Transient and persistent energy efficiency in the wastewater sector based on economic foundations

Stefano Longo[1, 2, 3, *], Mona Chitnis[2], Miguel Mauricio-Iglesias[1], Almudena Hospido[1]

[1] *Department of Chemical Engineering, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain*

[2] *Surrey Energy Economics Centre (SEEC), School of Economics, University of Surrey, Guildford, GU2 7XH, UK*

[3] *Gruppo Hera S.p.A., Bologna, Italy*

*Corresponding author (E-mail: stefano.longo@usc.es, stefano.longo@gruppohera.it)

Abstract Given the increasing importance of the wastewater sector in terms of energy usage, the understanding of the level of energy efficiency of wastewater treatment plants (WWTPs) is useful to both the industry itself as well as policy makers. Here, based on economic foundations, we apply a Stochastic Frontier Analysis (SFA) approach for energy demand modelling to estimate energy efficiency in the wastewater sector. Using specific SFA models and panel data from 183 Swiss WWTPs over the period 2001 to 2015, the paper illustrates that distinguishing between persistent and transient inefficiency is essential to deduce appropriate energy efficiency diagnosis in WWTPs. In this respect, persistent energy inefficiency is found to be more severe than transient energy inefficiency. Furthermore, it is shown that the age of the equipment influences the demand for energy and the energy savings due to technological innovation are quantified. Finally, economies of output density and scale are estimated demonstrating that for plants operating below optimal scale significant energy savings can be achieved if plants would be operated at higher size. Instead, our analysis reveals also that for plants larger than 100000 Population Equivalent, at least from an energy efficiency point of view, it would be no more beneficial to increase their scale.

Keywords Stochastic frontier analysis; energy efficiency; energy demand; benchmarking; wastewater treatment; water diagnosis

1 *Introduction*

Energy used to supply water to consumers is a source of demand that is set to grow rapidly over the coming decades. At European level, the data show values of around 90 kWh/year/person (EEA, 2014), or that for each person a 10 W light bulb is burning 24/7. Today, 4% of the global electricity consumption is used in the water sector, and over the period to 2040, this amount is projected to more than double (IEA, 2016). Around 30% of the electricity consumed for water, is used for wastewater treatment (EBC, 2016). Hence, energy-related considerations in the wastewater sector will receive increasing attention from an environmental and economic point of view, as also announced by the European Commission in the proposal for a revision of the European Union drinking water directive (EC, 2018). Assessing and improving energy efficiency is a valuable means to address these challenges (Huntington and Smith, 2011).

Recently, the first methodology specifically tailored to estimate energy efficiency at wastewater treatment plants (WWTPs) has been described (Longo et al., 2019). The methodology, for the first time, provides engineers, wastewater operators and decision-makers a method to obtain standardized and comparable efficiency information. However, a widespread concern in the wastewater sector is the extent to which efficiency estimates ignore external influences on performance (Guerrini et al., 2016). A development in Data Envelopment Analysis (DEA) application is the Robust Energy Efficiency DEA (REED), a methodology to systematize the inclusion of exogenous factors and to robustly estimate efficiency of WWTP (Longo et al., 2018). Although DEA is attractive as it easily handles multiple inputs/outputs and it does not require the specification of a functional form to define an efficiency frontier, a major drawback is that it attributes all deviations from the frontier to inefficiency. Yet, deviations from the frontier may be due to a number of factors other than inefficiency such as omitted variables and measurement

errors. Stochastic Frontier Analysis (SFA) represents an interesting framework to overcome the above limitations since it is able to separate the inefficient component from the statistical noise due to data errors and omitted variables (Boyd, 2008).

From an economic point of view, there exist a number of reasons why econometric techniques such as SFA can be useful apart from providing sound efficiency indicators. Even within a relatively non-competitive industry, such as the water industry, there may be public policy questions that could be considered to improve the operation. In particular, the ability to accommodate formal statistical testing makes econometric techniques such as SFA more attractive for both regulators and operators (Leth-Petersen, 2002). For example, knowing economic relationships between pollutants removed from water and energy demand is fundamental to understand/predict the trade-off between pollution control and energy footprint.

At micro-level, there are numerous critical questions that would benefit from this sort of analysis. For example, a key strategic question may be whether or not merging two adjacent WWTPs makes sense. Although there are multiple reasons for considering plants centralization, one of the key questions to answer is whether it will result in cost savings through economies of scale. Moreover, by using specific SFA models and panel data, wherein each plant is observed at different points of time, it is possible to examine whether inefficiency has been persistent over time or whether plant inefficiency is time-varying (Tsionas and Kumbhakar, 2014; Filippini and Greene, 2016). Distinguishing between persistent and transient inefficiency seems to be essential to deduce appropriate energy diagnosis and design useful energy efficiency strategies for WWTPs. Otherwise, water utilities may wrongly decide to invest in new equipment and infrastructure, while inefficiency arises from some applications of wrong operational strategies due to e.g. error in management of sludge age and return sludge, too infrequent sampling or inadequate evaluation of monitoring data, or vice versa. In order to answer to these questions, proper measure of efficiency and the effects of efficiency determinants are important.

Therefore, this study applies a SFA approach for energy demand modelling to determine energy efficiency in the wastewater sector based on economic foundations, for the first time to the best of our knowledge. An energy demand[1] frontier function for the wastewater sector is estimated using a panel data of 183 Swiss WWTPs for the period 2001 to 2015, explicitly controlling for technology, size and the removal of the main contaminants from wastewater, as well as for both observed and unobserved heterogeneity. Once these effects are controlled for, it is possible to better estimate a measure of energy efficiency for each plant and to take corrective energy-saving measures.

The objective of this paper is then to investigate how overall inefficiency of WWTPs can be decomposed into persistent and transient inefficiency. Moreover, in order to find out whether accounting for unobserved heterogeneity in the model significantly influences the results, the energy efficiency estimates obtained from three panel data models are compared, and context-specific considerations on the different models are done to explore their best use. In addition, the impact of technical progress on energy efficiency is studied in order to estimate the margin for improvements. Finally, to determine whether the analysed WWTPs are operated at optimal load and size, economies of output and scale are estimated.

The paper is organized as follows. Section 2 presents an econometric model for WWTPs energy demand and discusses the empirical specifications for estimating the level of efficiency in the use of energy. Section 3 describes data and the variables used in the model. The results of estimations and discussions are presented in section 4 and Section 5 concludes.

2. *Econometric model for WWTPs energy demand*

Wastewater treatment is a process used to produce an effluent that can be returned to the water cycle with minimal impact on the environment. The treatment process takes place in a WWTP, which is organized in different unit operations grouped together to provide various levels of

[1] The reader is informed that although we have used the general term "energy", the analysis in this research refers to electricity consumption at WWTPs, as all energy is in the form of electricity in this study.

treatment known as preliminary, primary, secondary, tertiary and sludge treatment (Metcalf and Eddy, 2003). Based on the function of the plant, WWTPs can produce different outputs, e.g. pumping wastewater, producing an effluent free of contaminants such as solids, chemical oxygen demand (COD), nitrogen (N), phosphorus (P) and pathogens, processing the sludge produced during treatment, recovering of energy and materials. The resources used for treatment process are the inputs, being electricity one of the main inputs in all the cases.

From an economic perspective, WWTPs energy efficiency can be discussed using the microeconomics theory of production framework. In this context, the production function can be described by a mathematical representation of a WWTP that converts input(s) (e.g. electricity) into output(s) (e.g. COD and other nutrients removed from wastewater). Considering the input-oriented nature of the problem, in this study we focus on WWTPs whose objective is to produce a given level of outputs with the minimum possible level of input. Although WWTPs can produce also undesirable outputs such as sewage sludge (which needs to be properly managed), and efficiency measurements (especially in non-parametric contexts) have been extended to include those into account (Färe et al., 2004; Kuosmanen, 2005), given the centrality of the water-energy nexus, the present paper will focus on energy efficiency as one of the priority areas of policy makers and water utilities.

WWTPs may be characterized by operating under particularly heterogeneous environment, e.g. under highly heterogeneous topography. Variation in operating environment that manifests as variation in energy use, if not controlled for, may be misinterpreted as efficiency differences. It is however virtually impossible to observe (or measure) all relevant aspects that may affect energy use at WWTPs. Thus, unless unobserved heterogeneity is properly taken into account the estimated inefficiencies are likely to be biased. This has been a pervasive problem in cross-sectional analysis (Arellano, 2003). If, however, panel data are available, this limitation can be overcome. Utilizing information on both the intertemporal dynamics and the individuality of the entities being investigated, panel data permit to control for the effects of unobserved variables (Hsiao, 2014).

Another important advantage of using panel data over cross-section is that it is possible to think of the inefficiency term as comprised by two components: persistent (i.e. time-invariant) and transient (i.e. time-varying) (Tsionas and Kumbhakar, 2014; Filippini and Greene, 2016). The persistent component is determined by the presence of structural problems such as inefficient equipment or design limitations that do not allow the plant to minimize the use of energy, and the transient component may be caused by the presence of non-systematic difficulties that can be solved in the short term such as adaption of wrong operational strategies due to e.g. too infrequent sampling.

The simplest frontier model that accounts for the stochastic effects, and extended for panel data, can be written as:

$$E_{it} = \alpha_0 + f(X_{it};\beta) + \varepsilon_{it}, \qquad \varepsilon_{it} = \nu_{it} - u_i, \qquad u_i \geq 0, i = 1, \dots N; t = 1, \dots, T, \tag{1}$$

where E denotes the energy consumption, X stands for the vector of explanatory variables influencing energy demand, including plant characteristics and exogenous factors, β is the vector of coefficients and α is the regression constant. Subscripts i and t stand for WWTP and time, respectively. In SFA models, the error term ε is composed by two random terms, ν and u. The first term, ν, is the error term capturing the effect of noise and assumed to be normally distributed. The other component, u, discussed in detail in Section 2.1, is interpreted as an indicator of the inefficient use of energy at the plant level.

Filippini and Hunt (2011) proposed that one way to econometrically estimate a measure of the efficient use of energy is to estimate an input demand frontier function, such as a demand function for energy.[2] In this paper, we use this approach and estimate a measure of energy efficiency based on the estimation of an input demand frontier function, i.e. the WWTPs demand function for energy (Eq. 2). The function represents the minimum or baseline energy demand of a WWTP that has highly efficient equipment, used in the most efficient way, to produce a given level of wastewater

[2] An overview of the theoretical background as well as the empirical methods for measuring the level of energy efficiency based on economic foundations can be found in Filippini and Hunt (2015).

treatment service. If a plant is not on the frontier, the distance from the frontier measures the level of inefficiency in the use of energy.

Apart from few recent attempts (Castellet-Viciano et al., 2018; Molinos-Senante et al., 2018), the literature in energy economics of wastewater sector regarding econometrics modelling enabling the analysis of energy consumption at WWTPs is scarce. Although previous studies have been very useful to understand the impact of some important variables on energy consumption such as the plant age and the volume of wastewater, the models presented in the literature are too simple for reflecting the complex structure of a WWTP. Given the discussion above and following the approach introduced by Filippini and Hunt (2011), and using a log-log functional form, it is assumed that there exists a WWTP energy demand function for panel data, as follows:

$$\ln E_{it} = \alpha_0 + \alpha_P \ln P_t + \alpha_{FLOW} \ln FLOW_{it} + \alpha_{CAP} \ln CAP_i + \alpha_{COD} \ln COD_{it} + \alpha_{NH4} \ln NH4_{it} + \alpha_{NO3} \ln NO3_{it} + \alpha_{TEMP} \ln TEMP_{it} + \sum_{j=1}^{6} \alpha_{TECH_j} TECH_{ij} + \alpha_{DEW} DEW_{it} + \varepsilon_{it}^{3} \quad (2)$$

where E is energy consumption (kWh/day), P is the real price of energy (CHF/kWh), $FLOW$ is the volume of wastewater treated (m^3/day), CAP is the plant capacity expressed as design flow rate (m^3/day), COD, $NH4$ and $NO3$ are the pollutants removed daily from wastewater (mg/L/day), $TEMP$ is the average yearly temperature in degrees Celsius, $TECH$ represents dummy variables to control for the effect of the type of secondary treatment whose value is 1 for plant having technology $j(j = 1,2,3,4,5,6)$ otherwise is 0 if the type of secondary treatment is Conventional Activated Sludge (CAS), DEW is a dummy indicating whether the plant carries out also dewatering of sludge, and ε is the random error term.

The ratio between the observed input and the optimal input demand on the frontier represents inefficiency. The energy efficiency is usually expressed in the following way:

[3] In order to investigate possible nonlinearity effects, a quadratic form of the equation was estimated by inclusion of $FLOW$ squared and CAP squared in the model, as well as an interaction term for $FLOW$ and CAP. As these terms were not significant, they were excluded in the final model.

$$EE_{it} = \frac{E_{it}^F}{E_{it}} \tag{3}$$

where E_{it} is the observed energy consumption and E_{it}^F is the minimum energy demand of plant i in year t on the frontier. An efficiency level of one indicates a plant on the frontier, thereby implying a 100% efficiency level. Plants that are not located on the frontier receive efficiency scores below one, implying the presence of inefficiency in plant energy consumption.

2.1 Estimation methodology

In panel data, unobserved heterogeneity can be taken into account by introducing a firm (unobservable) effect, noted as α_i, which is time-invariant and firm-specific (Wooldridge, 2010). Firm effect can be modelled as fixed effects when it is treated as a parameter to be estimated or as random effects when it is treated as a random variable. When modelling panel data, perhaps the first question the practitioner faces is whether to account for unit effects and, if so, whether to employ fixed effects or random effects. A common approach to answer this question is to employ the Hausman test (Greene, 2003), which is intended to detect violation of the random effects modelling assumption that the explanatory variables are orthogonal to the unit effects. A complete discussion on the choice between these approaches is given in Section 4.1.

In order to find out whether accounting for unobserved heterogeneity in the model significantly influences the results, in this paper the energy efficiency estimates obtained from three SFA models are compared. Our goal here is to investigate three panel data models that have different modelling approaches and different interpretations of the unobserved heterogeneity, as briefly summarized in Table 1.

Table 1. Econometric specifications of the stochastic cost frontier for *Model I*, *Model II* and *Model III*.

	Model I	***Model II***	***Model III***

Firm specific (random) effects α_i	α_i	α_i	$a_i = \mu_i - E(\eta_i)$
Full random error term ε_{it}	$\varepsilon_{it} = \nu_{it} + \alpha_i$	$\varepsilon_{it} = \nu_{it} - u_i - \tau_{it}$	$\varepsilon_{it} = \mu_i + \nu_{it} - \eta_i - \tau_{it}$
	$\varepsilon_{it} \sim i.i.d.(0, \sigma_\varepsilon^2)$	$\nu_{it} \sim i.i.d.N(0, \sigma_\nu^2)$	$\nu_{it} \sim i.i.d.N(0, \sigma_\nu^2)$
		$\tau_{it} \sim i.i.d.N(0, \sigma_\tau^2)$	$\tau_{it} \sim i.i.d.N^+(0, \sigma_\tau^2)$
			$\mu_i \sim i.i.d.N(0, \sigma_\mu^2)$
			$\eta_i \sim i.i.d.N^+(0, \sigma_\eta^2)$
Persistent inefficiency estimator	$u_i = \hat{\alpha}_i - \min_i\{\hat{\alpha}_i\}$	$u_i = \hat{\alpha}_i - \min_i\{\hat{\alpha}_i\}$	$\mathrm{E}(\eta_i \mid \varepsilon_{it})$
Transient inefficiency estimator	None	$\mathrm{E}(\tau_{it} \mid \varepsilon_{it})$	$\mathrm{E}(\tau_{it} \mid \varepsilon_{it})$

In *Model I* we consider the classic random-effects model proposed by Schmidt and Sickles (1984) :

$$E_{it} = \alpha_0 + f(X_{it}; \beta) + \nu_{it} + \alpha_i.^4 \quad (4)$$

The model in Eq. (4) can be estimated by the generalized least squares (GLS) technique (Kumbhakar, Wang, and Horncastle, 2015). *Model I* assumes efficiency to be plant-specific and time-invariant. That is, the efficiency levels may be different for different plants, but they do not change over time. Even if it may be a plausible assumption in non-competitive operating environment like the wastewater sector, it can be a rather limiting assumption, particularly in long panels.

From an engineering point of view, two cases where WWTPs are producing a wastewater treatment service without minimizing the use of energy can be identified: i) a plant employing modern equipment is utilizing energy in an inefficient way and ii) a plant is using relatively old equipment that does not allow the plant to minimize the use of energy. In both cases, the measures of energy efficiency represent a "waste" of energy. One, which is time-variant, associated e.g. with the efficient use of the equipment for the wastewater treatment, will depend on the technical skills of operators. The second, which is time-invariant, associated e.g. with the equipment used in the plant, will depend on the different level of technological innovation. A model that separates persistent and time-varying inefficiency can be formalized as follows:

[4] All variables except TECH and DEW are in natural logs.

$$E_{it} = \alpha_0 + f(X_{it};\beta) + \nu_{it} - u_i - \tau_{it}.^{5} \quad (5)$$

In Model II (Eq. 5) the error term ε_{it} is decomposed as $\varepsilon_{it} = \nu_{it} - u_{it}$, where u_{it} is the inefficiency and ν_{it} is statistical noise. The inefficiency part is further decomposed as $u_{it} = u_i + \tau_{it}$ where u_i is the persistent component and τ_{it} is transient component of inefficiency. The former is only plant-specific, while the latter is both plant- and time-specific. Estimation of *Model II* can be undertaken following a multistep procedure (Kumbhakar and Heshmati, 1995) and briefly summarised here for completeness: in step 1, consistent estimates of β are obtained using a standard random-effects panel data model (i.e. *Model I*); in step 2, persistent efficiency (PE) is estimated using the pseudo residual obtained in step 1; the parameters associated with the random components, ν_{it} and τ_{it}, are estimated in step 3 by maximum likelihood and assuming $\nu_{it} \sim i.i.d.\ N(0, \sigma_\nu^2)$ and $\tau_{it} \sim i.i.d.\ N(0, \sigma_\tau^2)$; finally, in step 4 transient efficiency (TE) is estimated using the estimates of inefficiency τ_{it} following the Battese and Coelli (1988) method (i.e. $TE = \exp(-\tau_{it})$). The overall efficiency (OE) is then obtained from the product of PE and TE, that is, $OE = PE \times TE$.

The main weakness of *Model II* is that it forces time-invariant unobserved plant-specific heterogeneity into the same term that captures persistent inefficiency. Consequently, this model does not have the ability to distinguish between time-invariant unobserved heterogeneity and (persistent) inefficiency; any unobserved heterogeneity effects are treated as inefficiency. To overcome the limitations of *Model I* and *Model II*, Kumbhakar, Lien, and Hardaker (2014) recently introduced a model that separates unobserved heterogeneity, persistent inefficiency and time-varying inefficiency. This model is specified as:

$$E_{it} = a_0 + f(X_{it};\beta) + \mu_i + \nu_{it} - \tau_{it} - \eta_i.^{6} \quad (6)$$

In *Model III* (Eq. 6) the error term is split in four components. The first component μ_i, captures firms' unobserved heterogeneity, which has to be disentangled from persistent inefficiency, u_{it} captures transient inefficiency, η_i captures persistent inefficiency, ν_{it} captures the standard error

[5] All variables except TECH and DEW are in natural logs.
[6] All variables except TECH and DEW are in natural logs.

term. This model can be estimated in three steps (Kumbhakar et al., 2014): in the first step the standard random-effects panel regression is used to estimate β; in step 2 transient efficiency is estimated by maximum likelihood using the overall error component predicted in step 1 and assuming $v_{it} \sim i.i.d.\ N(0, \sigma_v^2)$ and $\tau_{it} \sim i.i.d. N(0, \sigma_\tau^2)$; finally, in step 3 persistent efficiency is estimated by maximum likelihood using the random-error component predicted in step 1 and assuming $u_{it} \sim i.i.d.\ N(0, \sigma_\mu^2)$ and $\eta_i \sim i.i.d.\ N^+(0, \sigma_\eta^2)$. Again, the overall efficiency is obtained from the product of PE and TE.

All the estimation procedures used in this paper were carried out in STATA (version 15, Stata Corp, USA) following the implementation available in Kumbhakar et al. (2015).

2.2 *Exogenous determinants of inefficiency*

In addition to estimating inefficiency for each plant, it is of interest to evaluate which variables explain inefficiency at plant level such as for example the effect of technical progress on the inefficiency. In this context, *Model III* can be extended to discern the impact of determinants of inefficiency level for a given plant. The rationale is that plants constructed in different years will use equipment with different level of technical efficiency (which is an unobserved variable). Assuming that newer equipment tend to be more efficient, the year of construction (*CONSTR*) (used as a proxy for technical progress) is expected to correlate with the mean of the inefficiency term (i.e. persistent inefficiency). This approach consists of making the mean of the distribution on the inefficiency term depend on a set of exogenous variables z_i (Wang, 2002).

Formally, this specification is given by:

$$\eta_i^2 = \exp(z_i \delta) \tag{7}$$

where η is a non-negative parameter of the density function of a random variable η_i with an exponential distribution; z_i is a vector of variables (*CONSTR* in our case), including a constant, that influence the inefficiency of plant i; and δ is a vector of unknown parameters to be estimated. Since

the estimated coefficient of z_i is not directly interpretable due to the nonlinearity of the model, to analyse the effect of z_i we need to compute the marginal effect of z_i on the unconditional expectation of inefficiency, that is $\partial E\,(\eta_i)/\partial z_i$. This can be done by extending *Model III* (see section 2.1) to accommodate the z variables. These z variables can be used as determinants of inefficiency, from which one can compute marginal effects of these z variables in increasing efficiency. More details on this model can be found in Kumbhakar et al. (2015). The four-components model in Eq. (6) extended to accommodate determinants for persistent inefficiency is:

$$E_{it} = a_0 + f(X_{it};\beta) + \mu_i + \nu_{it} - \tau_{it} - \eta_i(z_i).^{7} \tag{8}$$

3 Data

We use a database of the Directorate General for the Environment (DGE)[8] of the canton of Vaud (Switzerland), responsible for the Swiss environment policy. The region considered in the sample is a small region accounting 3275 km^2 but with altitudes ranging from 372 m to summits above 3000 m. Hence, there are different average temperatures in the region despite its limited size. The database is composed by 183 WWTPs that on average treated the wastewater of about 9.3% of the Swiss population over the years 2001-2015. During the observation period, some of WWTPs have been dismissed and their influents were redirected to larger WWTPs in the area, others have been renovated[9], while 15 new WWTPs have been constructed. As a result, our final dataset is an unbalanced panel containing data on 201 ID and a total of 2136 observations over the years 2001-2015.

The WWTPs in the sample are operated for the removal of chemical oxygen demand (COD) and nitrogen (N) compounds (e.g. NH_4 and NO_3), covering a wide range of common treatment technologies, e.g. Conventional Activated Sludge (*CAS*), Medium/High load Conventional

[7] All variables except TECH and DEW are in natural logs.
[8] www.vd.ch/autorites/departements/dte/environnement/.
[9] Renovated plants that went through a complete substitution of mechanical equipment have been considered as a new WWTP (i.e. a new ID).

Activated Sludge (MH-CAS), Rotating Biological Contactor (RBC), Trickling Filter (TF), Trickling Filter - Conventional Activated Sludge (TF-CAS), Fluidized Bed Reactor (FBR), Fluidized Bed Reactor - Conventional Activated Sludge (FBR-CAS).[10]

Descriptive statistics of the variables included in the model are presented in Table 2. Electricity consumption (*E*) is equal to the average daily electricity demand. *FLOW* is equal to the daily influent wastewater flow rate. *CAP* is the plant capacity and expressed as design influent wastewater flow rate. *COD*, *NH4* and *NO3* represent the average concentration removal of pollutant COD, NH_4 and NO_3[11], respectively (i.e. the difference between the influent and effluent concentration).[12] *TEMP* is equal to the yearly average atmospheric temperature of the WWTP.[13] *PRICE* is the real price of Swiss electricity[14].[15] Apart for the type of the secondary treatment (*TECH*) WWTPs were classified based on the presence or absence of sludge dewatering using a dummy variable (*DEW*). Finally, *CONSTR* is equal to the year of plant construction.

Table 2. Descriptive statistics.

Variable	Definition	Obs.	Mean	SD	Min	Max
E	Electricity consumption (kWh/day)	2136	596	2685	1.4	36060
FLOW	Wastewater flow rate (m^3/day)	2136	1792	8435	4.0	122889
CAP	Plant capacity (design wastewater flow rate) (m^3/day)	2136	3576	17399	17.0	206250
COD	COD removal (mgCOD/L/day)	2136	437.5	229.0	20.0	1362
NH4	NH_4 removal (mgN/L/day)	2136	19.3	14.7	0.0	80.3
NO3	NO_3 removal (mgN/L/day)	2136	7.9	10.4	0.0	63.4

[10] An overview of the wastewater treatment technologies can be found in Metcalf and Eddy (2013).
[11] Given that NO_3 is the product of the process of ammonia oxidation, the influent concentration of NO_3 is equal to the difference between influent and effluent NH_4.
[12] When due to measurement errors effluent concentration was higher than influent concentration, to avoid problem with the regression analysis we have considered the following data adjustment: $C_kOUT = C_kIN \times 0.99$, where C_kOUT and C_kIN are effluent and influent concentration of pollutant k.
[13] Yearly average temperature was calculated from 45 representatives daily average temperature (i.e. 15 days in January, 15 days in May and 15 days in September) obtained from www.meteoblue.com.
[14] Electricity price data (low tension industrial electricity C4 category) obtained from the Swiss Federal Electricity Commission ElCom (www.vd.ch/autorites/departements/dte/environnement/) were adjusted for inflation (base year 2015).
[15] All the WWTPs in the dataset operate in the same Swiss region with same electricity price.

TEMP	Temperature (°C)	2136	8.7	1.3	4.3	12.2
PRICE	Real electricity price (CHF/kWh)	2136	0.13	0.91	0.12	0.16
DEW	DEW=1 if WWTP has sludge dewatering and DEW=0 otherwise	2136	0	0	0	0
CONSTR	Year of plant construction (year)	2136	1981	10.6	1961	2014
TECH0	Conventional Activated Sludge (CAS)	1225	0	0	0	0
TECH1	Medium/High load - CAS	449	0	0	0	0
TECH2	Rotating Biological Contactor (RBC)	55	0	0	0	0
TECH3	Trickling Filter (TF)	346	0	0	0	0
TECH4	TF - CAS	8	0	0	0	0
TECH5	Fluidized Bed Reactor (FBR)	33	0	0	0	0
TECH6	FBR - CAS	20	0	0	0	0

4 Results and discussion

4.1 Estimated energy demand model

As discussed previously, when selecting whether to use fixed or random effects the most frequent suggestions is to rely on the Housman test in order to assess whether there is a significant difference between the estimates of the two models. If there is no correlation between the independent variable(s) and the unit effects, then estimates of β in the fixed-effects model should be similar to estimates of β in the random-effects model. A significant difference, on the other hand, is taken as evidence of bias in the random-effects estimates, and the user is consequently guided to employ fixed effects instead. However, while the Hausman test has a role to play in comparing the estimates obtained from FE and RE models it is neither a necessary nor sufficient statistic for deciding between fixed and random effects. Instead there is a range of situations in which the random-effects model may be preferable to the fixed effects model for estimating β, regardless of whether the assumption of "random" effects holds (Kim and Schmidt, 2000; Kumbhakar and Lovell, 2003). For example, in the case in which variation is primarily across units, and there is an

intermediate amount of data[16], such as in our case, even if there is a moderate level of correlation between the unit effects and the regressors, the random effects estimator outperforms the fixed effects estimator (Clark and Linzer, 2015).[17]

In addition to these theoretical considerations, there are number of practical issues which should take into account when deciding between fixed and random effects estimator. An important advantage of assuming α_i being random is that important time-invariant characteristics of interest (such as plant size and type of secondary treatment) can be directly included in the model as explanatory variables and their effects can be estimated, while in the fixed-effects model these effects are viewed as unobserved heterogeneity and are captured by the firm-specific time-invariant term whose effects on energy demand cannot be estimated. Finally, another advantage of using random effects is the possibility to use the statistical model to make predictions about WWTPs not included in the dataset, which may be of interest for future energy benchmarking exercises. Given the composition of our data, we feel that the choice of random-effects model over fixed-effects is justifiable, based on the arguments presented above, and in accordance with previous energy demand SFA studies (Filippini and Hunt, 2011; Filippini et al., 2018; Lundgren et al., 2016).

The estimation results of the WWTPs energy demand model are given in Table 3. The selection of models relies upon the results of a series of regression diagnostic to evaluate the robustness of selected specifications, such as checking for the possible presence of unusual and influential data, normality of residuals, multicollinearity, serial correlation and heteroscedasticity. Since the three models presented in Section 2.1 differ only in the interpretation of the firms' time-invariant heterogeneity (or unobserved heterogeneity), the coefficients of the energy demand frontiers are the same for the three models. Moreover, since we use a log-log functional form for the electricity

[16] E.g. few units and many observations per unit, and vice-versa.

[17] The results of employing Hausman test to our dataset rejected the null hypothesis, which is indication of failure of the random-effects assumption. However, looking at the coefficients of both random and fixed effects models, only minor differences were present, suggesting a low level of correlation between unit effects and regressors. Hence, rather than rejection of the random effects assumption, the result of the Hausman test would seem more related to a caveat which is common to all significance tests when increasing sample size, namely, a small difference between the fixed effects and random effects estimates can lead to rejection of the null hypothesis even when the difference between estimates is substantively insignificant. We therefore decided to use random-effects in order to have direct information on the effect of time-invariant variables on energy demand.

demand and other continuous variables in the model, the estimated coefficients are interpreted as demand elesticities, i.e. the expected percentage change in electricity demand relative to a percent change in one of the regressors.

Table 3. Estimated WWTPs energy demand model.

Parameter	**Coefficient**	**Std. Err.**	**[95% Conf. Interval]**	
Constant	0.0985***	0.0334	0.0294	0.1642
FLOW	0.2779***	0.0375	0.2044	0.3514
CAP	0.6476***	0.0450	0.5592	0.7360
COD	0.0692***	0.0112	0.0471	0.0913
NH4	0.0472***	0.0113	0.0250	0.0694
NO3	-0.0408***	0.0105	-0.0614	-0.0202
TEMP	0.0522**	0.0239	0.0053	0.0914
PRICE	-0.0048	0.0042	-0.0132	0.0035
DEW=YES	0.0674***	0.0237	0.0209	0.1140
TECH1=MH-CAS	-0.1066	0.0766	-0.2569	0.0435
TECH2=RBC	-0.4472***	0.1321	-0.7062	-0.1881
TECH3=TF	-0.5671***	0.0592	-0.6831	-0.4510
TECH4=TF-CAS	-0.1608	0.2307	-0.2913	0.6131
TECH5=FBR	0.1369	0.2329	-0.3196	0.5935
TECH6=FBR-CAS	-0.0931	0.1359	-0.3596	0.1734
Persistent inefficiency determinants [a]				
Constant	-3.5585***	0.1455	-3.8437	-3.2733
CONSTR	-0.3975***	0.0788	-0.5521	-0.429
Variance parameters for the compound error				
σ_u	0.2977***			
σ_e	0.1678***			

ρ	0.7590***

[a] Applies only to *Model III*

*** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level.

The majority of the estimated coefficients have the expected signs and are seen to be statistically significant. Parameter ρ, which represents the relative contribution of the inefficiency term over the complete disturbance term (i.e. $\rho = \sigma_u^2/\sigma$), is also significant and relatively high indicating that 76% of the variance in the error term is caused by inefficiency differences rather than data noise.

The results suggest that WWTPs energy demand is price inelastic. Then, as expected, the volume of wastewater entering the plant (*FLOW*) has a positive and highly significant influence on the WWTP energy demand. This is the portion of the energy consumption that is proportional to the volume of wastewater treated and accounting e.g. to influent pumping.

Interestingly, a significant part of the electricity consumption is function of the plant capacity (*CAP*), which is a time-constant variable, implying that this portion of energy consumption is independent from the actual wastewater flow rate entering the plant. The coefficient for *CAP* is found to be positive and highly significant in addition to be the largest one. WWTPs are in fact characterized by a number of equipment that work continually independently of the daily or seasonal influent flow variations that exist in the WWTPs, such as screenings and grit removal in the preliminary treatment and mixers, sludge recirculation and decanters in the secondary treatment. Furthermore, coefficient of *FLOW* and *CAP* reveal evidence of moderate economies of scales, i.e. on average plants can reduce their energy demand by operating at higher scale. A more in deep analysis of economies of scale is given in Section 4.5.

Although it is widely acknowledged that the level of wastewater treatment intensity affects energy use (Rodriguez-Garcia et al., 2011), a clear relation between energy use and removal of each single pollutant, to our knowledge, has never been reported. For example, recently Molinos-Senante et al. (2018) have concluded that pollutant removal efficiency has a low impact on energy intensity after

regressing pollutants (among others technical variables) with the energy intensity of a large sample of WWTPs. Similarly, using Machine Learning techniques to model WWTPs energy cost, Torregrossa et al. (2018) have showed that although pollution load (COD, N and P) at the inlet has an high impact of the energy cost, the removal efficiency has minor importance. Our results highlight that an important part of the energy consumed by the plant is proportional to the intensity of the removal of contaminants; e.g. COD, NH_4 and NO_3 appear to have consistent and significant influence on WWTPs energy demand. In particular increasing of 1% the removal of COD and NH_4 will increase the energy consumption by about 0.069% and 0.047%, respectively. In contrast, the increasing of 1% of the removal of NO_3 will decrease the energy demand by 0.041%. Unlike COD and NH_4, the removal of NO_3 is not an oxygen demanding process since biological denitrification occurs in anoxic conditions and requires electron donors (i.e. a carbon source such as the readily biodegradable fraction of COD) to reduce the combined oxygen in NO_3. In particular, the removal of 1 kg of NO_3 requires 4-15 kg of COD (depending on the nature of the carbon source) (Kujawa and Klapwijk, 1999).[18] As a consequence, a plant implementing biological denitrification will consume part of the COD present in the wastewater for the removal of NO_3, which in turn will not be eliminated by energy consuming aeration. Thus, in plants carrying out conventional N removal (i.e. nitrification-denitrification), implementing efficient denitrification (e.g. by adjusting recirculation) has the possibility to offset at least partially the energy consumed for NH_4 removal.

A positive and highly significant relationship of *TEMP* with energy demand was also found. In particular the energy consumption will increase by 0.52% for every unit increase of the temperature. Although it has often been overlooked, temperature is an important factor that significantly affects energy consumption. In a previous contribution with a different dataset (Longo et al., 2018) we found the same effect, apparently demonstrating that the increase in aeration energy

[18] We do recognize that *COD* and *NO3* may be to a certain degree correlated, i.e. COD removal is totally independent from NO_3 level but NO_3 removal is partially dependent on the COD. VIF test however do not evidence severe collinearity. We therefore decided to leave *NO3* in the model in order to better reflect energy behaviour at WWTPs.

caused by lowered oxygen solubility prevails over the stimulated biological activity at higher temperatures, at least for the range of temperature studied (9-18 °C).

A significant portion of total plant energy consumption is normally accounted for sludge dewatering. The coefficient of *DEW* indicates that when dewatering is carried out in the plant the energy demand will increase by about 6.7%, in line with literature data (Longo et al., 2016).

Regarding the type of secondary treatment, using CAS as the reference, it appears that, *ceteris paribus*, TF is the least energy intensive technology consuming less than half the energy of CAS, followed by RBC whose consumption is 43% lower. On the contrary, we do not find the rest of technologies to be statistically significant. A comparison of about 2500 German WWTPs (DWA, 2014) has revealed that low-energy technologies such as TF and RBC consume about half the electricity of high-energy technologies, such as CAS, therefore in agreement with this study. In the United States, analysis of the energy consumption of existing wastewater technologies has produced data similar to ours (Crawford and Sandino, 2010). Historically, TF and RBC have been considered to have major advantages of using less energy than activated sludge treatment and being easier to operate, but have disadvantage of lower-quality effluent, especially in term of N (Metcalf and Eddy, 2003). This is also observable in our dataset (data not shown). Since, unlike CAS, TF and RBC can operate only under not stringent effluent requirements regarding N, the comparison of the two technologies should be limited to the common domain, i.e. for plants having low N effluent requirement TF and RBC perform better than CAS. In order to reach efficient N removal, those treatment systems have to be combined with activated sludge process, such as the TF-CAS system (i.e. combination of TF and CAS). The result is that energy demand of combined systems is statistically not significantly different from WWTPs designed and operated for the removal of both organic matter and N, such as CAS systems.

It should be noted that the type of secondary treatment is not usually decided based only on energy reasons, but instead is the result of complex considerations on effluent requirements, size of plant, footprint, ease of operation, robustness, location of the plant and others socio-economic

motivations. As an example, although CAS technology has been found to be more energy consuming in comparison with TF, it has the possibility to reach lower pollutants removal in comparison with TF. Therefore, for plants that need to treat wastewater at a higher level it is an obliged choice. On the other hand, even for plants without particular stringent effluent requirements, technologies with higher energy demand but with lower footprint such as CAS can be preferred to lower energy demand technology such as TF if space is a problem.

4.2 Estimated efficiency

The results of the econometric analysis reported in Table 3 can be used to estimate the level of energy efficiency using Eq. (3). Table 4 provides summary statistics of the estimated efficiency levels for the three models.

Table 4. Estimated energy efficiency scores for *Model I*, *Model II* and *Model III*.

	Efficiency type	Mean	Std. Dev.	Minimum	Maximum
Model I	Overall	0.408	0.121	0.105	1
	Transient	0.933	0.039	0.327	0.986
Model II	Persistent	0.408	0.120	0.105	1
	Overall	0.381	0.114	0.080	0.981
	Transient	0.933	0.039	0.327	0.986
Model III	Persistent	0.854	0.087	0.315	0.951
	Overall	0.796	0.088	0.239	0.933

Figure 1 shows the kernel density of estimated efficiency across the three models. The various models clearly produce different empirical distributions, in some cases markedly so. The estimated average energy efficiency resulting from *Model I* is found to be about 41%, with a minimum and maximum efficiency between 10% and 100%, respectively. Overall, a fair degree of variation among plants is established in energy efficiency estimates, indicating that there is considerable

room for improvement. *Model I*, however, does not enables to examine whether inefficiency has been persistent over time or whether the inefficiency of plant varied with time. Such decomposition is desirable from a diagnostic point of view because persistent and residual components of inefficiency have different engineering implications. Thus, for example, if a plant uses old and inefficient equipment its inefficiency will be repeated over time, hence its persistent inefficiency will be large. By contrast, if the transient inefficiency component for a plant is increasing over time it may be argued that inefficiency is caused by some application of wrong operational strategies due to e.g. error in management of sludge age and return sludge, too infrequent sampling or inadequate evaluation of monitoring data.

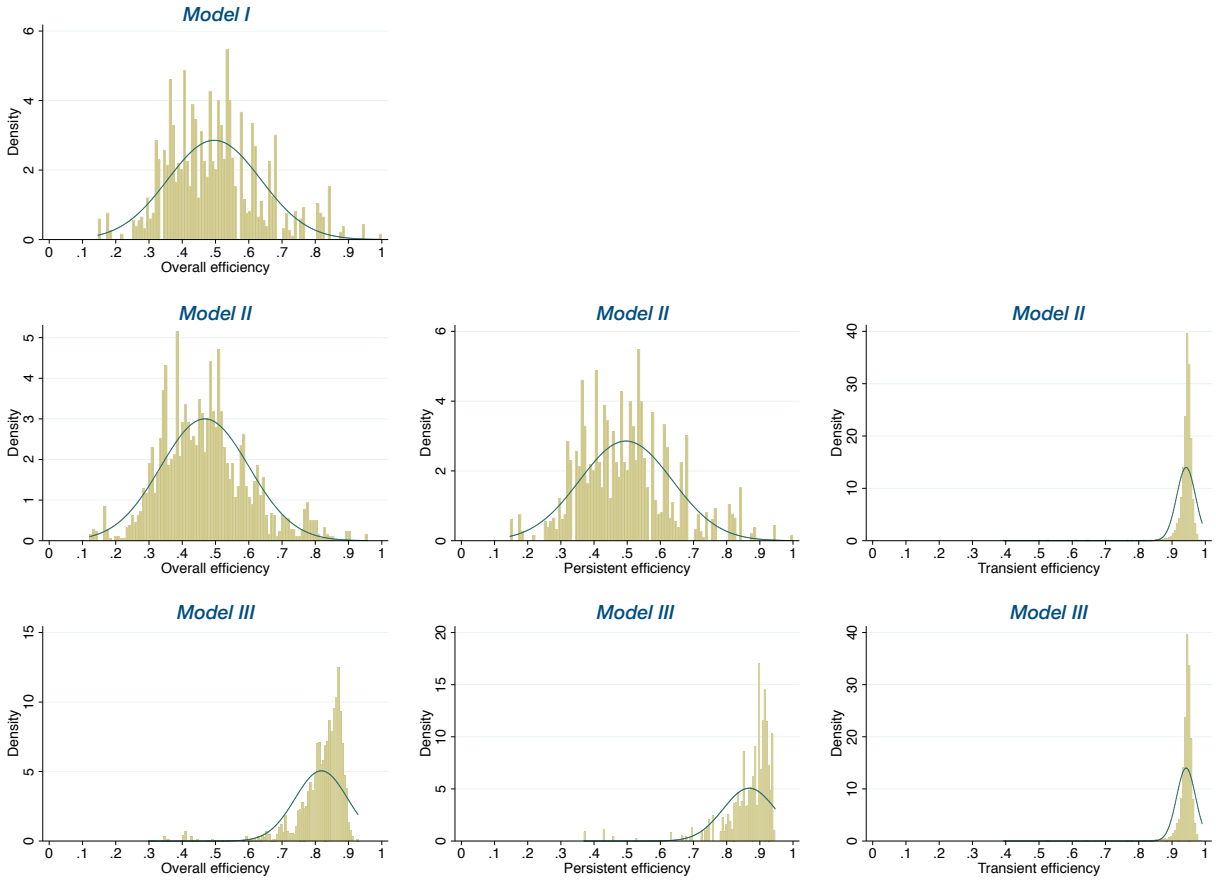

Figure 1. Estimated efficiency from *Model I*, *Model II* and *Model III.* From left to right: overall, persistent and transient efficiency.

To overcome this limitation *Model II* provides a way to decompose the overall energy efficiency component (Fig. 1). The transient part of the efficiency in WWTPs electricity consumption is found to be between 33% and 98%, while the persistent part of the efficiency ranges from 10% to 100% and has a mean value of 41%. Furthermore, the variation in the estimated transient energy efficiency is lower than the variation in the estimated persistent energy efficiency. Consequently, the majority of inefficiency is not caused by operational technical problems but instead to recurring (over the years) identical problems. Thus, unless there is a structural change in the operation of individual plants such as a change in mechanical equipment, it is very unlikely that the persistent inefficiency component will change.

Although the decomposition of the inefficiency term in *Model II* permits a better evaluation and diagnosis of plant energy efficiency one problem is that if some time-invariant unobserved heterogeneity exists, the variation in energy use due to heterogeneity will be captured by the persistent component of inefficiency, since *Model II* is not able to distinguish between inefficiency and unobserved heterogeneity. Therefore, for a plant having, for instance, higher energy requirement for pumping due to a particularly unfavourable topography, these unobserved heterogeneities would be labelled as inefficiency and consequently, having a time-invariant character, their persistent inefficiency would be unfairly inflated. *Model III* overcomes this problem by decomposing the time-persistent component of inefficiency into a time-invariant heterogeneity effect and a persistent inefficiency effect. The results are a mean level of transient efficiency equal to be 93% (like in *Model II*), while persistent efficiency is now 85% (not 41% as predicted by *Model II*), and the overall efficiency is now 80% (much higher than 38% predicated by *Model II*). These results indicate that a large part of the persistent inefficiency captured by *Model II* is considered time-invariant unobserved heterogeneity in *Model III* and suggest that differences in topography may affect significantly efficiency at WWTP.

The efficiency levels presented above indicate that there is a sufficient potential for the Swiss wastewater sector to save energy. Depending on the model employed, considering overall

efficiency estimates, WWTPs could save as much as 59%, 62% and 20% of their electricity usage, respectively for Model *I*, Model *II* and Model *III*.

The average estimated efficiency (across plants) from *Model I*, *Model II* and *Model III* are plotted over time in Figure 2.

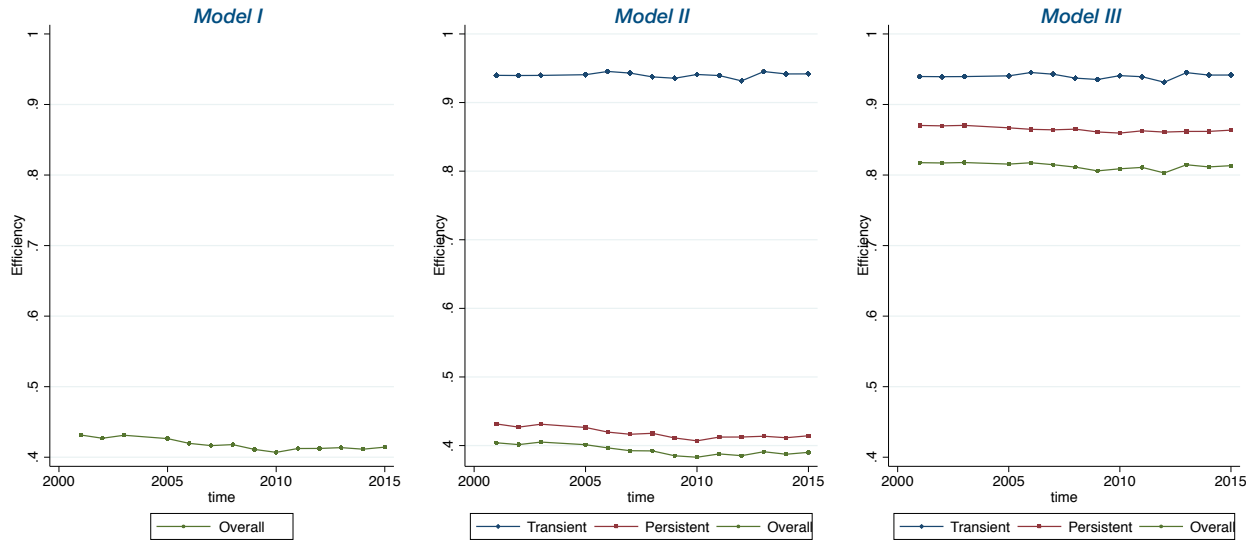

Figure 2. Efficiency over time for *Model I*, *Model II* and *Model III*.

Persistent efficiency in *Model II* coincides with the overall efficiency in *Model I*, which is not surprising given that *Model I* is a time-invariant efficiency model, i.e. inefficiency is represented by the entire time-invariant component. Note that even if persistent efficiency is time-invariant, it is not perfectly constant over the 15 observed years as the panel is unbalanced. Furthermore, it is interesting to see that the average transient efficiency has not apparently changed significantly over the 15 years. However, a closer look at the individual transient efficiency trends (Fig. 3) highlights that this stability is not the result of the general efficiency immobility among the plants, but rather the average result of plants that have improved their efficiency and others that have worsened their efficiency.

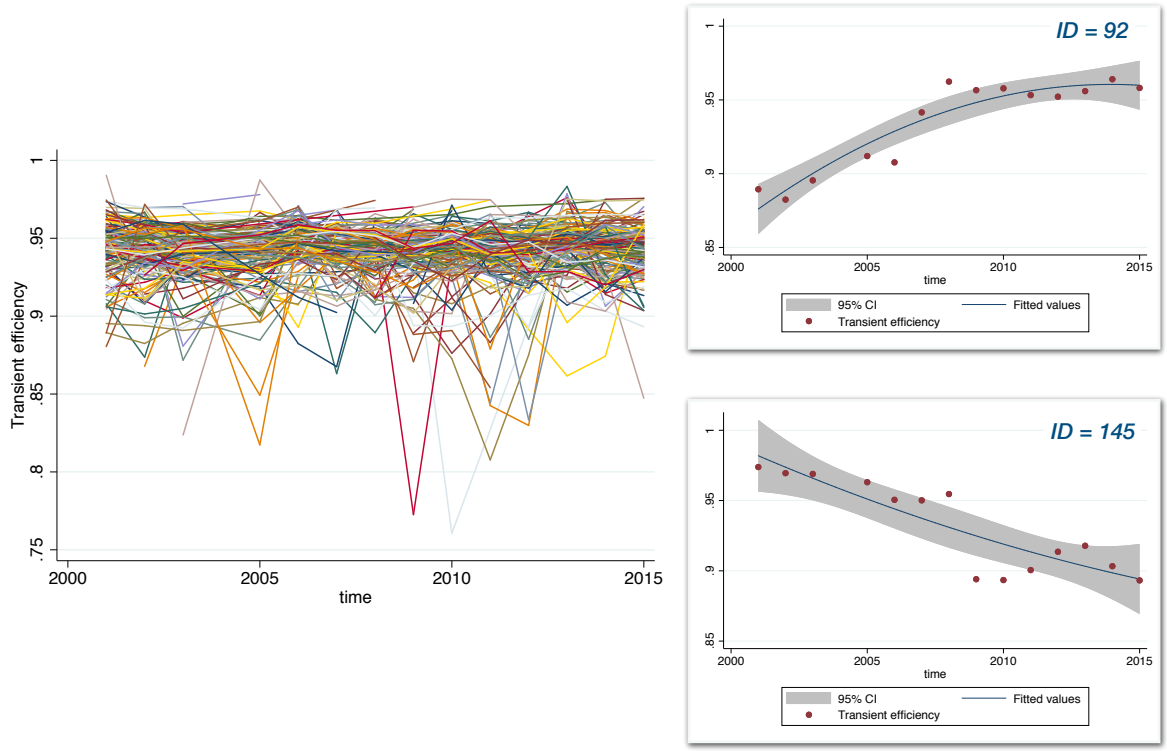

Figure 3. Transient efficiency over time (left part) and example of best and worst practice (right part).

Looking at individual plant time-varying efficiency results, it is possible to individuate plants that have considerable increased their efficiency over time (e.g. ID 92 in Fig. 3); this may result from the continuous operation efficiency improvement of some operators, due to e.g. an optimal evaluation of monitoring data, which is essential for highly efficient plant operation (Rieger and Olsson, 2012). By contrast, other plants have deteriorated considerably their efficiency (e.g. ID 145 in Fig. 3); this may result from the adaption of wrong operational strategies, due to poor data analysis, too infrequent sampling, inadequate controller settings (e.g. automatic control of dissolved oxygen), among others (Rieger and Olsson, 2012). As a consequence, the use of panel data associated with time-variant SFA models such as *Model II* and *Model III* have the advantage to be able to individuate operational best and/or worst practices, which is essential information for future development of better operation strategies.

4.3 Heterogeneity or inefficiency?

As the above results illustrate, the efficiency scores are, as expected, sensitive to model specifications. In particular, the results suggest that a large part of the persistent inefficiency captured by *Model II* is actually time-invariant unobserved heterogeneity in *Model III*. So, the question is: should one view the time-invariant heterogeneity as persistent inefficiency or as plant heterogeneity that captures the effects of external factors and cannot be tackled by plant management?

Examples of sources of such type of heterogeneity at WWTPs may derive from structural characteristics of WWTPs, such as differences in the hydraulic design of the initial pumping station. For example, a WWTP that was designed in a particularly unfavourable pumping condition, e.g. due to the specific topography of the area where the plant is located, is expected to have higher hydraulic requirement (and as a consequence higher energy demand for pumping) in comparison with a plant having a more favourable hydraulic design.

Plant operators may be interested in controlling for the impact on energy use due to the location and hence in omitting its effect from the inefficiency to obtain well-grounded comparisons of WWTPs. The reason is that without disentangling persistent inefficiency from heterogeneity (as done in *Model III*), plant operators are going to obtained efficiency estimates that can be potentially biased as inefficiency is assumed to be fully attributable to managerial decisions, while plant location is not under the operator's control, and for that virtually impossible to be eliminated. Although in principle it may be possible to dismiss a plant and build a new one in a place with better topography characteristics in order to reduce pumping requirements, this is very unlikely to happen. As a consequence, when benchmarking WWTPs, provided panel data is available, a SFA model able to take into account for unobserved heterogeneity such as *Model III* would be preferable.

An alternative solution to that would be to find a way to construct a control variable in order to control for topography among WWTPs. This may be implemented e.g. by considering geographical

information systems (GIS) data to quantify topography. Topography is not however the only reason for differences in pumping requirements but also plant design considerations have an impact. As a consequence, a more robust solution would be to include in the regression function a variable for geodetic pumping head (in meters), i.e. the actual physical difference in height between the wastewater level in pumping station and the highest point of the discharge or water level in the outlet. This information is normally provided in pump technical data sheet of each pump. We have tried to collect those data, but we were unable to obtain them.

4.4 Impact of technical progress on persistent efficiency

We have studied here the impact of technical progress on efficiency by estimating the marginal effects of the year of plant construction on persistent inefficiency. Figure 4 reports the scatter plot of the marginal effects against the values of *CONSTR*. The graph indicates that, for all the observations, the marginal effect of *CONSTR* is negative; meaning that renovating a plant effectively decreases the persistent inefficiency. Furthermore, the relation between technical progress and persistent inefficiency is not linear and depends on the value *CONSTR*. The marginal effect of the technical progress is higher for older plants indicating that, if renovated, older plants have a higher potential to reduce their inefficiency in comparison with newer plants, which is reasonable. In particular, the level of persistent inefficiency is reduced, on average, by 0.085% for each unit increase in *CONSTR* (i.e. one year). The convenience of renewing is highest for the oldest systems having the ability to eliminate up to 5.3% of their persistent inefficiency. As an example, the efficiency of a plant renovated in the year 2000 is 5.3% higher than a plant constructed in the year 1960 as a result of the different level of technical progress.

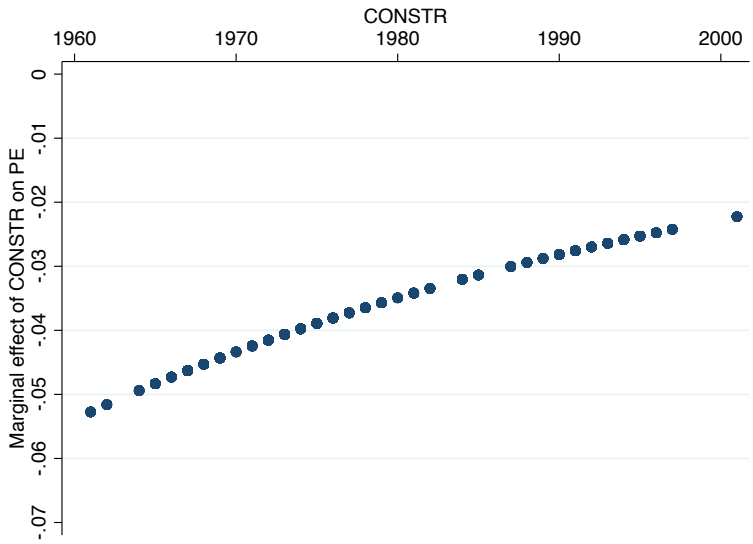

Figure 4. Marginal effects of year of plant construction (*CONSTR*) on persistent efficiency (PE).

4.5 Economies of scale

The existence of economies of scale at WWTPs, i.e. average energy consumption decrease with the increasing of the amount of treated wastewater, it is a well-known effect (Krampe, 2013; Molinos-Senante et al., 2018; Vaccari et al., 2018). However, there are two cases where a WWTP can increase its scale: 1) in case of residual capacity by simply directing higher volume of wastewater to an existing plant (i.e. increasing output density); 2) in case of no residual capacity by increasing proportionally the treated wastewater and plant capacity (i.e. increasing scale). Distinguish between both types of economies of scale is important because they have different economic interpretations. The estimation results in Table 3 can be used to compute the plants' level of economies of output density and scale following the work of Caves, Christensen, and Swanson (1981). Economies of output density (E_{OD}) measure the reaction of energy demand to an increase in output (i.e. the amount of treated wastewater), holding the rest of variables, e.g. plant capacity, constant and are obtained as follows:

$$E_{OD} = \left(\frac{\partial \ln E}{\partial \ln FLOW}\right)^{-1}. \tag{9}$$

Economies of scale (E_S) differs from E_{OD} in the assumption that an increase in output not only raises the volume of wastewater received by the plant, but to the same proportion also the plant capacity (i.e. by scaling up all the equipment as well as reactors volumes). Therefore, economies of scale can be written as:

$$E_S = \left(\frac{\partial \ln E}{\partial \ln FLOW} + \frac{\partial \ln E}{\partial \ln CAP}\right)^{-1}. \quad (10)$$

Economies of output density and scale exist if the respective values of E_{OD} and E_S are greater than 1. Analogously, values smaller than 1 indicate diseconomies of output density or scale.

Estimated economies of output density and scale can be found in Table 5. With respect to the design plant size (expressed as COD-basis Population Equivalent (PE)[19]) three types of representative WWTPs are chosen, e.g. very small WWTPs (PE < 1000), small WWTPs (1000 < PE < 10000), medium WWTPs (10000 < PE < 40000), medium-large WWTPs (40000 < PE < 100000), and the model re-estimated for the three size categories. Due to the fact that only one WWTP in the dataset has size bigger than 100000 PE this plant has been excluded from this analysis.

Economies of output density (E_{OD}) are present for all class of WWTPs with respect to size. A 1% of increase in energy demand is associated with a more than 1% increase in the amount of wastewater treated (*FLOW*), holding plant capacity (*CAP*) constant. This finding therefore suggests that it is beneficial for undersized plants to treat larger amount of wastewater into existing facilities in order to operate close to the design size and, as reported in a previous study, providing the compliance with effluent requirements, energy performance keeps increasing for over-dimensioned plants due to the generosity with which WWTPs are normally designed (Longo et al., 2018).

The economies of scale (E_S) are equal to the inverse of the percentage change in energy demand when the size increases by 1%. The results show that on average economies of scale are present, i.e.

[19] Population equivalent or unit per capita loading, (PE), in waste-water treatment is the number expressing the ratio of the sum of the pollution load produced during 24 hours to the individual pollution load in household sewage produced by one person in the same time.

E_S of all WWTPs is 1.08. It would be therefore rational for WWTPs to expand their services by merging adjacent treatment plants, which is in line with previous literature (Ganora et al., 2019). What has not been clarified before, to our knowledge, is that if this is true for all ranges of WWTPs plant size. The results of our analysis reveal that substantial economies of scale are present in very small, small and medium sized WWTPs. On the other hand, this is no longer the case of medium-large WWTPs, where constant return to scale is present, i.e. E_S is close to one. This is also an indication that the optimal size of WWTPs is between 40000 and 100000 PE. Thus, for WWTPs larger than 100000 PE, at least from an energy efficiency point of view, it would be no more beneficial to increase their scale. On the contrary, for plants smaller than 40000 PE (about 95% of plants), a significant potential for energy saving is present.

Table 5. Economies of output density (E_{OD}) and scale (E_S)

	Very small WWTPs (n. = 924)	**Small WWTPs (n. = 938)**	**Medium WWTPs (n. = 175)**	**Medium-large WWTPs (n. = 95)**	**All WWTPs**
E_{OD}	2.27	2.13	2.08	2.44	1.54
E_S	1.37	1.45	1.35	0.99	1.08

An illustration of the economies of scales is given in Figure 5. On the x-axis is represented the plant size (expressed as influent wastewater flow rate), while the y-axis represents the corresponding specific energy consumption (kWh/m^3 of treated wastewater) for each plant size level. In the chart the grey curves represent the specific energy consumption when varying the plant size for a given level of load factor (LF). As LF represents the actual flow rate of the plant compared to the design capacity (i.e. $FLOW/CAP$), it indicates whether the plant is oversized (LF<100%). The graph illustrates that WWTPs can decrease their specific energy consumption by increasing their size. Thus, grey curves represent WWTPs economies of scales, i.e. as plant expands its service the unit energy consumption decreases. However, in order to expand their services WWTPs need to increase

their capacity by scaling up all the equipment. For a given volume of wastewater treated, expanding plant capacity will increase specific energy consumption for the additional energy demand due to bigger equipment. Let us represent it by a thought experiment. The capacity of an operating WWTP is increased at constant flow rate by 5 times (i.e. from 24000 m^3/d to 120000 m^3/d, represented in the chart by the dotted arrow and a severe increase in specific energy consumption). Increasing the flow rate in this newly expanded plant will lead to a LF decrease (red curve) and of the unit consumption. This curve represents the economies of output density, i.e. the average decreasing energy consumption when output increases so that an increase in output results in a less than proportional increase in energy consumption. Furthermore, the results indicate that the effect of oversize on energy demand increase with decreasing plant size, as a result special attention should be given during WWTP designing phase in particular to small-medium size plant in order to reduce energy demand by avoiding extra and unnecessary reaction volumes.

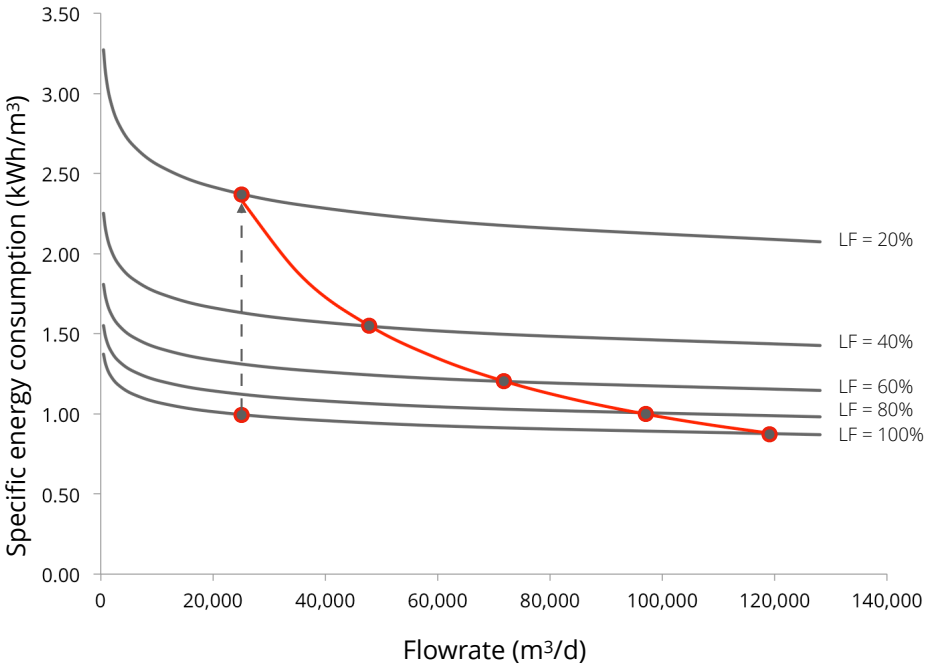

Figure 5. Average specific energy consumption as function of plant scale and load factor.

5 Conclusions

This paper proposes, for the first time to our knowledge, the use of SFA to estimate the energy efficiency of WWTPs. The proposed electricity demand specification controls for the price of energy, volume of wastewater treated, plant capacity, level of main pollutants (COD, NH_4 and NO_3) removed from wastewater, temperature, type of secondary treatment and the presence of sludge dewatering in order to obtain a measure of energy efficiency.

The paper illustrates that persistent energy inefficiency is more severe than transient energy inefficiency. Consequently, the majority of inefficiency is not caused by operational technical problems but instead to recurring (over the years) identical problems due to e.g. inefficient equipment. However, it is shown that a large part of the persistent inefficiency is due to time-invariant unobserved heterogeneity suggesting that unobserved factors such as the topography of the service area should be controlled for in order to obtain meaningful efficiency estimates.

The results show that the level of energy efficiency of equipment influences the demand for energy. Therefore, policies aimed at promoting technological innovation can lead to improvements in energy efficiency, provided that the equipment is used in an efficient way. Finally, the estimation of economies of output density and scale highlights that large energy savings can be achieved by directing higher volume of wastewater to the plant. Even if in general, design guidelines propose over-dimensioned WWTP designs in order to avoid malfunctions and non-compliance with effluent requirements, our results suggest that special attention should be given during WWTP designing phase in order reduce energy demand by avoiding extra and unnecessary reaction volumes.

Thanks to the possibility to take into account the presence of unobserved heterogeneity, to distinguish persistent from transient inefficiency and to take into account the statistical noise of data errors, the proposed approach, compared to previous research, is superior to deduce appropriate energy diagnosis in order to make inefficient WWTPs efficient.

In light of the above findings, we believe that this manuscript can be of interest for academics, policymakers and utilities in the evaluation and planning of actions to increase efficiency in wastewater treatment.

Acknowledgments

Stefano Longo, Almudena Hospido and Miguel Mauricio-Iglesias belong to the Galician Competitive Research Group GRC2013-032 and the CRETUS strategic partnership (AGRUP2015/02), co-funded by FEDER (EU). Besides, they are supported by 'ENERWATER' Coordination Support Action that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 649819. Stefano Longo would like to thank REGATA network (ED341D R2016/033), funded by the Galician Government, for the international short-term research visit grant. The authors are also immensely grateful to Dr. Laura Hospido (Banco de España) for her comments on an earlier version of the manuscript. We would also like to thank the reviewers for their valuable comments. All remaining errors are our own.

References

Arellano, M. (2003). *Panel data econometrics*. Oxford: Oxford University Press.

Battese, G. E., and Coelli, T. J. (1988). "Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data," *Journal of Econometrics* 38(3): 387-399.

Boyd, G. A. (2008). "Estimating plant level energy efficiency with a stochastic frontier," *The Energy Journal* 29(2): 23-43.

Castellet-Viciano, L., Hernández-Chover, V., Hernández-Sancho, F. (2018). "Modelling the energy costs of the wastewater treatment process: The influence of the aging factor," *Science of the Total Environment* 625: 363-372.

Caves, D. W., Christensen, L. R., and Swanson, J. A. (1981). "Productivity growth, scale economies, and capacity utilization in US railroads, 1955-74," *The American Economic Review* 71(5): 994-1002.

Clark, T.S., and Linzer, D.A. (2015). "Should I use fixed or random effects?" *Political Science Research and Methods* 3: 399-408.

Crawford, G., and Sandino, J. (2010). "Energy efficiency in wastewater treatment in north America: A compendium of best practices and case studies of novel approaches," WERF Report.

DWA. (2014). "25th performance comparison of municipal wastewater treatment plants in Germany," *Korrespondenz Abwasser, Abfall (KA), International Special Edition*: 15-21.

EBC. (2016). "Learning from international best practices," EBC Foundation report.

EC. (2018). "Proposal for a directive of the European parliament and of the council on the quality of water intended for human consumption (recast)," *Office for Official Publications of the European Communities*.

EEA. (2014). "Performance of water utilities beyond compliance - Sharing knowledge bases to support environmental and resource-efficiency policies and technical improvements," EEA Technical report no. 5/2014.

Filippini, M., and Greene, W. (2016). "Persistent and transient productive inefficiency: A maximum simulated likelihood approach," *Journal of Productivity Analysis* 45(2): 187-196.

Filippini, M., Geissmann, T., and Greene, W.H. (2018). "Persistent and transient cost efficiency—an application to the Swiss hydropower sector," *Journal of Productivity Analysis* 49: 65-77.

Filippini, M., and Hunt, L. C. (2011). "Energy demand and energy efficiency in the OECD countries: A stochastic demand frontier approach," *The Energy Journal* 32(2): 59-80.

Filippini, M., and Hunt, L. C. (2015). "Measurement of energy efficiency based on economic foundations," *Energy Economics* 52: S5-S16.

Färe, R., Grosskopf, S., and Hernandez-Sancho, F. (2004). "Environmental performance: an index number approach," *Resource and Energy Economics* 26: 343-352.

Ganora, D., Hospido, A., Husemann, J., Krampe, J., Loderer, C., Longo, S., Bouyat, L.M., Obermaier, N., Piraccini, E., and Stanev, S. (2019). "Opportunities to improve energy use in urban wastewater treatment: a European scale analysis," *Environmental Research Letters*.

Greene, W.H. (2003). *Econometric analysis*. Upper Saddle River: Pearson Education.

Guerrini, A., Romano, G., Mancuso, F., and Carosi, L. (2016). "Identifying the performance drivers of wastewater treatment plants through conditional order-m efficiency analysis," *Utilities Policy* 42: 20-31.

Hsiao, C. (2014). *Analysis of panel data.* Cambridge: Cambridge University Press.

Huntington, H., and Smith, E. (2011). "Mitigating climate change through energy efficiency: an introduction and overview," *The Energy Journal* 32(1): 1-6.

IEA. (2016). "*World Energy Outlook 2016 - Excerpt - Water-Energy Nexus,"* OECD/IEA report.

Kim, Y., and Schmidt, P. (2000). "A review and empirical comparison of Bayesian and classical approaches to inference on efficiency levels in stochastic frontier models with panel data," *Journal of Productivity Analysis* 14: 91-118.

Kuosmanen, T. (2005). "Weak disposability in nonparametric production analysis with undesirable outputs," *American Journal of Agricultural Economics* 87(4): 1077-1082.

Krampe, J. (2013). "Energy benchmarking of south Australian WWTPs," *Water Science and Technology* 67(9): 2059-2066.

Kujawa, K., and Klapwijk, B. (1999). "A method to estimate denitrification potential for predenitrification systems using NUR batch test," *Water Research* 33(10): 2291-2300.

Kumbhakar, S. C., and Heshmati, A. (1995). "Efficiency measurement in Swedish dairy farms: An application of rotating panel data, 1976–88," *American Journal of Agricultural Economics* 77(3): 660-674.

Kumbhakar, S. C., Lien, G., and Hardaker, J. B. (2014). "Technical efficiency in competing panel data models: A study of Norwegian grain farming," *Journal of Productivity Analysis* 41(2), 321-337.

Kumbhakar, S.C., and Lovell, C.K. (2003). *Stochastic Frontier Analysis*. Cambridge: University Press.

Kumbhakar, S. C., Wang, H., and Horncastle, A. P. (2015). *A practitioner's guide to stochastic frontier analysis using Stata.* Cambridge: Cambridge University Press.

Leth-Petersen, S. (2002). "Micro econometric modelling of household energy use: testing for dependence between demand for electricity and natural gas," *The Energy Journal* 23(4): 57-84.

Longo, S., Mauricio-Iglesias, M., Soares, A., Campo, P., Fatone, F., Eusebi, A.L., Akkersdijk, E., Stefani, L., and Hospido, A. (2019). "ENERWATER–A standard method for assessing and improving the energy efficiency of wastewater treatment plants," *Applied Energy* 242: 897-910.

Longo, S., Hospido, A., Lema, J. M., and Mauricio-Iglesias, M. (2018). "A systematic methodology for the robust quantification of energy efficiency at wastewater treatment plants featuring data envelopment analysis," *Water Research* 141: 317-318.

Longo, S., d'Antoni, B. M., Bongards, M., Chaparro, A., Cronrath, A., Fatone, F., et al. (2016). "Monitoring and diagnosis of energy consumption in wastewater treatment plants. A state of the art and proposals for improvement," *Applied Energy* 179: 1251-1268.

Lundgren, T., Marklund, P., and Zhang, S. (2016). "Industrial energy demand and energy efficiency–Evidence from Sweden," *Resource and Energy Economics* 43: 130-152.

Metcalf, E., and Eddy, H. P. (2003). *Wastewater engineering: Treatment and reuse.* New York: McGraw Hill.

Molinos-Senante, M., Sala-Garrido, R., and Iftimi, A. (2018). "Energy intensity modeling for wastewater treatment technologies," *Science of the Total Environment* 630: 1565-1572.

Rieger, L., and Olsson, G. (2012). "Why many control systems fail," *Water Environment and Technology* 24(6): 42-45.

Rodriguez-Garcia, G., Molinos-Senante, M., Hospido, A., Hernández-Sancho, F., Moreira, M. T., and Feijoo, G. (2011). "Environmental and economic profile of six typologies of wastewater treatment plants," *Water Research* 45(18): 5997-6010.

Schmidt, P., and Sickles, R. C. (1984). "Production frontiers and panel data," *Journal of Business and Economic Statistics* 2(4): 367-374.

Torregrossa, D., Leopold, U., Hernández-Sancho, F., and Hansen, J. (2018). "Machine learning for energy cost modelling in wastewater treatment plants," *Journal of Environmental Management,* 223: 1061-1067.

Tsionas, E. G., and Kumbhakar, S. C. (2014). "Firm heterogeneity, persistent and transient technical inefficiency: A generalized true Random‐Effects model," *Journal of Applied Econometrics,* 29(1): 110-132.

Vaccari, M., Foladori, P., Nembrini, S., and Vitali, F. (2018). "Benchmarking of energy consumption in municipal wastewater treatment plants–a survey of over 200 plants in Italy," *Water Science and Technology* 77(9): 2242-2252.

Wang, H. (2002). "Heteroscedasticity and non-monotonic efficiency effects of a stochastic frontier model," *Journal of Productivity Analysis* 18(3): 241-253.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* Cambridge: MIT Press.