases, based on: *i)* no overlap between GIs of the same product category in the same area; and *ii)* presence of control farms (i.e., farms producing the same product without the certification) in the GI area. Considering also the need for sufficiently large sample sizes, we selected two GIs: Mela Val di Non PDO (apple) and Riviera Ligure PDO (extra-virgin olive oil).

As mentioned in the previous section, a form of the CSA estimator for unbalanced panel has not been provided yet, thus we needed to balance the samples to conduct our analysis. Since FADN data cover a nine-years period, we created several balanced panels selecting different time spans and dropping units that were not observed in all the years included in the selected span. This balancing procedure will affect our results, because we are dropping treated units. However, as we discuss in the fifth section, this is not a concern for our purpose of comparing the two estimators. The balancing provides a data structure that complies with all the assumptions required by Callaway and Sant'Anna (2018) to implement their technique.

3.2 Impact analysis

In each sample, the treatment variable, GI_{it}, is the binary indicator showing whether a farm i produces the GI product in year t. Treated units are those farms for which $GI_{it} = 1$ in at least one year t, that is, farms that at some point in time certify their production as a GI[5]. In line with the CSA assumptions, once a farm adopts the certification, it is not sup-

---

and to 8000 euros afterward. The stratification is based on three levels: geographical location (European NUTS2 regions), economic size, and type of farming. Further details can be found at https://ec.europa.eu/agriculture/rica/index.cfm.

[5] It could be the case that some farms produce two versions of the same product certifying a part of the production and commercializing the remaining share without the GI sign. In these cases, the structure of the FADN dataset does not allow to distinguish between the two kinds of production. In the analysis, whenever a farm is reported to use the GI certification for a certain product, is considered to produce "only" GI-certified product. Therefore, farms that possibly has a "mixed" production (GI and non-GI) for the same crop are always considered as treated. It must be noted that this issue is probably more relevant for apple farms than for olive oil farms. In the Mela Val di Non PDO origin area the production of non-certifiable varieties is possible and common, while olive varieties grown in the Riviera Ligure PDO area are quite exclusively the ones admitted by the product specification.

posed to withdraw from the GI scheme, i.e., the treatment is irreversible[6]. On the other hand, a farm is included in the control group if it is never treated, i.e. $GI_{it} = 0$ in every year *t*. Control units are selected only among farms located in the same region (NUTS2 level for Riviera Ligure PDO and NUTS3 level for Mela Val di Non PDO), and producing the same product of treated farms (e.g., apple farms without the certification for the Mela Val di Non PDO sample). This allows us to perform our analysis in a sufficiently homogeneous socio-economic and legislative setting.

To measure whether the GI certification is actually able to increase the added value of the crop to which it applies, the crop gross margin per hectare is used as the outcome variable. The use of this variable has several advantages for our aim. In contrast to farm-level economic indicators, crop-level indicators are not affected by the economic performance of other processes or by the organization of the farm as a whole, and this allows to isolate the effect of the certification[7]. In addition, the crop gross margin indicator is defined as the difference between the total crop production and total variable costs. Measuring the certification impact on the crop gross margin thus allows to consider the effects of the certification both on the crop revenues (e.g., increased prices) and on the variable costs associated to that specific crop (e.g., inputs and certification costs). In turn, this definition of crop gross margin does not account for other EU subsidies that farms might benefit[8]. The exclusion of other subsidies from the indicator is important to isolate the effect of the GI certification from the possible effects of other CAP measures connected to product quality (e.g., second pillar measures).

Another possible option would have been to use farm prices to measure the effect of the certification, thus focusing on the expected ability of GIs to increase these prices, supposing that this is the main effect of the certification. However, it must be noted that the certification usually entails additional costs (e.g., the certification cost to be allowed to use the GI sign). Even if one assumes that those additional costs have just a minor importance with respect to the possible effects on farm prices, disregarding the cost side would inevitably lead to a bias in the estimation of the ability of the GI certification to generate an additional value.

With respect to the outcome variable, we decided to focus on relative performance improvements rather than on absolute ones. For this reason, since in the DID setting results are not scale invariant (Lechner, 2010), we use the crop gross margin per hectare in the logarithmic form.

The analysis proceeded creating two completely balanced panels, one for each sample. In each sample the effect was first estimated using the TWFE and then implementing

---

[6] In the original samples, few farms exit the certification scheme. In these cases, we dropped, before creating the balanced panels, the observations of receding farms from the year when the certification is removed onward. Similarly to the reduction in farms due to the balancing, we deem this is not an issue for our purpose of comparing the two estimators.

[7] Had the objective of the study been to measure the effect of the certification on farm profitability, the crop gross margin would have been a poor choice because it does not allow to attribute to the GI process the costs of factors shared between different farm processes (e.g., labor and capital). This indicator does in fact include the remuneration of these factors. However, the inclusion of these remunerations is exactly what one seeks in estimating the effects on the value added of the GI-certified crop, as in our case.

[8] In the FADN database, subsidies are included in the computation of farm-level indicators, such as farm gross margin, farm net value added or farm net income.

the CSA procedure. Initially, for each sample, we performed basic analysis using models without covariates. In a second stage, we included some independent variables to consider also the role of other factors that may confound the relationship between treatment and outcome. The identification of these factors was based both on previous studies investigating the determinants of GI adoption (van de Pol, 2017; Marongiu and Cesaro, 2018; Niedermayr, Kapfer and Kantelhardt, 2016) as well as on our knowledge of GI systems. We reported these factors in Table 2 (first column), where the type of each variable and their summary statistics are also shown.

In the last two columns of Table 2, we specified how, in the two methods of analysis that we compared (TWFE and CSA), we controlled for each factor. The unobservable factor (*Individual characteristics of the farmer*) is automatically controlled for by the DID structure of the two estimators (*Estimator structure* in columns 4 and 5 of Table 2), under the assumption that farmer's characteristics do not change over time (at least in the period considered in the analysis). The structure of the estimators accounts for the *Less favored area* (*LFA)* variable and for the *Year of observation* as well. The location of a farm in a less favored area does not change over time and the DID framework differences out its effect. On the other hand, the *Year of observation* is controlled by the time fixed effects in the TWFE estimator and by the within-year propensity score estimation in the CSA estimator. We controlled for the other observable factors in three different ways. Most of them are included in the TWFE equation as covariates and in the propensity score equation of the CSA estimator (*Covariate* and *Propensity score* in columns 4 and 5 of Table 2, respectively). On *Direct selling* and *Organic* a sort of direct matching is performed (*Direct matching* in Table 2). Because, in the two samples, none or very few treated farms adopt organic farming or directly sell their products, we dropped organic and/or direct selling farms from both the treated and control groups (this procedure explains the absence of sample variation for these variables in Table 2). Dropping organic and direct selling farms is like directly matching farms on a specific value (i.e., zero) of these variables. This strategy, therefore, allows to control for these factors without including them among the regressors of the TWFE model or in the CSA propensity score equation[9]. Finally, the definition of the control group (*Control group* in Table 2) allows to control for the *Type of GI product* variable, because control units are selected among farms that produce the same type of product of treated farms.

4. Results

The analysis were conducted on the two samples (Mela Val di Non PDO and Riviera Ligure PDO) for different time spans, first using basic models without covariates and then adding independent variables[10]. The first one is the Mela Val di Non PDO sample in a seven years period (from 2008 to 2014). In this sample, 15 farms join the certification system in 2009 and 13 farms enter the GI scheme in 2010. The control group consists of

[9] It should be noted that, in this way, only specific farm types are compared (i.e., non-organic and non-direct selling), which makes the results of the analysis not extendable to organic or direct selling farms. Again, the objective of our analysis makes this issue irrelevant.

[10] The whole analysis was performed using the statistical software R. Callaway and Sant'Anna (2018) provide a specific R command to implement their methodology.

Table 2. Factors to be controlled for in the models.

Factor	Type	Summary statistics[1]						Method (TWFE)	Method (CSA)
		Mela Val di Non PDO			Riviera Ligure PDO				
Age of the farmer	Continuous	[min;max]	Mean	St.dev	[min,max]	Mean	St.dev	Covariate	Propensity score
Farm located in a less favored area	Binary	[1.00;1.00]	1.00	0.00	[0.00;1.00]	0.71	0.45	Estimator structure	Estimator structure
Farm performing direct selling	Binary	[0.00;0.00]	0.00	0.00	[0.00;0.00]	0.00	0.00	Direct matching	Direct matching
Farm producing other GI products	Binary	[0.00;1.00]	0.15	0.35	[0.00;1.00]	0.16	0.37	Covariate	Propensity score
Farm with organic production	Binary	[0.00;0.00]	0.00	0.00	[0.00;0.00]	0.00	0.00	Direct matching	Direct matching
Farm utilized agricultural area	Continuous	[0.42;40.85]	5.56	4.73	[0.25;14.62]	1.94	1.70	Covariate	Propensity score
Individual characteristics of the farmer	Not observable	-	-	-	-	-	-	Estimator structure	Estimator structure
Labor intensity	Continuous	[0.08;4.04]	0.37	0.26	[0.07;3.87]	0.84	0.58	Covariate	Propensity score
Education of the farmer: none	Binary	[0.00;0.00]	0.00	0.00	[0.00;1.00]	0.02	0.14	Covariate	Propensity score
Education of the farmer: primary	Binary	[0.00;1.00]	0.09	0.29	[0.00;1.00]	0.13	0.34	Covariate	Propensity score
Education of the farmer: lower secondary	Binary	[0.00;1.00]	0.42	0.49	[0.00;1.00]	0.41	0.49	Covariate	Propensity score
Education of the farmer: upper secondary	Binary	[0.00;1.00]	0.42	0.49	[0.00;1.00]	0.42	0.49	Covariate	Propensity score
Education of the farmer: university	Binary	[0.00;1.00]	0.07	0.25	[0.00;1.00]	0.01	0.10	Covariate	Propensity score
Type of GI product	Categorical	-	-	-	-	-	-	Control group	Control group
Year of observation	Categorical	-	-	-	-	-	-	Estimator structure	Estimator structure

[1]Summary statistics were omitted, in addition to the not observable variable, for the type of product, because it is unique in the samples (either apple or olive oil) and for the year of observation, because the balanced structure of the panels that makes each level (year) equally represented.

Table 3. TWFE results for Mela Val di Non PDO and Riviera Ligure PDO.

Variable	Mela Val di Non PDO		Riviera Ligure PDO	
	Basic	Covariates	Basic	Covariates
GI	-0.35** (0.09)	0.02 (0.10)	0.15 (0.13)	0.31** (0.14)
UAA	-	0.00 (0.02)	-	-0.12 (0.13)
Age	-	-0.11** (0.02)	-	-0.01 (0.01)
Education (primary)	-	2.64** (0.65)	-	1.39 (0.161)
Education (lower secondary)	-	-	-	1.59 (1.10)
Education (upper secondary or university)	-	-	-	0.00 (0.56)
Other GI	-	0.21 (0.18)	-	-0.23* (0.13)
Labor/ha	-	0.17 (0.14)	-	0.24* (0.15)

Note: Asterisk (*) and double asterisks (**) denote group-time ATTs significant at 10% and 5% respectively.

17 farms that never use the GI certification in the observed period. Farms in the Riviera Ligure PDO sample are observed continuously for 5 years (from 2008 to 2012). Only one treated group is present (farms that start to certify in 2010), which consists of 17 units. The control sample is larger, including 91 farms.

The results of the impact analysis performed using the TWFE for the basic models (without covariates) and for the models with independent variables are reported in Table 3. In two of the four models, the GI certification has no statistically significant effect on the outcome variable. However, the GI certification has a negative impact on the crop gross margin per hectare in the Mela Val di Non model without covariates, while the GI effect is positive for Riviera Ligure olive oil when independent variables are included. In both cases, the parameters associated to the treatment variable are statistically significant at the usual 5% level.

In Table 4, we report the CSA group-time ATT estimates, which are also displayed graphically in Figures 3- 6, along with their 95% confidence intervals. According to the CSA estimator definition, groups refer to individuals that receive the treatment (i.e., adopt the GI certification) for the first time in year *g*. On the other hand, the *Year* column in Table 4 indicates the time at which the effect is estimated. In Figures 3-6 the estimates in red (*post* = 0 in the figures boxes) refer to pre-treatment ATTs and can be used to validate the extended parallel trend assumption. In all samples, the pre-treatment ATTs do not statistically differ from zero, therefore the assumption is not rejected. The standard errors were computed using the CSA bootstrap procedure. Referring to the same level of

Table 4. CSA group-time results for Mela Val di Non PDO (without covariates) and Riviera Ligure PDO (with covariates).

Mela Val di Non PDO				Riviera Ligure PDO			
Group[1]	Year	ATT (basic)	ATT (covariates)	Group[1]	Year	ATT (basic)	ATT (covariates)
2009	2009	-0.17* (0.09)	-0.29* (0.17)	2010	2009	0.15 (0.18)	0.37 (0.20)
2009	2010	-0.10 (0.11)	-0.15 (0.16)	2010	2010	0.33** (0.17)	0.37** (0.19)
2009	2011	-0.21 (0.16)	-0.33 (0.24)	2010	2011	0.06 (0.19)	0.07 (0.20)
2009	2012	0.49** (0.17)	0.78* (0.46)	2010	2012	-0.09 (0.16)	-0.08 (0.18)
2009	2013	0.07 (0.12)	0.28 (0.26)				
2009	2014	0.54** (0.25)	1.23** (0.44)				
2010	2009	-0.17 (0.12)	-0.17 (0.12)				
2010	2010	-0.20* (0.12)	-0.23 (0.17)				
2010	2011	-0.66** (0.21)	-0.69** (0.20)				
2010	2012	-0.01 (0.21)	-0.20 (0.25)				
2010	2013	-0.21 (0.21)	-0.31 (0.20)				
2010	2014	-0.25 (0.35)	-0.59* (0.32)				

[1]The column *Group* identifies farmers that enter the GI scheme in a specific year *g*. In the Mela Val di Non sample some farmers adopt the certification in 2009 and others in 2010. Conversely, all farmers in the Riviera Ligure PDO sample start certifying in 2010, therefore only one group is present.
Note: Asterisk (*) and double asterisks (**) denote group-time ATTs significant at 10% and 5% respectively.

statistical significance (5%), we note that all samples are characterized by few significant estimates, while the majority of the group-time ATTs are not statistically significant. The differences between the basic models and the models where covariates were included are minor. The effect of the certification, in the Mela Val di Non case, is positive and statistically significant at the 5% level, for the group first treated in 2009, in 2012 and 2014 (only in 2014 when covariates are considered). Conversely, in the same sample, the effect is negative, for the group first treated in 2010, in 2011. In the other sample, the only significant estimate (ATT(2010,2010)) shows a positive sign.

Figure 3. Group-time ATTs estimates (basic model) – Mela Val di Non PDO sample.

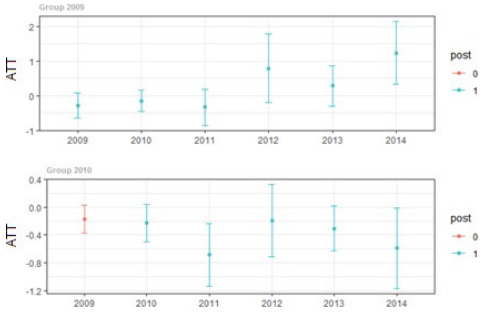


Figure 4. Group-time ATTs estimates (covariates model) – Mela Val di Non PDO sample.

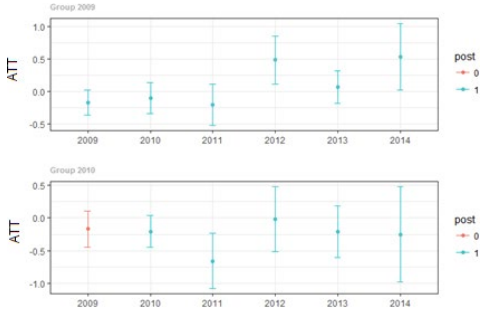


Figure 5. Group-time ATTs estimates (basic model) – Riviera Ligure PDO sample.

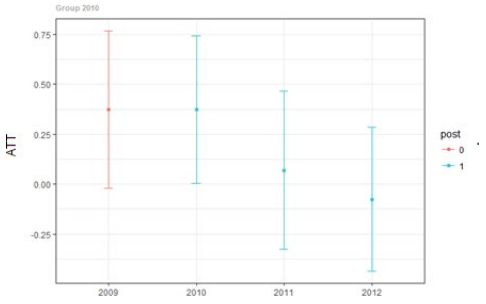
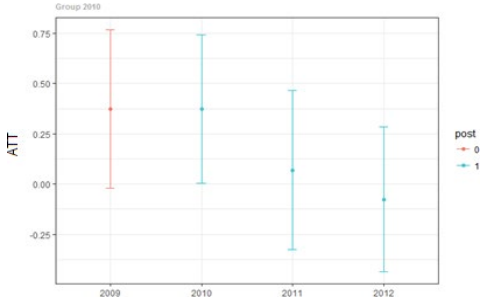


Figure 6. Group-time ATTs estimates (covariates model) – Riviera Ligure PDO sample.



Finally, in Table 5, we report the CSA summary measures. Similarly to what observed for the group-time ATTs, with the exception of the *Selective treatment timing* for the Mela Val di Non sample, the significance levels of the basic models estimate are similar to those of the models where covariates are considered. Because of the presence of only one group of treated units in the Riviera Ligure PDO sample, all the summary measures for this sample converge to the *Weighted average*. The *Weighted average* is the CSA counterpart of the TWFE impact estimate, and therefore the one in which we are most interested in for the comparison of the two estimators. In all samples this summary measure is not statistically different from zero. While this results are in line with the TWFE estimates for two models (Mela Val di Non PDO covariates model and Riviera Ligure PDO basic model), for the other two models the evidence is in contrast to what obtained from the TWFE estimation.

With respect to the other parameters, that can be estimated only in the Mela Val di Non PDO sample, the first-level measures that are statistically significant are usually dynamic or calendar effects but, for the covariates model, selective timing too. We must consider that both dynamic and calendar measures are obtained aggregating two group-time ATTs. In this way, each ATT has a considerable power in shaping the aggregated measure. With respect to the second-level of aggregation measures, none of them are sta-

Table 5. CSA summary measures for Mela Val di Non PDO and Riviera Ligure PDO.

Mela Val di Non PDO[1]											
Weighted average			Selective treatment timing			Dynamic treatment effects			Calendar time effects		
Summary measure	Basic	Covariates	Summary measure	Basic	Covariates	Summary measure	Basic	Covariates	Summary measure	Basic	Covariates
θ	-0.05 (0.13)	-0.02 (0.14)	$\theta_S(2009)$	0.10 (0.09)	0.25** (0.18)	$\theta_D(1)$	-0.18** (0.08)	-0.26** (0.12)	$\theta_C(2009)$	-0.17 (0.11)	-0.29 (0.18)
			$\theta_S(2010)$	-0.27 (0.18)	-0.40** (0.16)	$\theta_D(2)$	-0.36** (0.14)	-0.40** (0.13)	$\theta_C(2010)$	-0.15* (0.09)	-0.16 (0.11)
			θ_S	-0.07 (0.13)	-0.05 (0.14)	$\theta_D(3)$	-0.12 (0.13)	-0.27 (0.19)	$\theta_C(2011)$	-0.42** (0.14)	-0.50** (0.17)
						$\theta_D(4)$	0.16 (0.15)	0.28 (0.23)	$\theta_C(2012)$	0.25 (0.17)	-0.33 (0.28)
						$\theta_D(5)$	-0.08 (0.16)	-0.12 (0.21)	$\theta_C(2013)$	-0.06 (0.13)	0.01 (0.19)
						$\theta_D(6)$	0.54* (0.27)	1.23** (0.47)	$\theta_C(2014)$	0.17 (0.27)	0.39 (0.30)
						θ_D	-0.01 (0.11)	0.08 (0.19)	θ_C	-0.06 (0.10)	-0.04 (0.16)
Riviera Ligure PDO											
Weighted average			Selective treatment timing			Dynamic treatment timing			Calendar time effects		
Summary measure	Basic	Covariates	Summary measure	Basic	Covariates	Summary measure	Basic	Covariates	Summary measure	Basic	Covariates
θ	0.16 (0.13)	-0.03 (0.21)	-	-	-	-	-	-	-	-	-

[1] For the definition of each summary measure reported in this table refer to Table 1.
Note: Asterisk (*) and double asterisks (**) denote group-time ATTs significant at 10% and 5% respectively.

tistically significant, indicating that there is no trend of the effect due to selective, dynamic or calendar effects.

5. Discussion

The results of our analysis show that, in a European agricultural policy framework characterized by event study characteristics, the TWFE, the parametric technique that has been commonly used in literature to estimate the ATT in these contexts, might provide different estimates than a novel non-parametric estimator that accounts for treat-

ment effect heterogeneity. The main concern we observed is not the discrepancy in the magnitude of the estimated effects, which could be traced back to a cumbersome interpretation of the TWFE estimate (Imai and Kim, 2019; Athey and Imbens, 2018; Goodman-Bacon, 2018). Rather, in some samples, there is a substantial difference in the significance levels of the two estimates. Technical literature warns about the possibility that this eventuality may occur in contexts characterized by a differed administration of the treatment and by the continuation of the treatment over multiple periods, and attributes this fact to the possible occurrence of negative weights in the construction of the TWFE estimate (Abraham and Sun, 2018; Borusyak and Jaravel, 2017; de Chaisemartin and D'Haultfoeuille, 2019) when treatment effect heterogeneity is at stake. In two of our samples, evidences of group-time effect heterogeneity emerged. In the Mela Val di Non samples, the aggregate measures show that the effect varies over time. We must use caution in relying heavily on these measures, because of the few number of groups in the sample. Therefore, despite we found some hints of time heterogeneity, it is difficult to clearly attribute it to dynamic rather than to calendar effects. A stronger evidence of the presence of effect heterogeneity is provided by the single group-time ATT estimates, whose variability is observed in all samples used in the analysis. The presence of time heterogeneity is reliable given the structure of the GI policy. Especially under a *calendar* point of view, the economic effect of this policy may well depend on factors that varies over time (e.g. prices, level of production, demand). This variability might translate into an inter-annual effect variability. In light of this, the issue of negative weights pointed out by the literature, which may cause the TWFE estimator to be biased, may be relevant when estimating economic impacts in the GI context. Since the CSA estimator is specifically built to address the issue of negative weights when aggregating the single group-time ATTs, our results cast doubts about the reliability of the TWFE estimates in this policy context.

It should be noted, however, that our results are valid just for our scale of analysis, i.e. the farm level, and should not be extended to contexts where the analysis is performed at different scales or with a continuous treatment variable. For example, among the studies we referred to in the introduction, Raimondi et al. (2019) study the trade effects of the number of GIs in a given product line using decomposed bilateral trade flows at the HS 6-digit level as their units of analysis. In cases like this the CSA estimator simply cannot be computed in the current specification. In addition, a characterization of calendar and group effects when the treatment is continuous has not been developed yet.

A possible limitation of our study derives from the fact that we dropped several units from the samples we used in the analysis. This was done to balance the panels as well as to perform direct matching on some covariates. On the one hand, dropping observations increases the variance of the estimates (Caliendo and Kopeinig, 2008; Faries et al., 2020). On the other hand, similarly to what happens when trimming observations that lay out of the common support in matching studies, the reference population change (Yang et al., 2016). Especially the latter issue would be a relevant concern if the aim of the study was to provide a rigorous impact assessment of the GI policy, because results would not be externally valid. However, since our aim is to compare the two estimators using a real policy setting, these concerns are not relevant.

6. Conclusions

In this article, we explored the relevance and the possible effects of group-time treatment effect heterogeneity in impact analysis in the context of the European GI policy. As highlighted by a recent strand of literature, this kind of heterogeneity creates some problems for the estimation of the impact with traditional techniques. In line with these concerns, we observed that the standard parametric way to estimate the effect of the certification, the two-way fixed effects regression, provides different results from a recently developed estimator that accounts explicitly for group-time effect heterogeneity and its negative effects on the estimate unbiasedness. While these results are in line with the evidences reported in technical literature, our study represents, to our knowledge, a first application of a method that takes into account group-time treatment effect heterogeneity in the European agricultural economics context. Moreover, our results showed that this kind of heterogeneity might have important practical implications when measuring the impact of policies where the treatment is administered on a voluntary basis and whose effect may depend on factors that changes over time. In the case that we analyzed, i.e. the GI policy, the estimation through the two-way fixed effects estimator not only masks the underlying time effect heterogeneity, but also fails in providing an unbiased estimate of the average GI effect. Therefore, estimating the effect of this policy through the TWFE might provide biased results and wrong conclusions. Notably, time, but also group effects are particularly relevant for agricultural productions which are dependent on weather vagaries and related biotic factors which in turn impact on market equilibrium and resulting prices. Since we worked on logarithms of gross margin, wide changes in the level of prices for the baseline (i.e the conventional) product are likely to impact on the percentage premium of the GI counterpart.

This issue is particularly relevant because of the tendency of agricultural economists, so far, to estimate the impacts of this type of policies through the classical TWFE estimator. It is important to note that the use of the standard parametric estimator does not automatically lead to biased results, because both the presence and the extent of the bias depend on the structure of the weights associated to the single group-time ATTs. However, using an estimator that does not consider the possibility that the effect of these policies may vary over time and/or across groups of units can lead to misleading conclusions. This is particularly important whenever the outcome of a policy is affected by market conditions, in which case calendar effects are likely to arise.

The focus of our study was the European GI certification system, but several other policies can be found in the European agricultural body of legislation where treatment effect heterogeneity may be at stake, such as the Rural Development Programs and their measures that are developed in each CAP programming period, or other types of voluntary certifications (e.g., organic certification). In addition, in the EU political context, the assessment of the policies' performances, especially in the agricultural sector, is acquiring a leading position, which makes the concern of group-time treatment effect heterogeneity even more pressing. The evidence-based policy making course, undertaken in the current CAP programming period and confirmed and strengthened for the next one, needs, in fact, reliable evidence to show out its usefulness in building new policies or in improving old ones. The methodological accuracy of impact evaluation studies becomes thus fundamental to avoid inappropriate policy actions guided by misinterpreted evidence.

Bibliography

Abadie, A. (2005). Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies* 72(1):1–19.

Abraham, S., and Sun, L. (2018). Estimating Dynamic Treatment Effects in Event Studies With Heterogeneous Treatment Effects. *SSRN Electronic Journal*.

Anania, G., and Nisticò, R. (2004). Public Regulation as a Substitute for Trust in Quality Food Markets: What if the Trust Substitute cannot be Fully Trusted? *Journal of Institutional and Theoretical Economics* 160:681–701.

Athey, S., and Imbens, G.W. (2018). Design-based Analysis in Difference-in-Differences Settings with Staggered Adoption. Working paper No. 24963, National Bureau of Economic Research, Cambridge, MA, https://arxiv.org/abs/1808.05293.

Borusyak, K., and Jaravel, X. (2017). Revisiting Event Study Designs. Working paper, Harvard University, https://scholar.harvard.edu/borusyak/publications/revisiting-event-study-designs.

Caliendo, M., and Kopeinig S. (2008). Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys* 22(1):31–72.

Callaway, B., and Sant'Anna, P.H.C. (2018). "Difference-in-Differences With Multiple Time Periods and an Application on the Minimum Wage and Employment." *SSRN Electronic Journal*.

Cei, L., Stefani, G., Defrancesco, E., and Lombardi, G.V. (2018a). Geographical indications: A first assessment of the impact on rural development in Italian NUTS3 regions. *Land Use Policy* 75.

Cei, L., Defrancesco, E, and Stefani, G. (2018b). From Geographical Indications to Rural Development: A Review of the Economic Effects of European Union Policy. *Sustainability* 10:3745.

Cerulli, G. (2015). *Econometric Evaluation of Socio-Economic Programs. Theory and Applications*. Heidelberg: Springer.

de Chaisemartin, C., and d'Haultfoeuille, X. (2019). Two-way Fixed Effects Estimators with Heterogeneous Treatment Effects. Working Paper No. 25904, National Bureau of Economic Research, Cambridge, MA, https://arxiv.org/abs/1803.08807.

Dawson, P.J. (2005). Agricultural exports and economic growth in less developed countries. *Agricultural Economics* 33(2):145–152.

De-Magistris, T., and Gracia, A. (2016). Consumers' willingness to pay for light, organic and PDO cheese An experimental auction approach. *British Food Journal* 118(3):560–571.

Deselnicu, O.C., Costanigro, M., Souza-Monteiro, D.M., Mcfadden, D.T. (2013). A Meta-Analysis of Geographical Indication Food Valuation Studies: What Drives the Premium for Origin-Based Labels? *Journal of Agricultural and Resource Economics* 38(2):204–219.

European Commission (2018). *Proposal for a Regulation of the European Parliament and of the Council establishing rules on support for strategic plans to be drawn up by Member States under the Common agricultural policy (CAP Strategic Plans) and financed by the European Agricultural Guarantee Fund (EAGF) and by the European Agricultural Fund for Rural Development (EAFRD) and repealing Regulation (EU) No 1305/2013*

of the European Parliament and of the Council and Regulation (EU) No 1307/2013 of the European Parliament and of the Council, Bussells, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A392%3AFIN.

Faries, D., Zhang, X., Kadziola, Z., Siebert, U., Kuehne, F., Obenchain, R.L., and Haro, J.M. (2020). *Real World Health Care Data Analysis. Casual Methods and Implementation Using SAS*. SAS Institute.

Garavaglia, C., and Mariani, P. (2017). How Much Do Consumers Value Protected Designation of Origin Certifications? Estimates of willingness to Pay for PDO Dry-Cured Ham in Italy. *Agribusiness* 33(3):403–423.

Goodman-Bacon, A. (2018). Difference-in-Differences with Variation in Treatment Timing. Working paper No. 25018, National Bureau of Economic Research, Cambridge, MA, https://www.nber.org/papers/w25018.

Hayes, D.J., Lence, S.H., and Stoppa, A. (2004). Farmer-owned brands? *Agribusiness* 20(3):269–285.

Hirano, K., and Imbens, G.W. (2004). The propensity score with continuous treatments. In A. Gelman and Meng, X.L. eds. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. John Wiley & Sons Ltd, pp. 73–84.

Imai, K., and Kim, I.S. (2019). On the Use of Two-way Fixed Effects Regression Models for Causal Inference with Panel Data., MIT, http://web.mit.edu/insong/www/pdf/FEmatch-twoway.pdf.

Imai, K., and Ratkovic, M. (2013). Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation. *The Annals of Applied Statistics* 7(1): 443–470.

Josling, T. (2006). The war on terroir: Geographical indications as a transatlantic trade conflict. *Journal of Agricultural Economics* 57(3):337–363.

Khandker, S.R., Koolwal, G.B., and Samad, H.A. (2009). *Handbook on Impact Evaluation*. Washington, D.C.: The World Bank, http://documents.worldbank.org/curated/en/650951468335456749/Handbook-on-impact-evaluation-quantitative-methods-and-practices

Landi, C., and Stefani, G. (2015). Rent Seeking and Political Economy of Geographical Indication Foods. *Agribusiness* 31(4):543–563.

Lechner, M. (2010). The Estimation of Causal Effects by Difference-in-Difference Methods. *Foundations and Trends R in Econometrics* 4(3):165–224.

Leufkens, D. (2018). The problem of heterogeneity between protected geographical indications: a meta-analysis. *British Journal of Nutrition* 120(12): 2843–2856.

Lien, G., and Hardaker, J.B. (2001). Whole-farm planning under uncertainty: impacts of subsidy scheme and utility function on portfolio choice in Norwegian agriculture. *European Review of Agricultural Economics* 28(1):17–36.

Marette, S., Crespi, J.M., and Schiavina, A. (1999). The role of common labelling in a context of asymmetric information. *European Review of Agricultural Economics* 26(2):167–178.

Marongiu, S., and Cesaro, L. (2018). I fattori che determinano l'adozione delle indicazioni geografiche in Italia. *Agriregionieuropa* 52:1–8.

Menapace, L., Colson, G., Grebitus, C., and Facendola, M. (2011). Consumers' preferences for geographical origin labels: Evidence from the Canadian olive oil market. *European Review of Agricultural Economics* 38(2):193–212.

Menapace, L., and Moschini, G. (2012). Quality certification by geographical indications, trademarks and firm reputation. *European Review of Agricultural Economics* 39(4):539–566.

Menapace, L., and Moschini, G. (2014). Strength of Protection for Geographical Indications: Promotion Incentives and Welfare Effects. *American Journal of Agricultural Economics* 96(4):1030–1048.

Moran, W. (1993). Rural space as intellectual property. *Political Geography* 12(3): 263–277.

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.

Niedermayr, A., Kapfer, M., and Kantelhardt, J. (2016). Regional heterogeneity and spatial interdependence as determinants of the cultivation of an emerging alternative crop: The case of the Styrian Oil Pumpkin. *Land Use Policy* 58.

Nordin, M., and Manevska-Tasevska (2013). Farm-level employment and direct payment support for grassland use: A case of Sweden. AgriFood economic centre Working paper n.5.

Nordin, M. (2014). Does the Decoupling Reform Affect Agricultural Employment in Sweden? Evidence from an Exogenous Change. *Journal of Agricultural Economics* 65(3):616–636.

OECD. (1994). Performance management in government: performance measurement and results-oriented management.

Pawson, R., and Tilley, N. (1997). *Realistic evaluation*. London: SAGE Publications.

Perrier-Cornet, P. (1990). Les filières régionales de qualité dans l'agro-alimentaire. Etude comparative dans le secteur laitier en Franche-Comté, Emilie Romagne et Auvergne. *Économie Rurale* 195:27–33.

van de Pol, L. (2017). *Explaining the spatial distribution in the uptake of PDO and PGI in Europe*. MS thesis, Wageningen University.

Raimondi, V., Curzi, D., Arfini, F., Olper, A., and Aghabeygi, M. (2018). Evaluating Socio-Economic Impacts of PDO on Rural Areas. Paper presented at 7th AIEAA Conference Conegliano, Italy, 14-15 July.

Raimondi, V., Falco, C., Curzi, D., and Olper, A. (2019). Trade effects of geographical indication policy: The EU case. *Journal of Agricultural Economics*.

Rosenbaum, P.R., and Rubin, D.B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association* 79(387):516–524.

Sanderson, I. (2002). Articles Evaluation , Policy Learning and. *Public Administration* 80(1):1–22.

Santeramo, F.G., and Lamonaca, E.. (2020). Evaluation of geographical label in consumers' decision-making process: A systematic review and meta-analysis. *Food Research International* 131.

Shapiro, C. (1983). Premiums for High Quality Products as Returns to Reputations. *Source: The Quarterly Journal of Economics* 98(4):659–680.

Thiedig, F., and Sylvander, B. (2000). Welcome to the Club? - An Economical Approach to Geographical Indications in the European Union. *Agrarwirtschaft* 49(12):428–437.

Torres, J., Valera, D., Belmonte, L., and Herrero-Sanchez, L. (2016). Economic and Social Sustainability through Organic Agriculture: Study of the Restructuring of the Citrus Sector in the "Bajo Andarax" District (Spain). *Sustainability* 8: 918.

WTO. (1994). Agreement on Trade-Related Aspects of Intellectual Property Rights.

Yang, S., Imbens, G.W., Cui, Z., Faries, D.E., and Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics* 72(4):1055–1065.

Yin, R. (2013). *Case Study Research: Design and Methods*. SAGE Pubications.

Zago, A.M., and Pick, D. (2004). Labeling Policies in Food Markets: Private Incentives, Public Intervention, and Welfare Effects. *Journal of Agricultural and Resource Economics* 29(1):150–165.