Using Synthetic Farm Data to Estimate Individual Nitrate Leaching Levels

Konstantinos Mattas, [*] Michail Tsagris[†] and Vangelis Tzouvelekas[†]

February 6, 2024

Abstract

The paper develops a novel synthetic population generation scheme to deal with the NPS pollution problem of nitrate leaching from agricultural farms. The scheme relies upon estimation of the joint distribution of the variables using Bayesian network learning which, coupled with the use of non-parametric regression models facilitate the generation of realistic synthetic populations. Then building upon the sequential GME model suggested by Kaplan et al. (2003) in line with the multiple production relations model suggested by Murty et al. (2012) we obtain econometric estimates of both the production technology and nature's residual generating mechanism for the synthetic population of farms. These estimates are used to proxy a reliable optimal taxation scheme that corresponds to local environmental and economic conditions. The methodology is applied to the Greek FADN dataset for the Greek NUTS II region of Thessaly during the 2017-18 cropping year.

Keywords: nitrate leaching, multiple production relations, Generalized Maximum Entropy, synthetic population generation, Bayesian network learning

JEL Codes: C40, Q53, Q24, Q25.

[*]Department of Agricultural Economics, School of Agriculture, Aristotle University of Thessaloniki, Greece

[†]Department of Economics, School of Social Sciences, University of Crete, Greece. An earlier version of the paper has been developed within the context of the Research Project Agricore financed by the European Commission during the 2020-23 period.

Corresponding author: Vangelis Tzouvelekas, e-mail: v.tzouvelekas@uoc.gr.

Introduction

Modern agricultural practices have dramatically increased crop production, but have also put significant pressure on both groundwater and surface water pollution owing to reactive nitrogen. Since 1970, reactive nitrogen creation has increased considerably, largely driven by increased inorganic fertilizer application to meet growing global demand for agricultural commodities (Galloway et al., 2008).[1] Although nitrate concentrations have slightly decreased over the past decades in some European reservoirs, levels have remained high in others and, overall, nitrate levels in water stock have remained constant. Although some improvements have been made in reducing nutrient inputs from wastewater discharge, diffuse pollution of agricultural origin remains a major threat for water in the EU. From 2000 to 2016, nearly 40 per cent of the groundwater monitoring stations in the EU exceeded average values of 25mg NO_3^- l^{-1} and almost 50 per cent of the surface water monitoring stations exceeded average values of 10mg NO_3^- l^{-1}. Similar high levels occur in other parts of the world where high levels of chemical fertilizers are used (Eurostat, 2018).

Effectively managing the tradeoffs inherent in nitrogen use requires information on the true marginal benefits and costs to both private farmers and society. The benefits of nitrogen-based fertilizer application, measured in terms of improved crop yields, are easily quantified based on the market value of crop production. In contrast, the social costs of nitrogen are not captured in market prices for fertilizer or agricultural commodities and are incurred primarily by the public downwind or downstream of agricultural activities. Accounting for these costs in policies, payment schemes, or programs designed to influence land management, offers the potential to mitigate these tradeoffs and substantially improve environmental and social outcomes, especially in agriculturally dominated landscapes. However, translating environmental changes to damage costs requires an integrated approach that links specific interventions with the cascade of nitrogen related damages over space and time. Recent studies have attempted to fill this gap by monetizing nitrogen related damages for the EU (Brink et al., 2011), the US (Sobota et al., 2015), and China (Gu et al., 2015). These studies effectively highlight the potential magnitude of nitrogen damages and the urgent need to improve nitrogen cost accounting.

However, nitrate pollution is a typical non-source pollution (NPS) problem, as only the ambient concentration of nitrogen is observed, posing serious challenges in policy formation and regulation even when using the correct cost accounting. The main reasons are informational asymmetries between the regulator and the individual farmers, along with the coexisting uncertainty related to farm technologies and natural conditions. In policy formulation, these informational asymmetries induce moral hazard and adverse selection problems. Under moral hazard, as monitoring and

[1] Specifically, the amount of nitrogen in the environment has been increased by more than 100% above preindustrial levels, whereas on the other hand, CO_2 emissions have increased approximately by 40% (Keeler et al., 2016).

measurement of individual nitrate emissions is not possible, farmers can always increase their profits by choosing higher than the socially desirable nitrogen emissions levels. On the other hand, under adverse selection, individual farmers may have incentives not to reveal their specific characteristics or farming type to the regulator if this is profitable for them.[2] As the empirical evidence worldwide reveals, in such situations the standard environmental policy instruments cannot be used to internalize external damages or to obtain the *Pareto* optimal outcome. This failure had resulted in increasing effort to develop policy schemes appropriate for such problems. Recently, the focus of applied research is on the possibility of measuring individual emissions by applying either monitoring technologies or conventional econometric tools to estimate individual emissions from farm-level data in order to use standard policy instruments to regulate NPS pollution to some, or even to a full, extent.[3]

Nevertheless, an important limitation of these individual assessments is that one needs the population of farm operations in each specific region to approximate more accurate individual leaching levels. Using data from a representative sample of farmers, researchers and policy makers aim to approximate mean leaching levels that then can be applied to introduce conventional policy instruments for the population of farmers. However, as noted by several authors nitrogen is lost to aquatic, regional atmospheric, and global atmospheric pools in a variety of forms. These loss pathways are associated with damages that occur over heterogeneous spatial and temporal scales (Erisman et al., 2013). Valuing these damages requires tracking several forms of nitrogen across space to endpoints where the environment or the water resources are impacted. The mean value used to impose taxes or levies in several countries and regions across the globe will turn to a non-optimal outcome which intensifies the problem in certain occasions.

A promising and cost-effective alternative to proxy more accurately individual leaching levels is to use the existing farm surveys to construct a synthetic population of farmers covering the whole region under study incorporating all alternative farming techniques and simulating all possible natural conditions. Generally speaking, the task of synthetic population generation (SPG) refers to the process of generating a synthetic dataset that mimics the true population. In other words, it facilitates the inductive passage from the sample to the population, assuming the observed sample is quasi-representative of the true population. With an accurate process the researchers construct synthetic farm data that mimics the true population and hence provide more solid evidence to support their research claims. In addition, using a synthetic population of farmers ambient nitrate concentrations in water reservoirs are not neglected in the estimation of individual leaching levels as they reflect actually local climatic, environmental and hydrological conditions that determine

[2] For a more detailed discussion on these issues see Segerson (1988) and Shortle and Horan (2001).

[3] Xepapadeas (2011) provides a thorough review of all approaches developed so far to deal with NPS pollution problems.

the extent of externality for rural and urban dwellers leading to more optimal policy interventions.

Along these lines, our aim in this paper is to use synthetic farm data to estimate both crop technology and nature's nitrate residual generating mechanism to approximate individual nitrate leaching levels in the Greek NUTS II region of Thessaly during the 2017-18 cropping year. Our analysis is based on the FADN database that accurately collects individual farm data for the last 30 years across the EU. Then following Young et al. (2009) and Sun and Erath (2015), we use a Bayesian network learning approach taking into account the conditional distribution of input and output variables to formulate a consistent way to generate a synthetic population of farmers. The developed Bayesian network is coupled with the use of non-parametric regression models to facilitate the generation of synthetic population that mimic the observed ones to a high degree. Then, building upon the sequential GME approach suggested by Kaplan et al. (2003) and Farzin and Kaplan (2004) we impose into a generalized entropy filter a specific theoretical structure describing both crop production technology and nature's nitrogen residual generating mechanism. Finally, the theoretical model is based on the multiple production relations model developed by Murty et al. (2012) that identifies appropriately the features of by-production of pollution in intended output production activities which is adapted for nitrate leaching occasions in intensive crop production.

The remainder of the paper is structured as follows. Next section discusses the various approaches used in synthetic population generation process and presents the use of Bayesian learning networks. The following section presents the farm technology taking into account individual emissions arising from chemical fertilizer use, while section 4 presents the GME estimator applied. Section 5 discusses the practical problems and presents the econometric estimates, while the last section discusses policy implications and concludes the paper.

The Synthetic Population Generation Process

The current SPG techniques[4] can be divided into two broad categories, namely, *synthetic reconstruction* (SR) and *combinatorial optimization* (CO). The former approach first generates new observations on a set of variables and then exploiting those generated observations it fills the gaps and proceed to generate values for the next set of variables. This sequential process is continued until all population values are filled. Corner-stone to this process are some true population constraints, usually taken from available census data. A standard algorithm for this technique is the *Iterative Proportional Fitting* (IPF) which requires prior knowledge on the joint (bivariate) distribution of any two variables.[5] On the other hand, CO uses the publicly available data and samples

[4] Chapuis and Taillandier (2019) and Ramadan and Sisiopiku (2019) provide a brief review of the recent SPG methods.

[5] In the case of more than two variables, IPF considers pairs of variables conditioning on the values of the other variables.

from them (with replacement) until the value of a stress criterion is minimized. Similarly to SR, a list of constraints must also be satisfied (related to the stress criterion), but their difference lies in their generated output. In summary, SR simulates new values, whereas CO reproduces different combinations of the observed values. SR proceeds in a hierarchical fashion, simulating variables with a specific order, whereas CO uses all variables in an iterative process. CO starts from a randomly chosen set of observations it replaces an observation with a new one if the fit is improved, until the fit can not be further improved. The SR approach is evidently faster, but CO can yield a synthetic population that better fits some known constrains of the true population.

Ye et al. (2009) generalized the IPF to the *Iterative Proportional Updating* (IPU) algorithm in order to better capture the overall joint distribution of the variables solving the zero-cell and zero-marginal problems. However, in extreme cases such as when all individuals of certain types completely fall into a single observational type, IPU fails to converge. Further, IPU may reach to a solution that lies is outside the feasible region.[6] Gargiulo et al. (2010), on the other hand, proposed an iterative approach to generate statistically realistic populations matching few variables only. Their approach can be extended to the complete set of variables, but the order of generating these variables remains unknown. The advantage of their approach is that they consider no sample data, only the tabular information, which however, does not take into account the relationship among the variables.[7] The drawback of both approaches though is that the synthetic population is calibrated against some known tabular information which in practice may not be representative of the characteristics of the true population.

On a completely different direction, Casati et al. (2015) proposed a hierarchical generation using *Markov Chain Monte Carlo* (MCMC) simulation and in particular the *Gibbs sampler*. Keeping the existing hierarchy of the variables, Casati et al. (2015) carefully generated the values of the true population. For instance, in their case study the age and gender of a spouse are generated after the household owner is generated. Generation of a variable conditional of previous variables is accomodated using a regression model. In their case, a multinomial regression is fitted and the estimated probabilities are fed into the multinomial distribution that generates values for the variable of interest. According to the MCMC theory, hundreds of thousands, or perhaps millions of values must be generated and then only a small fraction of them is used. Further, the final synthetic population does not satisfy some known marginal constraints and a post-process of the generated data must be applied.

[6] In those cases, IPU will iterate until a corner solution is found.

[7] Lenormand and Deffuant (2013) compared the sample-free method of Gargiulo et al. (2010) to the IPU algorithm (Ye et al., 2009) in generating individuals and households in France and found that the differences between the two approaches are very small.

Bayesian Networks in SPG

Using this hierarchical approach, Young et al. (2009) and Sun and Erath (2015) proposed the use of *Bayesian Networks* (BNs) to formulate a suitable and theoretically consistent way to generate a synthetic population taking into account the conditional distribution of the variables characterizing individuals.[8] The rationale is to first construct a network of the variables that can be represented via a graph where all variables appear with nodes (vertices) and can either be connected indicating the direction of their relationship, or not connected at all. This yields two advantages over the previous SPG approaches: first, it provides information on the joint distribution of all variables and, second, it shows which variables depend upon which in an ordered fashion. The population is then hierarchically generated as in the SR approaches, but the estimated conditional distributions will be more accurate than in the MCMC based approach. The generated data following the BN approach will match to a high degree the observed data, when the variables are categorical. In continuous variables following an asymmetric distribution one should rely on non-parametric techniques in order to accurately sample values from the observed (conditional) distributions.[9]

Formally, a BN $B = \langle G, P \rangle$ is defined as a composite structure comprising a *Directed Acyclic Graph* (DAG) G applied to a set of vertices representing variables denoted as $\mathbf{V}$, along with a joint probability distribution denoted as P (Pearl, 1988; Spirtes et al., 2000). The relationship between P and G is governed by the *Markov* condition, which posits that each variable is conditionally independent of its non-descendants given its parental nodes.[10] Hence, the joint distribution P can be factorized into a product of conditional distributions:

$$P(V_1, \ldots, V_D) = \prod_{i=1}^{D} P\left(V_i | \text{Pa}(V_i)\right) \tag{1}$$

where, D represents the total count of variables, and $\text{Pa}(V_i)$ designates the set of parental nodes for V_i within the graph G. G, P are faithful to each other when G exclusively captures the conditional (in)dependencies in P, and when all conditional (in)dependencies in P are implied by G. This correspondence characterizes G as a perfect map of P (Neapolitan, 2003). A fundamental postulate underlying BN learning algorithms is causal sufficiency, which presumes the absence of latent or unobserved variables among the observed variables. It is evident that a comprehensive representation of the population's characteristics requires the inclusion of all pertinent and important variables.

[8] BNs have been used by Sebastiani and Ramoni (2001) to analyse data extracted from the British general household survey. More importantly, Zhang et al. (2017), Ilahi and Axhausen (2019), and Deeva et al. (2020) used BNs to generate synthetic privacy data, population synthesis and social media profiles data, respectively. Further, BNs have been successfully coupled with ABM models (Kocabas and Dragicevic, 2009, 2013)

[9] In our case study, we address this issue by either sampling from the kernel density estimate of the observed (unconditional) distribution or by employing the k-NN regression when the values should be sampled from the conditional distribution. The decision between them relies upon the learned BN structure.

[10] The parental nodes of a given variable V_i are constituted by the nodes directed towards that variable.

The BN learning offers two interconnected advantages: first, it encompasses the detection of statistically significant associations among variables and, second, it defines a topological order for the variables characterized by a tree structure. This topological structure is instrumental in generating values hierarchically from the variables which entails the formulation of the joint distribution of the data. According to the *Markov* condition, the joint distribution can be explicitly and sequentially expressed, facilitating the process of SPG. This generative process commences by determining values for variables devoid of parental nodes. Subsequently, these values are employed to generate values for their respective child variables, and this procedure iterates until values have been generated for all variables.

The MMHC BN learning algorithm

In our case study, for the BN learning process we have employed the *Max-Min Hill-Climbing* (MMHC) algorithm (Tsamardinos et al., 2006). The MMHC algorithm initially identifies statistically significant associations (edges) between the variables and subsequently employs a scoring method to establish the orientation or assignment of arcs in these relationships. In order to construct the skeleton structure of the BN, the algorithm employs the *Max-Min Parents and Children* (MMPC) variable selection algorithm (Tsamardinos et al., 2003; Tsamardinos and Brown, 2008) which effectively controls the false discovery rate, ensuring that a low proportion of non-significant variables are incorrectly selected. The MMPC algorithm is particularly well-suited for datasets with small sample sizes and a large number of variables, as the conditional independence tests it employs retain high statistical power, even in the presence of limited data. Subsequently, MMHC seeks to determine the optimal DAG where edges either transition into arrows or are removed to maximize a scoring metric.

In cases of continuous variables, the *Bayesian Information Criterion* (BIC) score is employed (Lam and Bacchus, 1994):

$$BIC(G, \Theta \mid \mathbf{V}) = \sum_{i=1}^{n} \log P(V_i \mid Pa(V_i), \Theta_{V_i}) - \frac{\log(n)}{2} |\Theta_{V_i}| \tag{2}$$

This scoring phase involves a greedy *Hill-Climbing* search, an iterative local search approach, within the space of BNs. In this process, edge deletions or direction reversals that yield the most substantial increase in the score within the BN space are applied. It is imperative to note that every time an edge is removed or an arrow direction is modified, a check for cycles is conducted. If cycles are introduced, the operation is aborted, even if it would otherwise enhance the score. The search process continues recursively in a similar manner.

The MMHC algorithm, like all BN learning algorithms, operates under the assumption of being

agnostic to the true underlying relationships within the input data. However, it is common for both practitioners and researchers to possess prior knowledge regarding the necessary directions (whether allowed or forbidden) of certain relationships among variables. In such cases, economic theory can contribute to enhancing the quality of the BN model by imposing or forbidding directions among specific sets of variables. This prior knowledge can be seamlessly integrated into the scoring phase of the MMHC algorithm, thereby reducing errors and producing more realistic BN structures. Failure to incorporate such *a priori* information could result in the development of an unrealistic BN structure, consequently yielding an unrealistic joint distribution that inadequately captures the true underlying joint distribution.

The strength of significant relationships detected by the BN is quantified by the reduction in the BIC score when a specific arrow (or arc or directed relationship) is removed, while keeping the BN structure stable. A larger reduction in the score indicates a stronger indication of the importance or strength of that particular directed relationship. This allows for the ordering of relationships based on their relative strength. Bootstrap resampling serves as an additional measure to assess the validity of the discovered (directed) relationships among the variables. In this approach, a set of observations is randomly sampled with replacement from the original dataset (comprising observed farms), and the BN is learned using the MMHC algorithm. This process is iterated 1,000 times, with the discovered arcs recorded for each repetition. The metric of interest is the proportion of times the observed directed relationships are replicated in the bootstrap samples. This metric serves as an indicator of the confidence and stability in the relationships of each discovered (directed) relationship within the original sample. In our case study, we remove the arrows corresponding to directed relationships that appear in less than 50% of the bootstrap samples, ensuring that we retain only the more robust and reproducible relationships.

Farm Production and Nitrogen Leaching Technology

Once the synthetic population of farmers is generated, we need to recover farm production and nitrate leaching technologies to proxy individual emission rates. Since individual leaching levels are unknown and we only know ambient concentration of nitrous oxide, we need to employ the population of farmers in the area to proxy those unknown values. Keeping technological representations simple, we consider a farm production process that uses a vector of variable inputs $x^v \in \Re_+^m$ together with chemical fertilizers $x^q \in \Re_+$ and irrigation water $x^w \in \Re_+$, to produce a single output denoted by $y \in \Re_+$. Chemical fertilization results in NO_3^- leaching (i.e., nitrous oxide), denoted by $q \in \Re_+$, that contaminates water reservoirs. Further, we assume that the extent of nitrate leaching into the water reservoirs depends on irrigation water application, the soil conditions of the farm denoted by the vector $s \in \Re_+^k$, and precipitation denoted by $r \in \Re_+$. Following the multiple

production relations model developed by (Murty et al., 2012)[11], as it was empirically applied by (Tsagris and Tzouvelekas, 2022), farm production technology and the nature's nitrogen residual generating mechanism can be represented by the following closed, non-empty sets:

$$\begin{aligned} T^y &= \{(x^v, x^q, x^w, y, q, s, r) : \ (x^v, x^q, x^w) \ \text{can produce} \ y\} \\ T^q &= \{(x^v, x^q, x^w, y, q, s, r) : \ x^q \ \text{can pollute by} \ q \ \text{for a given level of} \ (x^w, s, r)\} \end{aligned}$$

Variable inputs (including chemical fertilizers and irrigation water) and farm output are strongly disposable in farm production, whereas nitrogen generating technology satisfies *costly disposability* of nitrate emissions (Murty, 2010) (i.e., nitrate leaching is an output of farm production whose disposal is not free):

$$\text{if} \ (x^v, x^q, x^w, y, q, s, r) \in T^q \wedge \bar{q} \geq q \wedge \bar{x}^q \leq x^q \ \text{then} \ (x^v, \bar{x}^q, x^w, y, \bar{q}, s, r) \in T^q$$

The above monotonicity property implies that T^q is bounded from below. Any given level of chemical fertilizers application may create a minimal level of nitrate leaching but it can always generate a greater amount of leached nitrogen if farmers are ignorant about the nature's residual generating mechanism.[12]

Hence, overall farm technology may be described as the intersection of the two sub-technologies $T = T^y \cap T^q$ reflecting both the transformation of inputs into farm output and the nitrogen pollution generating mechanism resulting from chemical fertilization. According to Murty et al. (2012), the unified crop technology violates free disposability with respect to chemical fertilizers application, satisfies free disposability with respect to farm output and variable inputs use, and it satisfies cost-disposability with respect to nitrogen pollution. In effect, if farm production is inefficient, farmers can always decrease variable input use without changing fertilizer use that generates nitrogen pollution in the reservoirs. On the other hand, if nitrogen pollution is inefficient, then farmers can decrease nitrate leaching without altering variable input use and farm output by improving their knowledge about nature's pollution generating mechanism.

Using functional representations and assuming that farmers are technical inefficient in both farm production and nitrate leaching, then farm production technology[13] and nature's nitrogen

[11] Their framework builds on the factorially determined multi-output model developed by Frisch (1965), as it was further elaborated by Førsund (2009), that captures the physical process of generation of residuals allowing for some inputs and outputs that exhibit technological non-rivalness/jointness.

[12] Water contaminated with nitrogen does not harm crop growth and therefore, farm technology set does not impose any constraint on q.

[13] Using the implicit function theorem, Murty et al. (2012) proved that the marginal product of fertilizers is non-negative in farm production, while at the same time more fertilizers applied on field increase nitrate leaching for any given soil characteristics and irrigation water use.

generation mechanism may be defined as[14]

$$T = \{(x^v, x^q, x^w, y, q, s, r) : \ g\left(x^q, x^w, s, r\right)\theta^q = q \ \wedge \ f\left(x^v, x^q, x^w\right)\theta^y = y\}$$

where $f\left(x^v, x^q, x^w\right) : \Re_+^{m+2} \rightarrow \Re_+$ is a continuous and, strictly increasing, twice differentiable concave crop production function representing maximal farm output obtained from variable input, chemical fertilizers and irrigation water use. Similarly, $g\left(x^q, x^w, s, r\right) : \Re_+^{k+3} \rightarrow \Re_+$ is also a continuous and twice-differentiable convex *nitrate leaching* function providing minimum nitrate leaching levels attained from chemical fertilizer application and irrigation water use given soil characteristics and precipitation. It also holds that $q = 0$ if $x^q = 0$, that is when chemical fertilizers are not applied on field nitrate leaching is zero regardless of the other factors affecting by-production.[15]

Finally, $\theta^y \in (0, 1]$ represents the percentage of maximal output realized by farm households in the presence of technical inefficiency in farm production. Similarly, $\dfrac{\theta^q - 1}{\theta^q} \in (0, 1]$ represents the percentage of excess nitrates leached due to inefficiency in fertilizers and irrigation water application through nature's residual generation mechanism.[16] Except of wrong fertilizer application within farming activities, farmers who are unaware of the natural processes may further intensify water contamination through nitrate leaching. In other words, except of utilizing an appropriate input mix during farm production exploring fully the potential of farm technology, farmers should also be aware of the natural processes that trigger nitrate leaching in their own fields.

The GME Estimator

Since nitrate leaching is only detectable and measurable after it has entered the ecosystem we cannot apply conventional econometric tools to estimate individual emissions. To overcome this problem we utilize the *Generalized Maximum Entropy* (GME) method which is an information-theoretic approach initially devised for ill-posed problems of inference where the sample sizes are limited (Golan et al., 1996).[17] To make the model empirically operational, and to apply GME, we need to assume specific functional representations for crop production and nitrogen residual generating technology. Starting from farm production, we choose the following *transcendental logarithmic*

[14] As noted by Førsund (2018) scaling of y and q is necessary to avoid the intersection of the two sets to be empty. We resolve that in the econometric setup of the model.

[15] Nitrogen leaching may be non-zero in cases that farmers do not apply fertilizers at all due to the existing nitrogen stock in the soil. However, since our primary focus is on estimating individual leaching levels we do not take that into account.

[16] It holds that $\theta^q \geq 1$ as nitrate is leached in excess.

[17] On theoretical grounds, Golan and Perloff (2002) proved that GME, unlike Renyi-GME and Tsallis GME, satisfies the properties of completeness, transitivity and uniqueness, permutation invariance, scaling, as well as subset and system independence.

(translog) specification to approximate production technology:

$$\begin{aligned}\ln y_i &= \beta_0 + \sum_m \beta_m^v \ln x_{mi}^v + \frac{1}{2}\sum_m \sum_l \beta_{ml}^{vv} \ln x_{mi}^v \ln x_{li}^v + \ln h_i \left(\beta^h + \frac{\beta^{hh}}{2} \ln h_i + \sum_m \beta_m^{hv} \ln x_{mi}^v \right) \\ &+ \ln x_i^q \left(\beta^q + \frac{\beta^{qq}}{2} \ln x_i^q + \sum_m \beta_m^{qv} \ln x_{mi}^v + \beta^{qw} \ln x_i^w + \beta^{qh} \ln h_i \right) \\ &+ \ln x_i^w \left(\beta^w + \frac{\beta^{ww}}{2} \ln x_i^w + \sum_m \beta_m^{wv} \ln x_{mi}^v + \beta^{wh} \ln h_i \right) + \varepsilon_i^y \end{aligned} \tag{3}$$

where subscript $i = 1, \ldots, n$ indicates farms, β's are the associated parameters and, $\varepsilon_i^y = \epsilon_i^y - u_i^y$, is the composed error term in stochastic frontier terminology with ϵ_i^y denoting random disturbances, and u_i^y capturing technical inefficiency in crop production obtained from $\theta_i^y = \exp(-u_i^y)$.

Accordingly, following Knapp and Schwabe (2008) and Wang and Baerenklau (2014) we approximate nature's nitrate residual generation mechanism as:

$$q_i = \frac{-\delta_i^q x_i^q + \delta_i^{qq} (x_i^q)^2}{1 + \exp(-\delta_i^w x_i^w)} \exp(\varepsilon_i^q) \tag{4}$$

with

$$\delta_i = \alpha_0 + \sum_k \alpha_k^s s_{ki} + \alpha_h^h h_i \quad \text{and} \quad \varepsilon_i^q = \epsilon_i^q + u_i^q \tag{5}$$

where again subscript $i = 1, \ldots, n$ indicates farms, α's are the associated parameters, s_k is the k^{th} environmental characteristic affecting soil nitrate absorption, ϵ_i^q is the usual random term and u_i^q captures technical inefficiency in nitrate leaching obtained from $\theta_i^q = \exp(u_i^q)$.

The GME principle dictates that the k-th regression coefficient in (3) can be expressed in the form of a weighted combination of J plausible real values $\mathbf{z}^{\beta_k} = \left(z_1^{\beta_k}, z_2^{\beta_k}, \ldots, z_J^{\beta_k} \right)$ for β_k as:[18]

$$\beta_k = \mathbf{z}^{\beta_k} \mathbf{p}_k^{\beta_k}$$

such that $\beta_k \in \left[z_1^{\beta_k}, z_J^{\beta_k} \right]$ and the J *non-negative* weights $\mathbf{p}^{\beta_k} = \left(p_1^{\beta_k}, p_2^{\beta_k}, \ldots, p_J^{\beta_k} \right)'$ sum to unity, $\sum_j p_j^{\beta_k} = 1$. Accordingly, the vector of random disturbances in the production frontier is expressed:

$$\boldsymbol{\epsilon}^y = \mathbf{Z}^{\epsilon^y} \mathbf{p}^{\epsilon^y} Z \tag{6}$$

where $\mathbf{Z}^{\epsilon^y}$ is a $n \times nJ$ diagonal matrix with elements $\mathbf{z}_i^{\epsilon^y} = \left(z_{i1}^{\epsilon^y}, z_{i2}^{\epsilon^y}, \ldots, z_{ij}^{\epsilon^y} \right)$ refering to the support

[18] The subscript k collects all superscripts and subscripts of the production frontier. On the other hand J is the number of support values assumed for the regression coefficients.

values of the i^{th} random disturbance and $\mathbf{p}_i^{\epsilon^y} = \left(p_{i1}^{\epsilon^y}, p_{i2}^{\epsilon^y}, \ldots, p_{iJ}^{\epsilon^y}\right)'$. Similarly, we will denote by $\mathbf{Z}^{u^y}$ the matrix of support values of the technical inefficiencies in crop production. Accordingly we denote by $\mathbf{z}^{a_k^q}$, $\mathbf{z}^{a_k^{qq}}$ and, $\mathbf{z}^{a_k^w}$ the support values of the constant and slope coefficients that appear in the numerator and denominator of the nitrate leaching function. Finally, the matrix of support values of the random disturbances and inefficiencies of the nitrate leaching function will be denoted by $\mathbf{Z}^{\epsilon^q}$ and $\mathbf{Z}^{u^q}$ respectively.

The constraint that forms the unified farm technology can be expressed as $f_i(\cdot) \geq g_i(\cdot) \quad \forall\, i = 1, \ldots, n$. Since GME requires only equality constraints we introduce non-negative valued slack variables to turn the inequality above into equality. Expressing both the production and nitrate leaching functions in log-scale, the constraint becomes

$$\ln f_i(\cdot) - \ln g_i(\cdot) = s_i \quad \forall\, i = 1, \ldots, n. \tag{7}$$

Again, the slack variable can be expressed in a similar form to equation (6) and its matrix of support values will be denoted by $\mathbf{Z}^s$.

Assuming that individual nitrate leaching rates attributed to each farmer solely determine observed concentration of nitrates in the reservoir, i.e., $\sum_i q_i = Q^N$, the GME problem in our specification is to minimize *Shannon's* entropy $I(\mathbf{p})$ subject to three sets of equality constraints:

$$\begin{aligned}\min_{\mathbf{p}} I(\mathbf{p}) \;=\; \min_{\mathbf{p}} \Bigg\{ & \sum_k^K \sum_j^J p_j^{\beta_k} \ln p_j^{\beta_k} + \sum_{i=1}^n \sum_j^J p_{ij}^{\epsilon^y} \ln p_{ij}^{\epsilon^y} + \sum_{i=1}^n \sum_j^J p_{ij}^{u^y} \ln p_{ij}^{u^y} + \sum_l^L \sum_j^J p_j^{\alpha_k^q} \ln p_j^{\alpha_k^q} \\ + & \sum_l^L \sum_j^J p_j^{\alpha_k^{qq}} \ln p_j^{\alpha_k^{qq}} + \sum_j^J p_j^{\alpha_k^w} \ln p_j^{\alpha_k^w} + \sum_{i=1}^n \sum_j^J p_{ij}^{\epsilon^q} \ln p_{ij}^{\epsilon^q} + \sum_{i=1}^n \sum_j^J p_{ij}^{u^q} \ln p_{ij}^{u^q} + \sum_{i=1}^n \sum_j^J p_{ij}^s \ln p_{ij}^s \Bigg\}\end{aligned}$$

subject to[19]

$$\ln y_i = \ln f_i(\cdot) + \epsilon_i^y - u_i^y \quad \forall\, i = 1, \ldots, n \tag{8a}$$

$$Q^N = \sum_i^n g_i(\cdot) \exp\left(\epsilon_i^q + u_i^q\right) \tag{8b}$$

$$s_i = \ln f_i(\cdot) - \ln g_i(\cdot) \quad \forall\, i = 1, \ldots, n \tag{8c}$$

[19] We have omitted the summation to unity constraints of the probabilities as their Lagrangean multipliers vanish. However these constraints have been taken into consideration in the computations.

Introducing the Lagrangean multipliers $\lambda^y = (\lambda_1^y, \ldots, \lambda_n^y)$, λ^q and $\lambda = (\lambda_1, \ldots, \lambda_n)$, becomes:

$$\begin{aligned}\min_{\mathbf{p}} \quad I(\mathbf{p}, \lambda^y, \lambda^q, \lambda) = \min_{\mathbf{p}} \Bigg\{ & \sum_{k}^{K}\sum_{j}^{J} p_j^{\beta_k} \ln p_j^{\beta_k} + \sum_{i=1}^{n}\sum_{j}^{J} p_{ij}^{\epsilon^y} \ln p_{ij}^{\epsilon^y} + \sum_{i=1}^{n}\sum_{j}^{J} p_{ij}^{u^y} \ln p_{ij}^{u^y} + \sum_{l}^{L}\sum_{j}^{J} p_j^{\alpha_k^q} \ln p_j^{\alpha_k^q} \\ & + \sum_{l}^{L}\sum_{j}^{J} p_j^{\alpha_k^{qq}} \ln p_j^{\alpha_k^{qq}} + \sum_{j}^{J} p_j^{\alpha_k^w} \ln p_j^{\alpha_k^w} + \sum_{i=1}^{n}\sum_{j}^{J} p_{ij}^{\epsilon^q} \ln p_{ij}^{\epsilon^q} + \sum_{i=1}^{n}\sum_{j}^{J} p_{ij}^{u^q} \ln p_{ij}^{u^q} \\ & + \sum_{i=1}^{n}\sum_{j}^{J} p_{ij}^{s^1} \ln p_{ij}^{s^1} + \sum_{i=1}^{n}\sum_{j}^{J} p_{ij}^{s^2} \ln p_{ij}^{s^2} + \sum_{i=1}^{n} \lambda_i^y \left(\ln y_i - \ln f_i(\cdot) - \epsilon_i^y + u_i^y\right) \\ & + \lambda^q \left(\sum_{i}^{n} g_i(\cdot) \exp\left(\epsilon_i^q + u_i^q\right) - Q^N \right) + \sum_{i=1}^{n} \lambda_i \left(\ln f_i(\cdot) - \ln g_i(\cdot) - s_i\right) \Bigg\}\end{aligned}$$

The problem was solved with respect to the probabilities[20] using an iterative scheme. We begin with some initial values in the Lagrangean multipliers, we then compute the associated probabilities and solve each set of equations (8a)-(8c) serially. We update the multipliers and the associated probabilities and then we solve those equations again. We repeat this process until convergence[21] and then we use the probabilities to estimate all regression parameters. For the support values of the parameters we used $J = 5$ for each regression coefficient[22], random disturbances, inefficiencies and, slack variables. As a starting point for the farm production model, we centered the support values for the β coefficients coming from the stochastic frontier model, assuming a half normal distribution, and we added a small perturbation from left and right. We examined the range of the random disturbances to construct their support values, and used the estimated inefficiencies as initial values. Finally, we decided upon the support values of the nitrate leaching function parameters and of the slack variables based on trial-and-error as no prior information is available.

The Practical Problem

The Region of Thessaly

Our case study refers to the NUTS II Greek region of Thessaly where nitrate pollution from agricultural activities turned into a major problem the last 15 years. The region is located in Central Greece and accounts approximately for the 30% of total agricultural production in Greece (see Figure 1a). Indeed, data on nitrate pollution of the water reservoirs in Thessaly provided by the *Greek Ministry of Agriculture* indicate an extensive nitrate pollution of both surface water

[20] The detailed expressions of the model probabilities are presented in the Appendix Á.

[21] In our case convergence was achieved when the change between two successive vector of estimates of the production function was tiny, i.e., less than 0.001.

[22] Golan et al. (1996) after several Monte Carlo simulation concluded that 5 support values are sufficient enough.

reservoirs and underground aquifers. In total there are 62 sample points, depicted in Figure 2, capturing approximately the 95% of the water resources in Thessaly for the 2017-18 cropping year. Table 1 presents the stock of nitrous oxide measured by the personnel of the Greek Ministry of Agriculture in all 62 locations. As it is evident from this Table, nitrate pollution is severe in the south and south-east municipalities of the region where point measurements in 2018 exceed by far the *Drinking Water Directive* limit of 50mg NO_3^- l^{-1}.[23] In the majority of the sites though measured nitrates are beyond the recommended by the EU standards of 11.3mg NO_3^- l^{-1}. Maximum value is 123mg NO_3^- l^{-1} in the municipality of Karditsa and minimum value is 0.5mg NO_3^- l^{-1} in minor water reservoirs of Elassona and Rigas Ferraios municipalities. In the four mountainous municipalities nitrate pollution is not monitored by the Greek authorities as pollution is minor. It should be noted though that all water reservoirs used to supply water for domestic use in the major cities of the regions, nitrate pollution is beyond the limit of 50mg NO_3^- l^{-1} (Volos 61.0, Karditsa, 123.0, Farsala 96.5, Almyros 107.7).[24]

The FADN Dataset and BN Mapping

The FADN dataset used for the construction of joint probability distributions in the SPG framework refers to the 2017-18 cropping period provided by the *Greek Ministry of Agriculture*. Overall, there are 3,638 farms in the dataset located in the twelve Greek Nuts II regions.[25] From those farms, we focused on 509 holdings located in the region of Thessaly, the exact location of which along with the twenty-two municipalities of the region of Thessaly are depicted in Figure 1b. The FADN database contains very detailed information on crop production that cannot be applied effectively for the generation of the synthetic population of farmers. Therefore, initially we followed the *EU Regulation No1166/2008* that establishes a framework for European statistics at the level of agricultural holdings to aggregate across different output and cost items of crop and livestock production. Next, attributes of the collected data were clustered in five groups in order to provide a basis for the construction of the BN. These groups include: farm labour characteristics containing information on the farm manager and paid or unpaid labour related characteristics, crop and animal production including twenty crop and ten livestock products, other farm income and subsidies containing information on the values of other farm income sources as well as subsidies and grants grouped in four clusters (decoupled payments, crops and animals, exceptional support and rural development and subsidies on cost), farm assets including current and non-current assets, land, building and machinery and, variable inputs including twelve attributes measuring variable inputs

[23] Looking at the historical data there is a clear increasing trend in nitrate pollution since the beginning of 2000's.

[24] It should be noted that the WHO standard for drinking water is 50mg NO_3^- l^{-1} for short-term exposure, and 3mg NO_3^- l^{-1} for chronic effects.

[25] Regional distribution of the surveyed FADN farms is presented in Table B.1 in the Appendix.

cost.

Soil spatial and environmental data were obtained from the *European Soil Data Centre*, the *NASA's Earthdata* program and *Socioeconomic Data and Applications Center* and cover locations throughout Greece. Those measurements were matched with the available FADN data points using the *1-nearest neighbour* (1-NN). Climatic data were accessed from the six Meteorological Stations, located in the region of Thessaly and operated by the *Greek Meteorological Service*. Finally, total crop and livestock production data necessary for the synthetic reconstruction of the population of farmers in Thessaly were obtained from *Agricultural Census* of the *Greek Statistical Service*. Our final aggregation scheme includes 204 attributes-variables describing production relations among farmers in the sample used to construct the BN topological structure. Details on the data used in the analysis, the aggregation scheme adopted as well as on the definition of variables following the FADN coding are provided in Tables B2 through B11 in the Appendix.

As previously mentioned, BN learning algorithms are agnostic of the input data and require some prior knowledge to facilitate the generation of more realistic populations. A set of constraints must be imposed among these 204 variables. These refer to rationally forbidden directions between the pairwise relationships. We keep the number of constraints at a minimum in order to avoid restrictive structures in the resulting BN. The majority of the constraints refer either to crop or to livestock production. The detailed list of constraints is presented in Appendix C. Running the MMHC learning algorithm we discovered 220 statistically significantly associated relationships among farm attributes. All of them are presented in Table C1 in the Appendix along with their directions and their strength. For instance, the relationship, in the first row of the table, between G4 and G5 is directed from G4 to G5 and hence in the BN terminology this is denoted by G4→G5. The same is true for all relationships. The topological structure of the obtained BN is depicted in Figure 3.

The results of bootstrap validation also appear in the Table C1. The 149 out of the 220 (67.73%) identified directed relationships in the observed farms were observed more than 50% of the times in the bootstrap samples. This, rather low, number does not come by surprise as the data contain many variables with high proportions of zero values. When sampling with replacement, the percentage of unique values in the bootstrap sample is on average equal to $1 - (1 - 1/n)^n$, which in the current situation is equal to 63%. Hence, the bootstrap sample of 509 farms contains around 63% unique farms. Variables having more than 63% zeros may contribute only zeros to the bootstrap sample and hence no relationship can be discovered, even if there is one.

Synthetic Sample Generation

Before generating the synthetic population of farmers in Thessaly, we generated a synthetic sample that mimics the attributes of the surveyed farms in FADN database. The synthetic sample was used as an assessment tool of the constructed BN and as a validation guide in order to proceed to the SPG. The generation of random values from BNs using continuous data often results in values that follow a normal distribution. However, this may not accurately reflect real-world scenarios, particularly in cases where most variables exhibit strong right-skewness and contain numerous zero values. To address this issue, we employed a refined data generation scheme based on non-parametric regression, utilizing the BN structure learned from the variables of the observed farms. The generation process follows a sequential order so that each variable is generated conditionally upon its parent variable(s) following the existing BN structure depicted in Figure 3. In that way, the produced synthetic data align more closely with the distribution patterns observed in the real data, thereby providing a more realistic representation of the underlying data structure.

For variables with no parents, we computed the *kernel density estimate* (KDE) of the distribution of the non-zero values and generated non-zero values from this KDE, whereas zero values remained the same. Specifically, the KDE is given by

$$\tilde{f}(x) = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^{n} e^{-\frac{(x-x_i)^2}{2h^2}}$$

where $h = 0.9 \min(\tilde{\sigma}, IQR/1,34)\, n^{-1/5}$, $\tilde{\sigma}$ denotes the sample standard deviation and IQR is the interquartile range. Random number generation for a variable with no parents is obtained from $\tilde{x}_i = \bar{x} + (x_i - \bar{X} + \tilde{h} z_i)/\left(1 + \tilde{h}^2/\tilde{\sigma}^2\right)^{1/2}$, where z_i are random values generated from the standard normal distribution, $\tilde{h}$ is the estimated bandwidth and $\bar{x}$ and $\tilde{\sigma}^2$ denote the sample mean and variance, respectively, of the observed values (Silverman, 2018).

For variables with at least one parent, we employed the *k-Nearest Neighbors* (k-NN) algorithm which is a non-parametric kernel regression technique that considers only the values of the k nearest neighbors. Using information from the surveyed farms, the estimated values are obtained from $\tilde{x}_i = \sum_{j \in \mathcal{C}_k} x_j$ where, $\mathcal{C}_k$ represents the set of k closest neighbors determined from the *Euclidean* distance computed using the parent variables. To select the appropriate value of k, we utilized the observed variables to minimize the sum of squares of the errors between the observed and fitted values. Once values for a variable are generated, we performed data transformation to ensure that their mean matches the mean of the observed variable values. However, some post-generation refinement steps were considered necessary. Specifically, in the case of crop production, when the synthetic cultivated land for a crop is zero, the corresponding synthetic values for irrigated land and crop production were set to zero. If the irrigated area for certain crops exceeded the corresponding

cultivated land, we adjusted the irrigated area to be equal to the cultivated area. Similar refinement procedures were applied to animal products.

Evaluation of the Synthetic Sample

First, in order to assess the equality of variable distributions between the observed and synthetic farms, the energy distance-based test as described by Székely et al. (2004) was used. Specifically, the energy distance test was used to examine the equality of both the joint distributions and all univariate distributions between the observed and synthetic farms computed from:

$$e(S_1, S_2) = \frac{n_1 n_2}{n_1 + n_2} (2M_{12} - M_{11} - M_{22})$$

where $M_{ij} = \frac{1}{n_i n_j} \sum_{p=1}^{n_i} \sum_{q=1}^{n_j} \|V_{ip} - V_{jq}\|$, for $i, j = 1, 2$ and $\|.\|$ denotes the *Euclidean* norm. For attributes taking discrete values we applied the familiar χ^2 test:

$$\chi^2 = \sum_{k=1}^{K} \frac{\left(A_k - \hat{A}_k\right)^2}{\hat{A}_k}$$

where A_k and $\hat{A}_k$ denote the frequency of the k-th possible value of the attribute of the observed and of the synthetic farm, respectively and K is the number of possible values of the attribute. If $\chi^2 > \chi^2_{0.95,K-1}$ the equality of the distributions of the observed and the synthetic farms is rejected.

Next, we tested the cultivated area and crop production compositions between the true and the synthetic farms. We first normalized the data by dividing the cultivated area and production volume of each product by the total cultivated area and total farm production for each farmer, respectively. After expressing all variables in percentages,[26] we then applied the α-transformation (Tsagris et al., 2011) and applied the energy test for equality of the two joint distributions. For either testing procedure, if the p-value is less than 0.05 the H_0 is rejected and hence the distributions cannot be assumed statistically equal.

Additionally, we employed the k-NN algorithm for this purpose. In all cases, we utilized a 10-fold cross-validation protocol to measure the discrimination performance. Ideally, the two samples, namely the observed and synthetic farms, should exhibit minimal separability, with the separability measure approaching 50%. Finally, we applied *Principal Component Analysis* (PCA) to reduce the dimensionality of the data, facilitating visual inspection of the observed and synthetic farm samples in lower-dimensional space. The results showed that the energy test applied to the joint distributions produced a p-value equal to 1 and in 94.8% of the cases the energy test for equality of the univariate distributions of the attributes produced p-values larger than 0.05. The energy

[26] This type of data are termed compositional data (Aitchison, 1982).

test applied to the compositions of the cultivated areas and the compositions of the production produced p-values greater than 0.05 as well. The k-NN estimated an accuracy equal to 60% which is satisfactory.

The results of the PCA along with the density plots of each attribute are presented in the Appendix C (Figures C.1-C.3). Specfiically, Figure C.3 shows the combined samples of farms (observed and synthetic), projected onto the 2-D space spanned by the pairs of the first 4 principal components. The black and red dots indicate the observed and the synthetic sample of farms, respectively. Accordingly, Figures C.1 and C.2 present the density plots of the cultivated area for all crops as well as animal stock by the synthetic farmers. Evidently, both samples cannot be distinguished from one another. Evidently, statistical testing implies that the obtained BN reflects more than satisfactory the existing farm structures in Thessaly and it can be used to construct the synthetic population of farmers in the region.

Synthetic Population Generation

Upon the successful generation of our synthetic sample, we proceeded to address the SPG task using the same approach. Applying KDE for the variables with no parents we generate zeros and non-zero values of size equal to the number of farms in the region. The available information from the *Agricultural Census* was used to calibrate the generated values. For instance, the total synthetic cultivated area should equal the total true cultivated area and the same applies for the total production, the number of animals and so forth.[27] The values of the variables with parents was generated using the aforementioned k-NN algorithm using the same strategy as before and applying the calibration using census information. It should be noted that the calibration was not applied after all variables have been generated, but at each variable once its values have been generated.

A crucial constraint imposed on the SPG process was that the cumulative synthetic totals of cultivated land areas across the different crop and livestock products must align with the corresponding observed totals, which were obtained from *Agricultural Census* in Thessaly. To fulfill this constraint, a weighting generation scheme was required. Unfortunately, the existing representation weights based on the FADN farm representation weights did not meet this constraint.[28] Consequently, we undertook the task of estimating representation weights utilizing the empirical likelihood method (Owen, 2001). These estimates were conditioned on ensuring that the mean of the synthetic irrigated land areas equals the observed mean values (census total cultivated land divided by the number of farms), thereby aligning our generated data with the observed agricultural

[27] Table C.2 in the Appendix shows the calculations for crop and animal production.
[28] FADN does not include small size farms.

landscape.

The weights are obtained under the H_0 that $\boldsymbol{\mu} = \boldsymbol{\mu}_0$[29]. The goal is to maximize *Wilk's* log-ratio test statistic with respect to some weights

$$\max_{w_i} \left\{ \sum_{i=1}^{n} \log(nw_i) \,|\, \sum_{i=1}^{n} w_i \mathbf{x}_i = \boldsymbol{\mu}_0, w_i \geq 0, \sum_{i=1}^{n} w_i = 1 \right\}.$$

under the constraint that $\sum_{i=1}^{n} w_i \mathbf{x}_i = \boldsymbol{\mu}_0$. The introduction of Lagrangean parameters followed by some tedious algebra, yields the following form of the weights: $w_i = \frac{1}{n}\left[1 + \lambda^T (\mathbf{x}_i - \boldsymbol{\mu}_0)\right]^{-1}$, where λ is the Lagrangean parameter introduced and its value is computed via solving the equation $\sum_{i=1}^{n} \frac{1}{n} \frac{\mathbf{x}_i - \boldsymbol{\mu}_0}{1 + \lambda^T (\mathbf{x}_i - \boldsymbol{\mu}_0)} = \mathbf{0}$.

Upon computation of the weights w_i we can then generate random samples of farmers. If we denote by N the total farms in the population the representation weights are given by Nw_i, that is for the i-th observed farmer we will generate Nw_i synthetic farmers with similar characteristics. In total, we generated 34,853 synthetic farms for the region matching the acreage cultivated as well as crop and livestock production in Thessaly.

For the empirical approximation of crop and livestock technology, we consider one output and three variable inputs together with irrigation water (summary statistics of these variables are presented in Table 2). Different crops (including quantities sold off the farm and quantities consumed by the farm household during the crop year) are aggregated into a single aggregate *Tornqvist* output index with the revenue shares of each crop or livestock products defining the relevant weights. On average, total farm production for the synthetic sample is 11,761€ varying significantly among farms. Farm labor is defined as the total working hours devoted to supervision and organizational activities as well as to field activities such as harvesting, planting, fertilization, spraying and irrigation water application. Farm labor includes farm owner, family members and hired workers with either permanent or seasonal occupation status.[30] On average, farmers devote 2,283 hours in their holdings for all farming activities.

Land input includes total acreage (rented or owned) measured in stremmas. Given the diverse nature of farm activities, farms in Thessaly are large relative to the rest of Greece, with 122 stremmas on average. Chemical fertilizers include mostly ammonium nitrate and to a lesser extent urea or ammonium sulfate. The different categories of chemical fertilizers are aggregated into a single input index using again *Tornqvist* procedures with cost shares as weights. On average, farms in the sample applied 3,901 Kgs of chemical fertilizers in their holdings. Irrigation water is measured in m^3. During the whole cropping period, farmers in the sample used 668 m^3 of irrigation water.

[29] The $\boldsymbol{\mu}_0$ refers to the mean vector obtained from the census totals.

[30] Given the competitive local labor market conditions we assume that family and hired labor are perfect substitutes, implying that returns to farm and off-farm work are equal.

To avoid problems associated with units of measurement, all variables were converted into indices, with the basis for normalization being their maximum value for the production function and the aggregate level of nitrates into the aquifer for the nitrate leaching function. This way, all values are within the $(0, 1]$ range.

Finally, for the allocation of synthetic farms among water reservoirs we assume that each farm contributed to the pollution of its five nearest water resources in a manner inversely proportional to its distance from the resource.[31] Specifically, the *Euclidean* distance of the i-th observed farm, from each of the 62 water resources (d_{ij}, $i = 1, \ldots, 198$ unique locations, $j = 1, \ldots, 62$ water resource sample points) were computed and the 5 smallest distances, d^*_{ij} were kept. The proportion of the estimated pollution of the i-th farm to its j-th nearest water resource was set equal to $w_{ij} = \frac{1/d^{*2}_{ij}}{\sum_{j=1}^{5} 1/d^{*2}_{ij}}$.

Econometric Estimation of Synthetic Farm Technology

The estimated parameters of the translog production frontier along with those of the nitrate leaching function appear in Table 3 together with their corresponding standard errors obtained using non-parametric bootstrapping. The first-order parameters of variable inputs, chemical fertilizers and irrigation water are statistically significant at least at the 5 percent level with their magnitudes being bounded in the unity interval. The bordered *Hessian* matrix is found to be negative semi-definite at the point of approximation and for the 87 percent of observations. Hence, concavity of the production function is satisfied with respect to all variable inputs, fertilizer application and irrigation water, implying positive and diminishing marginal products among synthetic farmers in the sample.

Concerning nitrate leaching function the majority of the estimated parameters are statistically significant at least at the 5 percent significance level, having the anticipated sign and magnitude. The bordered *Hessian* matrix is found to be positive semi-definite for the 93% of the observations implying that it is convex with respect to both chemical fertilizers and irrigation water application. Overall unified synthetic farm technology appears on Figure 4, where the blue curve represents crop production technology and the red curve nature's nitrate residual generation mechanism with respect to chemical fertilizers application. As the parameter estimates of both models imply, the desired output (crop production) is concave and the by-product (leaching) is convex with respect to fertilizer application. What is evident from the graph is that synthetic farms in Thessaly are both production and emission inefficient with respect to both technologies. The latter is more evident in higher chemical fertilizers application.

Based on the parameter estimates of the translog production frontier, output elasticities of

[31] The results do not vary substantially when the ten nearest water resources were utilised.

variable inputs, chemical fertilizers and irrigation water are estimated and presented in Table 4. Irrigation water together with labour input are found to have the greatest percentage impact on farm's crop production with their corresponding mean output elasticities being 0.3385 and 0.3447, respectively. In contrast, production output is found to be less responsive to changes in chemical fertilizer use with a point estimate of 0.2824. Finally, the corresponding point estimate for utilized agricultural area is 0.3006, which is an expected result. In total, returns-to-scale were found to be increasing (1.2662 on the average), implying that synthetic farmers in Thessaly are operating before their optimal scale of production. In turn, this implies that the average farm size of 122 stremmas is lower than the farm size that would maximize their ray average productivity. This is an expected outcome given the peculiarities of the agricultural sector in Thessaly.

Estimated individual leaching ranges from a minimum of 0.0001mg NO_3^- l^{-1} to a maximum of 0.8146mg NO_3^- l^{-1} with an average value of 0.0482mg NO_3^- l^{-1} (see Figure 5). The frequency distribution of these values depicted in Figure 5 reveal a smooth pattern with the majority of farms exhibiting leaching levels up to 0.0100mg NO_3^- l^{-1}. However, there is a group of farms with severe nitrate emission problems exhibiting values above 0.1700mg NO_3^- l^{-1}. Specifically, a large portion of farmers (46% of sample participants) pollute by a relatively small degree, whereas the 10.1% of surveyed farms are responsible for more than 0.1700mg NO_3^- l^{-1} as depicted by the histogram. The majority of these synthetic farms belong either to the first or to the fourth profit quartile pollute more the water reservoir (average nitrate leaching levels are above sample mean). Small farms with lower profit margin are more keen to use excess chemical fertilizers and paying less attention to water contamination. This is also true for more profitable farms who take full advantage of the common water resource. Nitrate leaching elasticities reported in Table 4 show that chemical fertilizers are the foremost important factor (positively correlated with profit margins) especially on farms with cultivating in sloped plots with eroded soil. Our estimates are in line with the survey of Hansen and Djurhuus (1996) who found higher nitrate leaching rates associated with increased use of chemical fertilizers from large farms. Wrong irrigation schedules combined with excess fertilization and adverse soil conditions further intensifies the water contamination problem. Nitrate leaching elasticity of irrigation water application is 0.0122 on the average, with an increasing trend over nitrate leaching quartiles.

Point estimates of both crop production and nitrate emission technical efficiencies are also presented in Table 4 per estimated individual nitrate leaching quartile. First, crop production efficiency was found only 50.29% on average ranging from a minimum of 41.23% to a maximum of 68.16%. These values exhibit a clear decreasing trend over individual nitrate leaching quartiles indicating that abatement efforts directed to small farms should be accompanied by measures aimed to improve utilization of crop production. Concerning technical efficiency in nitrate leaching, average

value is higher 65.68% with less variation across synthetic farmers though. It is encouraging that farms are doing more efficient job in realizing nature's nitrate residual generation mechanism. Still though, farmers in Thessaly can decrease individual leaching levels by 35% without changing chemical fertilizer application as long as their know how on nature nature's nitrate residual generation mechanism is improved, e.g., more efficient irrigation schedule relative to fertilization, avoid fertilizer application in long dry periods. Finally, both estimated and efficient individual leaching rates per fertilizer application are presented in Figure 6. Leaching is considerably higher at increased use of chemical fertilizers exhibiting however, a variability among synthetic farms.

Individual Tax Rates

In the EU there are four countries introduced a nitrate fertilizer tax. Finland was the first country who introduced the tax scheme in 1976. The tax was abolished in 1994 when the country became a member state of the EU. Until 1992 the tax was applied to chemical fertilizers regardless their nutrient content. It ranged from 0.006 to 0.09€/kg of fertilizers. After that year the tax was uniform only on nitrate and phosphorus fertilizers and increased sharply to 0.44€/kg. The tax system in Finland is somewhat similar to that found for pesticides in that what was once an environmental charge (implemented on July 1st 1984 for all chemical fertilisers) has graduated into a tax (since July 1st 1995). Pre-dating the environmental charge was the introduction of a price regulation charge on nitrogen in commercial fertiliser. After 1995, Finish government introduced voluntary programs for reduced fertilizer use implemented mainly through the Agri-Environmental *EU Regulation 2078/92*. Unlike Finland, Sweden did not abolish their fertiliser tax upon accession in 1995. The government saw the tax as a good way to finance environmental projects. Apart from that, the tax as such was expected to have a positive effect on the environment. From that time on, the tax on fertiliser has been equivalent to about 20% of the price of fertiliser or 0.27€/kg. In 1992, an evaluation was carried out by the Swedish government on the effect of chemical fertilizer use in agriculture. This evaluation suggested that the tax had some impact on the use of chemical fertilizer, and thus directly on nitrate and phosphate emissions to water, but the main effect was indirect through the financing of action programs leading to a decrease in intensive use. This is an important finding considering the inefficient use of fertilizers by the synthetic farms in Thessaly.

In Austria, the government implemented a tax system on fertilizers in 1986. The tax was abolished in 1994, when Austria also joined the EU. There were no alternative/replacement policy instruments implemented even since. Austrian authorities found that after the introduction of the tax, there was a decrease in the use of nitrogen fertilisers of about 15% during 1986, following the same trend over the next years, indicating that farmers had revised their nitrogen application plans due to government intervention. The tax rate until 1991 was 0.25€/kg increased after that year

to 0.47€/kg very close to the Swedish value. Finally, in Netherlands the government introduced a levy system at a national level called *Minas* on the nitrogen and the phosphate surplus in 1998 to reduce emissions. In summary, certain levels of nitrogen surplus are allowed (the levy-free surplus), and these are lowered over time. For the surplus above this level the farmer has to pay a levy. The tax rates in 2003 was set to 2.3€/kg. According to the Dutch government the levy system brought approximately a 20% reduction in the use of nitrogen fertiliser during the period 1998-2006.

However, in order to introduce optimal policy instruments for environmental regulation, one should first introduce social values for the pollutants and then to take into account the complex nature of nitrate leaching from agricultural activities. The standard social planning problem is to maximize consumer plus producer surplus, using demand functions for the desirable outputs, and given (positive) input prices, to calculate input cost. Usually, pollutants are evaluated through a monetized damage function of the form $D = h\left(Q^k\right)$, where k stands for the k^{th} pollutant, for which it holds that $\frac{\partial h(\cdot)}{\partial Q^k} \geq 0$.[32] The damage function is a typical relationship that is used in environmental economics to capture the consumers' willingness to pay for environmental qualities. The latter are not assumed to have intrinsic values, but are evaluated by the man-made goods that are given up to achieve certain environmental qualities. Assuming competitive farm output prices and that existing variable-input prices w_m^v are capturing their social evaluation, the social planning problem is:

$$\begin{aligned} \max_{x^v, x^q, x^w} \Pi &= p_y y - \sum_m w_m^v x_m^v - w^q x^q - w^w x^w - D \\ s.t. \quad y &= f\left(x^v, x^q, x^w\right)\theta^y, \quad D = h(Q^N), \\ Q^N &= \sum_i q_i \quad \text{and} \quad q_i = g\left(x^q, x^w, s, r\right)\theta^q \end{aligned}$$

Assuming interior solutions for all inputs, the necessary first-order condition with respect to chemical fertilizers is:

$$p_y \frac{\partial f(\cdot)}{\partial x^q} = w^q + \frac{\partial h(\cdot)}{\partial Q^N}\frac{\partial g(\cdot)}{\partial x^q}\theta^q$$

where the second term in the right-hand side can be interpreted as the optimal input base *Pigouvian* tax for the pollutant that generates non-point source pollution. Since we have estimated $g(\cdot)$ we only need a proxy for $\frac{\partial h(\cdot)}{\partial Q^N}$. To overcome this problem we use the marginal monetary damage estimated by Keeler et al. (2016) in their study on the social cost of nitrogen in Minesota. Specifically, Keeler et al. (2016) measured the social cost of NO_3^- as the present value of the monetary damages caused by an incremental increase in emitted nitrogen arising from chemical fertilizer application. Obviously, the social cost of NO_3^- is not uniform across space and time. Instead, changes in

[32] A by-product becomes a pollutant when the partial derivative of the damage function turns positive.

management practices will result in different nitrogen-related costs depending on where NO_3^- moves and the location, vulnerability, and preferences of populations affected. Hence, the range reported by the authors of the estimated social cost of NO_3^- varies from 0.0 to 0.23 \$/kg of fertilizers across the state. In our study, assuming that the level of environmental quality required does not differ considerably among regions, we adopt the mean value reported by Keeler et al. (2016) to serve as a benchmark for policy purposes.

The results are depicted in Figure 7, where individual tax rates are presented per chemical fertilizer application rate by the synthetic farmers. Specifically, the *Pigouvian* tax ranges from a minimum of 0.0712 to a maximum of 0.1732€/kg (approximately, 60% increase between the lower and the maximum value). As expected, it follows an increasing trend with fertilizer application by synthetic farms, exhibiting however a rapid burst in farms using more than 32 tonnes of chemical fertilizers in their plots. Next, Table 5 presents tax rates and revenues per both fertilizer and irrigation water application rates for the synthetic farmers. As it is shown from the Table, on the average tax rate increases from 0.0918€/kg of chemical fertilizer to 0.1365€/kg for farmers using more than 30 tonnes of fertlizers. This change implies approximately a 32% increase in tax rate between different users. This difference is not negligible and underlines the need of a progressive rather than a constant nitrate tax rate.

Concerning irrigation water use, the increased trend is also evident but to a lesser extent. Specifically, on the average individual tax rate increases from 0.1029€/kg for users with less than 4.5 ths of m^3 to just 0.1153€/kg for those users applying more than 11 ths of m^3 in their fields. This difference accounts only for the 10% of the tax rate, three times lower than fertilizer application. Concerning the tax revenues collected by the local government, these are 1,613,291€ in total, depending on irrigation water and chemical fertilizer use, ranging from just 1.08€ to 1,558€ for individual farmers. It is interesting the fact that if local government uses these tax revenues to finance policy initiatives to improve farmers' know-how concerning the nitrate pollution generating technology, taxes paid by individual farmers would be considerably less reaching a maximum of only 735.3€. Calculating average values across municipalities it is obvious that the uniform tax rate does not reflect the actual social cost incurred by individual farmers throughout the region of Thessaly. As it shown from Table 6, tax rates vary from a minimum of 0.0862€/kg in Kalampaka to a maximum of 0.1543€/kg in Sofades. This variability does not coincide with taxes paid on the average by farmers in each municipality indicating that the uniform mean tax rate does reflect accurately local conditions. These findings imply that the uniform mean tax adopted by some European countries in the past does not capture adequately damages caused in the region by the use of chemical fertilizers.

Concluding Remarks

In this paper we develop a novel synthetic population generation scheme to approximate more accurately farm nitrate leaching levels in the Greek region of Thessaly. The model is based on the estimation of the joint distribution of the variables characterizing farm structures using a Bayesian network learning together with non-parametric regression models. Synthetic farm data were constructed using the FADN dataset that describes detail farm structures. Then we adapted the GME approach suggested by Kaplan et al. (2003) which is incorporated into a specific theoretical structure describing both crop production technology and nature's nitrate residual mechanism based on the multiple production relations model developed by Murty et al. (2012). The model assumes a specific parametric structure of both technologies using an extensive soil science literature and the model suggested by Knapp and Schwabe (2008) and Wang and Baerenklau (2014). Using this complex modeling structure we were able to convert the NPS pollution problem into a PS one approximating individual nitrate leaching levels for the synthetic population of farmers in the region

Our empirical results provide a good proxy of the unified synthetic farm technology accommodating appropriately both crop production and nature's nitrogen residual generating mechanism. Individual nitrate leaching levels in the region vary from a minimum of 0.0001 mg NO_3^- l^{-1} to a maximum of 0.8i46 mg NO_3^- l^{-1}. Farms in the sample, belonging to the lowest and highest profit quartiles, pollute the underground water resources more than the other quantiles, indicating the group of farmers where appropriate policy measures should be directed toward. However, good farming practices are observed among large farms that can be used as a benchmark to lessen nitrate leaching levels in the area. Using these estimates we also calculated individual tax rates which turn to be different according to fertilizer use and more importantly vary considerably across region reflecting differences in soil, environmental and climate conditions. The results suggest that a uniform tax rate as applied by several European countries is not appropriate and it should be rather determined according to both fertilizer and irrigation water application.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B 44*(2), 139–177.

Brink, C., H. van Grinsven, B. H. Jacobsen, and G. Velthof (2011). Costs and benefits of nitrogen in the environment. In *The European Nitrogen Assessment, M.A. Sutton et al., Eds.*, pp. 513–540. Cambridge University Press.

Casati, D., K. Müller, P. J. Fourie, A. Erath, and K. W. Axhausen (2015). Synthetic opulation generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking. *Transportation Research Record 2493*(1), 107–116.

Chapuis, K. and P. Taillandier (2019). A brief review of synthetic population generation practices in agent-based social simulation. In *SSC2019, Social Simulation Conference*.

Deeva, I., P. D. Andriushchenko, A. V. Kalyuzhnaya, and A. V. Boukhanovsky (2020). Bayesian networks-based personal data synthesis. In *Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good*, pp. 6–11.

Erisman, J., J. Galloway, S. Seitzinger, A. Bleeker, N. Dise, A. Petrescu, A. Leach, and W. de Vries (2013). Consequences of human modification of the global nitrogen cycle. *Philosophical Transactions of the Royal Society B: Biological Sciences 368*, 20130116.

Eurostat (2018). *Agri-Environmental Indicator: Nitrate Pollution of Water, Statistics Explained*. Eurostat Publications.

Farzin, Y. and J. Kaplan (2004). Nonpoint source pollution control under incomplete and costly information. *Environmental and Resource Economics 28*, 489–506.

Førsund, F. (2009). Good modelling of bad outputs: Pollution and multiple output production. *International Review of Environmental and Resource Economics 3*, 1–38.

Førsund, F. (2018). Multi-equation modelling of desirable and undesirable outputs satisfying the materials balance. *Empirical Economics 54*, 67–99.

Frisch, R. (1965). *Theory of production*. D. Reidel Publishing Company.

Galloway, J., A. Townsend, J. Erisman, M. Bekunda, Z. Cai, J. Freney, L. Martinelli, S. Seitzinger, and M. Sutton (2008). Transformation of the nitrogen cycle: recent trends, questions, and potential solutions. *Science 320*, 889–892.

Gargiulo, F., S. Ternes, S. Huet, and G. Deffuant (2010). An iterative approach for generating statistically realistic populations of households. *PLOS One 5*(1), e8828.

Golan, A., G. Judge, and D. Miller (1996). *Maximum entropy econometrics: Robust estimation with limited data*. John Wiley and Sons.

Golan, A. and J. Perloff (2002). Comparison of maximum entropy and higher-order entropy estimators. *Journal of Econometrics 107*(1-2), 195–211.

Gu, B., J. Ju, X. Chang, Y. Ge, and P. M. Vitousek (2015). Integrated reactive nitrogen budgets and future trends in China. *PNAS 112*, 8792–97.

Hansen, E. and J. Djurhuus (1996). Nitrate leaching as affected by long-term N fertilization. *Soil Use Management 12*, 199–204.

Ilahi, A. and K. W. Axhausen (2019). Integrating Bayesian network and generalized raking for population synthesis in Greater Jakarta. *Regional Studies, Regional Science 6*(1), 623–636.

Kaplan, J., R. Howitt, and Y. Farzin (2003). An information-theoretical analysis of budget-constrained nonpoint source pollution control. *Journal of Environmental Economics and Management 46*, 106–130.

Keeler, B., J. Gourevitch, S. Polasky, F. Isbell, C. Tessum, J. Hill, and J. Marshall (2016). The social costs of nitrogen. *Science Advances 2*, e1600219.

Knapp, K. and K. Schwabe (2008). Spatial dynamics of water and nitrogen management in irrigated agriculture. *American Journal of Agricultural Economics 90*(2), 524–539.

Kocabas, V. and S. Dragicevic (2009). Agent-based model validation using Bayesian networks and vector spatial data. *Environment and Planning B: Planning and Design 36*(5), 787–801.

Kocabas, V. and S. Dragicevic (2013). Bayesian networks and agent-based modeling approach for urban land-use and population density change: a BNAS model. *Journal of Geographical Systems 15*(4), 403–426.

Lam, W. and F. Bacchus (1994). Learning Bayesian belief networks: An approach based on the MDL principle. *Computational Intelligence 10*(3), 269–293.

Lenormand, M. and G. Deffuant (2013). Generating a synthetic population of individuals in households: Sample-free vs sample-based methods. *Journal of Artificial Societies and Social Simulation 16*(4), 12.

Murty, S. (2010). Externalities and fundamental nonconvexities. A reconciliation of approaches to general equilibrium externality modeling and implications for decentralization. *Journal of Economic Theory 145*, 331–353.

Murty, S., R. Russell, and S. Levkoff (2012). On modeling pollution-generation technologies. *Journal of Environmental Economics and Management 64*, 117–135.

Neapolitan, R. E. (2003). *Learning Bayesian networks*. Pearson Prentice Hall Upper Saddle River, NJ.

Owen, A. B. (2001). *Empirical likelihood*. CRC press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible reasoning*. Morgan Kaufmann Publishers, Los Altos.

Ramadan, O. E. and V. P. Sisiopiku (2019). A critical review on population synthesis for activity- and agent-based transportation models. In *Transportation Systems Analysis and Assessment*. IntechOpen London.

Sebastiani, P. and M. Ramoni (2001). On the use of Bayesian networks to analyze survey data. *Research in Official Statistics 4*(1), 53–64.

Segerson, K. (1988). Uncertainty and incentives for nonpoint pollution control. *Journal of Environmental Economics and Management 15*, 87–98.

Shortle, J. and R. Horan (2001). The economics of nonpoint source pollution. *Journal of Economic Surveys 15*, 255–89.

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.

Sobota, D., J. Compton, M. McCrackin, and S. Singh (2015). Cost of reactive nitrogen release from human activities to the environment in the United States. *Environmental Research Letters 10*, 25006–18.

Spirtes, P., C. N. Glymour, and R. Scheines (2000). *Causation, prediction, and search*. MIT press.

Sun, L. and A. Erath (2015). A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies 61*, 49–62.

Székely, G. J., M. L. Rizzo, et al. (2004). Testing for equal distributions in high dimension. *InterStat 5*(16.10), 1249–1272.

Tsagris, M., S. Preston, and A. Wood (2011). A data-based power transformation for compositional data. In *Proceedings of the 4rth Compositional Data Analysis Workshop, Girona, Spain.*

Tsagris, M. and V. Tzouvelekas (2022). Nitrate leaching and efficiency measurement in intensive farming systems: A parametric by-production technology approach. *Agricultural Economics 53*(4), 633–647.

Tsamardinos, I., C. F. Aliferis, and A. Statnikov (2003). Time and sample efficient discovery of Markov blankets and direct causal relations. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 673–678. ACM.

Tsamardinos, I. and L. E. Brown (2008). Bounding the false discovery rate in local Bayesian network learning. In *Association for the Advancement of Artificial Intelligence Conference*, pp. 1100–1105.

Tsamardinos, I., L. E. Brown, and C. F. Aliferis (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning 65*(1), 31–78.

Wang, J. and K. Baerenklau (2014). Crop response functions integrating water, nitrogen, and salinity. *Agricultural Water Management 139*, 17–30.

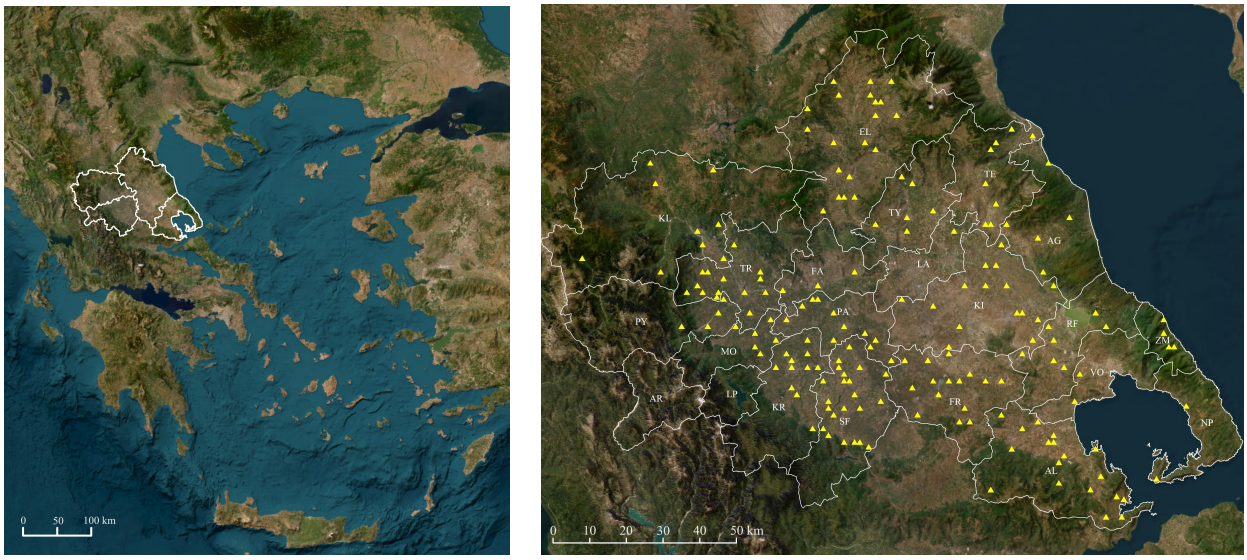
Xepapadeas, A. (2011). The economics of non-point-source pollution. *Annual Review of Resource Economics 3*, 355–73.

Ye, X., K. Konduri, R. M. Pendyala, B. Sana, and P. Waddell (2009). A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In *88th Annual Meeting of the Transportation Research Board, Washington, DC.*

Young, J., P. Graham, and R. Penny (2009). Using Bayesian networks to create synthetic data. *Journal of Official Statistics 25*(4), 549.

Zhang, J., G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao (2017). Privbayes: Private data release via Bayesian networks. *ACM Transactions on Database Systems (TODS) 42*(4), 1–41.

Tables and Figures



(a) (b)

Figure 1: (a) Map of Greece with the region of Thessaly highlighted. (b) Map of Thessaly with the coordinates of the surveyed farms in FADN database.

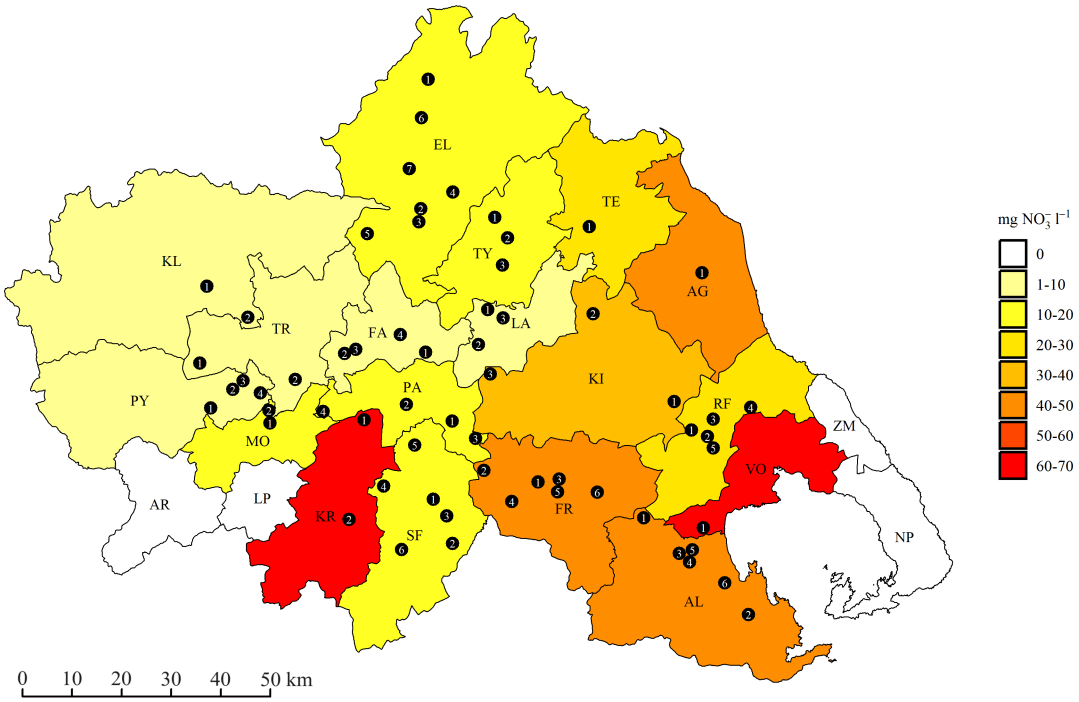


Figure 2: Map of Thessaly with the coordinates of the water reservoirs. Coloured areas indicate differences in the estimated total nitrate leaching from the synthetic population of farmers in each municipality (coding of municipalities appears in Table 1 next). Stock levels per water reservoir are presented in Table 1.

Table 1: Nitrate Stock in the Water Reservoirs of Thessaly

Municipality	Code	mg NO_3^- l^{-1}	Municipality	Code	mg NO_3^- l^{-1}
Agia	AG1	40.9	Larisa	LA1	4.0
Almyros	AL1	8.4	Larisa	LA2	5.5
Almyros	AL2	59.7	Larisa	LA3	4.0
Almyros	AL3	70.4	Mouzaki	MO1	12.6
Almyros	AL4	107.7	Mouzaki	MO2	13.3
Almyros	AL5	20.0	Palamas	PA1	7.7
Almyros	AL6	3.1	Palamas	PA2	20.3
Elassona	EL1	16.0	Palamas	PA3	13.8
Elassona	EL2	12.4	Palamas	PA4	10.2
Elassona	EL3	42.5	Pyli	PY1	0.5
Elassona	EL4	17.7	Pyli	PY2	0.5
Elassona	EL5	0.5	Pyli	PY3	13.0
Elassona	EL6	17.7	Pyli	PY4	9.3
Elassona	EL7	3.4	Rigas Ferraios	RF1	68.0
Farkadona	FA1	7.7	Rigas Ferraios	RF2	30.4
Farkadona	FA2	3.2	Rigas Ferraios	RF3	22.4
Farkadona	FA3	15.0	Rigas Ferraios	RF4	0.5
Farkadona	FA4	11.7	Rigas Ferraios	RF5	18.5
Farsala	FR1	30.7	Sofades	SF1	77.1
Farsala	FR2	17.7	Sofades	SF2	3.3
Farsala	FR3	96.5	Sofades	SF3	8.3
Farsala	FR4	31.8	Sofades	SF4	9.2
Farsala	FR5	38.1	Sofades	SF5	12.9
Farsala	FR6	58.1	Sofades	SF6	5.3
Kileler	KI1	21.4	Tempi	TE1	20.6
Kileler	KI2	26.9	Trikala	TR1	1.9
Kileler	KI3	63.4	Trikala	TR2	1.9
Kalampaka	KL1	1.3	Tyrnavos	TY1	9.8
Kalampaka	KL2	16.9	Tyrnavos	TY2	11.7
Karditsa	KR1	123.0	Tyrnavos	TY3	10.3
Karditsa	KR2	13.5	Volos	VO1	61.0

Source: Greek Ministry of Agriculture.

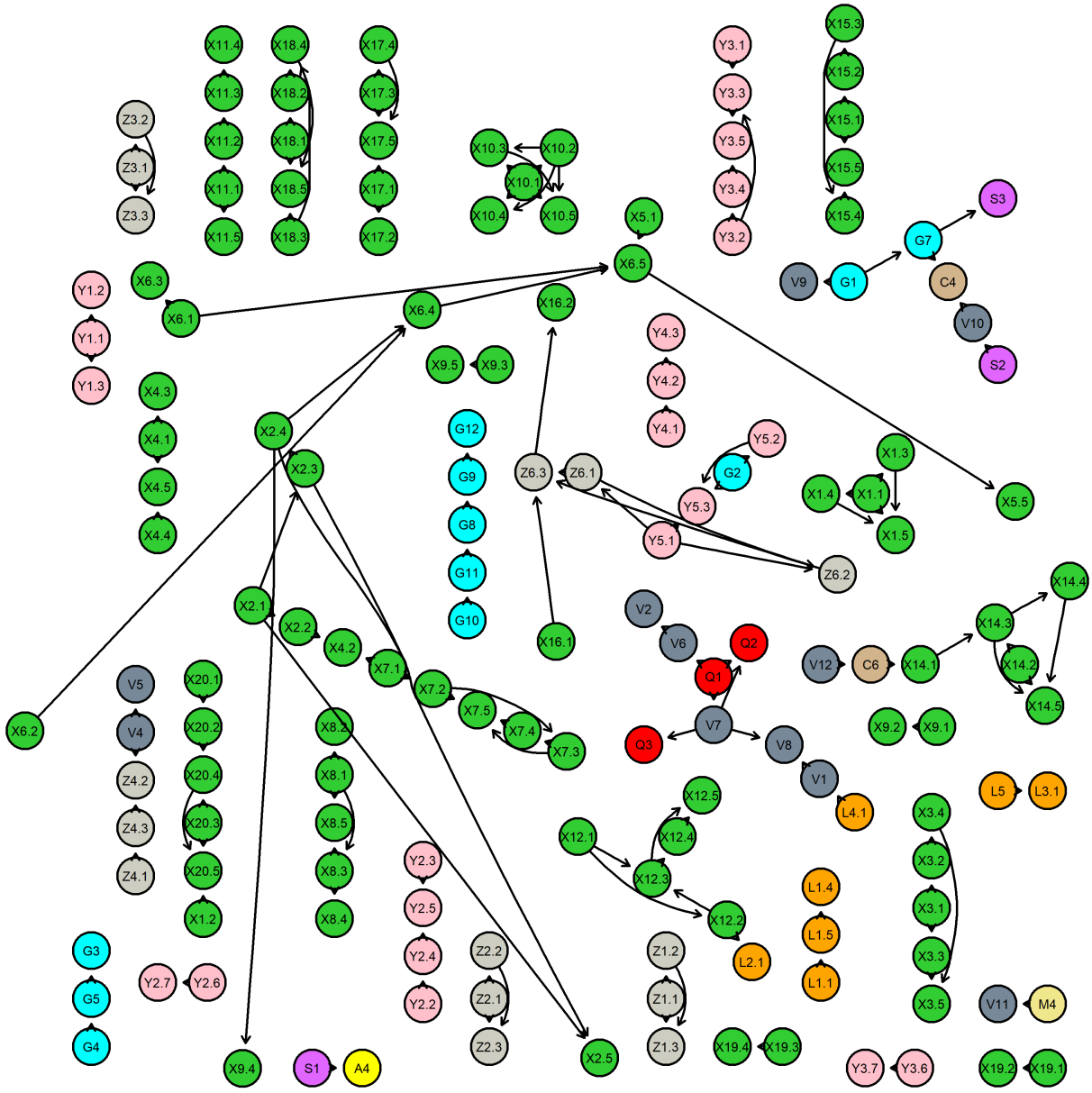


Figure 3: The BN structure of Thessaly (for the definition of the variables see Tables B2-B11 in Appendix B).

Table 2: Descriptive Statistics of the Synthetic Population of Farmers in Thessaly

Variable	Mean	Min	Max	St. Dev.
Farm Production (in €)	11,761	771	158,728	10,545
Land (in stremmas. 1 stremma equals 0.1 ha)	122	4.2	1,486	107
Labour (in hrs)	2,283	615	35,655	1,070
Irrigation Water (in m^3)	668	8	5093	667
Chemical Fertilizers (in Kgs)	3,901	17	33,829	3,653
Precipitation (in mm)	266	240	347	44
Slope (0%-70.2%. 100% is horizontal line)	21	4	61	10
Soil Erosion (% of land downgraded. Values 0-50.8)	5.862	0.044	41.22	9.121
Nitrogen levels in the reservoirs (in mg NO_3^- l^{-1})	24.69	0.500	123	23.79
No of water reservoirs		62		
No of surveyed farms		509		
No of farms in the synthetic population		34,853		

Table 3: Parameter Estimates of the Translog Production and Nitrate Leaching Frontiers for the Synthetic Population of Farmers in Thessaly

Parameter	Estimate	Std Error	Parameter	Estimate	Std Error
Crop Production Frontier			Nitrate Leaching Frontier		
β_0	0.3046	0.0685	δ_0^q	0.1000	2.9798
β_A^v	0.1966	0.0745	δ_R^q	0.0690	0.0540
β_L^v	0.1368	0.0525	δ_S^q	0.0385	0.0232
β^q	0.1467	0.1289	δ_E^q	0.0049	0.0342
β^w	0.1268	0.0870	δ_0^{qq}	1.3536	0.6759
β_{AA}^{vv}	-0.0227	0.0098	δ_R^{qq}	0.1879	0.1082
β_{LL}^{vv}	-0.0106	0.0007	δ_S^{qq}	0.1749	0.0687
β^{qq}	-0.0101	0.0048	δ_E^{qq}	0.1773	0.1606
β^{ww}	-0.0103	0.0018	δ_0^w	0.0705	0.0378
β_{AL}^{vv}	-0.0105	0.0017	δ_R^w	0.1499	0.0021
β_A^{qv}	0.0107	0.0041	δ_S^w	0.0489	0.0008
β_L^{qv}	-0.0109	0.0032	δ_E^w	0.0661	0.0093
β^{qw}	-0.0296	0.0029			
β_A^{wv}	0.0108	0.0104			
β_L^{wv}	-0.0296	0.0223			
Synthetic Population:			34,853 farms		

where A stands for area, L for labour, R for precipitation level, S for the slope of the land and, E for soil erosion. The corresponding standard errors are obtained using non-parametric bootstrap.

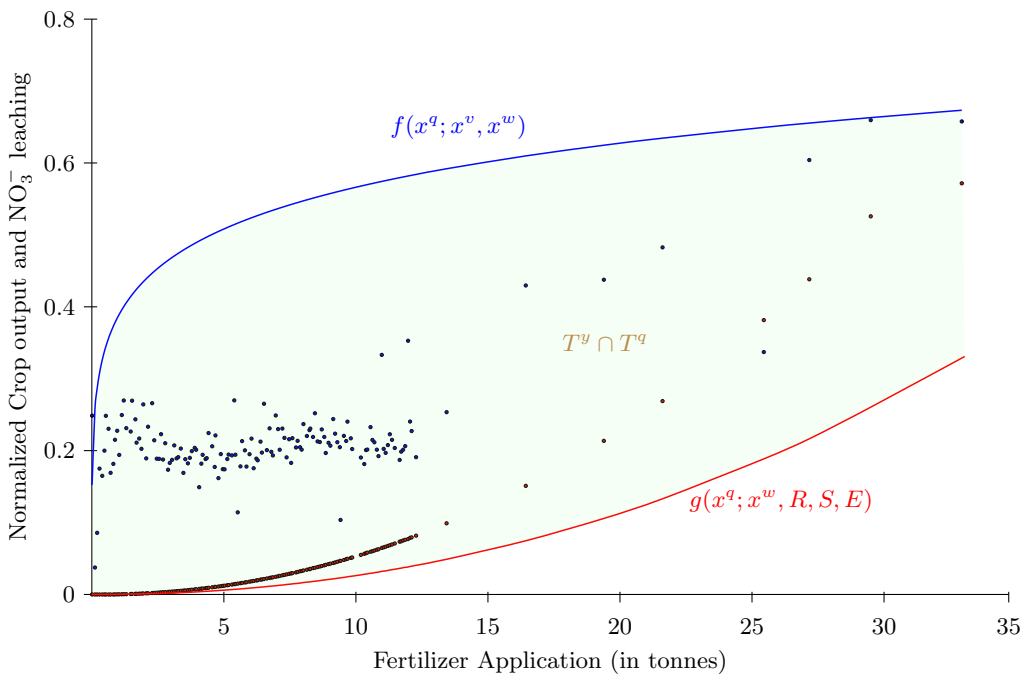


Figure 4: Production Frontier and Nitrate Leaching Functions

Table 4: Crop Output and Nitrate Leaching Elasticities, Returns-to-Scale and Farm Size per Nitrate Leaching Quartile for the Synthetic Farms

	Crop Output Elasticities					Leaching Elasticities		Farm Size	Efficiency in	
	Area	Labour	Fertilizers	Water	RTS	Fertilizers	Water	(in str)	Production	Leaching
1st Q	0.3101	0.3628	0.3050	0.3725	1.3504	0.0510	0.0001	72.9	52.86	66.63
2nd Q	0.3106	0.3485	0.2830	0.3396	1.2817	0.0523	0.0016	88.4	50.36	65.61
3rd Q	0.3003	0.3381	0.2733	0.3240	1.2357	0.0543	0.0076	125.6	49.17	65.21
4th Q	0.2815	0.3292	0.2682	0.3180	1.1970	0.0585	0.0393	203.0	48.78	64.17
Mean	0.3006	0.3447	0.2824	0.3385	1.2662	0.0540	0.0122	122.5	50.29	65.68
StdDev	0.0449	0.0422	0.0429	0.0403	0.1183	0.0037	0.0624	106.9	6.37	0.52
Median	0.2961	0.3381	0.2746	0.3337	1.2551	0.0529	0.0022	88.8	49.49	65.36

Elasticities are computed at the mean values of all exogenous variables and distortion parameters.

Table 5: Tax Rates and Revenues for the Synthetic Farms According to Water and Fertilizer Use

Water	Fertilizers (in tonnes)						
(in ths of m^3)	≤10	10-15	15-20	20-25	25-30	>30	Mean
Tax Rate (in €/kg):							
≤4.5	0.0856	0.0912	0.0970	0.1042	0.1125	0.1270	0.1029
4.5-11	0.0915	0.0976	0.1037	0.1104	0.1197	0.1371	0.1100
>11	0.0984	0.1035	0.1071	0.1139	0.1237	0.1454	0.1153
Mean	0.0918	0.0974	0.1026	0.1095	0.1186	0.1365	0.1094
Average Tax Paid under leaching inefficiency (in €/farm):							
≤4.5	3.6	10.0	21.0	37.4	74.7	148.1	49.1
4.5-11	3.6	10.7	22.5	39.7	79.1	161.6	52.9
>11	4.1	11.3	23.2	41.0	81.2	176.1	56.2
Mean	3.7	10.7	22.2	39.4	78.4	161.9	52.7
Total revenues (in €):	1,613,291		Min:	1.08	Max:	1,558	
Average Tax Paid under leaching efficiency (in €/farm):							
≤4.5	1.9	5.4	10.9	18.9	38.3	75.8	25.2
4.5-11	1.8	5.4	10.9	19.1	38.2	78.1	25.6
>11	2.0	5.5	11.0	19.2	38.0	83.4	26.5
Mean	1.9	5.4	10.9	19.1	38.1	79.1	25.8
Total revenues (in €):	788,674		Min:	0.08	Max:	735.3	

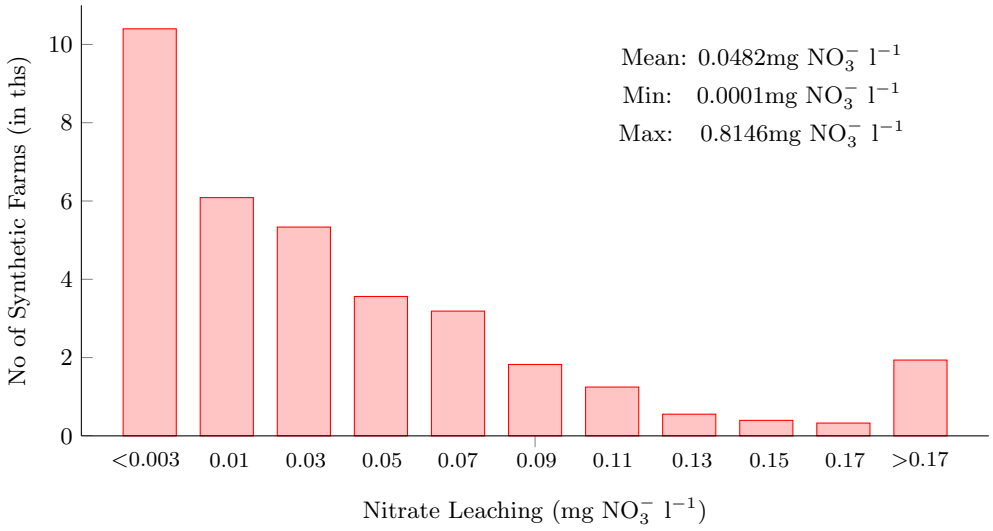


Figure 5: Frequency Distribution of Individual Nitrate Leaching Levels

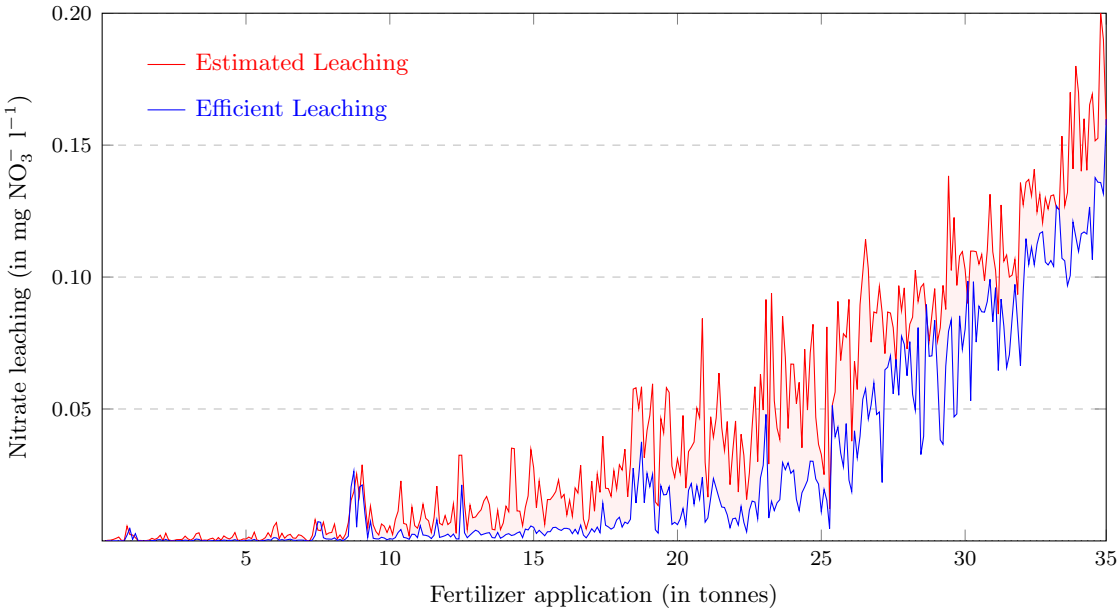


Figure 6: Estimated and Efficient Individual Nitrate Leaching Levels

Table 6: Average Tax Rates and Revenues in the Different Municipalities of Thessaly

Municipality	Tax Rate	Tax Paid	Municipality	Tax Rate	Tax Paid	Municipality	Tax Rate	Tax Paid
Agia	0.1099	49.3	Almyros	0.1424	28.4	Elassona	0.1087	45.4
Farkadona	0.0892	53.5	Farsala	0.1103	117.3	Kileler	0.1348	64.2
Kalampaka	0.0862	32.7	Karditsa	0.1359	101.0	Larisa	0.0987	71.7
Mouzaki	0.0955	67.8	Palamas	0.1093	68.0	Pyli	0.0875	35.4
Rigas Ferraios	0.1236	15	Sofades	0.1543	78.9	Tempi	0.1078	32.0
Trikala	0.0887	48.3	Tyrnavos	0.0966	28.4	Volos	0.0921	14.6

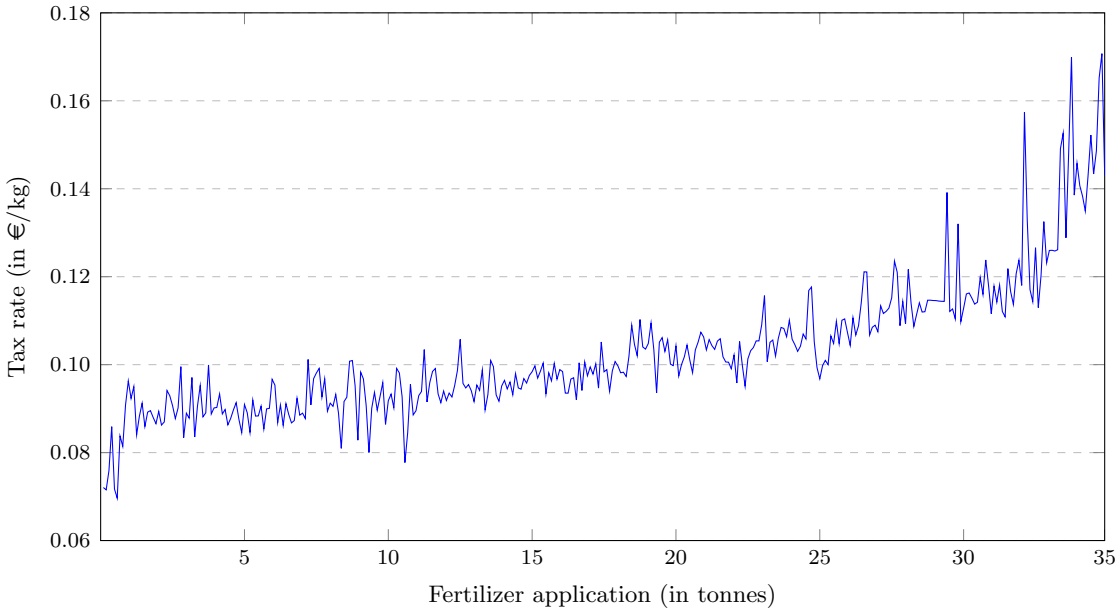


Figure 7: Tax Rate per Fertilizer Application