universität
innsbruck

Faculty of
Economics and Statistics

©eeecon

Informal employment from migration shocks

Marica Valente, Timm Gries, Lorenzo Trapani

Working Papers in Economics and Statistics

2023-09

University of Innsbruck
https://www.uibk.ac.at/EconStat/

University of Innsbruck
Working Papers in Economics and Statistics

The series is jointly edited and published by

- Department of Banking and Finance
- Department of Economics
- Department of Public Finance
- Department of Statistics

Contact address of the editor:
Faculty of Economics and Statistics
University of Innsbruck
Universitaetsstrasse 15
A-6020 Innsbruck
Austria
Tel: + 43 512 507 96136
E-mail: Dean-EconStat@uibk.ac.at

The most recent version of all working papers can be downloaded at
https://www.uibk.ac.at/fakultaeten/volkswirtschaft_und_statistik/forschung/wopec/

For a list of recent papers see the backpages of this paper.

Informal employment from migration shocks

Marica Valente[*] Timm Gries[†] Lorenzo Trapani[‡]

Abstract

We propose a new approach to detect and quantify informal employment resulting from irregular migration shocks. Focusing on a largely informal sector, agriculture, and on the exogenous variation from the Arab Spring wave on southern Italian coasts, we use machine-learning techniques to document abnormal increases in reported (vs. predicted) labor productivity on vineyards hit by the shock. Misreporting is largely heterogeneous across farms depending e.g. on size and grape quality. The shock resulted in a 6% increase in informal employment, equivalent to one undeclared worker for every three farms on average and 23,000 workers in total over 2011-2012. Misreporting causes significant increases in farm profits through lower labor costs, while having no impact on grape sales, prices, or wages of formal workers.

JEL: F22, J61, J43, J46, C53

Key words: Informal employment, Migration shocks, Farm labor, Machine learning

[*] Corresponding author: marica.valente@uibk.ac.at. University of Innsbruck, DIW Berlin, Innsbruck Digital Science Center (DiSC), Innsbruck Decision Sciences (IDS), and FBK Trento. We thank Jérôme Adda, Bernd Fitzenberger, Guido Friebel, Giovanni Mastrobuoni, Francesca Molinari, Paolo Pinotti, Julien Sauvagnat, Heiner Schumacher, Juan F. Vargas, and Jeffrey Wooldridge for helpful comments. We thank EU-FADN and CREA for providing the data. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

[†] Bocconi University.

[‡] University of Leicester.

Introduction

The informal sector, where irregular migrants are overrepresented, employs over sixty percent of the world's working population (ILO, 2023). How can we detect and quantify the impact of irregular migration on informal employment? Estimation is inherently difficult due to lack of data on both populations. Previous studies, like Warren and Passel (1987) and Borjas (2017), develop methods to estimate the number of irregular migrants from census and survey data, and study their labor supply. Yet, the effects of irregular migration on firms remain unexplored, and several important questions arise. Which firms are more likely to employ informal workers after a migration shock? What are the economic implications on formal employment, consumer prices, firm profits, and tax evasion? Answering these questions is of pivotal importance in order to inform immigration policy debates, design policies that foster formalization, and improve inspection targeting and audit effectiveness.

In this paper, we estimate the amount of informal employment arising from irregular migration at the firm level, and examine its impact on firm outcomes and fiscal revenues. To the best of our knowledge, this is the first paper to do this. Specifically, we develop a two-stage method to quantify the extent of labor underreporting. In the first stage, we estimate labor productivity functions in terms of firm characteristics, in the absence of a migration shock. Labor productivity of unskilled (substitutable) labor tends to be relatively stable over time,[1] and it is predictable using ensemble methods. As we show, these new techniques in machine learning allow to flexibly control for a large number of covariates, and achieve high predictive performance by combining predictions from multiple models. In the second stage, we infer labor underreporting after a migration shock from positive differences between reported and predicted ("true") labor productivity.

Our study uses data on vineyards in the Italian agriculture sector, which employs a large number of unskilled and, often, foreign harvest labor.[2] Focusing on the exogenous variation from the 2011 Arab Spring migration shock on southern Italian coasts, we find that the supply shock of irregular migrants caused large increases in reported (vs. predicted) labor productivity for farms exposed to the

[1] See, e.g., discussions in Jorgenson and Griliches (1967) and Lamouria et al. (1963).

[2] Italian farms employ around 270k foreign workers (32%), mainly seasonal and unskilled (INEA, 2012). Similar statistics are reported for e.g. Germany, Spain, France and Poland (EU, 2021).

shock. The average effect amounts to one undeclared worker each three farms, or 23,000 workers over 2011-2012. This figure is in line with the number of detected migrants that have disappeared from first aid centers. Compared to the amount of formal employment, the migration shock causes a 6% increase in informal employment. We detect the highest level of misreporting in small farms, located in remote areas, and producing grapes for quality wines. These findings indicate that farmers' benefits from misreporting and their perceived risk of detection could be key factors influencing the amount of informal employment. We find that misreporting increases farm profitability by 6.3% on average, and that it has no effect on grape sales, prices, and hourly wages of formal workers. Given that quantities are unaffected, our results imply a substitution of formal with informal labor within farms. Based on minimum wages, we estimate labor tax evasion for about 75 million euros, which is 5% of the total agricultural revenues collected in 2011-2012 (INEA, 2015).

Data and Context.– Our analysis uses a rich panel of Italian farms supplied by the European Farm Accountancy Data Network (EU-FADN), combined with a large agro-meteorological dataset (Agri4Cast). The FADN sample is representative of the economically relevant population of Italian farms due to stratified sampling and weighting methods developed in 2010 (CREA). The final database is a panel of 2,997 Italian vineyards over the sample period 2010-2012. In each year, the sample of farms represents about 100,000 vineyards located throughout Italy, of which 19,600 in Sicily and 19,000 in Apulia.

We focus on the rapid, large, and unexpected flows of Arab Spring migrants across the Mediterranean Sea into the southern Italian regions of Apulia, Calabria and Sicily.[3] Lacking resources for detection, aid and identification of thousands of migrants, Italy declared a state of emergency (Ghizzi, 2015). In line with other studies (e.g., Tumen, 2016), we treat these unexpected flows as an exogenous shock to the labor markets of the recipient regions. We expect to observe labor market effects in the grape growing sectors of Sicily and Apulia, because they often rely on informal labor supplied by migrants landing undetected.[4] Italy's informality rate is around 20%, similar to Southern and Eastern European countries like Spain,

[3] With 64,000 detected and further undetected landings, it was the largest wave of the previous decades crossing the central Mediterranean Sea (FRONTEX, 2016; INEA, 2014).

[4] We exclude Calabria as most of its vineyard holdings are too small to be covered in our data. Only 1% of Italian vineyards are in Calabria vs. 18% in Sicily and 16% in Apulia (quattrocalici.it).

Greece, and Poland, as well as other nations relying on seasonal migrant workers such as France, Austria, Germany, Ireland, and the UK (World Bank, 2011).

Empirical Strategy. – As mentioned above, we use a two-stage method to quantify the extent of misreporting and its confidence bounds. Firstly, we estimate the labor productivity functions in terms of farm and weather characteristics. To this end, we use data of exposed farms before the shock (or "treatment"), and data of unexposed farms before and after the shock. All farms in Sicily and Apulia – the only recipients of large flows of irregular migrants – are considered as "treated". Thus, we define the comparison groups (Sicily and Apulia vs. the rest) by treatment assigned rather than treatment received, as in intent-to-treat analyses.[5]

Secondly, we infer misreporting (i.e. the amount of treatment) from positive gaps between reported and predicted labor productivity; in the case of misreporting following a shock, reported outcomes will systematically deviate from predicted outcomes. Hence, our inference on misreporting is based on testing whether labor productivity gaps of treated farms are significantly larger than those of untreated farms. Our methodology hinges on having a good predictor for untreated outcomes. Individual prediction models may be subject to misspecification bias of unknown form, and provide unstable predictions (Stock and Watson, 2004). Therefore, we propose to combine several machine learning methods, building on an idea that dates back to Bates and Granger (1969). We use a least-squared based criterion known as Super Learner (SL) (Van der Laan et al., 2008). This procedure has been shown to lead to substantial improvements in prediction accuracy with little added computational costs (Bajari et al., 2015).

Finally, we explore heterogeneity in misreporting using causal forest algorithms (Athey et al., 2019). We construct confidence intervals for our estimates of heterogeneous effects using a novel test we develop which, differently than the standard bootstrap, is robust to multiple hypothesis testing issues and the algorithmic specification. Hence, we analyze the effects of the shock on farm profits, sales, costs, prices and hourly wages of misreporting farms (vs. non-treated farms) using the "residualization" approach by Chernozhukov et al. (2018), a common practice for confounding adjustment in machine learning applications. As farm covariates may predict both farm outcomes as well as the decision to misreport after the shock

[5] As vineyards are anonymized, we cannot use continuous treatment variables relying on, e.g., farms' distance to the landing areas.

(self-selection), we use lasso to partial out the effect of covariates in two separate regressions, and use the residuals as the new inputs for a final regression.

Identification.– Our main identifying assumption is that firms share common labor productivity functions and are unaffected by other major technological or labor market shocks, a common assumption in causal inference methods including synthetic controls (Abadie et al., 2015). In our analysis, we show accurate prediction of labor productivity for all farms (treated or not) in the absence of the shock. Finally, we quantify the amount of underreported labor from the estimated increases in productivity under the assumption that additional informal labor employed after the shock is equally productive as the existing formal labor. This is consistent with findings by La Porta and Shleifer (2014, 2008) and especially plausible for farm jobs with negligible returns on experience. We also evaluate possible spillover effects between exposed and unexposed farms, which could introduce a downward bias into our estimates. By conducting placebo tests, we ascertain that there are negligible effects on farms in unexposed regions, thus ensuring that we are not underestimating the impact of the shock. Finally, through leave-one-out estimation, we show that our estimates are unaffected by the inclusion of farms from any region in the non-treated sample.

Contribution.– This paper contributes to three strands of research. First, we build on the literature that aims to estimate the amount of the undocumented labor force using census and survey data (see, e.g., Borjas, 2017; Warren and Passel, 1987; Kelly, 1977). The methodology offered here thus shares with this research the goal to "count the uncountable" from observed statistics. Our contribution to this literature is threefold. First, we provide a new approach to detect and quantify informal employment arising from irregular migration at the firm level; in doing this, we overcome the inherent challenge of estimating this unobserved phenomenon. Second, we identify firm characteristics associated with the highest increase in informality. Third, we provide first evidence of the effects of irregular migration on firm outcomes, consumer prices, and fiscal revenues. Our approach can be applied to settings where there is no available data on irregular migrants as well as informal workers, or previous estimates thereof. Moreover, it can be extended to other labor-intensive sectors characterized by large demands for unskilled (substitutable) labor such as construction, manufacturing, and services.

Second, we add to a small but rapidly growing literature on machine learning for causal inference. This is the first study to show how ensemble methods can be used to detect data misreporting from observed statistics; prior work has focused on the use of similar Super Learning methods for demand estimation (Bajari et al., 2015), detection of cyber attacks (Rabbani et al., 2021), and prediction of users' movie rating on Netflix (Toescher et al., 2009). As a by-product, we also make a theoretical contribution by proposing a novel, alternative method to the bootstrap to compute confidence intervals for causal forests estimates (see Athey et al., 2019). This method relies on a completely different approach (thus allowing to check if the results obtained with the bootstrap are robust to the specifications of the resampling scheme), and it is designed to avoid the multiple hypothesis testing problem (see Section 2.3 and Appendix B for details).

Third, our study relates to the large literature on the effects of irregular migration on the labor market. Using existing estimates of informal migrant labor on farms, a body of work shows that migrant informal workers compete with, and substitute, formal workers due to the unskilled, homogeneous nature of farm labor (see, e.g., Venturini and Villosio, 2008; Vaiou and Hadjimichalis, 1997; Lianos et al., 1996). Further, a related set of contributions use survey data in order to analyze the labor market impacts of the refugee flows into Turkey caused by the 2011 Syrian civil war (see, e.g., Tumen, 2016; Balkan and Tumen, 2016), concluding that the combination of prevalent informal employment, along with a supply shock of irregular refugees, causes labor displacement. Altındağ et al. (2020) infer increased informality post-migration from abnormal increases in firm energy use, particularly among small firms operating in largely informal sectors like construction and services. For Italy, Labanca (2020) uses survey data around the Arab Spring events to show displacement of formal native workers by formal migrant workers from heavily affected countries, without effects on wages. Our results align with these studies, offering a novel methodology for quantifying effects on informal employment and conducting a comprehensive analysis of firm outcomes.

The paper is structured as follows: Section 1 provides background information, Section 2 outlines the empirical strategy, Section 3 describes the data, Sections 4 and 5 present the results and robustness checks, and Section 6 concludes.

1 Background

In this section, we discuss irregular migration and informal farm labor in Italy, focusing on the 2011 migration shock in Sicily and Apulia. We then explain migrants' incentives to enter informal channels and how they obtain informal jobs.

In Italy, over 400,000 informal farm workers of which, mostly, undocumented migrants are estimated to be part of *caporalato*, a widespread illegal system to recruit underpaid workforce through mediators known as *caporali* (Assosomm, 2016; FLAI, 2014).[6] The estimated value produced by these workers is around 4.8 billion euros,[7] and the estimated loss to national revenues is around 1.8 billion euros per year (FLAI, 2018). Overall, informal farm workers (in and outside *caporalato*) produce an estimated value of 15 billion euros (1% GDP) and losses for 7.2 billion euros per year (IDOS, 2019). Aggregated data on farm inspections show informal employment in more than half of the inspected farms (INL, 2012).[8] Informal employment in agriculture is a widespread phenomenon, especially in southern Italy. For instance, up to 70% of total agricultural labor in Sicily (114,000 workers) and Apulia (110,000 workers) is estimated to be informal as opposed to 30% at the national level (FLAI, 2012). Further, about one fourth of the total farm workers of Sicily and Apulia (58,000) are non-EU nationals who typically perform unskilled harvest labor in tree crops (grapes and olives), have no fully legal employment, and are paid wages below the legal floor (INEA, 2012). The 2011 migration shock in southern Italy might have further increased informal employment in these regions by raising the number of workers willing to work informally (either new informal workers or previous formal workers). An investigation of the National Institute of Agricultural Economics estimates large increases of foreign agricultural workers in Sicily and Apulia post-migration: the largest estimate is for Sicily where the increase amounts to 213% or 20,000 additional foreign workers of which only a small fraction is legally employed or resident (INEA, 2012).

Landed migrants have strong incentives to enter informal labor channels. Many migrants land on Italian shores undetected (INEA, 2014). Detected migrants ob-

[6] Wages vary between 1.60 and 3 euros per hour over a 12 to 16-hour working day. In addition, workers pay intermediaries for transportation, food, and shelter (Palmisano and Sagnet, 2016).

[7] This is 6% of the total value produced by all estimated informal workers in Italy – which are 3.5 millions and produce a value of 77 billion euros or 5% GDP (ISTAT, 2017).

[8] This data is not available before 2012.

tain first aid in emergency shelters called hotspots. In 2011, of the approximately 64,000 detected and further undetected[9] landings on southern Italian shores, only around 55,000 arrivals were recorded at the hotspots (SPRAR, 2011). Further 11,700 migrants seem to have disappeared in the initial months of 2011 (La Repubblica, 2011). Thus, undetected migrants from the 2011 wave amount to at least $20,700$. Most often, landed migrants want to reach other EU destinations, however, undocumented traveling causes higher risks of detection and expulsion. Under Italian law, migrant illegal entry and stay is a crime punished with detention and expulsion (Penal Code, 2009).[10] Consequently, irregular migrants are likely to remain in the landing area in the short run.[11] As a way to make money and obtain forged documents, undetected migrants often find informal jobs in the agricultural sectors of the landing regions through, e.g., caporalato (La Repubblica, 2017).

How can we envisage the transition between entering the country irregularly, leaving first aid camps, and finding informal employment on farms? When undocumented migrants leave camps or detention centers, they often seek protection in nearby communities of migrants sharing family ties and language. These communities, often located close to farmlands, provide informal labor to farmers, as well as transportation of and assistance to workers (e.g. accommodation, healthcare, food, and credit for remittances). The matching between supply and demand of informal labor is facilitated by mediators (caporali) who also live in these communities and are often migrants themselves. To provide an insider's perspective on this process, we report the firsthand account of an irregular migrant providing informal harvest labor in Apulia, interviewed in August 2011.

"When I landed in Lampedusa [Sicily], I was with other people from Gambia, and I heard everyone saying "Gambia, Gambia," so I also said "Gambia", thinking that I might obtain political asylum. So they sent me to the CARA [first aid camp] to Borgo Mezzanone [Apulia] where I applied for asylum. But it was rejected. [...] So I went to the Grand ghetto [a migrant community nearby]. [...] I did not have money, and one "caporale" from Senegal like me asked if I wanted to work for him. I accepted." [Translated from Sacchetto and Perrotta, 2012].

[9] As reported in several studies (see, e.g., INEA, 2012), the official number of landings underestimates the actual dimension of the 2011 migration shock.

[10] In summer 2011, Italy strengthened the law (129/2011) to forcibly remove irregular individuals and non-compliant migrants, while also allowing longer detention periods (from 6 to 18 months).

[11] We discuss the case of asylum seekers in the Appendix A.

2 Empirical Strategy

In this section, we discuss the main stages of our empirical strategy. We begin by laying out our model and assumptions (Section 2.1). We then discuss model estimation through the so-called "super learner"(Section 2.2). Finally, we show how to carry out inference on the local average treatment effects (Section 2.3).

2.1 Model and assumptions

We begin by defining the labor productivity $\{y_{i,t}, 1 \leq i \leq N, 1 \leq t \leq T\}$, for farm i at year t as

$$y_{i,t} = \frac{Y_{i,t}}{L_{i,t}}, \tag{1}$$

where $\{Y_{i,t}, 1 \leq i \leq N, 1 \leq t \leq T\}$ and $\{L_{i,t}, 1 \leq i \leq N, 1 \leq t \leq T\}$ denote total grape production in tons, and labor input in hours, respectively. Labor productivity is modelled as a function $\mu : \mathbb{R}^p \to \mathbb{R}$ of a p-dimensional vector of covariates $X_{i,t}$:

$$y_{i,t} = \mu\left(X_{i,t}\right) + u_{i,t}, \tag{2}$$

where $u_{i,t}$ is an error term representing the "natural" misspecification of the functional form $\mu\left(\cdot\right)$.

However, we entertain the possibility that farms may cheat by underreporting the true labor $L_{i,t}$, instead underreporting it as $L'_{i,t}$, in order to save labor costs in terms of social contributions and labor taxes. By the same token, farms may also misreport the true output $Y_{i,t}$ as $Y'_{i,t}$. Hence, we can assume that the observed labor productivity $y'_{i,t}$ deviates from $y_{i,t}$ by an additive random shock $\eta_{i,t}$,[12] which

[12] The additive form of equation (3) comes from a straightforward application of Taylor's expansion – indeed, it is easy to see that there exists a $c_0 > 0$ such that

$$\begin{aligned}
\frac{Y_{i,t}}{L_{i,t}} - \frac{Y'_{i,t}}{L'_{i,t}} &= \frac{Y_{i,t}}{L_{i,t}} - \frac{Y_{i,t}}{L'_{i,t}} + \frac{Y_{i,t}}{L'_{i,t}} - \frac{Y'_{i,t}}{L'_{i,t}} \\
&= \frac{Y_{i,t}}{L_{i,t}}\left(1 - \left(1 + \frac{L'_{i,t} - L_{i,t}}{L_{i,t}}\right)^{-1}\right) + \frac{Y_{i,t} - Y'_{i,t}}{L'_{i,t}} \\
&= \frac{Y_{i,t}}{L_{i,t}}\left(1 - \left(1 - c_0\frac{L'_{i,t} - L_{i,t}}{L_{i,t}}\right)\right) + \frac{Y_{i,t} - Y'_{i,t}}{L'_{i,t}} \\
&= c_0\frac{Y_{i,t}}{L_{i,t}}\frac{L'_{i,t} - L_{i,t}}{L_{i,t}} + \frac{Y_{i,t} - Y'_{i,t}}{L'_{i,t}} = -\eta_{i,t}.
\end{aligned}$$

captures baseline measurement errors and misreporting in productivity, viz.

$$y'_{i,t} = y_{i,t} + \eta_{i,t}. \tag{3}$$

Further, we allow for possible misspecification of the functional form $\mu(\cdot)$, and for the possibility that the covariates $X_{i,t}$ may be measured with an error, or not be reported, or be omitted, via

$$\begin{aligned} y'_{i,t} &= \mu'\left(X'_{i,t}\right) + \eta_{i,t} + u_{i,t} + \mu\left(X_{i,t}\right) - \mu'\left(X'_{i,t}\right) \\ &= \mu'\left(X'_{i,t}\right) + \varepsilon_{i,t} + \mu\left(X_{i,t}\right) - \mu'\left(X'_{i,t}\right) \\ &= \mu'\left(X'_{i,t}\right) + \varepsilon^*_{i,t}, \end{aligned} \tag{4}$$

where $X'_{i,t}$ is a p'-dimensional vector, $p' \leq p$, whose coordinates are (some of) the coordinates of $X_{i,t}$ plus a possible measurement error, and $\mu' : \mathbb{R}^{p'} \to \mathbb{R}$. We note, as a final remark, that in (4), $\mu\left(X_{i,t}\right) - \mu'\left(X'_{i,t}\right)$ is the misspecification of the functional form $\mu(\cdot)$ arising from errors and omissions in $X_{i,t}$.

As we mentioned above, in 2011 and 2012 vineyards in Sicily and Apulia experienced a sudden increase of irregular migration - i.e., an exogenous shock. Hence, we regard vineyards in Sicily and Apulia during 2011 and 2012 as *treated* as in intent-to-treat analyses; vineyards in Sicily and Apulia before 2011 are *pre-treated* units; finally, all other vineyards in all periods are *never-treated* (or *non-treated*). Evidence of treatment is provided in Section 1. Hence, model (4) is valid for the untreated sample, given by the union of all pre-treated and never-treated units; conversely, when treatment occurs, we denote this through the binary variable $D_{i,t}$ which is equal to 1 or 0 according as vineyard i at time t was exposed to the informal labor supply shock or not. This adds[13] an additional misreporting to (4)

$$y'_{i,t} = \mu'\left(X'_{i,t}\right) + \varepsilon^*_{i,t} + \delta_{i,t} D_{i,t}. \tag{5}$$

Heuristically, if treated farms underreport labor after the shock $D_{i,t}$, then the reported labor productivity $y'_{i,t}$ will be inflated, with $\delta_{i,t} > 0$. Indeed, if farms also underreported quantities, this would attenuate the impact of the shock $D_{i,t}$, i.e. it would attenuate $\delta_{i,t} > 0$. To ensure the absence of downward bias in our estimates,

[13] The additive form of (5) arises from the same arguments as for (3).

we analyze the impact of the shock on sales and prices of affected farms, and find no effects (see Section 4.4). We would like to point out that having $D_{i,t}$ equal to 1 only for the treated units, and zero otherwise, relies on the "canonical" assumption of no spillovers (more precisely, the SUTVA or Stable Unit Treatment Value Assumption as in Rosenbaum and Rubin 1983; Imbens and Rubin 2015). We evaluate possible spillover effects between exposed and unexposed farms, which would also attenuate the impact of the shock $D_{i,t}$, by conducting placebo tests within the non-treated sample. We ascertain that there are negligible effects of the shock on farms in unexposed regions, thus ensuring that we are not underestimating $\delta_{i,t}$. To show robustness of our estimates against the inclusion of farms from any region in the non-treated sample, we perform leave-one-out estimation (see Section 5).

Equation (5) is our main working model. Defining $y'_{0,i,t}$ and $y'_{1,i,t}$ as the outcomes of $y'_{i,t}$ in the absence and presence of treatment respectively, and noting that $y'_{0,i,t} = \mu'\left(X'_{i,t}\right) + \varepsilon^*_{i,t}$, an alternative formulation of (5) is the canonical potential outcome representation

$$y'_{i,t} = y'_{0,i,t} + \left(y'_{1,i,t} - y'_{0,i,t}\right) D_{i,t}. \tag{6}$$

We make the following assumptions.

Assumption 2.1. *It holds that: (i)* $E\left|\mu\left(X_{i,t}\right)\right| < \infty$ *and* $E\left|\mu'\left(X'_{i,t}\right)\right| < \infty$ *for all* i *and* t*; and (ii)* $E\left(\mu\left(X_{i,t}\right) - \mu'\left(X'_{i,t}\right)\right)$ *is constant across* i *and* t*.*

Assumption 2.2. *It holds that (i)* $\{\varepsilon_{i,t}, 1 \leq i \leq N, 1 \leq t \leq T\}$ *and* $\{D_{i,t}, 1 \leq i \leq N, 1 \leq t \leq T\}$ *are two mutually independent groups; (ii)* $\left\{\mu\left(X_{i,t}\right) - \mu'\left(X'_{i,t}\right), 1 \leq i \leq N, 1 \leq t \leq T\right\}$ *and* $\{D_{i,t}, 1 \leq i \leq N, 1 \leq t \leq T\}$ *are two mutually independent groups;*
(iii) $\{\delta_{i,t}, 1 \leq i \leq N, 1 \leq t \leq T\}$ *and* $\{D_{i,t}, 1 \leq i \leq N, 1 \leq t \leq T\}$ *are two mutually independent groups.*

Some comments on Assumptions 2.1 and 2.2 are in order. We begin by stressing that our main aim will be to estimate $\delta_{i,t}$ in (5), whereas we do not attempt – or even need – to estimate $\mu\left(\cdot\right)$ at any stage. Hence, our assumptions do not reflect the classical set-up which one would expect to have in the context of estimating a nonlinear model in the presence of covariates with measurement error (see Schennach, 2016 for a review on this important, but unrelated, topic).

Considering Assumption 2.1, the main requirement is – essentially – that the "residual" $\mu\left(X_{i,t}\right) - \mu'\left(X'_{i,t}\right)$ is, on average, homogeneous across i and t: indeed, we do not require that its mean be zero, but merely that it exists and that it is the same across all units, especially across treated and non-treated units alike. In other words, we require farms to have common labor productivity functions. This assumption is verifiable and, as we show, farm labor productivity can be accurately predicted for all farms (treated or not) in absence of the shock. Similarly, we do not require any form of independence across i and t at this stage; note that $T < \infty$, so – even as far as estimation is concerned – we do not require any restriction on serial dependence. Note that we do not require $E\left(\varepsilon_{i,t}\right) = 0$ in our procedure. Hence, we can allow for "cheating"; e.g., as mentioned above, farms may misreport productivity, whence $\eta_{i,t} > 0$ which, in turn, may result in $E\left(\varepsilon_{i,t}\right) > 0$.

In Assumption 2.2, we do not assume independence between the covariate $X'_{i,t}$ and the treatment $D_{i,t}$, which would be standard in ANCOVA. All we require is that the "residual" $\mu\left(X_{i,t}\right) - \mu'\left(X'_{i,t}\right)$ and the treatment dummy $D_{i,t}$ are independent. In order to shed more light on the plausibility of this assumption, note that, in general, the coordinates of $X'_{i,t}$ consist of labor productivity determinants such as e.g. weather and farm characteristics. In order to ensure the validity of our assumption, we have selected farm characteristics which can be regarded as exogenous, i.e. whose reporting is unlikely to be affected by a sudden migrant inflow. In other words, assuming no effect of the migration shock on capital, assets, and other farm covariates is plausible in the short run as the migrant wave was unexpected. By the same token, we point out that the reporting of weather data can be (plausibly) assumed to be independent of migration. Moreover, at the same time of migration, there was no other major technological or labor market shock that could create dependence between the treatment dummy and the residual productivity.

Let now $\widehat{\mu}'(\cdot)$ be an estimate of $\mu'(\cdot)$, and let

$$\begin{aligned} \widehat{y}'_{0,i,t} &= \widehat{\mu}'\left(X'_{0,i,t}\right), \\ \widehat{y}'_{1,i,t} &= \widehat{\mu}'\left(X'_{1,i,t}\right), \end{aligned}$$

where, with obvious notation, $X'_{0,i,t}$ and $X'_{1,i,t}$ are the explanatory variables for the (non-treated and treated) unit i and time t. In essence, $\widehat{y}'_{0,i,t}$ and $\widehat{y}'_{1,i,t}$ are the values of $y'_{0,i,t}$ and $y'_{1,i,t}$ predicted using $\widehat{\mu}'(\cdot)$ and neglecting the presence of treatment.

The next result offers an in-population justification of our estimators.

Theorem 2.1. *We assume that Assumptions 2.1-2.2 are satisfied, and that*

$$E\left[\widehat{\mu}'\left(X'_{0,i,t}\right) - \mu'\left(X'_{0,i,t}\right)\right] = E\left[\widehat{\mu}'\left(X'_{1,i,t}\right) - \mu'\left(X'_{1,i,t}\right)\right]. \tag{7}$$

Then it holds that

$$E\left(y'_{1,i,t} - \widehat{y}'_{1,i,t}\right) - E\left(y'_{0,i,t} - \widehat{y}'_{0,i,t}\right) = E\left(\delta_{i,t}\right),$$

where $E\left(\delta_{i,t}\right)$ is the Average Treatment Effect (ATE).

Theorem 2.1 suggests that $E\left(\delta_{i,t}\right)$ can be estimated as follows. Let $1 \leq i \leq N_1, 1 \leq t \leq T_1$ denote the cross-sectional and time series sizes of the treated sample, and $1 \leq i \leq N_0, 1 \leq t \leq T_0$ denote the cross-sectional and time series sizes of the non-treated sample. Similarly, let $1 \leq i \leq N_{PT}, 1 \leq t \leq T_{PT}$ denote the cross-sectional and time series sizes of the untreated sample (which also includes pre-treatment values of treated units). Also, denote with $\widehat{\mu}'_{PT}(\cdot)$ an estimate of $\mu'(\cdot)$ obtained using the untreated sample $1 \leq i \leq N_{PT}, 1 \leq t \leq T_{PT}$. We then denote the predicted values of $y'_{i,t}$ for the non-treated and treated units as

$$\widetilde{y}'_{0,i,t} = \widehat{\mu}'_{PT}\left(X'_{0,i,t}\right), \text{ and } \widetilde{y}'_{1,i,t} = \widehat{\mu}'_{PT}\left(X'_{1,i,t}\right).$$

Then, the sample analogue of Theorem 2.1 is

$$\widehat{\delta} = \frac{1}{N_1 T_1} \sum_{i=1}^{N_1} \sum_{t=1}^{T_1} \left(y'_{1,i,t} - \widetilde{y}'_{1,i,t}\right) - \frac{1}{N_0 T_0} \sum_{i=1}^{N_0} \sum_{t=1}^{T_0} \left(y'_{0,i,t} - \widetilde{y}'_{0,i,t}\right). \tag{8}$$

As a final step, we quantify the amount of underreported labor from the estimated increases in productivity. We do so under the assumption that additional informal labor employed after the shock is equally productive as the existing formal labor. We believe this is a plausible assumption for farm jobs with negligible returns on experience. More generally, this is consistent with findings by La Porta and Shleifer (2014, 2008) showing that differences in the human capital of formal and informal workers are small.

2.2 Estimation: the Super Learner

The first step of our analysis is to have an estimate of $\mu' (\cdot)$ which is “as good as it gets”. We estimate $\mu' (\cdot)$ using the untreated observations, and then apply it to predict $y'_{i,t}$ for both treated and non-treated farms. Hence, we need an estimate of $\mu' (\cdot)$ which is particularly good at forecasting. This is not a trivial task: whilst there is a plethora of non-parametric estimators (see e.g. the review of Fan and Gijbels, 2018), all these are notoriously unreliable as far as out-of-sample performance is concerned. This is further compounded by the fact that, in our case, the set of covariates $X'_{i,t}$ is large-dimensional.

In order to ensure predictive ability, we use an ensemble method known as Super Learner (SL) (Polley, 2021). Whilst the details are reported below, we summarize it as a two-stage algorithm. In the first stage, several user-chosen estimators of $\mu' (\cdot)$ are run; in order to enhance the predictive ability of the estimated functions, we heavily rely on cross-validation in this step. After performing this first part of the analysis, we construct an aggregate predictor, by combining the previous estimators into a weighted average. Weights are chosen so as to minimise a predictive loss function, thus ensuring an optimal combination of the estimators employed in the first stage. Essentially, the SL is based on forecast averaging, and thus its origins can be traced at least as far back as the seminal contribution by Bates and Granger (1969). Combining forecasts is understood to often deliver superior forecasting ability, and we refer to the paper by Elliott (2011) for a comprehensive review and analysis of the issue.

Although the details are in the paper by Van der Laan et al. (2008), implementing the SL requires some tuning, and we summarize hereafter how we have designed it. Recall that we denote the dimensionality of $X'_{i,t}$ as p', i.e. $X'_{i,t} \in \mathbb{R}^{p'}$; recall that we use the notation $1 \leq i \leq N_{PT}$, $1 \leq t \leq T_{PT}$ to denote the time and cross-sectional sample sizes of the untreated sample, given by the union of the non-treated units and the pre-treatment values of treated units - such union is henceforth denoted as S.

In the first step of the algorithm, we run $1 \leq j \leq l$ estimators of $\mu' (\cdot) : \mathbb{R}^{p'} \to \mathbb{R}$ (see below for a list thereof), and denote each estimator as $\widehat{\mu}'_j (\cdot)$. As mentioned above, each estimator is chosen through cross-validation. To do so, and considering the fact that N_{PT} is large, whereas T_{PT} is not, we do not partition across t, but only across i. Some notation is required at this stage. We use $1 \leq v \leq V$ folds,[14] and we denote $V(v)$ and $R(v)$ as the validation and training samples for fold v respectively. For each v, as is customary, it holds that $V(v) \cup R(v) = S$ and $V(v) \cap R(v) = \varnothing$; similarly, we construct $V(v)$, $1 \leq v \leq V$, as a partition, i.e. $V(v_i) \cap V(v_j) = \varnothing$ for all $i \neq j$, and $\bigcup_{i=1}^{V} V(v_i) = S$. We use the notation v_i to denote those units i which, for fold v, belong in the validation sample. Hence, for each estimator $\widehat{\mu}'_j(\cdot)$, we predict $y'_{i,t}$; note that the estimator $\widehat{\mu}'_j(\cdot)$, when predicting $y'_{i,t} \in V(v_i)$, is constructed using $R(v_i)$. Hence, we denote, with short-hand notation

$$\widehat{y}'_{(j),i,t} = \widehat{\mu}'_j\left(X'_{i,t}\right),$$

where it is understood that $X'_{i,t} \in V(v_i)$ and $\widehat{\mu}'_j(\cdot)$ is obtained using the training sample $R(v_i)$.

Defining the l-dimensional vector of predictions $z_{i,t} = \left(\widehat{y}'_{(1),i,t}, ..., \widehat{y}'_{(l),i,t}\right)'$, we note that this defines a mapping $\psi : \mathbb{R}^{p'} \to \mathbb{R}^l$ such that $z_{i,t} = \psi\left(X'_{i,t}\right)$. Finally, we define the *super learner* as the mapping $\phi : \mathbb{R}^l \to \mathbb{R}$ such that the predicted value $\widehat{y}'_{i,t}$ is given by

$$\widehat{y}'_{i,t} = \phi\left(z_{i,t}\right) = \phi \circ \psi\left(X'_{i,t}\right). \tag{9}$$

Since (9) is an aggregation of l predictions via ϕ, it makes sense to choose the mapping $\widehat{\phi}$ from a class of functions Φ such that

$$\widehat{\phi} = \arg\min_{\phi \in \Phi} \sum_{i=1}^{N_{PT}} \sum_{t=1}^{T_{PT}} L\left(y'_{i,t}, \widehat{y}'_{i,t}\right), \tag{10}$$

where $L\left(y'_{i,t}, \widehat{y}'_{i,t}\right)$ is a loss function. A common choice could be the L_2-norm loss (i.e. a quadratic loss function), viz.

$$L\left(y'_{i,t}, \widehat{y}'_{i,t}\right) = \left(y'_{i,t} - \widehat{y}'_{i,t}\right)^2. \tag{11}$$

Similarly, a computationally convenient choice for ϕ could be a linear mapping, so

[14] We have used, as suggested in the paper by Polley (2021), $V = 10$.

that (10) becomes

$$\widehat{w}_{N,T} = \underset{w\in[0,1]^l, s.t.\|w\|=1}{\arg\min} \sum_{i=1}^{N_{PT}} \sum_{t=1}^{T_{PT}} \left(y'_{i,t} - \sum_{j=1}^{p} w_j \widehat{y}'_{(j),i,t} \right)^2 . \tag{12}$$

The output of (12) is an l-dimensional vector $w = (w_1, ..., w_l)'$, whose coordinates represent the weight assigned to each individual predictor $\widehat{\mu}'_j \left(X'_{i,t} \right)$ in constructing the SL, which in this context has the interpretation of being a weighted average of the individual predictors. A consequence of using the SL is that $\widehat{w}_{N,T}$ has the "oracle property", in that it performs at least as well as the best of the $1 \leq j \leq l$ estimators.

As far as implementation is concerned, we have used the following "learners": a random forest, extreme gradient boosting, regularized regression, neural networks and support vector machines; we provide details on the implementation of these algorithms in the Appendix F.

2.3 Inference on the local ATEs

We now propose an estimate of the $\delta_{i,t}$'s for $1 \leq i \leq N_1$, i.e. for the treated units. Let, $\widetilde{\mu}'(\cdot)$ be the estimate of $\mu'(\cdot)$ based on the SL defined above, viz.

$$\widetilde{\mu}'(\cdot) = \sum_{j=1}^{l} w_j \widehat{\mu}'_j(\cdot),$$

and define, for short

$$\widetilde{y}_{i,t} = y'_{i,t} - \widetilde{\mu}'\left(X'_{i,t} \right),$$

i.e. the residual obtained through $\widetilde{\mu}'(\cdot)$. We now propose a "smoothed" estimate of the individual $\delta_{i,t}$, $1 \leq i \leq N_1$, based on equation (7) in Athey and Wager (2019):

$$\widehat{\delta}_{i,t} = \frac{\sum_{t=1}^{T} \sum_{i=1}^{N_0+N_1} \alpha_{i,t} \left[\widetilde{y}_{i,t} - \widetilde{\mu}'^{(-i,t)}\left(X'_{i,t} \right) \right] \left[D_{i,t} - \widehat{e}^{(-i,t)}\left(X'_{i,t} \right) \right]}{\sum_{t=1}^{T} \sum_{i=1}^{N_0+N_1} \alpha_{i,t} \left[D_{i,t} - \widehat{e}^{(-i,t)}\left(X'_{i,t} \right) \right]^2}. \tag{13}$$

In (13), the notation is as follows: $\alpha_{i,t}$ are the nearest-neighbor weights estimated by an auxiliary forest, as defined in equation (6) in Athey and Wager (2019), based on the frequency with which treated units fall in the same leaves as untreated units; $\widetilde{\mu}'^{(-i,t)}\left(X'_{i,t}\right)$ is the prediction of $y'_{i,t}$ based on the super learner $\widetilde{\mu}'(\cdot)$, estimated missing out the observations corresponding to the (i,t)-th unit from the estimation sample; and $\widehat{e}^{(-i,t)}\left(X'_{i,t}\right)$ is an estimate of the propensity score $P\left(D_{i,t}|X'_{i,t}\right)$, obtained missing out the (i,t)-th unit from the estimation sample. Using (13) serves the double purpose of getting rid of two types of endogeneous variations (confounding): the variation of residual due to $X'_{i,t}$, and differences between treated and untreated that can be ascribed to $X'_{i,t}$. This can be illustrated heuristically by noting that

$$\widetilde{y}_{i,t} = \left[\mu'\left(X'_{i,t}\right) - \widetilde{\mu}'\left(X'_{i,t}\right)\right] + \delta_{i,t}D_{i,t} + \varepsilon^*_{i,t},$$

and recalling that, by Assumption 2.2, $\varepsilon^*_{i,t}$ is independent of $X'_{i,t}$. Then, $\widetilde{y}_{i,t} - \widetilde{\mu}'^{(-i,t)}\left(X'_{i,t}\right)$ partials out the effect of $\left[\mu'\left(X'_{i,t}\right) - \widetilde{\mu}'\left(X'_{i,t}\right)\right]$ from $\widetilde{y}_{i,t}$ (thus, in essence, yielding an estimate of $\delta_{i,t}D_{i,t}$); and $D_{i,t} - \widehat{e}^{(-i,t)}\left(X'_{i,t}\right)$ removes the effect of $X'_{i,t}$ from $D_{i,t}$. Then, modulo the error term $\varepsilon^*_{i,t}$, $\widehat{\delta}_{i,t}$ in (13) may be viewed as a local, Least Squares estimate of $\delta_{i,t}$ after partialling out the effect of $X'_{i,t}$.

After "extracting" the $\delta_{i,t}$s via (13) for each treated unit $1 \leq i \leq N_1$ and period $1 \leq t \leq T_1$, we can carry out inference on $\delta_{i,t}$. We are interested in "significance tests" for the null

$$H_0 : \delta_{i,t} = 0, \tag{14}$$

for each $1 \leq i \leq N_1$ and $1 \leq t \leq T_1$; and we subsequently develop a methodology to construct confidence intervals for each estimate of $\delta_{i,t}$. When testing for (14), two potential issues need to be addressed. Firstly, confidence intervals for $\widehat{\delta}_{i,t}$ are computed using bootstrap (Athey et al., 2019). This requires some tuning, and therefore different specifications of the bootstrap algorithm may potentially result in different outcomes. This issue requires, at a minimum, some sensitivity analysis. Secondly and more importantly, in our empirical analysis we run tests for hundreds of treatment effects $\delta_{i,t}$; hence, our results are bound to be affected by the multiple testing problem. As a consequence, over-rejection of H_0 in (14) may occur, which, in turn, would spuriously indicate that too many $\delta_{i,t}$ are different from zero. This issue is more difficult to tackle than the previous one: some form of correction of the nominal level of the individual tests could be implemented, but using e.g. a Bonferroni correction may be overly conservative, thus understating the number of $\delta_{i,t}$ that are found to be nonzero. Hence, again, the need to carry out some further analysis to assess the sensitivity of our results to the specification of the nominal level of the tests.

Thus, as a way to ensure that our results are robust to both issues (sensitivity to the bootstrap specifications, and the multiple testing problem), we develop a novel methodology to obtain confidence intervals for $\widehat{\delta}_{i,t}$. Whilst the details of such a methodology are relegated to Appendix B to avoid overshadowing the main discussion, we would like to point out that results obtained by the bootstrap-based approach and by our method align, indicating that our empirical findings are robust, in that they do not depend on the methodology employed.

3 Data

Using data on Italian vineyards from the Farm Accountancy Data Network (EU-FADN) and gridded agro-metereological data (Agri4Cast), we construct a new farm level dataset with information on labor hours, grape quantity and type, labor productivity (main outcome) as well as physical, structural, economic and financial determinants of labor productivity. The EU-FADN database is accessible to authorized users through a formal request and confidentiality agreement. The EU-FADN provides representative, high quality and consistent datasets from individual countries, while maintaining anonymity of the data (Latruffe et al., 2017). It stands as the sole provider of microeconomic data that follows harmonised bookkeeping principles. This comprehensive dataset is extensively utilized in the field of agricultural economics (Ciaian et al., 2021). The database includes annual accounting information for commercial farms rotating over several years, typically five; therefore, the datasets are unbalanced panels.

Our dataset is representative of the economically relevant population of Italian farms in each year due to stratified sampling and weighting (CREA).[15] Stratification ensures high overall FADN coverage of the whole agricultural population not only in terms of output, but also area and farm labor (EU-FADN, 2018). The data contain 2,997 Italian vineyards over the sample period 2010-2012.[16] It includes 634 observations for Sicily and Apulia (the treated group) and 2,363 observations for

[15] Economic relevance is defined by farms' annual output equal to at least 8,000 euros. Stratification is based on selected covariates such as agricultural output, area, and labor units.

[16] We exclude farms with missing covariate data and years before 2010 because of a change in the way European farms are classified and selected (EU-FADN, 2018).

the rest of Italy. The stratified (and representative) sample of farms changes every year, however, the number of observations is stable across years and treatment group. For instance, treated farms are 209 in 2010, 223 in 2011, and 202 in 2012. The sample of treated farms represents about 19,600 vineyards in Sicily and 19,000 vineyards in Apulia each year. The sample of non-treated farms represents about 58,200 vineyards located throughout Italy.

The shock of informal labor supply, the treatment, takes place in 2011. As shown in Figure 3.1, there was a sudden, large rise in the number of irregular migrants arriving on the southern Italian coast during the spring of 2011. By the end of that year, a total of 64,000 border crossings were detected, along with additional undetected landings (INEA, 2012).

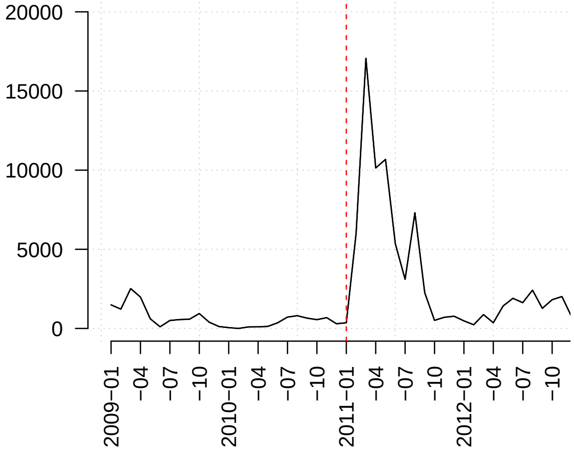


Figure 3.1: Detected illegal border crossings on the Central Mediterranean route (own illustration, source: FRONTEX, 2016)

3.1 Summary statistics

Table 3.1 compares key attributes of treated and non-treated farms over the sample period 2010-2012. The main outcome is labor productivity as tons of grape over labor hours. Full definitions and summary statistics of the included covariates are presented in Tables C.1 and C.2 in Appendix. Relevant predictors are discussed in the next section.

Table 3.1: Summary statistics for key attributes of treated vs. non-treated farms over the sample period 2010-2012. Observations equal 2,363 for non-treated farms and 634 for treated farms.

	Treated farms				Non-treated farms			
Obs. 2,997	Mean	Min	Max	SD	Mean	Min	Max	SD
labor prod	0.04	2e-04	0.34	0.03	0.02	2.1e-05	0.48	0.02
labor hours	4170	200	112285	8488	3427	320	58400	3275
grape quantity (tons)	157	0.3	4500	324	76.95	0.1	2836	135.32
land prod	16.88	0.01	55.15	10.18	11.53	0.02	35.60	5.18
vineyard land (ha)	11.48	0.51	415	25	6.91	0.01	171.48	10.86
capital intensity	3775	0	37224	4011	7376	0	348204	13348
table grapes only (0/1)	0.17	0	1	0.37	0.003	0	1	0.06
grapes for PDO wine only (0/1)	0.28	0	1	0.45	0.73	0	1	0.45

Labor productivity values of treated farms lie within the range of the non-treated farms. Labor productivity is slightly higher for treated farms on average. Treated farms are on average more labor intensive (employ more labor hours) and less capital intensive than non-treated farms. The latter produce on average about half of the grapes produced by treated farms. This is partly due to the fact that, compared to non-treated farms, a larger share of treated farms (17% vs. 0.3%) produce table grapes which are much heavier than wine grapes. Controlling for the types of grapes produced is, therefore, crucial to predict labor productivity differences. When considering grape quality, only 28% of treated farms cultivate grapes for quality wine with Protected Denomination of Origin (PDO), whereas 73% of non-treated farms do so. Average quantities and labor hours mask variation across farms, as indicated by large standard deviations relatively to the mean. However, this variability is not as pronounced when it comes to labor productivity values. Lastly, when it comes to land, treated farms exhibit larger vineyard areas on average and higher land productivity in comparison to non-treated farms.

Table 3.2 compares grape quantities and labor hours before and after the shock separately for treated and non-treated farms. While grape quantities remain stable on average in the two periods, we observe a decrease (increase) in reported labor hours for treated (non-treated) farms on average after the shock. The observed changes in reported labor hours in treated farms raise several questions. After partialling out the impact of weather and farm inputs, do we observe abnormal spikes in labor productivity in treated farms? What role do farm characteristics play in driving these gaps? Finally, how does labor misreporting affect farm economic outcomes, wages of formal workers, consumer prices, and labor tax revenues?

Table 3.2: Before-and-after comparison of grape quantities and labor hours of treated units vs. non-treated units over the sample period 2010-2012. Observations equal 2,363 for non-treated farms and 634 for treated farms.

		Before		After	
Obs. 2,997		Mean	Sd	Mean	Sd
Treated farms:	Grape tons	157.6	322.0	156.7	325.4
	Labor hours	4351.8	10022.1	4080.5	7632.1
Non-treated farms:	Grape tons	76.5	127.0	77.2	139.7
	Labor hours	3374.7	2764.4	3454.8	3522.0

3.2 Determinants of vineyard labor productivity

Determinants of labor productivity are possibly many and of different nature. Our dataset provides comprehensive information on physical, structural, economic, financial, and agro-metereological determinants of labor productivity. We include more than 400 variables chosen to be exogenous, i.e., unaffected by the sudden inflow of irregular migrants.

Farm size is measured by the total land area, which allows us to account for economies of scale. Additionally, we consider the proportion of land that is utilized and owned. Since farm size and managerial quality are often intertwined, the inclusion of land variables partially controls for management quality. Furthermore, the share of vineyard areas and other crops is taken into account to assess the level of specialization and control for economies of scope. The type of harvested grape is likely to have an impact on labor productivity. Specifically, grapes intended for quality wines with Protected Designation of Origin (PDO) require greater attention and care compared to grapes for other types of wines. Consequently, the harvest of PDO grapes can be more labor-intensive. To account for this variation, we include controls for the specific type of grapes produced, distinguishing between table grapes, grapes for PDO wine, grapes for non-PDO wine including table wine and wine with Protected Geographic Indication (PGI), and grapes for other wines. Additionally, the productivity of vineyard labor can be influenced by organic farming practices that involve reduced reliance on fertilizers and synthetic chemicals for vine disease prevention. Thus, we account for whether farms practice organic farming or are transitioning to organic methods. In terms of capital, we measure capital intensity by considering the book values of machinery per hectare. Capital intensity plays a crucial role in determining both labor productivity and the effectiveness of grower management techniques. These factors can impact labor productivity in activities such as grape harvesting and grapevine pruning (Lamouria et al., 1963). Furthermore, these variables also account for the variations in the use of grape harvesting machines, which can be more prevalent in certain regions. Finally, to address any unobserved differences in labor productivity across different regions and years, we incorporate region and year dummies into our analysis.[17]

Climate and weather conditions play a crucial role in determining grape yield and, consequently, the productivity of vineyard labor. Extensive research, as summarized by Ashenfelter and Storchmann (2016), supports this relationship. To address the impact of weather, we incorporate various weather-related variables across different time periods and locations. To control for weather effects over time and across farms, we consider factors such as temperature, wind speed, radiation, precipitation, and snow depth. Monthly and bimonthly averages, medians, and deviations are calculated to capture fluctuations in labor productivity caused by weather variations. Additionally, we account for specific weather events that significantly influence wine production. This includes identifying the first instance in a year when the temperature exceeds 10 degrees Celsius and recording the number of days in a year when the temperature falls below 0 degrees Celsius.

For a comprehensive description of all these variables, please refer to Table C.1 in the Appendix.

4 Results

We use the SL to estimate labor productivity functions based on training samples of untreated observations over 2010-2012.[18] The training sample is about 75% of the original sample, and it is built by drawing untreated observations at random. To increase prediction accuracy, untreated observations include 161 observations of treated farms before the shock. Similar to the synthetic control literature (see, e.g., Bueno and Valente, 2019; Abadie et al., 2015), estimation of counterfactual outcomes uses pre-treatment data to better approximate the data generating process

[17] These dummies do not preempt the treatment effect as the estimation process does not rely on outcomes that are directly affected by the treatment.

[18] We use the software R-4.0.3 and the SuperLearner package version 0.10.0 (Polley, 2021).

of treated outcomes in absence of the treatment. The resulting training sample contains 1,946 untreated observations. As mentioned in Section 2.2, the SL is trained using five different base learners: penalized regressions (glmnet), random forests (RF), gradient boosting (xgboost), support vector machines (SVM), and neural nets (nnet). The best predictive algorithms are glmnet (lasso) and xgboost followed by RF which receive, respectively, a coefficient of 0.50, 0.47 and 0.03.[19] SVM and nnet are assigned coefficients equal to zero. The resulting SL achieves an out-of-sample prediction accuracy of 99% for labor productivity of non-treated farms and 98% for farms in Sicily and Apulia. As estimation uses mostly data of non-treated farms, this means that non-treated farms can well approximate labor productivity of treated farms in absence of the treatment.

The optimal combination of algorithms building the SL may give an insight into the true functional form of vineyard labor productivity. Typically, SVM and nnet work well when the function to approximate is complex (e.g., highly nonlinear) and the data is unstructured (e.g., text and image recognition). Differently, xgboost, RF and, especially, glmnet work well when the target function is rather simple. This seems to be the case for vineyard labor productivity which can be well predicted by the included set of farm characteristics and weather factors.

4.1 Migration effects on misreporting at regional level

In the previous step we have obtained accurate predictions of labor productivity for all farms unexposed to the migration shock. Now we use the estimated labor productivity functions to predict labor productivity for farms in Sicily and Apulia after the migration shock, i.e., counterfactual outcomes of treated farms had migration not happened. If farms employed more informal labor after shock, we expect an increase of unreported labor hours and, thus, a positive gap between reported and predicted labor productivity. For robustness, we predict outcomes also for non-treated farms, expecting no significant differences between reported and predicted labor productivity. For this, we use the test sample including 1,051 observations of which 578 for non-treated farms and 473 for treated farms.

[19] As altering default tuning parameters does not yield any remarkable increase in prediction accuracy, we proceed with the default values of the individual base learners. For RF, we test number of trees = {number of variables/3 [default], 10, 50, 200, and 500 percent of the default value}, and node size = {5 [default], 2, 3, 10, 25}. For glmnet, we test penalty term ={lasso [default], ridge}, and penalty strength = {100 [default], 10, 50, 200, 500}.

Table 4.1 shows that reported labor productivity is higher than predicted labor productivity in the treated group in terms of mean and distribution. Differences are statistically significant after the shock but not before the shock. Non-treated farms, instead, show no significant differences between reported and predicted labor productivity neither before nor after the shock (Table D.1 in Appendix).

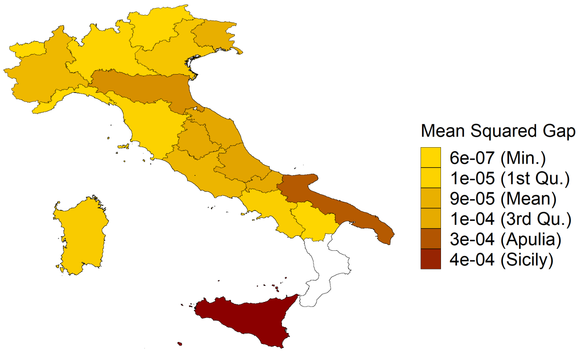
Table 4.1: Reported versus predicted labor productivity (y versus $\hat{y}$) for the treated group. Gaps are computed using observations in the test sample (not used for training the SL). After the migration shock in 2011, all observations are in the test sample.

Year (obs.)	Mean y	Mean $\hat{y}$	K-S test (two-sided)	K-S test (one-sided)	t-test	Wilcoxon test
2010 (n = 48)	0.034	0.032	0.853	0.472	0.520	0.740
2011 (n = 223)	0.037	0.030	2.28e-04***	1.14e-04***	0.003***	2.45e-04***
2012 (n = 202)	0.036	0.030	0.0012***	8.58e-04***	9.21e-04***	0.028**
Notes:	***p=.01; **p=.05; *p=.1					

Columns 4-7 in Table 4.1 report results (p-values) of statistical tests. The two-way Kolmogorov-Smirnov (K-S) test rejects the hypothesis of equality of empirical cumulative distribution functions (eCDF), while the one-way K-S test rejects the hypothesis that eCDFs of reported labor productivity are below predicted labor productivity's eCDF. T-tests indicate significantly greater means of reported than predicted labor productivity. Further, medians of reported labor productivity are generally higher than medians of predicted labor productivity. We use a non-parametric Wilcoxon rank sum test to test for a shift in location (mean ranks). While before the shock we find no significant difference in location, after the shock the distribution of reported labor productivity is significantly shifted (to the right) with respect to the distribution of reported labor productivity. In sum, statistical tests provide evidence of abnormally high reported labor productivity of farms in Sicily and Apulia after the migration shock.

Figure 4.1 plots post-migration labor productivity gaps, i.e., estimated differences between reported and predicted labor productivity aggregated by region. The plot shows that mean squared labor productivity gaps in Sicily and Apulia (treated group) are larger than in the other Italian regions (non-treated group). Differently, gaps are rather similar across treatment groups before the shock. Results are summarized in Table D.2 in Appendix.

Figure 4.1: Mean squared labor productivity gaps. Estimated differences between reported and predicted labor productivity post-migration aggregated by region.



We proceed comparing distributions of labor productivity gaps. Summary statistics show that gaps are larger for treated farms than for non-treated farms after the shock. Statistics are, instead, comparable across treatment groups before the shock. Table 4.2 reports means, medians, and higher percentiles of labor productivity gaps.

Table 4.2: Summary statistics for labor productivity gaps by treatment group and year. Mean gap, median gap, and highest percentiles (perc.).

Year	Group (obs.)	Mean gap	Median gap	75^{th} perc.	85^{th} perc.	95^{th} perc.
2010	Treated (n = 48)	0.002	5.27e-04	0.011	0.014	0.024
	Non-Treated (n = 188)	0.003	6.89e-04	0.006	0.010	0.020
2011	Treated (n = 223)	0.007	0.005	0.014	0.019	0.031
	Non-Treated (n = 215)	0.002	3.36e-04	0.004	0.008	0.020
2012	Treated (n = 202)	0.004	0.003	0.012	0.018	0.030
	Non-Treated (n = 175)	7.79e-04	-9.14e-05	0.003	0.006	0.013

For a visual representation, Figure D.1 in Appendix plots the distribution of standardized gaps for treated and non-treated farms before and after the shock. From Figure D.1 and Table 4.2 we observe similar distributions before the shock. Differently, after the shock, distribution of gaps for treated farms is more skewed to the left compared to non-treated farms, and shows more extreme gaps in the positive domain.

We test for differences in labor productivity gaps between treated and non-treated farms statistically. Table 4.3 reports the results.

Table 4.3: Differences in labor productivity gaps between treated and non-treated farms by year.

Year (obs.)	K-S test (two-sided)	K-S test (one-sided)	t-test	Wilcoxon test
2010 (n = 236)	0.070	0.100	0.825	0.978
2011 (n = 438)	3.93e-10***	0.654	2.94e-04***	2.04e-04***
2012 (n = 377)	5.84e-07***	0.214	0.005***	0.002***
Notes:	***p=.01; **p=.05; *p=.1			

K-S tests show that gap distributions (eCDFs) of treated and non-treated farms (two-sided test) differ statistically, and the eCDF of non-treated farms is not above the eCDF of treated farms (one-sided test). T-tests show unequal means after but not before the shock. Non-parametric Wilcoxon rank-sum tests reject equality of mean ranks, showing that the location of the two distributions is shifted.

We estimate by how much gaps in Sicily and Apulia abnormally increase post-migration, on average. We use a difference-in-differences regression. We are interested in the effects of the interaction term between a treatment dummy (Treat=1 for Sicily and Apulia) and a post-migration dummy (Post=1 for 2011 and 2012) on labor productivity gaps. Model (1) in Table 4.4 shows that gaps in the landing regions are about 23% higher than gaps in other regions post-migration.[20] Model (2) shows that this increase is slightly higher for farms in Sicily (24%) than in Apulia (22%).

[20] Since some gaps are negative, we take $\log(y) - \log(\hat{y})$. Since $y/\hat{y} \approx 1$, we can interpret coefficients as percent changes, in particular, as $\exp(0.206) - 1 = 0.23$ or 23%.

Table 4.4: Results from difference-in-differences regression by treatment group (1) and treated region (2). Results are robust to the inclusion of year and region fixed effects.

	Gaps (log)	
Obs. 1,051	(1)	(2)
Treat*Post	0.206^{***} (0.069)	
Apulia*Post		0.201^{**} (0.082)
Sicily*Post		0.218^{**} (0.105)
Adjusted R^2	0.04	0.04
F Statistic	16.3^{***}	9.8^{***}
Notes:	**p<0.05; ***p<0.01	

We conclude that there is statistical evidence of abnormally high labor productivity gaps in the landing regions post-migration. We will now assess which of these individual gaps is statistically significant and what drives gap heterogeneity or, in other terms, which type of farm misreports the most post-migration.

4.2 Migration effects on misreporting at farm level

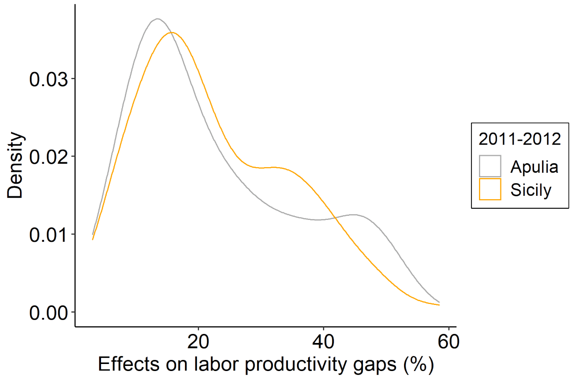
In line with our theoretical predictions, reported labor productivity of farms in Sicily and Apulia shows abnormal increases after the migration shock. In particular, the difference-in-differences estimation showed higher labor productivity gaps for treated farms on average. We now turn to a farm-level analysis of labor productivity gaps. Which treated farm does actually misreport labor input? And which are the characteristics of farms misreporting the most?

We pin down significant differences in labor productivity gaps after matching treated with non-treated farms based on their characteristics. A number of farm characteristics may drive differences in misreporting among treated farms. Plausibly, farms misreport when the benefit from reducing labor costs through informal employment are higher than the costs of detection. This depends on the detection probability. For instance, smaller farms located in remote areas face a lower probability to be inspected. Also, the benefit from reducing labor costs is larger for farms producing products of higher marginal value. Thus, we match on 16 variables capturing farm size and location as well as profitability measures unaffected by the amount of misreporting such as land productivity, type of harvested grapes, and the share of land used to cultivate other crops.

We estimate differences in misreporting at the farm level using causal forests (Athey et al., 2019). These algorithms estimate conditional average causal effects (local ATE's) using matching weights assigned to each treated farm based on their resemblance to non-treated farms. We refer to Section 2.3 for a coarse description of this estimator, and to Valente (2023) for implementation details.

We begin the analysis by estimating migration causal effects on labor productivity gaps for farms in Sicily and Apulia. Figure 4.2 presents the distribution of the estimated percent increases in labor productivity gaps post-migration for all treated farms in the sample.[21] The hypothesis of effect homogeneity (zero variance) is rejected statistically (Levene, 1960).

Figure 4.2: Farm level estimates of migration causal effects on labor productivity gaps in Sicily and Apulia over 2011-2012. Effects are measured as percent increases in labor productivity gaps post-migration.



On average, we find that labor productivity gaps are 23% higher post-migration compared to their counterfactual. This result aligns with the average effect obtained in Table 4.4, which indicates that controlling for a large number of covariates and removing confounding via a three-step algorithm[22] does not invalidate the difference-in-differences estimator. Figure 4.2 also shows that the estimated

[21] For the farm level analysis, we truncate the sample at the 5^{th} and 95^{th} percentiles to remove outliers. Treated observations drop from 425 to 356.

[22] See Section 2.3 for more details. This estimator builds on residualization (Robinson, 1988) and is also known as R-learner (Nie and Wager, 2021).

causal effects on labor productivity gaps post-migration are *all positive*, i.e., labor productivity gaps are always larger than their counterfactual for all treated farms.

Pointwise confidence intervals show that causal effects on labor productivity gaps post-migration are statistically significant for about 80% of treated farms. We interpret these abnormal (statistical significant) increases on labor productivity gaps as labor misreporting caused by the migration shock. Figure D.2 in Appendix plots the estimated causal effects with their 95% confidence intervals for farms in Sicily and Apulia over 2011-2012.

We estimate significant heterogeneity in misreporting across farms. In order to understand what drives larger misreporting, we linearly project farm level estimates of misreporting (gap increases) on farm characteristics using lasso regression. We include all 16 farm characteristics and their interaction terms, for a total of 136 covariates (120 interactions plus 16 variables). Lasso selects variables that largely explain differences in misreporting and drops unimportant variables. To obtain unbiased coefficients, we run OLS on individual variables selected by lasso.[23] Table 4.5 shows OLS coefficients of variables selected via post-lasso.

Table 4.5: OLS coefficients of farm characteristics explaining misreporting (after selection via lasso). Misreporting is defined by statistically significant increases in labor productivity gaps. Reference category for table and quality wines only is other wines only.

Obs. 277	Causal Effects	Std. Errors
utilized agricultural area (uaa, log ha)	-0.245^{***}	(0.007)
share land uaa for vine crops	0.146^{***}	(0.031)
share land uaa for other crops	0.087^{***}	(0.020)
altitude zone: high (>600m)	0.077^{***}	(0.030)
table grapes only (0/1)	0.006	(0.015)
grapes for PDO wine only (0/1)	0.045^{***}	(0.013)
n. grape types (table, PDO)	-0.027^{**}	(0.012)
Adjusted R^2	0.85	
F Statistic	221.7^{***}	

Notes: *p<0.1; **p<0.05; ***p<0.01

In our findings, we observe a negative relationship between misreporting and farm

[23] The selected interactions cause multicollinearity issues in the OLS regression post-lasso. We include individual variables selected as interactions and we exclude highly collinear variables.

size as measured by utilized agricultural area, all else being equal. It's important to note that agricultural land on farms can serve multiple purposes, such as woodland or other non-crop cultivation. We find that misreporting tends to be higher for farms that allocate larger shares of agricultural areas for growing vines or other crops like olives and vegetables. This suggests a potentially greater need for labor during harvest activities. Furthermore, misreporting is relatively more pronounced for remote farms, particularly those situated at higher altitudes (above 600 meters) categorized as the high altitude zone. Lastly, the type of grape being harvested also plays a role, with greater misreporting for farms that produce a single product, especially grapes for quality wines with Protected Designation of Origin (PDO). Since grapes for quality wine are worth more, this may indicate that farms have stronger incentives to misreport when the marginal income from misreporting is higher.

While the tendency for informal firms to be smaller in size has already been documented in previous studies (La Porta and Shleifer, 2014, 2008), the finding that economies of scope, remoteness, and product value correlate with the amount of informal employment are novel in the literature.

We perform *placebo tests* to show robustness of our estimates. We use causal forests to estimate migration causal effects on labor productivity gaps of farms in Sicily and Apulia pre-migration (2010). We find negligible and statistically insignificant effects for all farms, as expected. Figure D.3 in Appendix plots the estimated farm level causal effects in 2010 with their 95% confidence intervals.

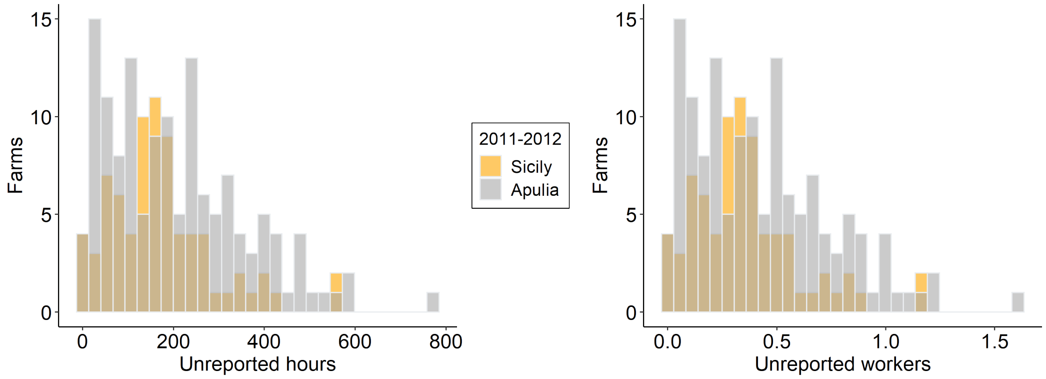
4.3 Migration effects on labor hours at farm level

We transform our estimated increases in labor productivity gaps into estimates of unreported labor hours employed on Sicily's and Apulia's vineyards. To do so, we focus on positive labor productivity gaps of treated farms for which we estimated statistically significant increases post-migration. We perform simple back of the envelope calculations in the following way. We calculate the estimated number of unreported hours for each misreporting farm as $\frac{\hat{\delta}(y-\hat{y})}{\hat{y}}h$, where $(y-\hat{y})$ is the estimated labor productivity gap, $\hat{\delta}$ is the forest-based estimate of the (statistically significant) percent gap increase caused by the migration shock, and

h is the number of reported labor hours.[24]

Figure 4.3 shows the distribution of unreported hours and workers in misreporting farms.

Figure 4.3: Farm level estimates of informal labor employment post-migration in Sicily and Apulia.



On average, farms misreport about 193 labor hours or 0.4 workers. Lower and upper bounds of the 95% confidence intervals for these estimates equal to, respectively, 88 and 313 unreported hours or 0.2 and 0.6 workers on average. As misreporting is significant in about 80% of farms exposed to the shock, our average estimate translates into 154 labor hours or 0.32 workers per farm in Sicily and Apulia. In other terms, the informal labor supply shock caused one in three farms to employ one worker informally. Mean estimates are on average higher in 2011 than in 2012 (219 vs. 165 hours) and higher in Apulia than in Sicily (206 vs. 171 hours), though differences are small.

Our data contains farm weights that make our sample of farms representative of the whole sector. The estimated number of misreporting farms equals to around 13,700 in Apulia and 11,200 in Sicily on average each year, with higher figures in 2011 (26,400) than in 2012 (23,000). We weigh the distribution of unreported labor hours at the farm level to obtain estimates at the sector level. We estimate a total of 11 millions unreported hours, of which 3.3 millions in Sicily and 7.7 millions in Apulia over 2011-2012. The share of unreported hours over reported hours is 6%. These estimates translate into about 23,000 workers informally employed in total over 2011-12, of which 14,000 workers in 2011 and 9,000 workers in 2012. This is

[24] This calculation is described in Table D.3 in Appendix.

6% (in 2011) and 4% (in 2012) of the total number of formal agricultural workers (224,000) in Sicily and Apulia reported by official national statistics, which is in line with the estimated shares using our data (INEA, 2012).[25] All upper and lower bound estimates of unreported hours and workers at farm and sector level for each region and year are reported in Appendix D.4.

These numbers may seem large compared to the official number of migrant landings in 2011 (64,000). Yet several studies, including INEA (2012), indicate a much larger scale for the migration shock which involved thousands of undetected migrants and at least 20,700 who escaped from aid camps and detention centers. Moreover, increased competition from migrant workers may have led many formally contracted workers to opt for informal employment.

Assuming minimum agricultural wages (net 7 euros per hour) and a 60-day 8-hour harvesting season, we estimate tax evasion for unpaid income taxes and social contributions for about 75 million euros.[26]

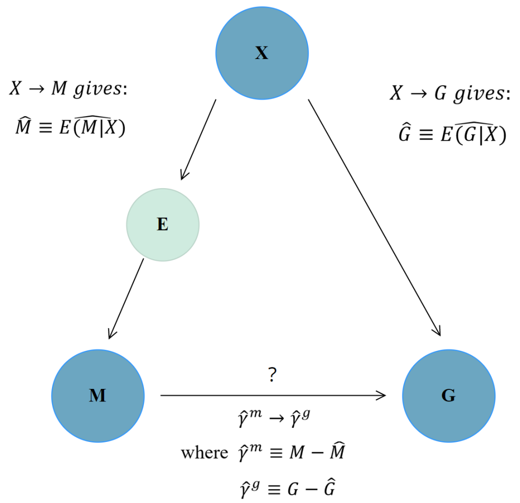
4.4 Migration effects on farm outcomes

We analyze changes in farms' economic outcomes caused by the increase of misreporting post-migration. Informal employment reduces farm production costs through saved labor taxes and net wages likely below the legal floor. Lower production costs may drive consumer prices down. In this case, part of the margin generated with informal employment would be redistributed to consumers. Differently, in the presence of some monopoly power, we expect producers to keep setting the same prices. This may also occur if producers are rather price takers because they sell mostly wholesale and not via private negotiations.

We estimate changes in farm outcomes causal to the migration shock using a residualized regression approach. For an intuitive illustration of this approach, consider the following causal diagram which graphically summarizes our data generating process:

[25] 95% confidence intervals around these estimates equal to [13k; 35k] total workers over 2011-2012.

[26] We calculate income taxes as the difference between gross and net salaries which amount to, respectively, 10 and 7 euros per hour, social contributions as 31% of total gross salaries, and complementary social contributions as total gross salaries divided by 13.5 (Danea, 2011).


X
X → M gives:
$\widehat{M} \equiv E(\widehat{M|X})$
X → G gives:
$\hat{G} \equiv E(\widehat{G|X})$
E
M
?
$\hat{\gamma}^m \rightarrow \hat{\gamma}^g$
where $\hat{\gamma}^m \equiv M - \widehat{M}$
$\hat{\gamma}^g \equiv G - \hat{G}$
G

- X = *(observed)* exogenous covariates: farm level (land, capital, altitude, etc.) + year and region fixed effects (to capture e.g. regions' inspection policy);
- E = *(unobserved)* farm expected payout (net benefit) from misreporting: expected benefits of misreporting net of expected costs of misreporting based on farmers' subjective probability of being caught;
- M = *(estimated in a previous step)* decision to misreport labor (increase in informal employment);
- G = *(observed)* farm gains (profitability).

The causal diagram shows that farm gains (G) are determined by both misreporting (M) and exogenous covariates (X) which include farm characteristics, region and year effects. Further, as shown by the heterogeneity analysis in the previous section, misreporting can be largely explained by farm characteristics. Realistically, X affect M through E, the unobservable expected payout from misreporting. E is how profitable a farmer might reasonably expect misreporting to be before hiring labor informally.[27] Given this diagram, the next steps are pretty clear. We need to get rid of the variation of G and M due to X in order to isolate just the variation we need.

[27] As mentioned in the Results, the marginal benefit of misreporting is higher if, e.g., land is more productive and farmers grow grapes for quality wine. The marginal cost of misreporting is based on the probability that farmers assign to being caught, which depends on factors such as location and the region's inspection policy.

The residualized regression approach consists in partialling out the confounding effects of X in three steps. First, we estimate propensities to misreport by regressing the (binary) treatment variable M on X and saving the residuals, $\hat{\gamma}^m \equiv M - \hat{M}$. As we set $M = 1$ for all misreporting farms in the treated group and $M = 0$ for non-treated farms, we can interpret $\hat{M}$ as propensity scores.[28] Second, we estimate the conditional outcome mean by regressing farm gains G on X and saving the residuals, $\hat{\gamma}^g \equiv G - \hat{G}$. Third, we regress residualized outcomes $\hat{\gamma}^g$ on the residualized treatment variable $\hat{\gamma}^m$ to remove the endogenous variation of G and M due to X. This method is based on Robinson (1988) and the Frisch-Waugh-Lovell theorem. In high-dimensional settings, this approach to estimate average causal effects has been generalized by Chernozhukov et al. (2018) under the name of "double machine learning".

We study the effects of misreporting labor hours on farm profitability in terms of Return On Assets (ROA, profits over assets). Dividing profits by fixed assets is a standard approach to capture differences in profits due to farm size. Further, since assets are fixed and cannot be altered in the short term, we can exclude migration effects on this outcome. To understand what drives changes in profits, we decompose the effect on ROA into an effect on sales and costs over assets. The effect on costs can be further decomposed into an effect on labor costs (total wages) and other costs.

Summary statistics of farm outcomes are reported in Table D.5 in Appendix. Compared to pre-migration, post-migration outcomes of misreporting farms and non-treated farms change in the same direction, with the median ROA growing by 3 percent points for misreporting farms and 1 percent point for non-treated farms. As shown by the causal diagram above, the question is to what extent this increase in ROA is causal to the migration shock once we control for the endogenous variation of outcomes and misreporting decisions with farm characteristics and fixed effects.

We predict farm outcomes using the full high-dimensional set of covariates (see Table C.1 in Appendix). We predict the treatment variable (=1 for misreporting farms post-migration) using the set of farm characteristics that possibly explain heterogeneity in misreporting (see Section 4.2). We control for fixed variation across years and regions with year and region fixed effects. As suggested by Cher-

[28] As before, misreporting farms are defined by statistically significant increases in labor productivity gaps after migration compared to their predicted gaps in absence of migration.

nozhukov et al. (2018), we use lasso for prediction.[29] This approach has been successfully applied to a variety of settings (see, Chernozhukov et al., 2016, and references therein).

Table 4.6 shows the OLS coefficients of the third-stage regression of the residualized outcome $\hat{\gamma}^g$, on the residualized treatment variable $\hat{\gamma}^m$.[30]

Table 4.6: Effects of misreporting post-migration ($\hat{\gamma}^m$) on farm outcomes. Third-stage model estimates (2010-2012). Models include high-dimensional set of covariates, year and region fixed effects. Robust double clustered standard errors (Driscoll and Kraay). The sample includes non-treated farms and misreporting farms in Sicily and Apulia observed in all periods.

Obs. 840	$\frac{Profit}{Assets}$	$\frac{Sales}{Assets}$	$\frac{Costs}{Assets}$	$\frac{Labor\ Costs}{Assets}$	$\frac{Other\ Costs}{Assets}$	$Price(log)$	$Wage/h$
$\hat{\gamma}^m$	0.063^{***}	0.0002	-0.071^{**}	-0.061^{*}	-0.012^{***}	0.010	-0.114
	(0.024)	(0.013)	(0.035)	(0.036)	(0.005)	(0.023)	(0.116)
Notes:	***p=.01; **p=.05; *p=.1						

Results shows that the migration shock had significant and positive effects on profits for misreporting farms in the landing regions. This effect is mostly driven by lower labor costs, rather than by higher sales or grape prices. In particular, misreporting significantly increases farm profitability by on average 6.3 percent points, ceteris paribus. While sales remain unaltered (+0.02 percent points, insignificant), costs significantly decrease by 7.1 percent points. This effect is largely explained by lower labor costs (-6.1 percent points).

These results seem plausible in light of how grape prices are set. The price fixation in wholesale markets is undertaken by the chamber of commerce based on production costs of landholdings. Thus, farmers are price takers when selling wholesale. Differently, farmers fix prices in private negotiations when trading, especially, higher quality (less substitutable) grapes. As we find that misreporting is higher for farms producing grapes for quality wine, ceteris paribus, farmers plausibly keep similar prices also when misreporting their labor inputs. In conclusion, we find that misreporting only benefits farm producers, without any positive effect on consumer prices.

Finally, we look at a possible declining effect on hourly wages. In fact, competition between formal and informal labor may lower wages of formal workers. In this direction, we find a small insignificant decrease in hourly wages of 11 euro

[29] We use the software R-4.0.3 and the hdm package version 0.3.1 (Chernozhukov et al., 2016).

[30] As for two-stage OLS, we do not report R^2 as it has no statistical meaning.

cents. This small effect is not surprising as most wages in the grape growing sector already touch the minimum floor admitted by law (INPS, 2011).[31]

5 Robustness checks

5.1 Placebo tests

Do we obtain comparable results if, instead of Sicily and Apulia, we consider other regions as treated? Inspired by the synthetic control literature (Abadie et al., 2015), we perform placebo tests by applying the SL to each non-treated region in the sample. If the estimated labor productivity gap in the treated regions is large relative to the one estimated for a non-treated region chosen at random, it is possible to conclude that the migration shock had a significant impact on misreporting in the treated regions. The probability of finding migration effects as high as in the treatment regions is reported as the fraction of non-treated units for which mean labor productivity gaps post-migration are larger or equal to the mean gap of the treated regions. Placebo tests account for the prediction accuracy of the estimated labor productivity before the shock, in particular, we do not include control regions for which the estimated mean gap is higher than the one obtained for the treated regions (in absolute terms). Figure 5.1 shows the density of the distribution of the estimated mean labor productivity gaps after the migration shock for the actual treated regions and for each placebo test.

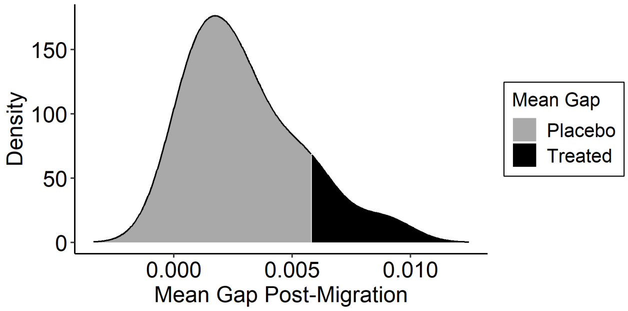


Figure 5.1: Placebo mean labor productivity gaps post-migration (grey) versus (black) actual gap for the treated regions (in absolute values).

Reassuringly, we find that the treated regions exhibit the largest estimated mean

[31] Hourly wages in our sample are mostly below 7.7 euros (third quartile).

labor productivity gap among all placebo tests. Figure D.4 in Appendix plots post-migration gaps against pre-migration gaps, showing that no control unit assigned to treatment has a higher gap post-migration as well as a lower gap pre-migration.

Additionally, we run placebo difference-in-differences (as in Table 4.4) using the estimated labor productivity gaps before and after the shock for farms in control regions assigned to treatment. We find that the coefficient of the treatment dummy post-migration is not significant in any of the placebo regressions.[32]

Overall, placebo tests provide evidence that vineyard labor productivity in other Italian regions is not affected by the migration shock as, instead, is the case for Sicily and Apulia.

5.2 Leave-one-out estimations

To evaluate the importance of including farms from a specific region outside of Sicily and Apulia when predicting treated outcomes following the shock, we perform Leave-One-Out estimations (LOO). This involves re-estimating labor productivity gaps for Sicily and Apulia while excluding farms from one non-treated region at a time.

Statistical tests show no significant differences between LOO estimates and actual estimates of labor productivity gaps in any year, neither in terms of mean (t-tests, p-values>0.3) nor distribution (K-S tests, p-values>0.4). Figure D.5 in Appendix plots LOO estimates versus actual estimates, revealing a strong overlap.

In conclusion, the results indicate that the inclusion of farms from any region in the non-treated sample does not influence our findings.

5.3 Alternative to bootstrap inference on misreporting

Whilst the bootstrap approach used in Section 2.3 is arguably the workhorse methodology for this type of applications, results obtained with it are potentially sensitive to the bootstrap algorithm specifications. Hence, as mentioned at the end of Section 2.3, we also estimate confidence intervals for the labor productivity gaps using an approach based on inverting randomised tests (Horváth and Trapani, 2019) – we refer to Appendix B for the full-blown description of this methodology. We find that our estimates of misreported labor hours are robust

[32] Results of all placebo regressions are available upon requests.

to the specific methodological choice. In particular, results from randomised tests show that causal effects on labor productivity gaps post-migration are statistically significant for about 85% of treated farms (vs. 80% using bootstrap confidence intervals). Full results with upper and lower bounds of 95% confidence intervals constructed via de-randomised inference are reported in Table D.6. These findings are in line with our main estimates in Table D.4.

6 Conclusions

This paper provides a new approach to detect and quantify informal employment at the firm level resulting from irregular migration shocks. Our estimates provide novel insights into the economic implications of irregular migration on formal employment, firm outcomes, consumer prices, and labor tax evasion.

By employing machine learning techniques and balance sheet data of Italian vineyards, we find that the deviations in reported labor productivity compared to predicted values for vineyards affected by the migration shock are significant. Our estimates reveal that the shock resulted in a 6% increase in informal employment, equivalent to one undeclared worker for every three farms on average and a total of 23,000 additional workers over 2011-2012. The underreporting of labor hours leads to lower labor costs and higher farm profitability, without affecting grape sales, prices, or hourly wages of formal workers. We estimate that labor tax evasion amounted to approximately 75 million euros, representing around 5% of the total agricultural revenues collected during 2011-2012. In terms of policy relevance, we believe that our methodology, which does not require data on irregular migrants or informal workers, could be useful for developing more effective policies that promote formalization in a wide set of labor-intensive sectors beyond agriculture (e.g., manufacturing, construction, and services).

References

Abadie, A., A. Diamond, and J. Hainmueller (2015). Comparative politics and the synthetic control method. *American Journal of Political Science 59*(2), 495–510.

Agri4Cast. Gridded Agro-Meteorological Data in Europe. *Data retrieved from Agri4Cast. Available at https://agri4cast.jrc.ec.europa.eu/DataPortal/Index.aspx.* Accessed: June 2023.

Altındağ, O., O. Bakış, and S. V. Rozo (2020). Blessing or burden? Impacts of refugees on businesses and the informal economy. *Journal of Development Economics 146*, 102490.

ANCI (2011). Accoglienza ai profughi provenienti dal Nord Africa (in Italian). *National Association of Italian Municipalities - Anci. Available at https://anci.lombardia.it/dettaglio-circolari/20115301746-circolare-59-2011/anci.lombardia.it.* Accessed: June 2023.

Ashenfelter, O. and K. Storchmann (2016). Climate change and wine: A review of the economic implications. *Journal of Wine Economics 11*(1), 105–138.

Assosomm (2016). Attiviamo lavoro. Le potenzialitá del lavoro in somministrazione per il settore dell'agricoltura (in Italian). *Report of The European House-Ambrosetti for the Italian Association of Labor Agencies (Assosomm).*

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Athey, S. and S. Wager (2019). Estimating treatment effects with causal forests: An application. *Observational Studies 5*(2), 37–51.

Bajari, P., D. Nekipelov, S. P. Ryan, and M. Yang (2015). Machine learning methods for demand estimation. *American Economic Review 105*(5), 481–85.

Balkan, B. and S. Tumen (2016). Immigration and prices: Quasi-experimental evidence from Syrian refugees in Turkey. *Journal of Population Economics 29*(3), 657–686.

Bates, J. M. and C. W. Granger (1969). The combination of forecasts. *Journal of the Operational Research Society 20*(4), 451–468.

Bickel, P. J. and D. A. Freedman (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics 9*(6), 1196–1217.

Borjas, G. (2017). The labor supply of undocumented immigrants. *Labour Economics 46*, 1–13.

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

Breiman, L. (2017). *Classification and regression trees.* Routledge.

Bueno, M. and M. Valente (2019). The effects of pricing waste generation: A synthetic control approach. *Journal of Environmental Economics and Management 96*, 274–285.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, 01). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., C. Hansen, and M. Spindler (2016). Hdm: High-dimensional metrics. *R Journal 8*(2), 185–199.

Ciaian, P., E. Baldoni, d. Kancs, and D. Drabik (2021). The capitalization of agricultural subsidies into land prices. *Annual Review of Resource Economics 13*(1), 17–38.

Corradi, V. and N. R. Swanson (2006). The effects of data transformation on common cycle, cointegration, and unit root tests: Monte Carlo and a simple test. *Journal of Econometrics 132*, 195–229.

CREA. Come funziona (in Italian). *Available at https://rica.crea.gov.it/come-funziona-726.php*. Accessed: June 2023.

Danea (2011). Costo del personale: Composizione e calcolo (in Italian). *Danea Soft. Available at http://www.danea.it/blog/costo-aziendale-dipendente/*. Accessed: June 2023.

DPC (2011). Piano per l'accoglienza dei migranti (12 April 2011, in Italian). *Department of Civil Protection (DPC) of the Presidency of the Council of Ministers*.

Driscoll, J. C. and A. C. Kraay (1994). Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics 80*(4), 549–560.

Drucker, H., C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik (1996). Support vector regression machines. *Advances in Neural Information Processing Systems 9*.

Elliott, G. (2011). Averaging and the optimal combination of forecasts. *Working Paper, University of California, San Diego*.

EU (2021). Migrant seasonal workers in the European agricultural sector. *EPRS, European Parliamentary Research Service*.

EU-FADN. European Farm Accountancy Data Network - DG AGRI. *Available at https://agriculture.ec.europa.eu/data-and-analysis/farm-structures-and-economics/fadn_en*. Accessed: June 2023.

EU-FADN (2018). EU Farm Economics Overview FADN.

Fan, J. and I. Gijbels (2018). *Local polynomial modelling and its applications.* New York: Routledge.

FLAI (2012). Agrimafie e Caporalato, First Report (in Italian). *Osservatorio Placido Rizzotto Flai-Cgil.*

FLAI (2014). Agrimafie e Caporalato, Second Report (in Italian). *Osservatorio Placido Rizzotto FLAI.*

FLAI (2018). Agrimafie e Caporalato, Fourth Report (in Italian). *Osservatorio Placido Rizzotto Flai-Cgil.*

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1.

FRONTEX (2016). Migratory routes map. *European Border and Coast Guard Agency. Available at https://frontex.europa.eu/we-know/migratory-map/.* Accessed: June 2023.

Geyer, C. J. and G. D. Meeden (2005). Fuzzy and randomized confidence intervals and p-values. *Statistical Science 20*(4), 358–366.

Ghizzi, E. (2015). L'accoglienza dei richiedenti e titolari di protezione internazionale in Italia. Aspetti giuridici e sociologici (in Italian). *La Rivista, ADIR Research Center, University of Firenze. Available at http://www.adir.unifi.it/rivista.* Accessed: June 2023.

Giangrande, A. (2017). *Caporalato Ipocrisia e Speculazione (in Italian)*, Volume 165. StreetLib. Available at Mondadori Store.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning.* Springer.

Hoerl, A. E. and R. W. Kennard (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 42*(1), 80–86.

Horváth, L. and L. Trapani (2019). Testing for randomness in a random coefficient autoregression model. *Journal of Econometrics 209*(2), 338–352.

IDOS (2019). Dossier statistico immigrazione (in Italian). *Centro Studi e Ricerche IDOS.*

ILO (2023). Intervention model for extending social protection to migrant workers in the informal economy. *International Labour Organization.*

Imbens, G. and D. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences - Chapter 14 Assessing Overlap in Covariate Distributions*, Volume III. Cambridge University Press.

INEA (2012). Indagine sull'impiego degli immigrati in agricoltura in Italia (in Italian). *Report of the National Institute of Agricultural Economics, INEA.*

INEA (2014). Indagine sull'impiego degli immigrati in agricoltura in Italia (in Italian). *Report of the National Institute of Agricultural Economics, INEA.*

INEA (2015). Le imposte sulle imprese agricole: Un'analisi quantitativa (in Italian). *Agriregionieuropa and National Institute of Agricultural Economics, INEA. Available* https://agriregionieuropa.univpm.it/it/content/article/31/43/le-imposte-sulle-imprese-agricole-unanalisi-quantitativa. Accessed: June 2023.

INL (2012). Rapporto annuale attivitá di tutela e vigilanza (in Italian). *Ispettorato Nazionale del Lavoro (INL).*

INPS (2011). Rilevazione delle retribuzioni contrattuali degli operai a tempo determinato e degli operai a tempo indeterminato del settore agricolo, in vigore alla data del 30.10.2011, per la determinazione delle medie salariali (in Italian). *INPS, Direzione Centrale Entrate, Circolare n. 152 of 06-12-2011.*

Interior Ministry (2012). Emergenza Nord Africa - Procedura informatizzata Vestanet C3 - gestione Nord Africa (in Italian). *Act of 30 October 2012, n. 5426.*

ISTAT (2017). Economia non osservata nei conti nazionali (in Italian). *ISTAT National Statistics.*

Jorgenson, D. W. and Z. Griliches (1967). The explanation of productivity change. *Review of Economic Studies 34*(3), 249–283.

Kelly, C. B. (1977). Counting the uncountable: Estimates of undocumented aliens in the United States. *Population and Development Review 3*(4), 473–481.

La Porta, R. and A. Shleifer (2008). The unofficial economy and economic development. *Brookings Papers on Economic Activity 2008*, 275–352.

La Porta, R. and A. Shleifer (2014). Informality and development. *The Journal of Economic Perspectives 28*(3), 109–126.

La Repubblica (2011). Immigrati, il 2011 anno record di arrivi: Aumentano le richieste d'asilo, piú 102% (in Italian). *La Repubblica. Available at* https://www.repubblica.it/solidarieta/immigrazione/2011/12/30/news/immigrati_il_2011_anno_record_di_arrivi_aumentano_le_richieste_d_asilo_pi_102_-27412805/#:~:text=ROMA%20%2D%20Si%20chiude%20l'anno,allo%20stesso%20periodo%20del%202010. Accessed: June 2023.

La Repubblica (2017). Caporalato, in Salento 4 imprenditori condannati a 11 anni: Migranti schiavizzati (in Italian). *La Repubblica. Available at* https://bari.repubblica.it/

cronaca/2017/07/13/news/caporalato_in_salento_4_imprenditori_condannati_a_11_anni_ridussero_in_schiavitu_-170708432/. Accessed: June 2023.

Labanca, C. (2020). The effects of a temporary migration shock: Evidence from the Arab Spring migration through italy. *Labour Economics 67*, 101903.

Lambruschi, P. (2012). Integrazione alla prova. La Caritas: Non rimandare all'inferno chi chiede asilo (in Italian). *Avvenire*.

Lamouria, L., H. Studer, and H. Brewer (1963). Improving the productivity of pruning labor in the vineyard. *California Agriculture 17*(3), 2–3.

Latruffe, L., B. E. Bravo-Ureta, A. Carpentier, Y. Desjeux, and V. H. Moreira (2017). Subsidies and technical efficiency in agriculture: Evidence from European dairy farms. *American Journal of Agricultural Economics 99*(3), 783–799.

L'Espresso (2012). Scandalo profughi (in Italian). *L'Espresso, Available at https://www.meltingpot.org/app/uploads/2012/10/Scandalo_profughi_ESPRESSO.pdf*. Accessed: June 2023.

L'Espresso (2015). Migranti, in centomila sono scomparsi (in Italian). *L'Espresso. Available at http://espresso.repubblica.it/plus/articoli/2015/01/21/news/migranti-la-grande-fuga-1.195858*. Accessed: June 2023.

Levene, H. (1960). Robust tests for equality of variances. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*.

Lianos, T. P., A. H. Sarris, and L. T. Katseli (1996). Illegal immigration and local labour markets: The case of Northern Greece. *International Migration 34*(3), 449–484.

Massacci, D., L. Sarno, and L. Trapani (2021). Factor models with downside risk. *Available at SSRN 3937321*.

Massacci, D. and L. Trapani (2022). High dimensional threshold regression with common stochastic trends. *Available at SSRN 4133488*.

MPP (2013). Dalla protezione temporanea alla protezione internazionale: Quale accoglienza? (in Italian). *Melting Pot Project*.

Nie, X. and S. Wager (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika 108*(2), 299–319.

Palmisano, L. and Y. Sagnet (2016). *Ghetto Italia. I braccianti stranieri tra capolarato e sfruttamento (in Italian).* Fandango Editore.

Penal Code (2009). Law n. 94 of 15 July 2009 (in Italian). *Italian Penal Code*.

Polley, E. (2021). Superlearner. R package version 2.0-28. Available at http://CRAN.R-project.org/package=SuperLearner.

quattrocalici.it. Superficie vitata delle regioni italiane (in Italian). *Available at https://www.quattrocalici.it/articoli/superficie-vitata-delle-regioni-italiane/*. Accessed: June 2023.

Rabbani, M., Y. Wang, R. Khoshkangini, H. Jelodar, R. Zhao, S. Bagheri Baba Ahmadi, and S. Ayobi (2021). A review on machine learning approaches for network malicious behavior detection in emerging technologies. *Entropy 23*(529).

Robinson, P. (1988). Root-n-consistent semiparametric regression. *Econometrica 56*(4), 931–954.

Rosenbaum, P. and D. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). Learning representations by back-propagating errors. *Nature 323*(6088), 533–536.

Sacchetto, D. and D. Perrotta (2012). Il ghetto e lo sciopero: Braccianti stranieri nell'Italia meridionale (in Italian). *Sociologia del lavoro* (128), 152–166.

Salzer, H. E., R. Zucker, and R. Capuano (1952). Table of the zeros and weight factors of the first twenty hermite polynomials. *Journal of Research of the National Bureau of Standards 48*, 111–116.

Schennach, S. M. (2016). Recent advances in the measurement error literature. *Annual Review of Economics 8*, 341–377.

SPRAR (2010). Rapporto annuale del sistema di protezione per richiedenti asilo e rifugiati (in Italian). *Report 2009/2010 of the System for the Protection of Asylum Seekers and Refugees (SPRAR).*

SPRAR (2011). Rapporto annuale del sistema di protezione per richiedenti asilo e rifugiati (in Italian). *Report 2010/2011 of the System for the Protection of Asylum Seekers and Refugees (SPRAR).*

SPRAR (2012). Rapporto annuale del sistema di protezione per richiedenti asilo e rifugiati (in Italian). *Report 2011/2012 of the System for the Protection of Asylum Seekers and Refugees (SPRAR).*

Stock, J. and M. Watson (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting 23*(6), 405–430.

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

Toescher, A., M. Jahrer, and R. M. Bell (2009). The BigChaos solution to the Netflix Grand Prize. *Netflix Prize Documentation*, 1–52.

Tumen, S. (2016). The economic impact of Syrian refugees on host countries: Quasi-experimental evidence from Turkey. *American Economic Review 106*, 456–60.

Vaiou, D. and C. Hadjimichalis (1997). *With the sewing machine in the kitchen and the Poles in the fields. Cities, regions and informal work.* Athens: Exandas.

Valente, M. (2023). Policy evaluation of waste pricing programs using heterogeneous causal effect estimation. *Journal of Environmental Economics and Management 102755.*

Van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2008). Super Learner. *Statistical Applications in Genetics and Molecular Biology 6*(1).

Venturini, A. and C. Villosio (2008). Labour-market assimilation of foreign workers in Italy. *Oxford Review of Economic Policy 24*(3), 517–541.

Wang, L. (2005). *Support vector machines: Theory and applications*, Volume 177. Springer Science & Business Media.

Warren, R. and J. S. Passel (1987). A count of the uncountable: Estimates of undocumented aliens counted in the 1980 United States Census. *Demography 24*(3), 375–393.

World Bank (2011). Informal workers across Europe: Evidence from 30 European countries. *Policy Research WPS n. 5912 Washington, D.C.: World Bank Group*.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B 67*(2), 301–320.

Online Appendix

A Background Appendix

In 2011, the Italian regions of Sicily, Apulia and Calabria hosted a large number of undetected migrants as well as asylum seekers. Asylum seekers are kept in Centers for First Assistance (aka CDA) and Centers for Assistance of Asylum Seekers (aka CARA). About 95% of asylum seekers in CDA and CARA are in Sicily, Apulia, and Calabria (SPRAR, 2012). As a consequence of the 2011 migration wave, Sicily and Apulia each hosted around 10,000 asylum seekers more than in 2010. No such increase was observed in other Italian regions.[33] Therefore, we may expect that, despite surveillance, only migrants hosted in asylum centers of Sicily, Apulia, and Calabria may leave camps in large numbers and enter informal labor channels in these regions. This is further reinforced by anecdotal evidence indicating many cases of migrant fleeing asylum centers (see, e.g., L'Espresso, 2015).

The relocation of asylum seekers may cause shocks to the local formal and/or informal agricultural labor market of other Italian regions. However, official statistics reveal that Italian regions did *not* experience abnormal arrivals or departures of asylum seekers in their SPRAR[34] centers over the years 2010-2012. Table A.1 shows the share of asylum seekers hosted in SPRAR centers in each Italian region over 2010-2012.

[33] Some regions (e.g., Lazio, Marche, and Friuli Venezia Giulia) even reduced the number of hosted migrants from around 2,100 on average in 2010 to 1,600 on average in 2011 (SPRAR, 2012).

[34] SPRAR is the System for Protection of Asylum Seekers and Refugees.

Table A.1: Share of asylum seekers hosted in SPRAR reception centers in each Italian region over 2010-2012 (SPRAR, 2010, 2011, 2012)

	2010	2011	2012		2010	2011	2012
North-East				**North-West**			
Friuli V.G.	4.8	4.6	4.2	Aosta V.	0.0	0.0	0.0
Veneto	4.7	5.8	4.0	Piedmont	4.6	5.3	4.5
Trentino A.A.	0.6	0.6	0.4	Lombardy	16.5	5.7	16.8
Emilia R.	6.2	7.8	6.8	Liguria	2.7	3.0	2.2
Center				**South**			
Tuscany	4.4	4.5	4.6	Abruzzo	0.5	0.5	0.6
Marche	4.2	4.5	3.5	Molise	0.5	0.6	0.6
Lazio	22.4	26.2	21.2	Campania	2.9	3.2	2.0
Umbria	2.0	2.7	2.0	Basilicata	0.7	0.6	0.7
Islands				Apulia	7.1	8.0	6.2
Sicily	11.4	11.3	14.6	Calabria	3.5	4.7	4.9
Sardinia	0.4	0.4	0.3	**Total**	7056	7598	7823

In addition to SPRAR centers, Italy distributed a total of 17,859 asylum seekers in hotels and apartments of each region according to its population (DPC, 2011).[35] As a result, each region received a relatively low number of migrants. Table A.2 shows that the relocated migrants per region are around 1,000 on average. We claim that asylum seekers and refugees have no incentives to put their status at risk by working illegally: hosted migrants cannot take unjustified daily leaves and are not allowed to work.[36] Also, in addition to board and lodging, asylum seekers receive pocket money and temporary documents to access health care services (ANCI, 2011). Therefore, it does *not* seem plausible that relocated migrants illegally worked on vineyards in other parts of Italy.

Finally, approved asylum seekers may cause spillovers by working in other Italian regions. In this respect, most of the approved asylum seekers seem to have left Italy after receiving a positive verdict (Labanca, 2020). However, as of November 2012, many pending and rejected asylum seekers did not leave migrant facilities whilst waiting for asylum decisions or appeals (Lambruschi, 2012). Thus, the Italian government offered an accelerated procedure to regularize asylum applicants through concession of a humanitarian visa (Interior Ministry, 2012). Thereby, from 2013 onward, rejected asylum seekers, in particular, may have entered illegal labor

[35] I.e., 10,000 migrants for 100,000 inhabitants.

[36] While asylum seekers are officially prohibited from working for six months after submitting their request, obtaining work permits remains challenging in practice, even up to 12 months following the initial request (MPP, 2013).

channels to avoid expulsion (Giangrande, 2017). For this reason we do not extend our empirical analysis after 2012.

Table A.2: Number of Arab Spring asylum seekers and refugees relocated across Italian regions in apartments and hotels (L'Espresso, 2012).

	N. People	Share	Pop.*		N. People	Share	Pop.*
North-East				**North-West**			
Friuli V.G.	397	0.02	1.2	Aosta V.	20	0.00	0.1
Veneto	1274	0.07	4.9	Piedmont	1549	0.09	4.4
Trentino A.A.	172	0.01	1.1	Lombardy	2548	0.14	10
Emilia R.	1585	0.09	4.5	Liguria	540	0.03	1.6
Center				**South**			
Tuscany	1141	0.06	3.7	Abruzzo	11	0.00	1.3
Marche	462	0.03	1.5	Molise	116	0.01	0.3
Lazio	1790	0.10	5.9	Campania	2155	0.12	5.8
Umbria	338	0.02	0.9	Basilicata	200	0.01	0.6
Islands				Apulia	1071	0.06	4.1
Sicily	1110	0.06	5.1	Calabria	956	0.05	2.0
Sardinia	424	0.02	1.6	**Total**	17,859		60.5

* Total regional population in million people

B Methodology Appendix

As mentioned in Section 2.3, confidence intervals for $\widehat{\delta}_{i,t}$ have been constructed using the bootstrap approach presented in Athey et al. (2019); hence, results may be sensitive to the specifications of the algorithm, and also be affected by the multiple testing problem.

In order to tackle both issues, we propose the following approach, which is based on randomised tests, and which has been developed in a series of contributions albeit in different contexts (we refer in particular to Horváth and Trapani, 2019, Massacci et al., 2021, and Massacci and Trapani, 2022). Consider the individual estimates $\widehat{\delta}_{i,t}$ defined in (13), and consider a sequence s_{N_1} such that

$$\lim_{N_1 \to \infty} s_{N_1} = \infty.$$

When estimating $\widehat{\delta}_{i,t}$ through (13), we follow Athey and Wager (2019), and use a Generalised Random Forest where trees are grown on subsamples of size $m = O\left(N^{\beta}\right)$, where $\beta < 1$ is chosen according to equation (13) in Athey et al. (2019). Hence, we make the following assumption on s_{N_1}.

Assumption B.1. *It holds that* $\lim_{N_1\to\infty} s_{N_1} = \infty$ *with*

$$s_{N_1} = O\left(\left(\frac{N_1}{m}\right)^{1/2-\varepsilon}\right),$$

for a user-chosen $\varepsilon > 0$*.*

Heuristically, note that Theorem 5 in Athey et al. (2019) entails that

$$\widehat{\delta}_{i,t} - \delta_{i,t} = O_P\left(\sqrt{\frac{m}{N_1}}\log^p\left(\frac{N_1}{m}\right)\right),$$

for some $p > 0$. Hence, by Assumption B.1, it follows immediately that

$$s_{N_1}\left(\widehat{\delta}_{i,t} - \delta_{i,t}\right) = o_P\left(1\right).$$

More importantly, when testing for

$$\begin{aligned} H_0 &: \delta_{i,t} = \delta_{i,t}^0, \\ H_A &: \delta_{i,t} \neq \delta_{i,t}^0, \end{aligned}$$

Assumption B.1 entails that

$$\begin{aligned} \lim_{N_1\to\infty} P\left(s_{N_1}\left|\widehat{\delta}_{i,t} - \delta_{i,t}^0\right| = 0\right) &= 1 \text{ under } H_0, \\ \lim_{N_1\to\infty} P\left(s_{N_1}\left|\widehat{\delta}_{i,t} - \delta_{i,t}^0\right| = \infty\right) &= 1 \text{ under } H_A. \end{aligned}$$

(Significance) testing for $H_0 : \delta_{i,t} = \delta_{i,t}^0$

We now consider the construction of the significance tests for H_0 of (14). To this end, we define the sequences

$$\phi_{i,t}\left(N_1\right) = \exp\left(\left(\frac{s_{N_1}\left|\widehat{\delta}_{i,t} - \delta_{i,t}^0\right|}{\sigma_\delta}\right)^{-1}\right) - 1, \tag{15}$$

where σ_δ is a scaling factor which we discuss later on. Based on Assumption B.1, it is easy to see that

$$\begin{aligned}\lim_{N_1\to\infty} P\left(\phi_{i,t}\left(N_1\right)=\infty\right) &= 1 \text{ under } H_0,\\ \lim_{N_1\to\infty} P\left(\phi_{i,t}\left(N_1\right)=0\right) &= 1 \text{ under } H_A;\end{aligned}$$

that is, we have a statistic, $\phi_{i,t}\left(N_1\right)$, which diverges under the null and drifts to zero under the alternative. In order to use it in a test, we propose the following randomisation algorithm (see also Corradi and Swanson, 2006).

Step 1 Generate an artificial sample $\{\xi_{i,t,m}, 1\leq m\leq M\}$, *i.i.d.* across i,t,m with $\xi_{i,t,1}\sim N\left(0,1\right)$.
Step 2 Define the Bernoulli random variable $\zeta_{i,t,m}\left(u\right)=I\left(\sqrt{\phi_{i,t}\left(N_1\right)}\times\xi_{i,t,m}\leq u\right)$.
Step 3 Define the test statistic

$$S_{i,t}\left(M,N_1\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left[\varsigma_{i,t}\left(u\right)\right]^2\exp\left(-\frac{1}{2}u^2\right)du, \tag{16}$$

where

$$\varsigma_{i,t}\left(u\right)=\frac{2}{M^{1/2}}\sum_{m=1}^{M}\left(\zeta_{i,t,m}\left(u\right)-\frac{1}{2}\right).$$

Let P^* denote the conditional probability with respect to the sample $\{\left(y'_{i,t},X'_{i,t}\right),1\leq i\leq N,1\leq t\leq T\}$; we use the notation "$\overset{D^*}{\to}$" and "$\overset{P^*}{\to}$" to define conditional convergence in distribution and in probability according to P^*. Finally, χ^2_1 denotes a chi-square with one degree of freedom.

Theorem B.1. *We assume that Assumption B.1 is satisfied and that* $M=O\left(N_1\right)$. *If* H_0 *holds, then, as* $\min\left(N_1,M\right)\to\infty$, *it holds that* $S_{i,t}\left(N_1,M\right)\overset{D^*}{\to}\chi^2_1$, *with probability tending to* 1, *for all* $1\leq i\leq N_1$ *and* $1\leq t\leq T_1$. *Under* H_A, *it holds that* $M^{-1}S_{i,t}\left(N_1,M\right)\overset{P^*}{\to}c_0>0$, *with probability tending to* 1, *for all* $1\leq i\leq N_1$ *and* $1\leq t\leq T_1$.

Theorem B.1 reports the limiting distribution of the test statistics $S_{i,t}\left(N_1,M\right)$ under the null $H_0:\delta_{i,t}=\delta^0_{i,t}$. In essence, the theorem states that, when carrying out tests for $H_0:\delta_{i,t}=\delta^0_{i,t}$, the value of $S_{i,t}\left(N_1,M\right)$ should be contrasted with a critical value from a χ^2_1 distribution. Also, according to the theorem, tests based on $S_{i,t}\left(N_1,M\right)$ will reject the null for large values of the test statistic.

Whilst the proof of the theorem is reported later on, here we offer a heuristic description of how the test works. We begin by considering what happens under the null. Since in this case $\phi_{i,t}(N_1)$ diverges, the Bernoulli random variable $\zeta_{i,t,m}(u)$ has, for every u, success probability 1/2, as well as being independent across m by construction. Thus, by the CLT, $\varsigma_{i,t}(u)$ should behave approximately as a standard normal as $M \to \infty$, and consequently its square, $S_{i,t}(N_1, M)$, follows a chi-squared distribution with one degree of freedom. Under the alternative, $\phi_{i,t}(N_1)$ drifts to zero, and therefore $\zeta_{i,t,m}(u)$ does not have success probability 1/2. Therefore, when constructing $\varsigma_{i,t}(u)$, there is a bias term which grows proportionally to $M^{1/2}$, whence the divergence of $S_{i,t}(N_1, M)$ at rate M.

Technically, we note that the mode of convergence of $S_{i,t}(N_1, M)$ is the same as in the case of the bootstrap (see Bickel and Freedman, 1981), i.e. $S_{i,t}(N_1, M)$ converges to χ_1^2 in distribution "in probability conditional on the sample".

De-randomised inference and confidence intervals

In (20), a crucial choice is the level of the individual tests, α. In our analysis, we want to check if $E(\delta_{i,t}) > 0$ for all treated units. In this case, we need to control the family-wise rejection rate of our procedure, and using a "traditional" nominal level such as e.g. $\alpha = 0.05$ would lead to over-rejection and, consequently, to a spurious detection of significant treatment effects. In our case, we can correct this upon noting that, by construction (and conditionally on the sample), the test statistics $S_{i,t}(N_1, M)$ are *i.i.d.* across i, t. Thus, we can propose a Bonferroni correction, whereby - in order to ensure control of the family-wise rejection rate - we use

$$\alpha_{N_1} = \frac{c}{N_1},$$

where c is the desired family-wise rejection rate. Then, letting c_{α,N_1} be defined as $P(\chi_1^2 \geq c_{\alpha,N_1}) = \alpha_{N_1}$, $\alpha_{N_1} \in (0,1)$, it is immediate to see that, if $\delta_{i,t} = 0$ for all $1 \leq i \leq N_1$ and $1 \leq t \leq T_1$, then it follows that

$$\lim_{\min(N_1,M)\to\infty} P^* \left(\max_{i,t} S_{i,t}(N_1, M) \geq c_{\alpha,N_1} \right) = c, \tag{17}$$

a.s. conditional on the sample. The test statistic $S_{i,t}(N_1, M)$ can be used directly to test for

$$H_0 : \delta_{i,t} = 0,$$

and to construct confidence intervals for $\widehat{\delta}_{i,t}$.

We would like to point out however that $S_{i,t}\left(N_1, M\right)$ is constructed using a randomisation which does not vanish asymptotically, and therefore, under the null, different researchers using the same data will obtain different values of $S_{i,t}\left(N_1, M\right)$ and, consequently, different p-values. Although reproducibility can be guaranteed if the researchers reported the seed of their random number generator, we also propose to "de-randomise" $S_{i,t}\left(N_1, M\right)$, in a similar way as proposed in Horváth and Trapani (2019). This can be done as follows. Each researcher, instead of computing $S_{i,t}\left(N_1, M\right)$ just once, will compute the test statistic B times, at each iteration b using a sequence $\left\{\xi_{i,t,m}^{(b)}, 1 \leq m \leq M, 1 \leq b \leq B\right\}$ independent across i, t, m and b, thence defining, for the generic null hypothesis $H_0 : \delta_{i,t} = \delta$

$$Q_{i,t}\left(\delta; \alpha_{N_1}\right) = B^{-1} \sum_{b=1}^{B} I\left[S_{i,t,b}\left(N_1, M\right) \leq c_{\alpha_{N_1}}\right]. \tag{18}$$

The function $Q_{i,t}\left(\delta; \alpha_{N_1}\right)$ is related to the notion of "fuzzy confidence interval" studied in Geyer and Meeden (2005). Massacci and Trapani (2022) show that

$$\begin{array}{ll} \lim_{\min(B,M,N_1)\to\infty} P^*\left(Q_{i,t}\left(\delta; \alpha_{N_1}\right) = 1 - \alpha_{N_1,T_1}\right) = 1 & \text{for } \delta_{i,t} = 0, \\ \lim_{\min(B,M,N_1)\to\infty} P^*\left(Q_{i,t}\left(\delta; \alpha_{N_1}\right) = 0\right) = 1 & \text{for } \delta_{i,t} \neq 0. \end{array} \tag{19}$$

Equation (19) stipulates that, as $B \to \infty$, averaging across b in (18) washes out the added randomness in $Q_{i,t}\left(\delta; \alpha_{N_1}\right)$: all researchers using this procedure will obtain the same value of $Q_{i,t}\left(\delta; \alpha_{N_1}\right)$, thereby ensuring reproducibility. The function $Q_{i,t}\left(\delta; \alpha_{N_1}\right)$ corresponds to (the complement to one of) the "fuzzy decision", or "abstract randomised decision rule" reported in equation (1.1a) in Geyer and Meeden (2005). Geyer and Meeden (2005) provide a helpful discussion of the meaning of $Q_{i,t}\left(\delta; \alpha_{N_1}\right)$: the problem of deciding in favour or against H_0 may be modelled through a random variable, say D, which can take two values, namely "do not reject H_0" and "reject H_0". Such a random variable has probability $Q_{i,t}\left(\delta; \alpha_{N_1}\right)$ to take the value "do not reject H_0", and probability $1 - Q_{i,t}\left(\delta; \alpha_{N_1}\right)$ to take the value "reject H_0". In this context, (19) states that (asymptotically), the probability of the event $\{\omega : D = \text{"reject } H_0\text{"}\}$ is α when H_0 is satisfied, for all researchers – corresponding to the notion of *size* of a test; see also the quote from Corradi and Swanson (2006) reported above. Conversely, under H_1, the probability of the event $\{\omega : D = \text{“reject } H_0\text{”}\}$ is 1 (asymptotically), corresponding to the notion of *power*.

Based on $Q_{i,t}(\delta;\alpha_{N_1})$, it is possible to propose a decision rule to decide in favour or against H_0, based e.g. on thresholding $Q_{i,t}(\delta;\alpha_{N_1})$. A possible threshold, suggested in Massacci and Trapani (2022), is based on the Law of the Iterated Logarithm

$$\begin{array}{ll} Q_{i,t}(\delta;\alpha_{N_1}) \geq 1-\alpha_{N_1}-\sqrt{2\alpha_{N_1}(1-\alpha_{N_1})\frac{\ln\ln B}{B}} & \text{do not reject } H_0 \\ Q_{i,t}(\delta;\alpha_{N_1}) < 1-\alpha_{N_1}-\sqrt{2\alpha_{N_1}(1-\alpha_{N_1})\frac{\ln\ln B}{B}} & \text{reject } H_0 \end{array}, \qquad (20)$$

which corresponds to (marginally) conservative confidence intervals. Alternatively, repeating the proof of Theorem 3.3 in Massacci and Trapani (2022), it is easy to see that, upon using $B = o(M)$, the confidence intervals defined as

$$C_{\alpha_{N_1}}(\delta) = \left\{\delta : Q_{i,t}(\delta;\alpha_{N_1}) \geq 1-\alpha_{N_1}-\widetilde{c}_{\alpha_{N_1}}\sqrt{\frac{\alpha_{N_1}(1-\alpha_{N_1})}{B}}\right\},$$

where $P\left(Z \geq \widetilde{c}_{\alpha_{N_1}}\right) = 1-\alpha_{N_1}$ and $Z \sim N(0,1)$, satisfy

$$\lim_{\min(B,M,N_1)\to\infty} P^*\left(\delta_{i,t} \in C_{\alpha_{N_1}}(\delta)\right) = 1-\alpha_{N_1}.$$

For the sake of reproducibility, we now describe in detail how we have implemented our tests and the construction of confidence intervals.

In the computation of the statistics (15), we use $s_{N_1} = \ln\ln N_1$, and we estimate σ_δ as suggested in Athey et al. (2019). Tests are always carried out at a family-wise nominal rejection level of $\alpha_{N_1} = \frac{0.05}{N_1}$.

Randomisation is implemented with $M = \left\lfloor N_1^{1/2} \right\rfloor$, and $B = \lfloor M/\ln\ln M \rfloor$. We would like to point out that the choice of M is quite conservative (according to Theorem B.1, M could even be proportional to N_1), which is likely to result in a conservative test, less prone to over-rejecting the null that $\delta_{i,t} = 0$. This, combined with the Bonferroni correction proposed above, should ensure that no spurious detection of significant $\delta_{i,t}$ occurs. We compute the integral in (16) by using a

Gauss-Hermite quadrature with

$$S_{i,t}\left(M, N_1\right) = \sum_{s=1}^{n_S} w_s \varsigma_{i,t}(\sqrt{2} z_s),$$

where the z_ss, $1 \leq s \leq n_S$, are the zeros of the Hermite polynomial $H_{n_S}\left(z\right)$ and the weights w_s are defined as

$$w_s = \frac{2^{n_S - 1}\left(n_S - 1\right)!}{n_S\left[H_{n_S - 1}\left(z_s\right)\right]^2}.$$

Thus, when computing $\varsigma_{i,t}(u)$ in Step 2 of the algorithm, we construct n_S of these statistics, each using $u = \pm\sqrt{2} z_s$. The values of the roots z_s, and of the corresponding weights w_s, are tabulated e.g. in Salzer et al. (1952). In our case, we have used $n_S = 4$, which corresponds to $w_1 = w_4 = 0.05$ and $w_2 = w_3 = 0.45$, and $u_1 = -u_4 = 2.4$ and $u_2 = -u_3 = 0.75$.

Our code is written in GAUSS 21, and random numbers are generated with seed equal to 5^{13}.

C Data Appendix

Table C.1: Variable description.

Variable name	Variable description
Farm characteristics	All farms classified as vineyards according to the European classification (i.e. the standard gross production from vineyards exceeds two thirds of the total)
grape quantity (tons)	Total grape quantity in tons
labor hours	Labor inputs in hours
labor prod	Grape quantity (tons) / labor hours
vineyard land (ha)	Area in ha used to produce grape products (log transformed for estimation)
land prod	Grape quantity / vineyard land (log transformed for estimation)

Continued on next page

Table C.1 – *Continued*

Variable name	Variable description
utilized agricultural area (uaa)	Total utilized agricultural area (uaa) in ha (log transformed for estimation). Does not include areas used for mushrooms, land rented for less than one year, woodland and other farm areas
share of vineyards / total land	Vineyard area in ha / (land uaa + land non uaa)
share land uaa for vine crops	Vineyard area in ha / land uaa
share land uaa for other crops	Non-vineyard area for other field crops in ha (e.g., cereals, olives, vegetables, flowers, other crops) / land uaa
share land uaa owned	utilized agricultural area in owner occupation in ha / land uaa
capital intensity	Machinery in EUR / land uaa (ha)
n. grape types (table, PDO)	Number of grape types produced: 2 (1) if both (either) table grapes and (or) grapes for PDO wines, 0 otherwise
table grapes only	Dummy=1 if farm produces only table grapes
grapes for PDO wine only	Dummy=1 if farm produces only grapes for quality wine (PDO, Protected Designation of Origin)
grapes for other wines only	Dummy=1 if farm produces only grapes for other wines than table wines, PGI and PDO wines
grapes for non-PDO wine only	Dummy=1 if farm produces only grapes for non-PDO wine including table wines, PGI and other wines
organic farming (1/2/3)	Categorical variable for not organic, organic, or converting into organic farming
altitude zone (1/2/3)	Categorical variable for the altitude zone: low (<300 meters), medium (300-600m), or high (>600m)
Weather characteristics	*Monthly and bimonthly means/medians, and deviations thereof*
max temp	Maximum air temperature (°C)
min temp	Minimum air temperature (°C)
temp	Mean air temperature (°C)
wind speed	Mean daily wind speed at 10m (m/s)
precipitation	Sum of precipitation (mm/day)
radiation	Total global radiation (KJ/m2/day)
snow depth	Snow depth (cm)
month pass 10 degree	First month in the year with temperature >10 °C
mean days below 0 degree	Mean number of days with temperature < 0 °C
dev (...)	Deviations of monthly and bimonthly means/medians

Table C.2: Descriptive statistics for treated and untreated farms over 2010-2012.

	Treated farms				Untreated farms			
Variable name	Mean	Min	Max	SD	Mean	Min	Max	SD
Farm characteristics								
grape quantity (tons)	157	0.3	4500	324	76.95	0.1	2836	135.32
labor hours	4170	200	112285	8488	3427	320	58400	3275
labor prod	0.04	2e-04	0.34	0.03	0.02	2.1e-05	0.48	0.02
vineyard land (ha)	11.48	0.51	415	25	6.91	0.01	171.48	10.86
land prod	16.88	0.01	55.15	10.18	11.53	0.02	35.60	5.18
utilized agricultural area (uaa)	0.51	484	36.74	13.08	0.27	358.39	23.15	
share of vineyards / total land	0.65	0.05	1	0.27	0.64	0.01	1.00	0.30
share land uaa for vine crops	0.67	0.09	1	0.27	0.64	0.01	1.00	0.30
share land uaa for other crops	0.48	0	1.63	0.42	0.41	0	1.71	0.34
share land uaa owned	0.93	0	1	0.23	0.70	0	1	0.40
capital intensity	3775	0	37224	4011	7376	0	348204	13348
n. grape types (table, PDO)	0.57	0	2	0.53	0.84	0	2	0.39
table grapes only (0/1)	0.17	0	1	0.37	0.003	0	1	0.06
grapes for PDO wine only (0/1)	0.28	0	1	0.45	0.73	0	1	0.45
grapes for other wines only (0/1)	0.44	0	1	0.50	0.17	0	1	0.38
grapes for non-PDO wine only (0/1)	0.44	0	1	0.50	0.17	0	1	0.38
organic farming (1/2/3)	1.13	1	3	0.40	1.04	1	3	0.21
altitude zone (1/2/3)	1.15	1	3	0.41	1.25	1	3	0.48
Weather characteristics								
mean temp march	11.67	8.97	12.72	0.83	9.65	-0.37	13.55	1.76
mean max temp march	15.72	12.27	17.4	1.15	14.5	2.76	18.97	2.37
mean min temp march	5.16	8.5	0.81	4.81	-3.53	9.88	1.75	
mean precipitation march	2.14	0.94	4.94	0.88	1.84	0.01	5.77	1.2
mean radiation march	14657	12321	16679	1102	13012	10933	17645	1743
mean wind speed march	4.04	2.69	5.41	0.54	2.52	1.39	6.25	0.67
median temp march	11.9	9.2	12.9	0.71	9.97	0.2	13.4	1.64
median max temp march	16.03	12.7	18	1.14	14.92	2.8	19.4	2.31
median min temp march	7.88	5.4	8.6	0.81	4.97	-2.9	9.9	1.79
median precipitation march	0.01	0	0.2	0.04	0	0	0.2	0.01
median radiation march	15249	11783	17987	1368	13644	10395	18014	1734
dev mean wind speed march	3.65	2.1	4.6	0.6	2.17	1.3	5.6	0.61
dev mean temp march	0.07	-0.55	1.56	0.55	0.26	-1.62	3.75	1.28
dev mean max temp march	0.14	-0.79	2.12	0.84	0.29	-2.26	4.86	2.07
dev mean min temp march	-0.01	-1.2	1	0.56	0.23	-1.11	3.24	0.68
dev mean precipitation march	0.09	-1.36	2.35	0.88	-0.12	-2.37	3.23	1.19
dev mean radiation march	313.3	-759.16	2365	916.87	198.77	-1993	3698	1729
dev mean wind speed march	-0.14	-1.74	1.32	0.47	-0.03	-0.92	1.79	0.38
dev mean snow depth march	9.64	-45.69	55.79	32.29	-18.2	-105.46	50.6	24.24
month pass 10 degree	2.26	1	4	0.89	3.49	1	6	0.71
mean days below 0 degree	0	0	0.17	0.02	0.35	0	20.32	1.65
. . . up to about 400 variables								

D Results Appendix

D.1 Migration effects on misreporting at regional level

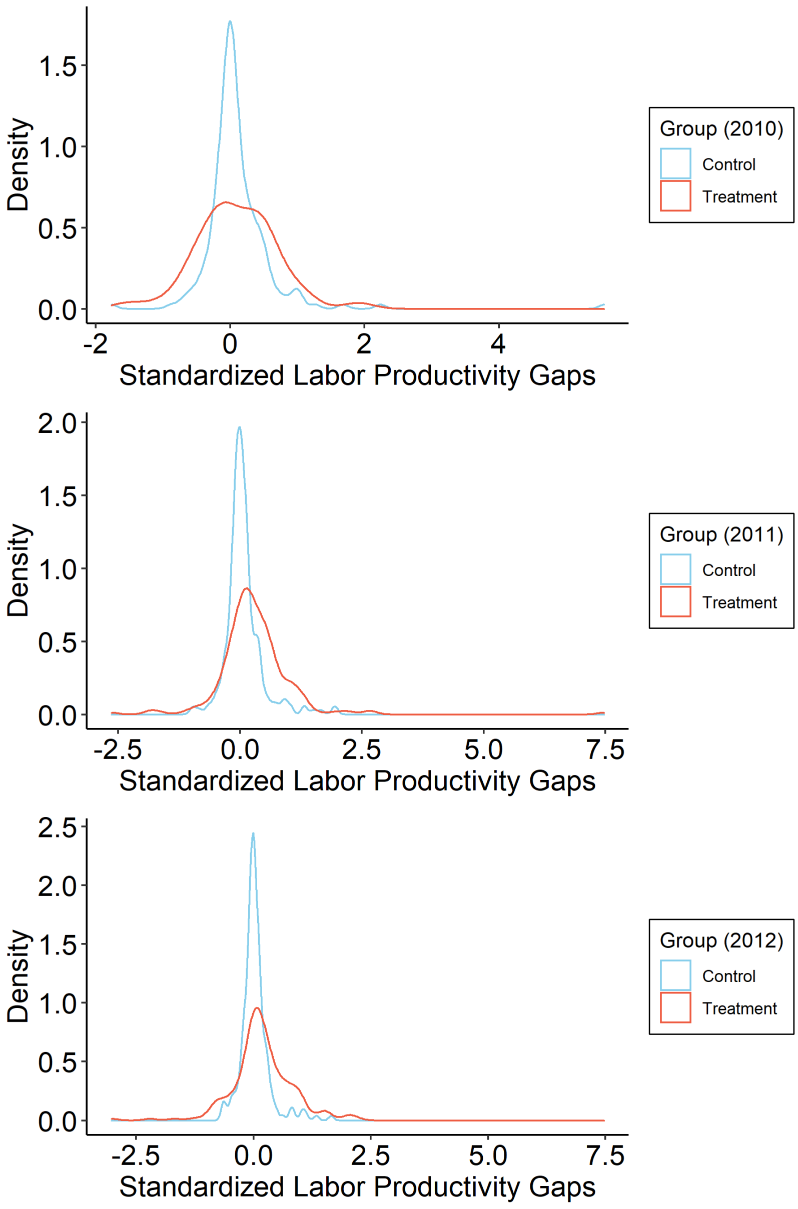
Table D.1: Reported versus predicted labor productivity for non-treated farms (y versus $\hat{y}$).

Year (obs.)	Mean y	Mean $\hat{y}$	K-S test (two-sided)	K-S test (one-sided)	t-test	Wilcoxon test
2010 (n = 188)	0.022	0.019	0.504	0.256	0.134	0.497
2011 (n = 215)	0.022	0.020	0.672	0.351	0.250	0.744
2012 (n = 175)	0.021	0.020	0.692	0.400	0.593	0.967
Notes:	***p=.01; **p=.05; *p=.1					

Table D.2: Mean squared labor productivity gaps estimated for treated and non-treated farms in the test set.

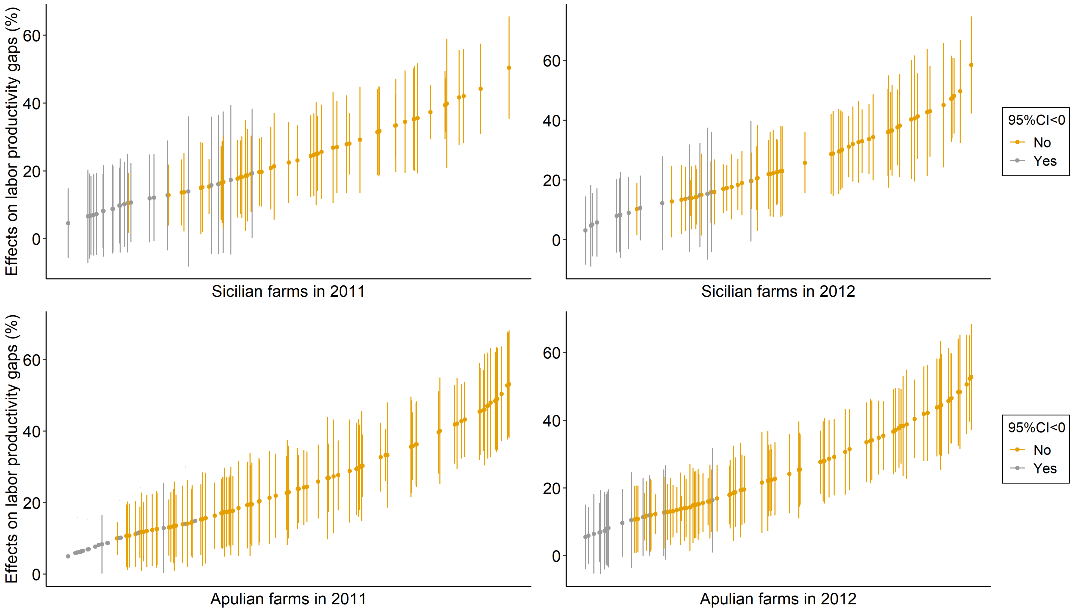
Year	Group (obs.)	MSE
2010	Treated (n = 48)	2.10e-04
	Non-Treated (n = 188)	1.94e-04
2011	Treated (n = 223)	4.23e-04
	Non-Treated (n = 215)	9.33e-05
2012	Treated (n = 202)	2.71e-04
	Non-Treated (n = 175)	5.77e-05

Figure D.1: Labor productivity gaps by treatment group before (2010) and after the shock (2011, 2012). Gaps are standardized, i.e., divided by their standard deviation.



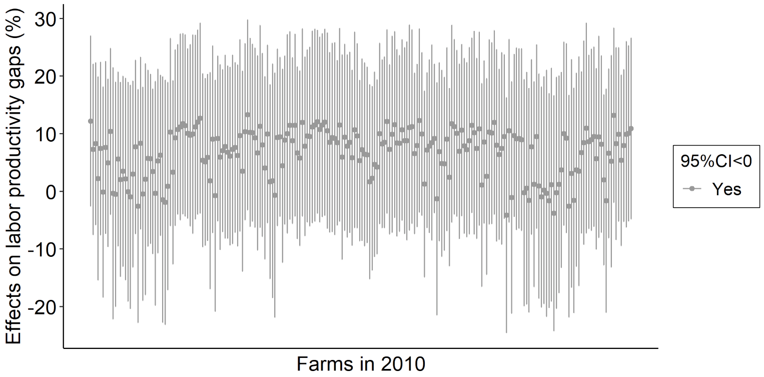
D.2 Migration effects on misreporting at farm level

Figure D.2: Farm level estimates of migration causal effects on labor productivity gaps for farms in Sicily and Apulia over 2011-12. Effects are measured as percent increases in labor productivity gaps post-migration. The length of the error bars is a 95% confidence interval for the point estimates. About 80% of the estimated effects are statistically significant at the 5% level.



Notes. Causal effects estimates for each farm on the x-axis are ranked from the lowest to the largest. The estimated effects are statistically significant for about 80% of farms in Sicily and Apulia (highlighted in orange). Wilcoxon rank-sum test, K-S tests and t-tests show that causal effects estimates on labor productivity gaps do not statistically differ across regions and years.

Figure D.3: Farm level estimates of migration causal effects on labor productivity gaps in Sicily and Apulia in 2010 (placebo test). Effects are measured as percent increases in labor productivity gaps post-migration. The length of the error bars is a 95% confidence interval for the point estimates.



Notes. In 2010 (pre-migration) the estimated effects are small and statistically insignificant for all farms in Sicily and Apulia.

D.3 Back of the envelope calculations

Table D.3: Variable definition to transform labor productivity gap estimates into estimates of the unreported (illegal) hours employed on Sicily's and Apulia's vineyards.

Variable name	Variable description
Q	reported grape quantity
Q^T	true grape quantity
$y(L)$	true labor productivity of legal labor
$y(IL)$	true labor productivity of illegal labor
$h(L)$	reported hours of legal labor
$h(IL)$	estimated hours of illegal labor
$\hat{y}$	predicted labor productivity
y	reported labor productivity

Notes. Total grape output Q^* is a function of labor productivity y and labor input h such that $Q^T = y(L)^*h(L) + y(IL)^*h(IL)$. We observe $Q = y^*h(L)$ and $Q = Q^T$ (no additionally unreported grape quantity). Solving for $h(IL)$ delivers the illegal input as a function of the observed values of output and legal labor input: $h(IL) = \frac{Q-y(L)^*h(L)}{y(IL)}$. We then substitute each element for either its observed or estimated counterpart. First, we substitute for $Q = y^*h(L)$ and factor out $h(L)$. We obtain $h(IL) = \frac{y-y(L)}{y(IL)}h(L)$. Second, as the true labor productivity of legal and illegal input is not observed, we substitute $y(L)$ for the estimated true labor productivity of legal labor input. This is given by predicted labor productivity in absence of the migration shock, $\hat{y}$. Further we assume $y(IL) = y(L)$ in order to obtain estimates of the number of illegal labor hours as $\hat{h}(IL) = \frac{y-\hat{y}}{\hat{y}}h(L)$. Misreporting is defined by statistically significant increases in labor productivity gaps, $\hat{\delta}$. As a result, we estimate significant increases of illegal labor hours post-migration as $\hat{h}(IL) = \frac{\hat{\delta}(y-\hat{y})}{\hat{y}}h(L)$.

D.4 Migration effects on labor hours at farm level

Table D.4: Summary statistics for the estimated number of unreported labor hours for each farm in Sicily and Apulia post-migration. Upper and lower bounds of 95% confidence intervals use standard errors estimated via causal forests (Athey et al., 2019).

			Farm			Sector
	Region	Year	Mean	Median	SD	Total
Point estimates	Apulia	2011	242.4	237.5	163.3	5,049,184.9
	Sicily	2011	164.2	147.1	117.4	1,642,333.6
	Apulia	2012	156.4	113.9	122.6	2,665,125.0
	Sicily	2012	175.9	158.4	116.4	1,651,688.2
Upper bounds	Apulia	2011	388.4	368.0	245.7	7,595,784.3
	Sicily	2011	280.7	242.5	201.0	2,669,199.1
	Apulia	2012	246.4	190.5	186.4	4,068,767.2
	Sicily	2012	290.5	260.9	181.5	2,711,767.7
Lower bounds	Apulia	2011	113.7	93.9	105.1	2,821,602.2
	Sicily	2011	61.8	41.7	69.0	740,855.7
	Apulia	2012	77.1	51.9	71.0	1,444,772.3
	Sicily	2012	75.0	63.2	67.9	728,309.6

Notes. Results come from 214 observations with positive labor productivity gaps over 2011-12. At the sector level, these observations represent about 13,700 farms in Apulia and 11,200 farms in Sicily on average each year.

D.5 Migration effects on farm outcomes

Table D.5: Descriptive statistics of farm outcomes for misreporting farms and non-treated farms. Misreporting farms are defined by statistically significant increases in labor productivity gaps after migration compared to their predicted gaps in absence of migration.

		Before		After	
Obs. 840		**Median**	**Sd**	**Median**	**Sd**
Misreporting farms:	Grape profit/assets	0.05	0.10	0.08	0.18
	Input costs/assets	0.04	0.09	0.04	0.07
	Grape sales/assets	0.09	0.14	0.12	0.23
	Grape prices	285.46	190.66	335.68	173.60
	Hourly wages	6.50	1.37	6.98	2.15
Non-treated farms:	Grape profit/assets	0.05	0.24	0.06	0.56
	Costs/assets	0.02	0.23	0.03	0.80
	Grape sales/assets	0.07	0.41	0.09	0.50
	Grape prices	366.2	300.8	428.9	344.9
	Hourly wages	9.11	3.84	9.61	4.61

Notes. Grape profits are sales minus input costs (labor + crop production). Observations equal 840 (262 treated + 578 non-treated farms).

D.6 Robustness checks

Figure D.4: Placebo tests for treated and non-treated regions. Mean gaps are reported in absolute values. The shaded area shows that, compared to the treated, no region assigned to treatment has a higher gap post-migration (y-axis) as well as a lower gap pre-migration (x-axis).

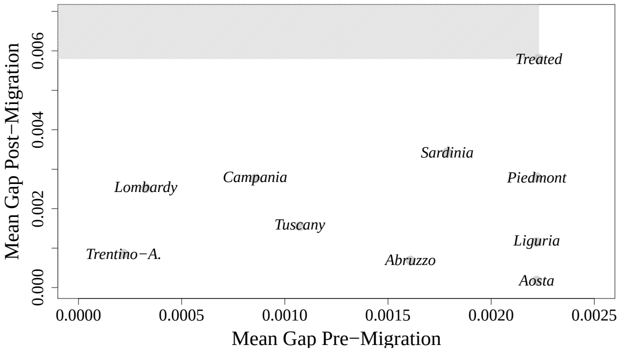


Figure D.5: LOO estimates of labor productivity gaps versus actual estimates before (2010) and after the shock (2011, 2012). Gaps are standardized, i.e., divided by their standard deviation.

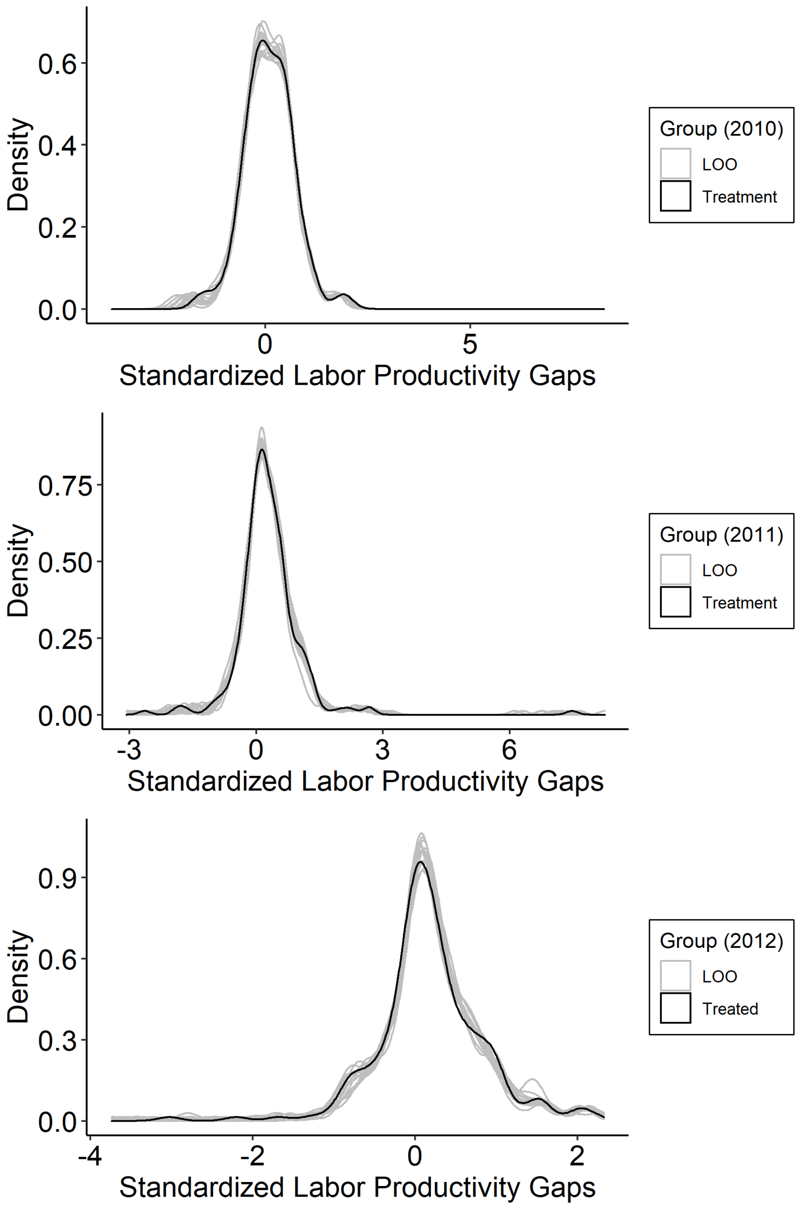


Table D.6: Summary statistics for the estimated number of unreported labor hours for each farm in Sicily and Apulia post-migration. Upper and lower bounds of 95% confidence intervals are estimated via de-randomised inference (as described in Appendix B).

			Farm			Sector
	Region	Year	Mean	Median	SD	Total
Point estimates	Apulia	2011	232.8	216.3	161.8	5,078,300.6
	Sicily	2011	168.4	148.4	115.0	1,718,049.9
	Apulia	2012	154.1	113.9	121.3	2,684,274.8
	Sicily	2012	184.6	159.8	124.4	1,681,282.5
Upper bounds	Apulia	2011	312.9	296.0	201.3	6,626,648.5
	Sicily	2011	268.0	215.9	260.5	2,547,455.2
	Apulia	2012	203.3	156.5	154.9	3,481,070.9
	Sicily	2012	273.4	221.6	223.2	2,442,836.6
Lower bounds	Apulia	2011	137.1	125.8	125.7	3,635,061.9
	Sicily	2011	74.4	56.8	77.5	1,024,942.7
	Apulia	2012	91.2	63.2	89.0	1,869,938.7
	Sicily	2012	94.2	79.3	84.2	1,019,102.2

Notes. Results come from 227 observations with positive labor productivity gaps over 2011-12. At the sector level, these observations represent about 13,900 farms in Apulia and 11,250 farms in Sicily on average each year.

E Proofs and derivations

Proof of Theorem 2.1. The proof is fairly standard. We begin by noting that, by definition

$$E\left(y'_{1,i,t}-\widehat{y}'_{1,i,t}\right)=E\left(\varepsilon_{i,t}\right)+E\left(\mu\left(X_{1,i,t}\right)-\mu'\left(X'_{1,i,t}\right)\right)+E\left(\widehat{\mu}'\left(X'_{1,i,t}\right)-\mu'\left(X'_{1,i,t}\right)\right)+E\left(\delta_{i,t}\right),$$

and

$$E\left(y'_{0,i,t}-\widehat{y}'_{0,i,t}\right)=E\left(\varepsilon_{i,t}\right)+E\left(\mu\left(X_{0,i,t}\right)-\mu'\left(X'_{0,i,t}\right)\right)+E\left(\widehat{\mu}'\left(X'_{0,i,t}\right)-\mu'\left(X'_{0,i,t}\right)\right).$$

Hence

$$\begin{aligned}
& E\left(y'_{1,i,t}-\widehat{y}'_{1,i,t}\right)-E\left(y'_{0,i,t}-\widehat{y}'_{0,i,t}\right)\\
= & E\left(\mu\left(X_{1,i,t}\right)-\mu'\left(X'_{1,i,t}\right)\right)-E\left(\mu\left(X_{0,i,t}\right)-\mu'\left(X'_{0,i,t}\right)\right)\\
& +E\left(\widehat{\mu}'\left(X'_{1,i,t}\right)-\mu'\left(X'_{1,i,t}\right)\right)-E\left(\widehat{\mu}'\left(X'_{0,i,t}\right)-\mu'\left(X'_{0,i,t}\right)\right)\\
& +E\left(\delta_{i,t}\right).
\end{aligned}$$

Using Assumption 2.1*(ii)* and (7), the desired result follows. We also note that, by (6)

$$\begin{aligned}
& E\left(y_{i,t}|D_{i,t}=1\right)-E\left(y_{i,t}|D_{i,t}=0\right) \\
= \;& E\left(\varepsilon_{i,t}|D_{i,t}=1\right)+E\left(\mu\left(X_{1,i,t}\right)-\mu^{\prime}\left(X_{1,i,t}^{\prime}\right)|D_{i,t}=1\right)+E\left(\delta_{i,t}|D_{i,t}=1\right) \\
& -E\left(\varepsilon_{i,t}|D_{i,t}=0\right)-E\left(\mu\left(X_{0,i,t}\right)-\mu^{\prime}\left(X_{0,i,t}^{\prime}\right)|D_{i,t}=0\right) \\
= \;& E\left(\varepsilon_{i,t}\right)+E\left(\mu\left(X_{1,i,t}\right)-\mu^{\prime}\left(X_{1,i,t}^{\prime}\right)\right)+E\left(\delta_{i,t}\right) \\
& -E\left(\varepsilon_{i,t}\right)-E\left(\mu\left(X_{0,i,t}\right)-\mu^{\prime}\left(X_{0,i,t}^{\prime}\right)\right),
\end{aligned}$$

having used Assumption 2.2 in the last passage, and the mean-independence of $\delta_{i,t}$ on $D_{i,t}$ in the previous one. Recalling Assumption 2.1*(ii)*, we finally have

$$E\left(y_{i,t}|D_{i,t}=1\right)-E\left(y_{i,t}|D_{i,t}=0\right)=E\left(\delta_{i,t}\right), \tag{21}$$

which proves that $E\left(\delta_{i,t}\right)$ is the ATE.

□

Proof of Theorem B.1. The proof is the same as that of Theorems 3 and 4 in Horváth and Trapani (2019), with one major technical difference: whilst in Horváth and Trapani (2019) the sequence to be randomised diverges under the null and drifts to zero under the alternative *almost surely*, in our case we have a weaker result.

We let $S_{i,t}\left(N_1,M\right)=S_{N_1,M}$ and $\exp\left(-\frac{1}{2}u^2\right)=F\left(u\right)$ for short, and denote the distribution function of the standard normal as $G\left(\cdot\right)$. We begin by studying the behaviour of $S_{N_1,M}$ under the null. The CLT in Theorem 5 in Athey et al. (2019) entails that

$$\liminf_{N_1\rightarrow\infty} P\left(\left(\frac{N_1}{m}\right)^{\epsilon^{\prime}-1/2} s_{N_1}\left|\widehat{\delta}_{i,t}-\delta_{i,t}^{0}\right|=0\right)=1, \tag{22}$$

for all $\epsilon^{\prime}<\varepsilon$. By elementary arguments, this entails that

$$\liminf_{N_1\rightarrow\infty} P\left(\exp\left(-\left(\frac{N_1}{m}\right)^{\varepsilon+\epsilon^{\prime}}\right)\phi_{i,t}\left(N_1\right)=0\right)=1. \tag{23}$$

We now have

$$\begin{aligned}
& \frac{1}{\sqrt{2\pi}}\left|\int_{-\infty}^{\infty}\left|\frac{2}{M^{1/2}}\sum_{m=1}^{M}\left(I\left(\xi_{i,t,m}\left(u\right)\right)-\frac{1}{2}\right)\right|^{2}dF\left(u\right)\right. \\
& \left.-\int_{-\infty}^{\infty}\left|\frac{2}{M^{1/2}}\sum_{m=1}^{M}\left(I\left(\xi_{i,t,m}\left(0\right)\right)-\frac{1}{2}\right)\right|^{2}dF\left(u\right)\right| \\
\leq \; & c_0\int_{-\infty}^{\infty}\left|\frac{2}{M^{1/2}}\sum_{m=1}^{M}\left(G\left(\frac{u}{\phi_{i,t}\left(N_1\right)}\right)-\frac{1}{2}\right)\right|^{2}dF\left(u\right) \\
& +c_0\int_{-\infty}^{\infty}\left|\frac{2}{M^{1/2}}\sum_{m=1}^{M}\left(\left(I\left(\xi_{i,t,m}\left(u\right)\right)-I\left(\xi_{i,t,m}\left(0\right)\right)\right)-\left(G\left(\frac{u}{\phi_{i,t}\left(N_1\right)}\right)-\frac{1}{2}\right)\right)\right|^{2}dF\left(u\right) \\
= \; & I+II.
\end{aligned}$$

It holds that

$$\begin{aligned}
I & \leq c_1 M\int_{-\infty}^{\infty}\left|\left(G\left(\frac{u}{\phi_{i,t}\left(N_1\right)}\right)-\frac{1}{2}\right)\right|^{2}dF\left(u\right) \\
& \leq c_1 m_G\frac{M}{\phi_{i,t}^{2}\left(N_1\right)}\int_{-\infty}^{\infty}u^{2}dF\left(u\right)
\end{aligned}$$

using the fact that $G\left(x\right)\leq m_G<\infty$ for all $-\infty<x<\infty$. Now (23) immediately entails that

$$\liminf_{M,N_1\to\infty}P\left(\frac{M}{\phi_{i,t}^{2}\left(N_1\right)}=0\right)=1. \tag{24}$$

Similarly, we note that the random variable

$$\left(I\left(\xi_{i,t,m}\left(u\right)\right)-I\left(\xi_{i,t,m}\left(0\right)\right)\right)-\left(G\left(\frac{u}{\phi_{i,t}\left(N_1\right)}\right)-\frac{1}{2}\right)$$

has variance

$$\left(G\left(\frac{u}{\phi_{i,t}\left(N_1\right)}\right)-\frac{1}{2}\right)\left[1-\left(G\left(\frac{u}{\phi_{i,t}\left(N_1\right)}\right)-\frac{1}{2}\right)\right]\leq\left|G\left(\frac{u}{\phi_{i,t}\left(N_1\right)}\right)-\frac{1}{2}\right|,$$

so that ultimately

$$\begin{aligned}
& E^{*}\int_{-\infty}^{\infty}\left|\frac{2}{M^{1/2}}\sum_{m=1}^{M}\left(\left(I\left(\xi_{i,t,m}\left(u\right)\right)-I\left(\xi_{i,t,m}\left(0\right)\right)\right)-\left(G\left(\frac{u}{\phi_{i,t}\left(N_{1}\right)}\right)-\frac{1}{2}\right)\right)\right|^{2}dF\left(u\right) \\
\leq & \int_{-\infty}^{\infty}\left|G\left(\frac{u}{\phi_{i,t}\left(N_{1}\right)}\right)-\frac{1}{2}\right|dF\left(u\right) \\
\leq & c_{0}\frac{m_{G}}{\phi_{i,t}\left(N_{1}\right)}\int_{-\infty}^{\infty}\left|u\right|dF\left(u\right)=o_{P^{*}}\left(1\right),
\end{aligned}$$

on account of (23). Putting all together we receive

$$\begin{aligned}
& \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left|\frac{2}{M^{1/2}}\sum_{m=1}^{M}\left(I\left(\xi_{i,t,m}\left(u\right)\right)-\frac{1}{2}\right)\right|^{2}dF\left(u\right) \\
= & \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left|\frac{2}{M^{1/2}}\sum_{m=1}^{M}\left(I\left(\xi_{i,t,m}\left(0\right)\right)-\frac{1}{2}\right)\right|^{2}dF\left(u\right)+o_{P^{*}}\left(1\right),
\end{aligned}$$

and the desired result now follows from the CLT for Bernoulli random variables.

Under the alternative, it is easy to see that

$$\liminf_{N_{1}\rightarrow\infty}P\left(\left(\frac{N_{1}}{m}\right)^{\epsilon'-1/2}s_{N_{1}}\left|\widehat{\delta}_{i,t}-\delta_{i,t}^{0}\right|=\infty\right)=1, \tag{25}$$

for all $\epsilon'<\varepsilon$, and therefore

$$\liminf_{N_{1}\rightarrow\infty}P\left(\exp\left(-\left(\frac{N_{1}}{m}\right)^{\varepsilon+\epsilon'}\right)\phi_{i,t}\left(N_{1}\right)=1\right)=1. \tag{26}$$

Now we write

$$I\left(\xi_{i,t,m}\left(u\right)\right)-\frac{1}{2}=I\left(\xi_{i,t,m}\left(u\right)\right)-G\left(u\right)+G\left(u\right)-\frac{1}{2},$$

and we therefore have

$$\begin{aligned}
& E^{*}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left|\frac{2}{M^{1/2}}\sum_{m=1}^{M}\left(I\left(\xi_{i,t,m}\left(u\right)\right)-\frac{1}{2}\right)\right|^{2}dF\left(u\right) \\
= & E^{*}\left(I\left(\xi_{i,t,m}\left(u\right)\right)-G\left(u\right)\right)^{2}+M\left(G\left(u\right)-\frac{1}{2}\right)^{2}.
\end{aligned}$$

Since $E^{*}\left(I\left(\xi_{i,t,m}\left(u\right)\right)-G\left(u\right)\right)^{2}<\infty$, by the Markov inequality it follows that

$$\int_{-\infty}^{\infty}\left[M^{-1/2}\sum_{m=1}^{M}\left(I\left(\xi_{i,t,m}\left(u\right)\right)-G\left(u\right)\right)\right]^{2}dF(u)=O_{P^{*}}(1),$$

with probability going to 1. Hence the proof of Theorem B.1 is complete. □

F Algorithms

The validity of our analysis depends on the prediction quality of the super learning model, which is partly determined by the selection of base learners. Therefore, we choose learners that have demonstrated good prediction performance across various settings and prediction competitions. To leverage the benefits of combining multiple algorithms, we select learners with different characteristics, as outlined below. This diversification allows us to capture a wide range of potential underlying data generating processes. The selected algorithms include random forests (RF), extreme gradient boosting (xgboost), penalized regression (glmnet), neural nets (NN), and support vector machines (SVMs).

While we choose the input variables based on economic theory, algorithms such as RF and xgboost perform well regardless of the form of the production function, as they can account for non-linearities and interactions between input variables. In the following, we provide a brief description of the selected methods.

We use random forests (RF) based on regression trees, which are non-parametric supervised methods that estimate regression functions (Breiman, 2001). Regression trees are grown by recursively splitting the covariate space to minimize the distance between the values of dependent variables (targets) within each resulting split (Breiman, 2017). This approach is well-suited to model complex interactions, including non-linear functions and interaction effects (Hastie et al., 2009). It allows us to remain agnostic about the specific relationship between labor productivity and covariates.[37] To prevent overfitting, RF incorporates mechanisms such as random subsampling of data and covariates, out-of-bag estimation, and parameter tuning. We estimate the forests using the randomForest R package.

[37] For example, the effect of fertilizer on labor productivity may vary conditional on precipitation. Trees can model such non-linearities if they are present in the data without the need to specify them a priori.

Similarly to random forests, xgboost is a tree-based method. However, instead of minimizing variance within leaves, xgboost aims to maximize the similarity scores within leaves. Specifically, each recursive split in xgboost aims to maximize the gain in similarity scores, grouping similar observations in the same leaves. Xgboost builds on multiple trees consecutively, with each new tree reducing residuals from previous trees. This allows us to replace the initial target (labor productivity) with parts of the target that were not predicted by previous trees. These algorithms offer properties that facilitate quick data processing and handling of missing data. More details can be found in Chen and Guestrin (2016). We utilize the R package xgboost to estimate this base learner.

For penalized regression estimation, we employ the glmnet package. This package fits a generalized linear model using regularization. The computation is expedited through coordinate descent. Friedman et al. (2010) provide further details on the procedure. Regularization penalizes the size of coefficients, leading to a shrunk vector of coefficients resulting from the regression optimization problem. We opt for regularized regression due to the relatively large covariate space, where standard linear regression is prone to collinearities and overfitting. The glmnet package provides options for lasso regularization (Tibshirani, 1996), ridge regression (Hoerl and Kennard, 2000), and elastic net, which combines both approaches (Zou and Hastie, 2005). While ridge regression is a continuous shrinkage method, the lasso performs variable selection, which means it sets individual coefficients of predictors that are collinear or do not (substantially) help to predict the target to zero. We perform cross-validation to determine the appropriate strength of penalization for improved prediction accuracy. A priori, we do not know which method performs best, and we test that empirically. We select the lasso, as it outperforms the other methods.

Neural networks (NN) model dependent variables as non-linear functions of linear combinations of independent variables (predictors). Back-propagation is employed to re-estimate network parameters and enhance prediction accuracy (Rumelhart et al., 1986). By leveraging their ability to learn from data, NN can approximate and generalize complex functions with high accuracy. To mitigate overfitting, we penalize (regularize) the weights assigned in the network, similar in spirit to the regularization in glmnet. Hastie et al. (2009) provide a comprehensive treatment of NN. We implement this method using the nnet package in R.

Support vector machines (SVMs) are another type of non-parametric machine learning algorithm that can effectively approximate complex functions. While they differ in architecture and operation from neural networks, SVMs are capable of learning and representing nonlinear relationships between inputs and outputs. While neural networks have gained popularity for their ability to handle large-scale and highly complex problems, SVMs have been widely used and studied for their strong theoretical foundations and ability to handle various data types. They are known for their generalization capabilities and robustness to overfitting. More specifically, SVMs work by mapping the input data (predictors) into a high-dimensional feature space and finding an optimal hyperplane that maximally groups similar targets (labor productivity) into the same subspaces. Predictions are made based on the separation achieved by the hyperplane. SVMs can utilize transformations of the covariates to achieve this hyperplane separation. SVMs were introduced for regression by Drucker et al. (1996). For a comprehensive overview, refer to Wang (2005).[38]

In summary, we employ a diverse set of estimators that are expected to yield accurate predictions while accommodating different assumptions about the underlying data generating process.

[38] Scaling the data can affect the solution obtained by support vector machines (e.g., Hastie et al. 2009). However, in our case, we do not employ data scaling as we use the same variables for the estimation of all base learners and the SL.

University of Innsbruck - Working Papers in Economics and Statistics
Recent Papers can be accessed on the following webpage:

https://www.uibk.ac.at/eeecon/wopec/

2023-09 **Marica Valente, Timm Gries, Lorenzo Trapani:** Informal employment from migration shocks

2023-08 **Natalie Struwe, Esther Blanco, James M. Walker:** No response to changes in marginal incentives in one-shot public good experiments

2023-07 **Sebastian Bachler, Sarah Lynn Flecke, Jürgen Huber, Michael Kirchler, Rene Schwaiger:** Carbon Pricing, Carbon Dividends and Cooperation: Experimental Evidence

2023-06 **Elisabeth Gsottbauer, Michael Kirchler, and Christian König-Kersting:** Climate Crisis Attitudes among Financial Professionals and Climate Experts

2023-05 **Xiaogeng Xu, Satu Metsälampi, Michael Kirchler, Kaisa Kotakorpi, Peter Hans Matthews, Topi Miettinen:** Which income comparisons matter to people, and how? Evidence from a large field experiment

2023-04 **Tanja Hoertnagl, Rudolf Kerschbamer, Regine Oexl, Rudi Stracke, and Uwe Sunde:** Heterogeneity in Rent-Seeking Contests with Multiple Stages: Theory and Experimental Evidence

2023-03 **Melissa Newham, Marica Valente:** The Cost of Influence: How Gifts to Physicians Shape Prescriptions and Drug Costs

2023-02 **Natalie Struwe, Esther Blanco, James M. Walker:** Determinants of Financial Literacy and Behavioral Bias among Adolescents

2023-01 **Marco Aschenwald, Armando Holzknecht, Michael Kirchler, Michael Razen:** Determinants of Financial Literacy and Behavioral Bias among Adolescents

2022-20 **Silvia Angerer, Daniela Glätzle-Rützler, Philipp Lergetporer, and Thomas Rittmannsberger:** Beliefs about social norms and (the polarization of) COVID-19 vaccination readiness

2022-19 **Edward I. Altman, Marco Balzano, Alessandro Giannozzi, Stjepan Srhoj:** Revisiting SME default predictors: The Omega Score

2022-18 **Johannes Diederich, Raphael Epperson, Timo Goeschl:** How to Design the Ask? Funding Units vs. Giving Money

2022-17 **Toman Barsbai, Vojtěch Bartoš, Victoria Licuanan, Andreas Steinmayr, Erwin Tiongson, and Dean Yang:** Picture This: Social Distance and the Mistreatment of Migrant Workers

2022-16 **Andreas Steinmayr, Manuel Rossi:** Vaccine-skeptic physicians and COVID-19 vaccination rates

2022-15 **Stjepan Srhoj, Alex Coad, Janette Walde:** HGX: The Anatomy of High Growth Exporters

2022-14 **Martin Obradovits, Philipp Plaickner** Price-Directed Search, Product Differentiation and Competition

2022-13 **Utz Weitzel, Michael Kirchler** The Banker's Oath And Financial Advice

2022-12 **Julian Granna, Wolfgan Brunauer, Stefan Lang:** Proposing a global model to manage the bias-variance tradeoff in the context of hedonic house price models

2022-11 **Christoph Baumgartner, Stjepan Srhoj and Janette Walde:** Harmonization of product classifications: A consistent time series of economic trade activities

2022-10 **Katharina Momsen, Markus Ohndorf:** Seller Opportunism in Credence Good Markets ? The Role of Market Conditions

2022-09 **Christoph Huber, Michael Kirchler:** Experiments in Finance ? A Survey of Historical Trends

2022-08 **Tri Vi Dang, Xiaoxi Liu, Florian Morath:** Taxation, Information Acquisition, and Trade in Decentralized Markets: Theory and Test

2022-07 **Christoph Huber, Christian König-Kersting:** Experimenting with Financial Professionals

2022-06 **Martin Gächter, Martin Geiger, Elias Hasler:** On the structural determinants of growth-at-risk

2022-05 **Katharina Momsen, Sebastian O. Schneider:** Motivated Reasoning, Information Avoidance, and Default Bias

2022-04 **Silvia Angerer, Daniela Glätzle-Rützler, Philipp Lergetporer, Thomas Rittmannsberger:** How does the vaccine approval procedure affect COVID-19 vaccination intentions?

2022-03 **Robert Böhm, Cornelia Betsch, Yana Litovsky, Philipp Sprengholz, Noel Brewer, Gretchen Chapman, Julie Leask, George Loewenstein, Martha Scherzer, Cass R. Sunstein, Michael Kirchler:** Crowdsourcing interventions to promote uptake of COVID-19 booster vaccines

2022-02 **Matthias Stefan, Martin Holmén, Felix Holzmeister, Michael Kirchler, Erik Wengström:** You can't always get what you want–An experiment on finance professionals' decisions for others

2022-01 **Toman Barsbai, Andreas Steinmayr, Christoph Winter:** Immigrating into a Recession: Evidence from Family Migrants to the U.S.

2021-32 **Fanny Dellinger:** Housing Support Policies and Refugees' Labor Market Integration in Austria

2021-31 **Albert J. Menkveld, Anna Dreber, Felix Holzmeister, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Sebastian Neusüss, Michael Razen, Utz Weitzel and et al:** Non-Standard Errors

2021-30 **Toman Barsbai, Victoria Licuanan, Andreas Steinmayr, Erwin Tiongson, Dean Yang:** Information and Immigrant Settlement

2021-29 **Natalie Struwe, Esther Blanco, James M. Walker:** Competition Among Public Good Providers for Donor Rewards

2021-28 **Stjepan Srhoj, Melko Dragojević:** Public procurement and supplier job creation: Insights from auctions

2021-27 **Rudolf Kerschbamer, Regine Oexl:** The effect of random shocks on reciprocal behavior in dynamic principal-agent settings

2021-26 **Glenn E. Dutcher, Regine Oexl, Dmitry Ryvkin, Tim Salmon:** Competitive versus cooperative incentives in team production with heterogeneous agents

2021-25 **Anita Gantner, Regine Oexl:** Respecting Entitlements in Legislative Bargaining - A Matter of Preference or Necessity?

2021-24 **Silvia Angerer, E. Glenn Dutcher, Daniela Glätzle-Rützler, Philipp Lergetporer, Matthias Sutter:** The formation of risk preferences throughsmall-scale events

2021-23 **Stjepan Srhoj, Dejan Kovač, Jacob N. Shapiro, Randall K. Filer:** The Impact of Delay: Evidence from Formal Out-of-Court Restructuring

2021-22 **Octavio Fernández-Amador, Joseph F. Francois, Doris A. Oberdabernig, Patrick Tomberger:** Energy footprints and the international trade network: A new dataset. Is the European Union doing it better?

2021-21 **Felix Holzmeister, Jürgen Huber, Michael Kirchler, Rene Schwaiger:** Nudging Debtors to Pay Their Debt: Two Randomized Controlled Trials

2021-20 **Daniel Müller, Elisabeth Gsottbauer:** Why Do People Demand Rent Control?

2021-19 **Alexandra Baier, Loukas Balafoutas, Tarek Jaber-Lopez:** Ostracism and Theft in Heterogeneous Groups

2021-18 **Zvonimir Bašić, Parampreet C. Bindra, Daniela Glätzle-Rützler, Angelo Romano, Matthias Sutter, Claudia Zoller:** The roots of cooperation

2021-17 **Silvia Angerer, Jana Bolvashenkova, Daniela Glätzle-Rützler, Philipp Lergetporer, Matthias Sutter:** Children's patience and school-track choices several years later: Linking experimental and field data

2021-16 **Daniel Gründler, Eric Mayer, Johann Scharler:** Monetary Policy Announcements, Information Schocks, and Exchange Rate Dynamics

2021-15 **Sebastian Bachler, Felix Holzmeister, Michael Razen, Matthias Stefan:** The Impact of Presentation Format and Choice Architecture on Portfolio Allocations: Experimental Evidence

2021-14 **Jeppe Christoffersen, Felix Holzmeister, Thomas Plenborg:** What is Risk to Managers?

2021-13 **Silvia Angerer, Daniela Glätzle-Rützler, Christian Waibel:** Trust in health care credence goods: Experimental evidence on framing andsubject pool effects

2021-12 **Rene Schwaiger, Laura Hueber:** Do MTurkers Exhibit Myopic Loss Aversion?

2021-11 **Felix Holzmeister, Christoph Huber, Stefan Palan:** A Critical Perspective on the Conceptualization of Risk in Behavioral and Experimental Finance

2021-10 **Michael Razen, Alexander Kupfer:** Can increased tax transparency curb corporate tax avoidance?

2021-09 **Changxia Ke, Florian Morath, Anthony Newell, Lionel Page:** Too big to prevail: The paradox of power in coalition formation

2021-08 **Marco Haan, Pim Heijnen, Martin Obradovits:** Competition with List Prices

2021-07 **Martin Dufwenberg, Olof Johansson-Stenman, Michael Kirchler, Florian Lindner, Rene Schwaiger:** Mean Markets or Kind Commerce?

2021-06 **Christoph Huber, Jürgen Huber, and Michael Kirchler:** Volatility Shocks and Investment Behavior

2021-05 **Max Breitenlechner, Georgios Georgiadis, Ben Schumann:** What goes around comes around: How large are spillbacks from US monetary policy?

2021-04 **Utz Weitzel, Michael Kirchler:** The Banker's Oath And Financial Advice

2021-03 **Martin Holmen, Felix Holzmeister, Michael Kirchler, Matthias Stefan, Erik Wengström:** Economic Preferences and Personality Traits Among Finance Professionals and the General Population

2021-02 **Christian König-Kersting:** On the Robustness of Social Norm Elicitation

2021-01 **Laura Hueber, Rene Schwaiger:** Debiasing Through Experience Sampling: The Case of Myopic Loss Aversion.

University of Innsbruck

Working Papers in Economics and Statistics

2023-09

Marica Valente, Timm Gries, Lorenzo Trapani

Informal employment from migration shocks

Abstract
We propose a new approach to detect and quantify informal employment resulting from irregular migration shocks. Focusing on a largely informal sector, agriculture, and on the exogenous variation from the Arab Spring wave on southern Italian coasts, we use machine-learning techniques to document abnormal increases in reported (vs. predicted) labor productivity on vineyards hit by the shock. Misreporting is largely heterogeneous across farms depending e.g. on size and grape quality. The shock resulted in a 6 % increase in informal employment, equivalent to one undeclared worker for every three farms on average and 23,000 workers in total over 2011-2012. Misreporting causes significant increases in farm profits through lower labor costs, while having no impact on grape sales, prices, or wages of formal workers.

ISSN 1993-4378 (Print)
ISSN 1993-6885 (Online)