UNIVERSITÀ DEGLI STUDI FIRENZE

Dipartimento di Statistica, Informatica, Applicazioni "Giuseppe Parenti"

Eccellenza 2023-2027

A GAMLSS-based Optimal Quantile estimator for Stochastic Frontiers

Francesco Vidoli, Elisa Fusco

DISIA WORKING PAPER
2025/12

A GAMLSS-based Optimal Quantile estimator for Stochastic Frontiers

Francesco Vidoli[a], Elisa Fusco[b]

[a] *Department of Economics, Society, Politics. University of Urbino Carlo Bo, Italy*
[b] *Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Italy*

Abstract

Efficiency in public services is an equity issue: inefficiency diverts resources from vulnerable populations who depend on public provision, while inaccurate measurement risks confounding structural disadvantage with managerial failure. To reply these issues, this paper proposes a new stochastic frontier estimator that combines Generalized Additive Models for Location, Scale and Shape (GAMLSS) with a data-driven optimal quantile criterion. By modelling the full conditional distribution of production outputs/costs, the approach captures non-linearity, heteroskedasticity and asymmetric inefficiency that traditional parametric frontier models cannot accommodate. Monte Carlo experiments, spanning linear, non-linear and endogenous inefficiency designs, show that the GAMLSS optimal quantile estimator systematically outperforms standard SFA and Fan-type corrections. An application to municipal waste management in Italy confirms its empirical advantages, revealing substantial heterogeneity in cost levels and dispersion. Results demonstrate that distributional flexibility is essential for fair benchmarking and targeted policy design in heterogeneous public service sectors.

Keywords: Stochastic Frontier Analysis, Quantile Regression, Generalized Additive Models for Location, Scale and Shape, Municipal Waste Management
JEL Codes: C14, C23, D24, Q53

1. Introduction

Efficiency analysis in public services is often framed as a managerial concern, reducing costs to ease fiscal pressure or satisfy budget constraints imposed by austerity policies. This technocratic view, dominant in New Public Management discourse, risks obscuring a more fundamental ethical dimension: in the provision of essential public services, inefficiency is not merely a technical failure but an issue of distributive justice.

When public resources are wasted through poor management, inadequate technology, or organizational slack, the opportunity cost is borne disproportionately by those who depend most on public provision, low-income households, geographically isolated communities, and citizens without access to private alternatives (Hirschman, 1972). Unlike private markets where inefficiency can trigger "exit" through consumer choice, public services operate under conditions of "voice" alone, and voice is unevenly distributed across social strata (Le Grand, 2009). From a Rawlsian perspective, institutions should be arranged to maximize the welfare of the least advantaged (Rawls, 1971). Inefficient public service delivery violates this principle by diverting resources away from areas where they could generate the greatest social benefit and undermines effective access to essential services, which is a constitutive element of citizens' substantive freedoms (Sen, 2001).

In municipal waste management, a service characterized by universal provision, geographic monopoly, and substantial environmental externalities, inefficiency translates directly into higher user charges, reduced service quality, or environmental degradation, all with regressive distributional consequences. Waste management inefficiency constrains "capabilities" by imposing hidden taxes through higher tariffs, environmental burdens through inadequate collection or disposal, and civic exclusion through unequal service standards across municipalities. Accurate measurement of efficiency is, therefore, not a neutral technical exercise but a prerequisite for equitable resource allocation and accountable governance.

Traditional stochastic frontier models (Aigner et al., 1977; Battese and Coelli, 1992) struggle to accommodate the heteroskedastic, asymmetric and heavy-tailed cost structures observed in sectors such as municipal waste management, whereas GAMLSS (Rigby and Stasinopoulos, 2005), combined with the quantile-based interpretation of the frontier (Jradi et al., 2019), provides a more coherent way to model the full conditional distribution and identify the relevant frontier quantile.

Our aim is to combine these insights in an empirical application to municipal waste management in Italy, a sector characterised by pronounced heterogeneity, outliers and asymmetric cost structures (Simões and Marques, 2012; Guerrini et al., 2017). Building on the full distributional flexibility of GAMLSS, we model the conditional moments as functions of demographic, infrastructural and institutional characteristics and identify the frontier through the optimal quantile implied by Jradi et al. (2019). In doing so, we introduce a new estimator (*GAMLSS_ OptQuantile*) which integrates distributional regression with a theoretically grounded quantile selection rule. This approach allows us to distinguish genuine managerial inefficiency from cost variation driven by unavoidable structural factors, an essential distinction for designing fair and effective policy interventions.

Our contribution is threefold. First, we extend frontier estimation to a fully distributional framework in which multiple parameters of the conditional production/cost distribution vary with covariates. This allows us to distinguish structural heterogeneity from inefficiency more effectively than mean-based or parametric approaches, thereby supporting more equitable performance assessment. Second, we empirically demonstrate, through Monte Carlo simulations and an application to Italian municipalities, that *GAMLSS_-OptQuantile* estimator closely approximates the true frontier under nonlinear, heteroskedastic, and endogenous inefficiency designs, outperforming ex-post Fan-type corrections and traditional stochastic frontier estimators in such settings. Third, we show that modelling the dispersion parameter explicitly reveals institutional and organizational factors that affect not only the level of production/costs but also their variability, providing richer insights for policy design than mean-based models can deliver. In a sector where service quality and environmental outcomes are critical public goods, and where vulnerable populations depend on reliable and affordable provision, these methodological refinements carry direct implications for distributive justice.

The remainder of the paper is structured as follows. Section 2 reviews the literature on stochastic frontier analysis, quantile methods, and distributional approaches, situating our contribution within the broader effort to develop flexible and equitable efficiency measurement tools. Section 3 presents the methodological rationale and outlines the integration of GAMLSS with the optimal quantile frontier approach, emphasizing the theoretical coherence of the method under heterogeneous production environments. Section 4 reports

Monte Carlo simulations assessing the performance of the proposed estimator under various data-generating processes, including scenarios with endogenous inefficiency that challenge traditional approaches. Section 5 presents an empirical application to municipal waste management costs in Italy, demonstrating how the GAMLSS framework uncovers layers of heterogeneity obscured by parametric models and how the optimal quantile criterion identifies a cost frontier consistent with both economic theory and distributional evidence. Finally, Section 6 concludes with implications for frontier estimation, public service benchmarking, and policy design in heterogeneous service sectors where efficiency and equity are inseparable objectives.

2. Literature

The classical stochastic frontier literature, originating from Aigner et al. (1977) and Meeusen and van den Broeck (1977), focuses on modelling the conditional mean of production outputs or costs. While effective for describing the central tendency of the technology, this mean-based approach cannot capture heteroskedasticity, asymmetry or tail behaviour in the conditional distribution. As modern productivity analysis increasingly recognises the importance of such distributional features, attention has shifted toward methods that go beyond expectations.

Quantile linear regression (Koenker and Bassett, 1978) and subsequent nonlinear extension (Koenker and Park, 1996; Chen, 2007) provide a natural framework for describing covariate effects across the entire distribution. However, applying it directly to frontier analysis is non-trivial, because a frontier is an absolute efficiency bound rather than a conditional quantile. Identifying the quantile that corresponds to the frontier therefore requires an explicit link between the composed error structure and the distribution of the dependent variable. Several contributions have explored this connection. Early work such as Behr (2010) considered quantile-based efficiency measures under simplified settings, while more recent research has moved toward full-distribution approaches. Notably, Tsionas (2020) and Bernstein et al. (2023) develop distributional stochastic frontier models that recover the entire conditional distribution of inefficiency, and the work of Wang et al. (2014) emphasises the importance of flexible, nonparametric modelling of both the technology and the disturbance components.

A decisive breakthrough is provided by Jradi et al. (2019), who establish a formal mapping between the Normal-Half Normal composed error and the conditional quantile at which the stochastic frontier lies. Their result resolves the identification challenge that had previously limited the use of quantile methods in frontier analysis and provides a rigorous foundation for quantile-based frontier estimation, even in non-standard quantile approaches (Fusco et al., 2023).

While these contributions move frontier analysis towards a fully distributional perspective, most stochastic frontier applications still introduce heterogeneity only through the mean function, typically by allowing observed covariates to shift the inefficiency term (Kumbhakar et al., 1991; Battese and Coelli, 1995; see also the critique in Wang, 2002). In contrast, many empirical settings display systematic variation not only in location but also in scale, skewness and tail behaviour, so that the entire shape of the conditional distribution changes with the covariates. Generalized Additive Models for Location, Scale and Shape (GAMLSS; Rigby and Stasinopoulos, 2005) are specifically designed to address this issue by allowing μ, σ, skewness and kurtosis parameters to depend flexibly on covariates through additive predictors. This framework is therefore well-suited for frontier problems in which both the magnitude and the dispersion of costs or outputs, as well as the degree of asymmetry, are driven by observed characteristics (see *e.g.* Ferrara and Vidoli, 2017).

Recent work by Papadopoulos and Parmeter (2022) further highlights the interaction between quantile methods and stochastic frontiers, showing that classical approaches may perform poorly in the presence of complex heteroskedasticity and misspecified error structures. Their analysis reinforces the need for models that can accommodate rich distributional heterogeneity when linking frontiers to conditional quantiles.

A related concern in stochastic frontier estimation is the potential endogeneity arising from correlation between inputs and the composed error. However, a large body of research (Mutter et al., 2013) argues that what is often interpreted as endogeneity is more plausibly the result of unobserved heterogeneity or functional form misspecification. As emphasised by Greene (2005), the main challenge in frontier models is disentangling inefficiency from latent heterogeneity, since inadequate modelling of heterogeneity can induce biases that mimic endogeneity. Similarly, recent work on unobserved heterogeneity in frontier settings shows that flexible distributional modelling can substantially reduce these biases, especially when compared with restrictive parametric frontiers such as the Cobb–Douglas or Translog (see, *e.g.*, Greene, 2005; Kumbhakar et al., 2007; Parmeter et al., 2014).

Given these premises, our use of GAMLSS provides a natural distributional regression backbone for the Jradi optimal-quantile criterion, extending the emerging literature on distributional stochastic frontiers by modelling not only the conditional mean, but the full conditional distribution, and by using this information to select an economically meaningful frontier quantile rather than fixing it ex ante.

Moreover, in contrast to approaches that correct endogeneity through instrumental variables, such as Karakaplan and Kutlu (2017) or the IV-based methods surveyed by Tran and Tsionas (2015), our framework tackles the problem from a different angle. By modelling the full conditional distribution of costs through GAMLSS, we allow heteroskedasticity, asymmetry and tail behaviour to vary with observable covariates, capturing forms of latent heterogeneity that would otherwise contaminate the error structure and be misinterpreted as endogeneity. This distributional flexibility, combined with the Jradi optimal-quantile criterion, mitigates the risk of confounding inefficiency with unmodelled heterogeneity and provides a robust frontier estimator without imposing instruments or strong exclusion restrictions.

Finally, the Italian waste management sector provides an ideal setting for illustrating the proposed approach, as it is widely recognised for its pronounced heterogeneity, outliers and asymmetric cost structures (Simões and Marques, 2012). Existing empirical studies for the Italian case, such as Agovino et al. (2018, 2020); Abrate et al. (2014), document substantial spatial and institutional variation in municipal performance and cost, often relying on spatial econometric models or on parametric SFA specifications. Similarly, Guerrini et al. (2017) and related contributions highlight the presence of strong dispersion in costs and service quality across utilities, confirming that the sector exhibits characteristics that challenge mean-based frontier estimators.

However, most of this literature focuses on average behaviour, providing insights into the "typical" municipality, but offering limited information on the extremes of the cost distribution. By modelling conditional quantiles, our approach allows the analysis to extend beyond the mean and sheds light on how the determinants of costs differ between best-performing and worst-performing municipalities. This is particularly relevant in a sector where technological, demographic and regulatory factors generate wide distributional heterogeneity that classical methods tend to obscure.

3. GAMLSS and the Optimal Quantile Stochastic Frontiers

As specified before, our proposal builds on the traditional stochastic frontier framework (SFA) proposed by Aigner et al. (1977) and Battese and Coelli (1992), which estimates cost or production frontiers by decomposing the composite error term into a symmetric noise component v_i and a one-sided inefficiency term $u_i \geq 0$.[1] Formally, the standard specification is

$$\ln Y_i = f(X_i) + v_i \pm u_i,$$

where $f(X_i)$ denotes the deterministic part of the frontier and the composed error $v_i \pm u_i$ captures random shocks and inefficiency[2].

As noted by many authors, a key limitation of this formulation is that it models only the deterministic part of the frontier and assumes that all remaining heterogeneity operates through the mean of the composed error. As a result, SFA models cannot accommodate non-linearities, heteroskedasticity, skewness or heavy tails in the conditional distribution of costs/production nor can they allow these features to vary with covariates.

Generalised Additive Models for Location, Scale and Shape (GAMLSS; Rigby and Stasinopoulos, 2005) provide a flexible framework for modelling the entire conditional distribution of municipal costs:

$$Y_i \mid X_i \sim D\big(\mu(X_i), \sigma(X_i), \nu(X_i), \tau(X_i)\big),$$

where μ, σ, ν, τ denote location, scale, skewness and tail thickness, allowing the frontier to be identified as an appropriate *conditional quantile* of the response variable.

In this setting, GAMLSS, therefore, not only provide a highly flexible estimator of the conditional mean of Y_i, but also deliver the entire family of conditional quantiles implied by the fitted distribution. This feature is crucial for frontier analysis: rather than limiting the analysis to $E(Y_i \mid X_i)$, the approach allows the estimation of lower-tail quantiles of the conditional cost distribution (or upper-tail quantiles of the production distribution), as

[1] In cost frontiers the inefficiency term enters with a positive sign, while in production frontiers it enters with a negative sign.

[2] In most empirical applications, the noise component is assumed to follow a Normal distribution ($v_i \sim N(0, \sigma_v^2)$), while inefficiency is modelled as a Half-Normal random variable ($u_i \sim N^+(0, \sigma_u^2)$).

already suggested by Vidoli and Ferrara (2015); Ferrara and Vidoli (2017). In other terms, GAMLSS naturally embed the concept of a stochastic frontier as a conditional quantile, offering a unified and distributionally coherent way to estimate efficiency without imposing a fixed functional form or constant error structure.

Given these premises, the key question becomes how to identify the *appropriate* conditional quantile, that is, the quantile that truly corresponds to the economic notion of a stochastic frontier. In a cost setting, this requires selecting the quantile that reflects the *minimum attainable* cost compatible with the Normal random component v_i and, therefore, attributable only to random shocks rather than to inefficiency[3].

Not all low (or high) quantiles satisfy this requirement: only the quantile that correctly separates the random noise component from the systematic inefficiency term can be interpreted as the true stochastic frontier as explicitly acknowledged in the frontier literature (Liu et al., 2008; Behr, 2010).

So, this identification problem is central, and it is precisely addressed by the quantile-based reinterpretation of the composed error proposed by Jradi et al. (2019) that provides a solution by proving that the frontier corresponds to a specific conditional quantile that can be recovered from the data via an iterative algorithm.

Their central insight is that the stochastic frontier corresponds to the quantile τ^* such that

$$P[\varepsilon_i \leq 0] = \tau^*,$$

where $\varepsilon_i = v_i - u_i$ is the composed error (for production) or $\varepsilon_i = v_i + u_i$ (for cost).

Based on the properties of the Skew-Normal distribution, they derive the following closed-form relationship:

$$\tau^* = 0.5 + \frac{\arcsin\left(-E[\varepsilon]/E[|\varepsilon|]\right)}{\pi}.$$

Since both expectations can be estimated nonparametrically from quantile regression residuals, the optimal quantile $\widehat{\tau}^*$ is recovered as follows:

1. Estimate a grid of conditional quantiles τ_c (*e.g.* 0.01 to 0.50 in a cost model or 0.50 to 0.99 in a production model).

[3] Conversely, in a production setting, the relevant frontier corresponds to the *maximum attainable* output conditional on X_i.

2. For each τ_c, estimate the quantile regression model and compute residuals $\widehat{\varepsilon}_i$.

3. Compute $\widehat{\tau}^*$ from the Jradi et al. (2019) formula.

4. Select the τ_c that minimizes (for cost) or maximizes (for production) $|\widehat{\tau}^* - \tau_c|$.

We refer to Appendix A for a formal derivation demonstrating that, under the assumption of a conditional Normal-Half Normal composed error, the theoretical properties of the Jradi et al. (2019) criterion are preserved and consistently generalized within the flexible GAMLSS framework. Specifically, we show that the bias-minimizing condition extends to the heteroskedastic setting, where the optimal quantile is no longer static but determined locally by the conditional signal-to-noise ratio $\lambda(x) = \sigma_u(x)/\sigma_v(x)$, allowing the selected global quantile to act as an effective average frontier level across the sample.

In other terms, embedding the Jradi et al. (2019) optimal quantile criterion within the GAMLSS framework still provides a theoretically grounded way to estimate the stochastic cost frontier. By modelling $\mu(X)$, $\sigma(X)$, $\nu(X)$ and $\tau(X)$ as flexible functions of the covariates, GAMLSS yields a full conditional distribution for Y_i whose quantiles naturally adjust to heteroskedasticity, skewness and departure from normality, features that are pervasive in real-world cost and production data. This allows the frontier to be identified as the *data-driven* conditional quantile implied by the Normal-Half-Normal structure, rather than through an arbitrarily pre-selected quantile level.

The following sections illustrate this property in two complementary settings. We first conduct a series of Monte Carlo experiments based on *production* frontiers, in order to assess the behaviour of the proposed GAMLSS-based optimal quantile estimator under increasing levels of non-linearity, heteroskedasticity and endogeneity. We then apply the same framework to real *cost* data from the municipal waste sector, showing that the method remains fully coherent and operational in a structurally different empirical context. This dual evidence highlights the flexibility of the GAMLSS approach: by identifying the frontier through the appropriate conditional quantile, rather than through a fixed functional form, the method adapts seamlessly to both production and cost environments.

4. Simulations

We now present a sequence of three Monte Carlo experiments that are designed to assess the performance of the proposed frontier estimators under increasing degrees of complexity. The first design considers a linear and homoskedastic stochastic production frontier that is fully consistent with the classical Normal-Half Normal structure. The second design introduces a non-linear frontier with heteroskedastic inefficiency, while the third and most demanding design adds endogeneity of inefficiency with respect to the regressor. In all cases, in addition to SFA, we compare two ways of extracting a stochastic production frontier from a flexible GAMLSS specification, namely a Fan-type ex post correction (as suggested by Ferrara and Vidoli, 2017, see Appendix B) and the optimal quantile approach in the spirit of Jradi et al. (2019) proposed in Section 3.

4.1. Linear and homoskedastic case

The first Monte Carlo experiment is presented as a baseline illustration using a linear and homoskedastic data-generating process, so as to isolate the performance of the proposed frontier estimators in a controlled setting. For transparency, we begin with a single representative design and subsequently generalise the analysis through Monte Carlo repetitions and error metrics. In this simple case, the goal is to compare the two frontier constructions when the distributional assumptions of both methods hold exactly.

With this aim, we generate a cross section of $n = 500$ observations.
The single explanatory variable is drawn from a uniform distribution, $x_i \sim U(0, 1)$ and the model follows the standard Normal-Half Normal structure of the standard SFA literature. Inefficiency is generated as a half normal random variable $u_i \sim N^+(0, 1)$ and the symmetric noise term is $v_i \sim N(0, \sigma_v^2)$ with $\sigma_v = 0.25$ has been chosen such that the signal to noise ratio is in a range typically observed in empirical applications. The true production frontier is here hypothesized as linear,

$$\ln P_i^* = 5 + 5x_i,$$

and observed productions follow the usual production frontier form,

$$y_i = 5 + 5x_i + v_i - u_i,$$

where u_i decrease the endogenous variable as in a production frontier.
To mimic realistic settings where the researcher does not know the true functional form, we estimate μ using flexible splines rather than imposing linearity

with a local polynomial basis function, implemented as a penalised B-spline term $pb(\cdot)$ of order 3 (Eilers and Marx, 1996):

$$\mu_i = pb(x_i,\, df = 3),$$

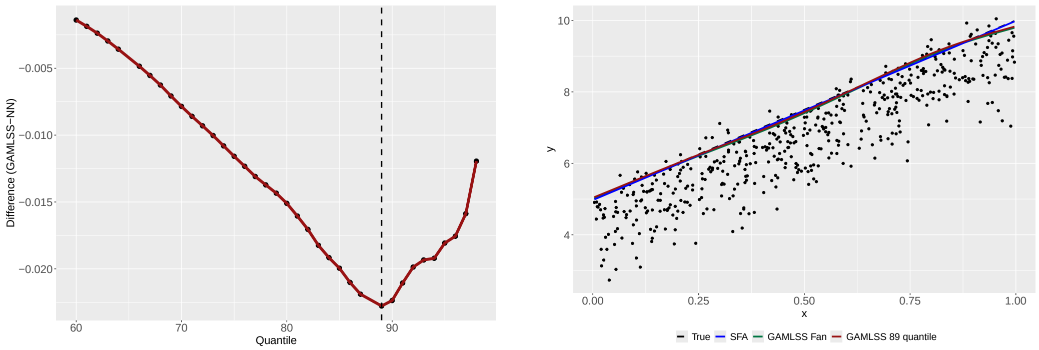
where $pb(\cdot)$ denotes a smooth function of x_i estimated using penalised B-splines. In this baseline specification we model only the location parameter μ_i, while the other distributional moments are, for the moment, kept constant.

So, the fitted conditional mean from this model, $\widehat{E}(y_i \mid x_i) = \widehat{\mu}_i$, is here used as the starting point for both GAMLSS based frontier constructions.

The first frontier estimator applies a Fan-type ex post correction to the GAMLSS mean fit. Specifically, the correction imposes the Normal-Half Normal structure on the residuals and shifts the estimated conditional mean by the implied mean inefficiency, following the profiling approach of Fan (1996). The resulting estimator, which we label *GAMLSS Fan*, preserves the flexible first-step mean function while relying on a global parametric adjustment consistent with classical SFA. Technical details are provided in Appendix B.

The second frontier estimator, presented in Section 3, does not rely on an ex post correction. Instead, it builds on the fact that, under one-sided inefficiency, the stochastic frontier coincides with a specific *conditional quantile* of the outcome distribution rather than with its conditional mean. We refer to this estimator as the *GAMLSS_ OptQuantile*. Like the Fan-corrected version, it preserves the flexible functional form estimated in the GAMLSS first step, but it locates the frontier at the quantile (equal to 0.89, see Figure 1a) that is, as demonstrated in the Appendix A, formally consistent with the Normal-Half Normal structure of the composed error.

In this simple linear and homoskedastic setting, all frontier estimators (see Figure 1b), the GAMLSS Fan, the GAMLSS optimal quantile, and the parametric SFA benchmark, closely reproduce the true frontier over the entire support of x.



(a) Estimated optimal quantile

(b) Estimated frontiers

Figure 1: Estimated optimal quantile and fitted frontiers - linear and homoskedastic case

In this benchmark case, where the distributional assumptions of the Fan (1996) and Jradi et al. (2019) methods are exactly satisfied, the two GAMLSS based frontiers are very similar, also with respect to SFA, over the entire support of x. This is reassuring, because it shows that the quantile based frontier does not contradict the classical SFA logic when the model is correctly specified. At the same time, the Jradi et al. (2019)-like approaches have the advantage of providing an explicit frontier quantile, which can be compared across designs and extended to more complex, heteroskedastic or non-linear settings considered in subsequent simulations.

To validate the robustness of the proposed approach, we, finally, performed a Monte Carlo simulation with 1,000 replications varying sample size ($N \in \{200, 500, 1000\}$), inefficiency variance ($\sigma_u \in \{0.5, 1.0, 1.5\}$), and noise variance ($\sigma_v \in \{0.2, 0.5, 0.8\}$). The "optimal quantile" was selected in each iteration by minimizing the distance between the quantile frontier and the data relative to the theoretical expected inefficiency.
Figure 2, although showing very small errors for all estimators, clearly indicates a relative advantage of the proposed method[4], except in the scenario characterised by a large variance of the symmetric noise and an almost negligible inefficiency component. In this case, observations are widely dispersed around the frontier due to stochastic noise rather than inefficiency, making accurate frontier recovery inherently difficult for any estimator.

[4] Simulation R code and additional error measures are available from the authors upon request.

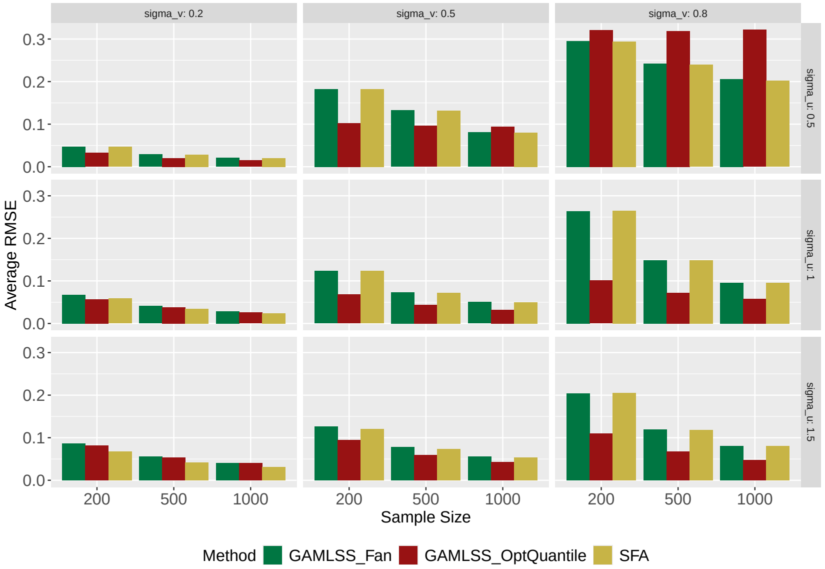


Figure 2: Average RMSE with respect to the true frontier varying sample size, σ_v and σ_u by method - 1,000 Monte Carlo replications

4.2. Non-linear and heteroskedastic case

The second Monte Carlo experiment departs from the linear homoskedastic design and considers a non-linear and heteroskedastic setting. This case is intended to stress the flexibility of the GAMLSS based frontier estimators and to highlight the differences between an ex post Fan type correction and the optimal quantile approach in a more realistic environment.

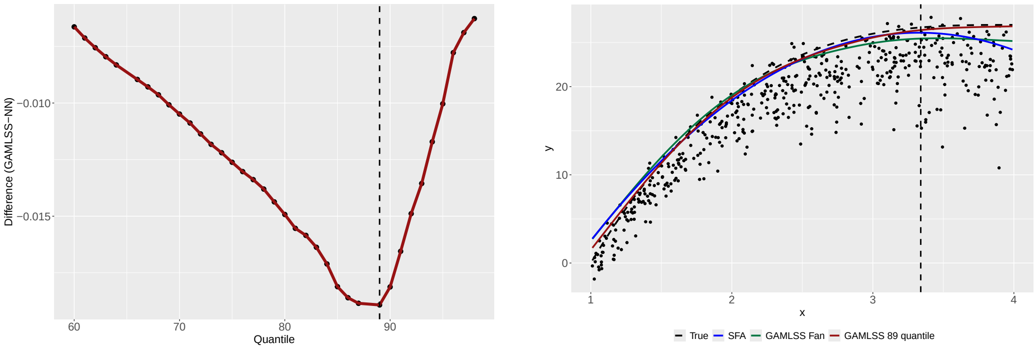
The basic setting remains the same, while the deterministic production frontier is now specified in logarithms as a cubic polynomial with the aim to introduce substantial curvature while maintaining tractability,

$$\ln P_i^* = x_i^3 - 12x_i^2 + 48x_i - 37.$$

In this setting the composed error follows the usual SFA structure and the estimation step is identical in spirit to the previous case.

Also in this case, Figure 3b compares four curves: the true frontier, the *GAMLSS_ Fan*, the *GAMLSS_ OptQuantile* (also in this case, Figure 3a indicates an optimal quantile equal to 0.89.), together with the SFA estimated frontier.

In this case, unlike the previous one, a key strength of the proposed approach becomes evident: the fact that it is able to follow the true frontier closely across the entire domain of x. By contrast, the GAMLSS-Fan and the SFA estimators reproduce the true frontier reasonably well only in the central region of the support, but tend to over and underestimate production respectively in the lower and upper parts of the domain. This behaviour reflects the inability of a global mean shift to adapt to the heteroskedastic pattern of inefficiency.



(a) Estimated optimal quantile

(b) Estimated frontiers

Figure 3: Estimated optimal quantile and fitted frontiers - non-linear and heteroskedastic case

This becomes particularly evident in the diagnostic results reported in Table 1: the estimation gains of the proposed method are especially pronounced for values of x above its 80th percentile.

Method	RMSE	MAE	BIAS	Max Error	Median AE
Full support of x					
SFA	0.9671	0.7557	-0.3743	2.7958	0.5384
GAMLSS-Fan	1.0666	0.9120	-0.3880	2.4404	0.9142
GAMLSS-Optimal-Quantile	0.3968	0.3385	-0.1914	1.4211	0.2950
Upper tail, $x \geq q(x, 80)$					
SFA	1.6007	1.4554	-1.4554	2.7958	1.3347
GAMLSS-Fan	1.5507	1.5407	-1.5407	1.8298	1.5440
GAMLSS-Optimal-Quantile	0.2438	0.2407	-0.2407	0.3012	0.2453

Table 1: Diagnostic error measures for the nonlinear frontier simulation

Figure 4, as before, summarises the Monte Carlo results for the non-linear and heteroskedastic design, reporting the average RMSE across 1,000 replications for different combinations of sample size, noise variance σ_v, and inefficiency variance σ_u. The results show a clear and systematic pattern: the *GAMLSS_ OptQuantile* estimator consistently attains the lowest RMSE across almost all parameter configurations.
The performance gap widens as either σ_v or σ_u increases: the *GAMLSS_ OptQuantile* estimator continues to displays substantially lower RMSE, thanks to its ability to adapt to the conditional distribution. By contrast, both SFA and the Fan-type correction show increasing error, particularly when inefficiency variability rises.

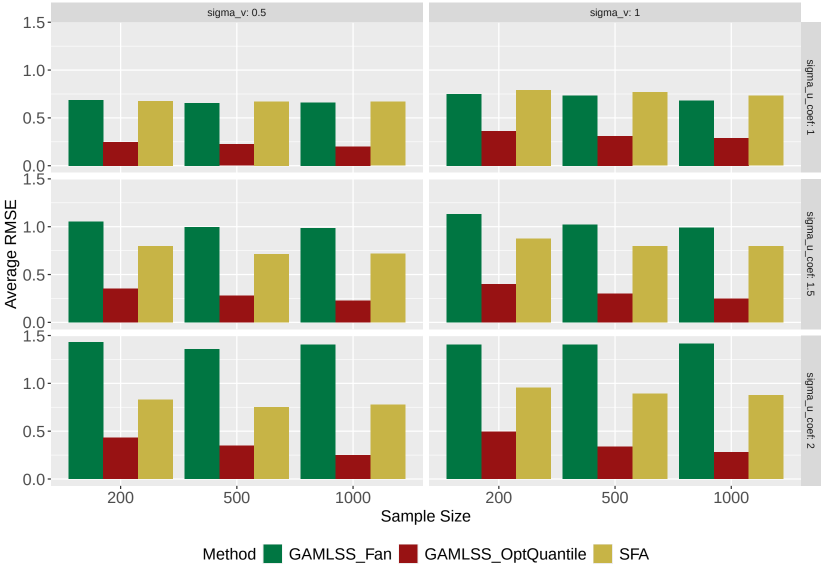


Figure 4: Average RMSE with respect to the true frontier varying sample size, σ_v and σ_u by method - 1,000 Monte Carlo replications

Overall, this non-linear and heteroskedastic experiment shows that in a setting where inefficiency varies with x_i, the optimal quantile approach based on GAMLSS provides a more flexible and locally accurate representation of the stochastic frontier than a global Fan-type correction, while remaining consistent with the classical SFA framework.

4.3. Non-linear, heteroskedastic and endogenous inefficiency

The third simulation design is the most demanding scenario and represents the final step in our progression from simple to complex data generating processes. In this case the production frontier is non-linear, inefficiency is heteroskedastic, and the inefficiency term is endogenous with respect to the regressor.

Also in this case, the basical setting remain the same and the deterministic frontier in logs is specified as a cubic polynomial,

$$\ln P_i^* = x_i^3 - 12x_i^2 + 48x_i - 37.$$

Also in this case, the symmetric noise component is homoskedastic ($v_i \sim N(0,1)$), while the inefficiency term is now both heteroskedastic and endogenous. So, we first generate a latent heteroskedastic component

$$u_i^{\text{latent}} = |z_i|, \qquad z_i \sim N(0,\, 1.5x_i),$$

so that the variance of inefficiency increases with x_i.
To introduce endogeneity, we construct an auxiliary variable

$$w_i = x_i + 0.05\,\eta_i, \qquad \eta_i \sim N(0,1),$$

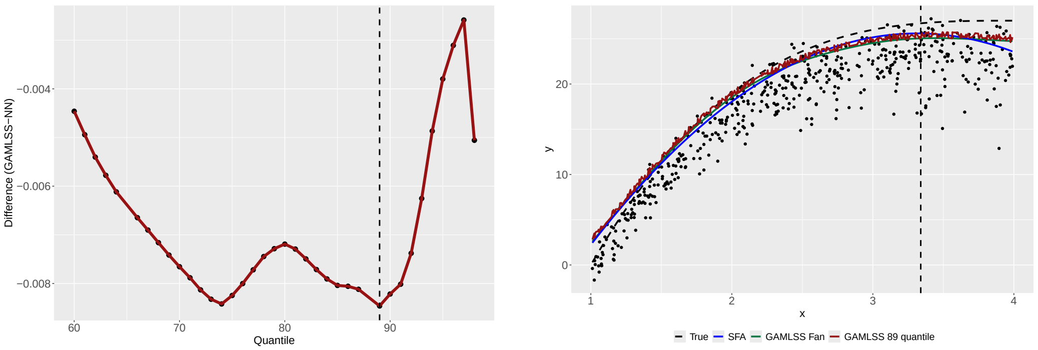
which is positively correlated with x_i. The observed inefficiency is then defined as a convex combination of w_i and the latent heteroskedastic component u_i^{latent}, weighted by the parameter ρ:

$$u_i = \rho\, w_i + (1-\rho)\, u_i^{\text{latent}}, \qquad u_i > 0,$$

with $\rho = 0.2$. As a result, u_i is positively related to x_i through both channels, inducing $E[u_i|x_i]$ to violate the standard SFA independence assumption, and the strength of this endogeneity is governed by the parameter ρ.

The two frontier estimators *GAMLSS_Fan* and *GAMLSS_ OptQuantile* are constructed exactly as in the previous designs, starting from $\widehat{\mu}_i$ and $\widehat{\sigma}_i$ and applying, respectively, the Fan type ex post correction and the Jradi based optimal quantile selection algorithm. We therefore omit the algebraic details.

Figure 5 compares the true frontier - a smooth, concave curve that increases sharply for small values of x_i and then flattens as x_i approaches four - with other three estimators. It is worth noting that, in this setting, the optimal-quantile search exhibits two local minima (Figure 5a), a consequence of the non-standard structure imposed on the inefficiency component. Nevertheless, the absolute minimum is attained at 0.89.



(a) Estimated optimal quantile

(b) Estimated frontiers

Figure 5: Estimated optimal quantile and fitted frontiers - non-linear, heteroskedastic and endogenous case

The *GAMLSS-Fan* frontier provides a reasonable fit in the central region but tends to lie systematically below the true frontier for higher values of x_i, reflecting the fact that a constant shift cannot capture the endogenous increase in inefficiency, while our approach seems to be more stable. Moreover, the estimated frontier is no longer smooth but noticeably rougher, as it now depends explicitly on the variable w. This can also be observed in Figure 6, which shows that the effect of w becomes stronger as the quantile increases, a pattern that can be captured only by a native quantile-based estimator such as the one proposed here. In fact, as the quantile increases, the frontier must adjust more strongly to the endogenous component w_i, which GAMLSS captures through its flexible quantile structure.

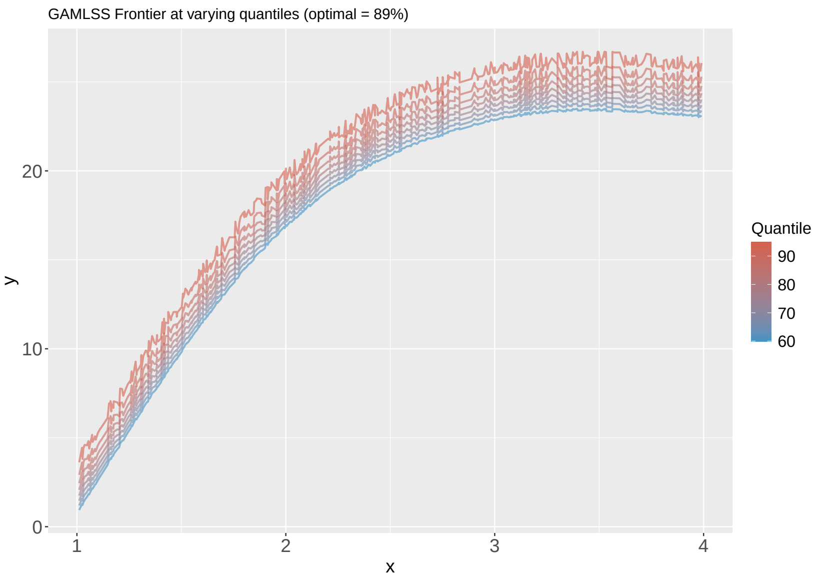


Figure 6: Estimated quantile GAMLSS frontier by quantile

Even in this case (Table 2), the optimal quantile frontier fully exploits the conditional distribution estimated by GAMLSS along all domain of x, allowing the frontier to adjust locally as a function of both x_i and the auxiliary variable w_i that drives heteroskedasticity and endogeneity.

Method	RMSE	MAE	BIAS	Max Error	Median AE
		Full support of x			
SFA	1.2577	1.0842	-0.8148	3.3957	0.9354
GAMLSS-Fan	1.3034	1.1343	-0.8235	2.3103	1.1976
GAMLSS-Optimal-Quantile	1.1446	0.9714	-0.5090	2.8337	0.9422
		Upper tail, $x \geq q(x, 80)$			
SFA	2.1296	2.0147	-2.0147	3.3957	1.8940
GAMLSS-Fan	1.9691	1.9603	-1.9603	2.2614	1.9658
GAMLSS-Optimal-Quantile	1.6631	1.6357	-1.6357	2.3771	1.6380

Table 2: Diagnostic error measures for the non-linear, heteroskedastic and endogenous frontier simulation

Finally, also in this case, we tested the approach in a Monte Carlo setting of *endogenous inefficiency*. Here, the inefficiency term u_i is not independent but correlated with the regressor x_i via an auxiliary variable $z_i \approx x_i$. Figure 7 illustrates that when the correlation parameter is set to $\rho = 0$, all estimators perform relatively similarly, although the optimal–quantile frontier already attains the lowest average RMSE. As ρ rises to 0.4 and 0.8, thereby strengthening the dependence between inefficiency and the regressor, the advantages of the proposed method become progressively more pronounced. Both SFA and the *GAMLSS_ Fan* correction deteriorate sharply under stronger endogeneity, whereas the *GAMLSS_ OptQuantile* estimator remains comparatively stable and continues to track the true frontier with substantially lower error.

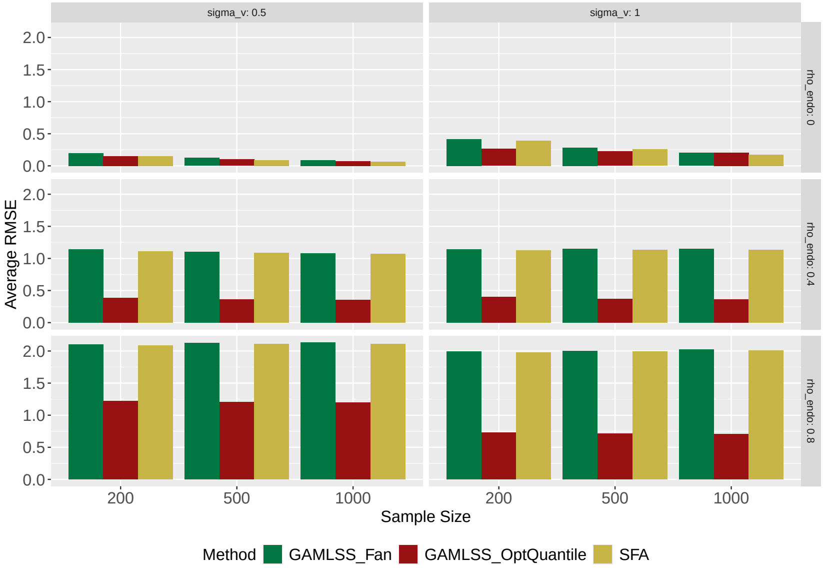


Figure 7: Average RMSE with respect to the true frontier varying sample size, σ_v and σ_u by method - 1,000 Monte Carlo replications

The results highlight that while standard SFA estimates become biased due to the violation of the independence assumption, the *GAMLSS_ OptQuantile* estimator adapts to the local distributional shifts induced by the endogeneity, maintaining lower RMSE levels across all correlation strengths and adapting to local changes in the conditional distribution.

In summary, this most demanding simulation design confirms that the GAMLSS based optimal quantile method is more robust than Fan type ex post corrections when the frontier is non-linear, inefficiency is heteroskedastic and endogenous, and the researcher relies on flexible first step distributions rather than fully parametric SFA specifications. This is the scenario that most closely resembles our empirical application (in this case within a cost-frontier framework), and it is precisely in this setting that the proposed method delivers the clearest gains.

5. An application to Municipal Waste Management Costs

The simulation results in Section 4 demonstrate that the GAMLSS-optimal quantile estimator provides accurate frontier estimates when inefficiency is heteroskedastic, non-linear, and endogenous, precisely the conditions likely to arise in heterogeneous service sectors. We now apply the method to municipal waste management costs in Italy, where substantial variation in demographic, infrastructural, institutional, and territorial conditions generates heterogeneity in both cost levels and dispersion. As noted by Simões and Marques (2012), Abrate et al. (2014), Guerrini et al. (2017) and Fusco and Allegrini (2020) ignoring such distributional features can hinder accurate frontier identification and bias efficiency assessments.

To provide a benchmark for comparison and to illustrate these issues within a parametric setting, we first estimate a stochastic cost frontier model using the official dataset developed by SOSE for the assessment of municipal standard needs in the waste management sector[5]. The database combines administrative records, survey data and ISPRA information on regional treatment capacity, and covers Italian municipalities for the years 2010, 2013, 2015 and 2016. It includes detailed measures of municipal waste expenditure, quantities generated, recycling shares, population and territorial characteristics, distances to treatment facilities, and organisational features such as associated management arrangements. All variables are defined at the municipal level, yielding a large cross-sectional dataset suitable for analysing the technological and institutional determinants of waste management costs.

The SFA model (Table 3) identifies strong and significant effects of demographic factors, infrastructural constraints, and regional treatment policies, while simultaneously quantifying substantial inefficiency. The high value of γ confirms that the frontier model is appropriate in this context and that inefficiency is a central component of cross-municipal cost variation.

[5] https://www.opencivitas.it/en/open-data.

Variable	**Estimate**	**Std. Error**	**z-value**
Intercept	7.5155	0.0258	291.113***
Population	6.76e-07	5.73e-08	11.794***
Total Buildings	0.09112	0.00187	48.656***
Population Density (per km^2)	3.54e-05	3.07e-06	11.528***
Urban Waste per Capita	-1.3784	0.0289	-47.669***
Urban Waste per Capita (sq.)	0.57025	0.01869	30.507***
Distance to Plants (weighted)	0.001885	9.20e-05	20.480***
Regional Landfills (number)	-0.002959	0.000520	-5.691***
Regional Compost Treatment Share	-0.008753	0.000187	-46.860***
Regional Incineration Treatment Share	-0.003973	0.000125	-31.793***
Associated Management Dummy	0.01457	0.00449	3.243**
σ^2	0.17242	0.00259	66.541***
γ	0.79178	0.00717	110.495***

Log-likelihood: -4159.357

Observations: 24,323 (cross-section)

Mean efficiency: 0.765

Table 3: Stochastic Cost Frontier Estimates (Battese & Coelli, 1992)

All the main explanatory variables are statistically significant and display the expected signs for a municipal cost function. Demographic and structural variables such as population, total buildings, and population density increase costs, reflecting the scale and complexity of service provision in larger or denser municipalities. The non-linear relationship in per-capita waste generation is strongly confirmed: the negative coefficient on the linear term and the positive coefficient on the squared term indicate a U-shaped pattern. At low levels of waste generation, economies of density reduce costs, while at higher levels the pressure on collection and treatment systems raises costs.

On the supply side, the distance to treatment plants significantly increases costs due to higher transportation requirements. The number of landfills has a small but negative effect, which may signal the presence of disposal alternatives that reduce upward pressure on treatment costs. Regional treatment mix variables are among the strongest predictors. A higher regional composting share reduces local costs, consistent with the lower cost structure of composting. In contrast, a higher share of incineration reduces costs in this specification, suggesting that access to thermal treatment capacity may alleviate bottlenecks or exploit economies of scale.

Finally, the associated management dummy has a positive and significant effect, indicating that municipalities participating in an associated management structure (about 32% of the municipalities) exhibit slightly higher frontier costs once inefficiency is controlled for.

The SFA model yields a mean efficiency of 0.765, indicating potential cost reductions of approximately 23%. More importantly, the estimate $\gamma = 0.79$ reveals that nearly 80% of the residual variance stems from inefficiency rather than noise, confirming that management performance differences dominate cost dispersion across municipalities.

However, the SFA specification imposes strong restrictions: a parametric functional form, homoskedastic noise, and constant inefficiency variance. These assumptions may be violated in a sector where cost dispersion, distributional asymmetry, and tail behaviour vary systematically with demographic and infrastructural characteristics. The high γ value itself suggests that the lower tail of the conditional cost distribution approximates the efficient frontier, motivating a quantile-based approach that can capture such heterogeneity without global parametric corrections.

We then re-estimate the model within a GAMLSS framework, obtaining the full conditional distribution of costs and identifying the optimal frontier quantile via the Jradi et al. (2019) criterion, without relying on restrictive functional forms or constant error assumptions.

In cost frontier analysis, efficient units lie in the lower tail of the conditional distribution, as they achieve minimum expenditure given their operating conditions. Unlike production settings where the frontier corresponds to high quantiles (maximum output), cost efficiency requires identifying low quantiles that reflect best-practice performance.
Applying the iterative Jradi et al. (2019) algorithm over the range $\tau_c \in [0.10, 0.50]$, Figure 8 identifies the optimal cost frontier at $\tau^* = 0.34$. This value, substantially below the median, is consistent with the high inefficiency variance ($\gamma = 0.79$) estimated under SFA and indicates that approximately two-thirds of municipalities operate above the efficient frontier.

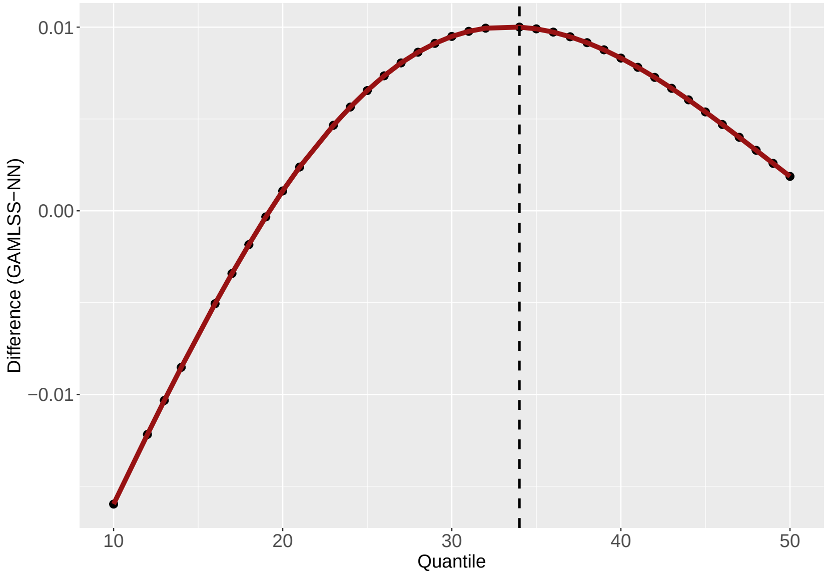


Figure 8: Estimated optimal cost quantile

Variable	**Estimate**	**Std. Error**	**t-value**
Location parameter μ (identity link)			
Intercept	7.7430	0.0272	285.224***
Population	7.38e-07	6.84e-08	10.801***
Total Buildings	0.09597	0.00196	49.049***
Population Density (per km^2)	2.85e-05	3.19e-06	8.927***
Urban Waste per Capita	-1.4254	0.0312	-45.690***
Urban Waste per Capita (sq.)	0.6132	0.0208	29.464***
Distance to Treatment Plants (weighted)	0.001827	9.67e-05	18.886***
Regional Landfills (number)	-0.004657	0.000514	-9.052***
Regional Compost Treatment Share	-0.008524	0.000184	-46.316***
Regional Incineration Share	-0.004634	0.000126	-36.640***
Scale parameter σ (log link)			
Intercept	-1.3500	0.00904	-149.319***
Regional Code (categorical)	0.01335	0.00086	15.578***
Associated Management Dummy	0.09400	0.01124	8.362***
Observations: 24,323 AIC: 9432.6 SBC: 9546.0			

Table 4: GAMLSS Estimates for Municipal Waste Management Costs, $\tau^* = 0.34$

The GAMLSS estimates (Table 4) provide a richer characterization of the determinants of municipal waste management costs by modelling simultaneously the conditional mean (μ) and the conditional variance (σ). This allows the analysis to distinguish between factors that influence the expected cost level and factors that drive heterogeneity in the dispersion of costs across municipalities, supporting more equitable policy targeting by distinguishing structural constraints from managerial inefficiency.

The results for the location parameter μ closely mirror those obtained in SFA estimates, but with greater numerical stability and remarkably high statistical significance. Population size, total buildings and population density all increase costs, consistent with scale and urbanisation effects. The non-linear pattern in waste generation per capita (negative linear term and positive quadratic term) confirms the presence of economies of density at lower waste levels and rising marginal costs beyond a threshold.

Supply-side factors also play an important role. Distance to treatment facilities significantly raises costs, while the coefficient for regional landfills is negative, suggesting that additional disposal options may relieve pressure on local treatment costs. The treatment mix variables have very large and precisely estimated effects: a higher regional share of composting reduces costs markedly, while a higher share of incineration also reduces costs in the GAMLSS specification. This latter sign, already emerging in the SFA frontier, likely reflects the role of thermal treatment capacity in stabilising the waste cycle and reducing upstream collection or transfer costs once inefficiency is separated from systematic variation.

From a policy perspective, therefore, the location parameter $\mu(X)$ absorbs unavoidable cost drivers (distance to treatment plants, regional infrastructure deficits, demographic pressures) that should inform compensatory transfers or infrastructure investment rather than trigger accountability mechanisms. Conversely, units with costs substantially above their estimated conditional quantile $\widehat{Q}_{\tau*}(Y_i|X_i)$, after controlling for these structural factors, represent candidates for performance incentives, technical assistance, or regulatory oversight. This decomposition prevents the systematic misclassification of structurally disadvantaged municipalities as inefficient, a risk inherent in mean-based benchmarks that conflate heterogeneity with poor management.

The explicit modelling of the dispersion parameter $\sigma(X)$, instead, reveals a second dimension for intervention design. Associated management arrangements, while potentially achieving scale economies, increase conditional cost variance by 9.4%, signalling greater operational unpredictability that may reflect coordination challenges, contractual heterogeneity, or incomplete risk-sharing mechanisms.

The positive and significant coefficient on the regional indicator implies that the heterogeneity of costs differs systematically across regions (approximately 1.3% across regions), even after conditioning on observables in the mean equation. This is consistent with institutional fragmentation and differing regulatory or infrastructural contexts.

The positive effect of the associated management dummy on σ suggests that municipalities belonging to joint-management arrangements experience greater dispersion in costs. This may reflect differences in contractual structures, service standards or economies of coordination across consortia.

Policymakers should recognize the trade-off between μ and σ drivers: municipalities in high-variance institutional settings require different support, such as risk-pooling mechanisms, clearer governance protocols, or enhanced monitoring than those in low-variance contexts facing pure efficiency deficits. By adapting both the frontier benchmark and the dispersion structure to local conditions, the GAMLSS optimal quantile approach ensures that accountability and support are allocated where they can genuinely improve performance, rather than penalizing circumstances beyond managerial control.

6. Final remarks

This paper began with the premise that efficiency measurement in public services is not a neutral technical exercise, but an issue of distributive justice. When essential services are delivered inefficiently, the costs, whether through higher tariffs, reduced quality, or environmental degradation, fall disproportionately on those who depend most on public provision and have the least voice in demanding accountability.

Given these premises, this paper develops a new frontier estimation approach that embeds the optimal-quantile criterion of Jradi et al. (2019) within a GAMLSS distributional framework. By modelling the full conditional distribution of costs or outputs, rather than only the conditional mean, the method provides a flexible, data-driven frontier estimator that remains valid under nonlinearity, heteroskedasticity, asymmetry and endogeneity in inefficiency.

Simulation results show that the *GAMLSS_ OptQuantile* estimator consistently recovers the true frontier across increasingly complex designs. Its advantages become particularly evident in scenarios with heteroskedastic or endogenous inefficiency, where traditional SFA and Fan-type mean-shift corrections deteriorate because they rely on global parametric adjustments that cannot adapt to local variation in the conditional distribution. In contrast, the proposed estimator leverages the full set of GAMLSS parameters, allowing the frontier to adjust locally to changes in both location and scale.

The empirical application to municipal waste management confirms these insights. By modelling both the mean and the dispersion of costs, the GAMLSS framework uncovers multiple layers of structural heterogeneity that parametric SFA cannot accommodate. The estimated optimal quantile identifies a lower-tail cost frontier consistent with the large inefficiency component detected by standard SFA, providing a more robust benchmark for evaluating municipal performance.

Overall, our results highlight the usefulness of full-distribution methods for productivity analysis in heterogeneous service sectors, demonstrating that the equity ethical imperative links with concrete technical implications: accurate frontier estimation requires methods capable of distinguishing genuine managerial inefficiency from structural heterogeneity that reflects unavoidable cost drivers beyond administrative control. Embedding a theoretically grounded quantile frontier within a flexible GAMLSS structure offers a practical and conceptually coherent alternative to classical SFA, particularly when the underlying technological and institutional environment generates asymmetric, heavy-tailed or heteroskedastic cost structures.
Future research may extend this approach to panel data, alternative distributional families and multi-output settings.

References

Abrate, G., Erbetta, F., Fraquelli, G., Vannoni, D., 2014. Costs and convergence in the italian municipal solid waste sector. Utilities Policy 30, 12–19.

Agovino, M., Cerciello, M., Garofalo, A., 2020. Behavioral and socioeconomic determinants of municipal waste recycling: Evidence from italy. Journal of Cleaner Production 254, 120776.

Agovino, M., Ferrara, M., Garofalo, A., 2018. Waste management performance in italian provinces: Efficiency and spatial effects. Ecological Indicators 89, 680–693.

Aigner, D., Lovell, C.K., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production function models. Journal of Econometrics 6, 21–37.

Battese, G.E., Coelli, T.J., 1992. Frontier production functions, technical efficiency and panel data: With application to paddy farmers in india. Journal of Productivity Analysis 3, 153–169.

Battese, G.E., Coelli, T.J., 1995. A model for technical inefficiency effects in a stochastic frontier production function for panel data. Empirical Economics 20, 325–332.

Behr, A., 2010. Quantile regression for stochastic frontier models. Empirical Economics 39, 593–608.

Bernstein, D., Parmeter, C.F., Tsionas, E.G., 2023. Distributional stochastic frontier models. Journal of Econometrics 232, 356–380.

Chen, C., 2007. A finite smoothing algorithm for quantile regression. Journal of Computational and Graphical Statistics 16, 136–164.

Eilers, P.H., Marx, B.D., 1996. Flexible smoothing with b-splines and penalties. Statistical Science 11, 89–102.

Fan, S., 1996. On the estimation of mean inefficiency in stochastic frontier models. Journal of Productivity Analysis 7, 137–160.

Ferrara, G., Vidoli, F., 2017. Semiparametric stochastic frontier models: A generalized additive model approach. European Journal of Operational Research 258, 761–777.

Fusco, E., Allegrini, V., 2020. The role of spatial interdependence in local government cost efficiency: An application to waste italian sector. Socio-Economic Planning Sciences 69, 100681.

Fusco, E., Benedetti, R., Vidoli, F., 2023. Stochastic frontier estimation through parametric modelling of quantile regression coefficients. Empirical Economics 64, 869–896.

Greene, W., 2005. Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. Journal of Econometrics 126, 269–303.

Guerrini, A., Romano, G., Campedelli, B., 2017. Efficiency and environmental factors in the italian waste management sector. Waste Management 67, 442–452.

Hirschman, A.O., 1972. Exit, voice, and loyalty: Responses to decline in firms, organizations, and states. Harvard university press.

Jradi, S., Parmeter, C.F., Ruggiero, J., 2019. Quantile estimation of the stochastic frontier model. Economics Letters 182, 15–18.

Karakaplan, M.U., Kutlu, L., 2017. Handling endogeneity in stochastic frontier analysis. Economics Bulletin 37, 889–901.

Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46, 33–50.

Koenker, R., Park, B.J., 1996. An interior point algorithm for nonlinear quantile regression. Journal of Econometrics 71, 265–283.

Kumbhakar, S.C., Ghosh, S., McGuckin, J.T., 1991. A generalized production frontier approach for estimating determinants of inefficiency in us dairy farms. Journal of Business & Economic Statistics 9, 279–286.

Kumbhakar, S.C., Park, B.U., Simar, L., Tsionas, E.G., 2007. Nonparametric stochastic frontiers: A local maximum likelihood approach. Journal of Econometrics 137, 1–27.

Le Grand, J., 2009. The other invisible hand: Delivering public services through choice and competition. Princeton university press.

Liu, S.M., Wang, H.J., Lee, L.F., 2008. Quantile estimation of a heteroscedastic stochastic frontier model. Journal of Econometrics 147, 349–364.

Meeusen, W., van den Broeck, J., 1977. Efficiency estimation from cobb–douglas production functions with composed error. International Economic Review 18, 435–444.

Mutter, R.L., Rosko, M.D., Greene, W.H., Wilson, P., 2013. The impact of unobserved heterogeneity on stochastic frontier estimates of efficiency. Health Economics 22, 189–205.

Papadopoulos, A., Parmeter, C.F., 2022. Quantile methods for stochastic frontier analysis. Foundations and Trends® in Econometrics 12, 1–107.

Parmeter, C.F., Sun, K., Henderson, D.J., Kumbhakar, S.C., 2014. Estimation and inference in semiparametric stochastic frontier models. Journal of Applied Econometrics 29, 919–947.

Rawls, J., 1971. A theory of justice. Harvard Press, Cambridge.

Rigby, R.A., Stasinopoulos, D.M., 2005. Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society: Series C (Applied Statistics) 54, 507–554.

Sen, A., 2001. Development as Freedom. Number 9780192893307 in OUP Catalogue. none ed., Oxford University Press.

Simões, P., Marques, R.C., 2012. Literature review on the measurement of water and waste utilities' efficiency. Omega 40, 894–905.

Tran, K., Tsionas, E.G., 2015. Endogeneity in stochastic frontier models: A bayesian perspective. Econometric Reviews 34, 825–849.

Tsionas, M.G., 2020. Quantile stochastic frontiers. European Journal of Operational Research 282, 1177–1184.

Vidoli, F., Ferrara, G., 2015. Analyzing italian citrus sector by semi-nonparametric frontier efficiency models. Empirical economics 49, 641–658.

Wang, H.J., 2002. Model formulation, specification, and estimation of the stochastic frontier model. Journal of Econometrics 108, 85–106.

Wang, Q.L., Parmeter, C.F., Racine, J.S., 2014. Regression estimators with nonparametric generated regressors. Journal of Econometrics 180, 92–106.

Appendix A. Validity of the Jradi optimal quantile under GAMLSS

The optimal quantile frontier approach of Jradi et al. (2019) is derived under the standard Normal–Half–Normal stochastic frontier structure. In this Appendix, we formally extend this result to a GAMLSS setting where distributional parameters vary flexibly with covariates, we clarify the relationship between the general GAMLSS framework and the composed-error structure, and we discuss the estimator's robustness to departures from the baseline assumptions.

Appendix A.1. The classical Jradi result

Consider the standard stochastic frontier model

$$Y_i = f(X_i) + \varepsilon_i, \qquad \varepsilon_i = v_i - u_i, \tag{A.1}$$

where $f(\cdot)$ is the cost frontier, $v_i \sim N(0, \sigma_v^2)$ is symmetric noise, $u_i = |z_i|$ with $z_i \sim N(0, \sigma_u^2)$ is one-sided inefficiency, and $v_i \perp u_i$. Under these assumptions, Jradi et al. (2019) show that the cost frontier corresponds to a specific quantile of the distribution of Y_i, and the associated quantile level τ can be expressed as a function of the first two moments of the composed error:

$$\tau = 0.5 + \frac{1}{\pi} \arcsin\left(\frac{m_1}{m_A}\right), \tag{A.2}$$

where $m_1 = -E[\varepsilon_i] = E[u_i] > 0$ and $m_A = E|\varepsilon_i|$.

The key insight of Jradi et al. (2019) is that, under the Normal–Half–Normal structure, the stochastic frontier can be identified as the τ-quantile of Y_i rather than through a location shift of the conditional mean. This provides a distributional interpretation of frontier estimation and motivates quantile-based approaches.

Appendix A.2. Conditional Normal–Half–Normal structure

We now extend this framework to allow the parameters of the error distribution to vary with covariates. In our setting, the distribution of Y_i is modelled conditionally on X_i as

$$Y_i \mid X_i = x = f(x) + \varepsilon_i(x), \qquad \varepsilon_i(x) = v_i(x) - u_i(x), \tag{A.3}$$

where both the symmetric noise and the inefficiency components are now indexed by the conditioning value x. The analysis is therefore conducted at the conditional level, allowing for heteroskedasticity in both error components.

Assumption 1. *For each x in the support of X_i, the composed error satisfies*

$$v_i(x) \sim N\big(0, \sigma_v^2(x)\big), \qquad u_i(x) = |z_i(x)|, \quad z_i(x) \sim N\big(0, \sigma_u^2(x)\big), \tag{A.4}$$

with $v_i(x) \perp u_i(x)$. Define the conditional moments

$$m_1(x) = -E\big[\varepsilon_i(x) \mid X_i = x\big] = E\big[u_i(x) \mid X_i = x\big], \qquad m_A(x) = E\big[|\varepsilon_i(x)| \mid X_i = x\big]. \tag{A.5}$$

Assumption 1 generalizes the classical Normal–Half–Normal structure by allowing the variance parameters $\sigma_v^2(x)$ and $\sigma_u^2(x)$ to vary with x, while maintaining the local independence between noise and inefficiency. This accommodates heteroskedastic inefficiency and heteroskedastic noise, both of which are empirically relevant in many applications.

Remark 1 (Relationship to GAMLSS). *In Section 3, we specify a GAMLSS model of the form*

$$Y_i \mid X_i \sim NO\big(\mu(X_i), \sigma^2(X_i)\big),$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ are smooth functions of covariates estimated via flexible regression techniques (e.g., neural networks, penalized splines). At first glance, this appears to model $Y_i \mid X_i$ as a simple normal distribution, which would be inconsistent with the composed-error structure in Assumption 1.

However, the two representations are compatible when interpreted correctly. The GAMLSS specification models the marginal conditional distribution *of Y_i given X_i, which, under the composed error $\varepsilon_i(x) = v_i(x) - u_i(x)$, is* not *exactly normal, but can be well approximated by a normal distribution with location $\mu(x)$ and scale $\sigma(x)$ for the purpose of estimating conditional quantiles. In other terms, GAMLSS does not impose a parametric structure on the composed error. It provides a flexible approximation of the conditional distribution of observed costs, from which quantiles can be computed. The Jradi quantile criterion is then applied to this estimated distribution. Specifically:*

- *The GAMLSS location parameter $\mu(x)$ estimates $E[Y_i \mid X_i = x] = f(x) - E[u_i(x) \mid X_i = x]$, which includes both the frontier $f(x)$ and the expected inefficiency.*
- *The GAMLSS scale parameter $\sigma(x)$ captures the total conditional dispersion, $Var(Y_i \mid X_i = x) = \sigma_v^2(x) + \sigma_u^2(x)$.*
- *By modelling $\mu(x)$ and $\sigma(x)$ flexibly, GAMLSS provides a data-driven approximation to the conditional distribution of Y_i without requiring explicit parametric forms for $f(x)$, $\sigma_v(x)$, or $\sigma_u(x)$.*

Assumption 1 is then invoked when we apply the Jradi criterion to extract the frontier from the estimated conditional distribution. In other words, we use GAMLSS to model the conditional moments $\mu(x)$ and $\sigma(x)$ non-parametrically*, but we rely on the Normal–Half–Normal structure to justify the quantile-based identification of the frontier. This approach separates the flexible estimation of the conditional distribution (via GAMLSS) from the structural interpretation of the frontier (via the Jradi mapping).*

In practice, even if the true conditional distribution of $Y_i \mid X_i$ is not exactly normal (due to the presence of one-sided inefficiency), the normal approximation provided by GAMLSS is sufficiently accurate for computing conditional quantiles, especially when $\sigma_u(x)$ is not too large relative to $\sigma_v(x)$. Our simulation results in Sections 4 confirm that this approximation works well across a range of designs.

Please note that we do not claim that the true composed error strictly follows a conditional Normal-Half Normal structure; instead, we use the Jradi mapping as a pseudo-structural criterion that identifies the quantile minimising bias under this family of mixtures, which our simulations show to be robust even under deviations from the classical model.

Remark 2 (Notation: $f(x)$ vs. $\mu(x)$). *To avoid confusion, we clarify the distinction between $f(x)$ and $\mu(x)$:*

- *$f(x)$ denotes the* cost frontier*, i.e., the minimum attainable cost for a unit with characteristics x operating at full efficiency. This is the object of interest in frontier analysis.*
- *$\mu(x) = E[Y_i \mid X_i = x]$ denotes the* conditional mean *of observed costs, which includes both the frontier and the expected inefficiency:*

  $$\mu(x) = f(x) + E[v_i(x) \mid X_i = x] - E[u_i(x) \mid X_i = x] = f(x) - E[u_i(x) \mid X_i = x],$$

  where $E[u_i(x) \mid X_i = x] > 0$ represents the expected inefficiency at x.

In the presence of one-sided inefficiency, $\mu(x) \neq f(x)$. The GAMLSS framework estimates $\mu(x)$ (and $\sigma(x)$) directly from the data, and we then use the Jradi optimal quantile criterion to extract $f(x)$ from the estimated conditional distribution. This avoids the need for parametric corrections (such as Fan-type adjustments) that shift $\hat{\mu}(x)$ by an estimated mean inefficiency.

Appendix A.3. Pointwise optimal quantile: Main result

We now establish that the Jradi optimal quantile criterion extends naturally to the conditional framework under Assumption 1.

Proposition 1. *Under Assumption 1, for each x in the support of X_i, there exists a conditional quantile level $\tau(x) \in (0,1)$ such that the stochastic frontier can be written as*

$$f(x) = Q_{\tau(x)}\big(Y_i \mid X_i = x\big), \tag{A.6}$$

where $Q_{\tau(x)}(\cdot \mid X_i = x)$ denotes the $\tau(x)$-th conditional quantile function of Y_i given $X_i = x$. Moreover, the quantile level $\tau(x)$ satisfies

$$\tau(x) = 0.5 + \frac{1}{\pi}\arcsin\left(\frac{m_1(x)}{m_A(x)}\right), \tag{A.7}$$

i.e., the Jradi mapping between the first two moments of the composed error and the frontier quantile holds pointwise in x.

Proof 1. *Fix any x in the support of X_i. By Assumption 1, conditionally on $X_i = x$ we have*

$$Y_i \mid X_i = x = f(x) + \varepsilon_i(x), \qquad \varepsilon_i(x) = v_i(x) - u_i(x),$$

with $v_i(x) \sim N\big(0, \sigma_v^2(x)\big)$, $u_i(x) = |z_i(x)|$, $z_i(x) \sim N\big(0, \sigma_u^2(x)\big)$, and $v_i(x) \perp u_i(x)$.

For this fixed value of x, the conditional model is therefore a standard Normal–Half–Normal stochastic frontier model with parameters $\sigma_v^2(x)$ and $\sigma_u^2(x)$. The distribution of the composed error $\varepsilon_i(x) \mid X_i = x$ is thus a convolution of a normal and a (negative) half-normal random variable, which is well-defined and has finite first and absolute first moments.

Denote by $F_{\varepsilon(\cdot|x)}$ the cumulative distribution function of $\varepsilon_i(x) \mid X_i = x$, and by $Q_{\varepsilon(\cdot|x)}(\cdot)$ the corresponding quantile function. Let

$$m_1(x) = -E\big[\varepsilon_i(x) \mid X_i = x\big] = E\big[u_i(x) \mid X_i = x\big], \qquad m_A(x) = E\big[|\varepsilon_i(x)| \mid X_i = x\big].$$

Since the law of $\varepsilon_i(x) \mid X_i = x$ follows a Normal–Half–Normal structure with finite first absolute moment, the derivations in Jradi et al. (2019) apply verbatim to the conditional model at x. In particular, their Theorem 1 establishes that there exists a unique $\tau(x) \in (0,1)$ such that the stochastic frontier can be expressed as the $\tau(x)$-quantile of $Y_i \mid X_i = x$, i.e.,

$$f(x) = Q_{\tau(x)}\big(Y_i \mid X_i = x\big),$$

and the quantile level satisfies

$$\tau(x) = 0.5 + \frac{1}{\pi}\arcsin\left(\frac{m_1(x)}{m_A(x)}\right).$$

To make this last step explicit, note that the conditional distribution of $Y_i \mid X_i = x$ *is obtained by a location shift of* $\varepsilon_i(x) \mid X_i = x$*, so that for any* $q \in (0,1)$

$$Q_q\big(Y_i \mid X_i = x\big) = f(x) + Q_q\big(\varepsilon_i(x) \mid X_i = x\big).$$

Hence $f(x)$ *coincides with the* q*-th conditional quantile of* Y_i *if and only if* 0 *coincides with the* q*-th quantile of* $\varepsilon_i(x) \mid X_i = x$*. Theorem 1 in Jradi et al. (2019) states that, for a Normal–Half–Normal composed error, the unique quantile level* q *such that* $Q_q(\varepsilon_i) = 0$ *is given by*

$$q = 0.5 + \frac{1}{\pi} \arcsin\left(\frac{m_1}{m_A}\right),$$

where $m_1 = -E[\varepsilon_i]$ *and* $m_A = E|\varepsilon_i|$*. Applying this relationship to* $\varepsilon_i(x) \mid X_i = x$ *yields precisely*

$$\tau(x) = 0.5 + \frac{1}{\pi} \arcsin\left(\frac{m_1(x)}{m_A(x)}\right),$$

with $m_1(x)$ *and* $m_A(x)$ *as defined in (A.5).*

This proves the existence of $\tau(x)$ *and the claimed expression for the corresponding quantile level, and completes the argument for the fixed value* x*. Since* x *was arbitrary, the result holds for every* x *in the support of* X_i*, which establishes the proposition.*

Proposition 1 shows that the Jradi optimal quantile has a natural conditional extension under GAMLSS: as long as the composed error retains a Normal–Half–Normal structure given $X_i = x$, the frontier remains a conditional quantile of Y_i and the corresponding quantile level is characterised by the same moment-based mapping as in the unconditional case.

Remark 3 (Dependence of $\tau(x)$ on x). *The quantile level* $\tau(x)$ *in equation (A.7) depends on* x *through the ratio of moments* $m_1(x)/m_A(x)$*, which in turn depends on the local parameters* $\sigma_v(x)$ *and* $\sigma_u(x)$*. Specifically, it can be shown that* $\tau(x)$ *is an increasing function of the signal-to-noise ratio* $\lambda(x) = \sigma_u(x)/\sigma_v(x)$*:*

- *When* $\lambda(x)$ *is small (i.e., inefficiency variance is small relative to noise variance),* $\tau(x)$ *is close to* 0.5*, meaning the frontier lies near the median of the conditional distribution.*

- *When* $\lambda(x)$ *is large (i.e., inefficiency dominates),* $\tau(x)$ *approaches* 1*, reflecting the fact that most observations are shifted upward by inefficiency and the frontier lies in the lower tail.*

In the special case where $\lambda(x) = \lambda$ is constant across x, the quantile level $\tau(x) = \tau$ does not depend on x, and the frontier is a single, global conditional quantile of Y_i, exactly as in Jradi et al. (2019). When $\lambda(x)$ varies with x, the frontier can be viewed as a quantile function with x-specific order $\tau(x)$. *This is a natural generalization that accommodates heterogeneous production environments, and it motivates the empirical strategy described in the next subsection.*

Appendix A.4. From pointwise optimal quantiles to a global estimator

Proposition 1 establishes that, for each value x, the frontier corresponds to a conditional quantile with level $\tau(x)$. In general, $\tau(x)$ varies with x whenever the ratio $\lambda(x) = \sigma_u(x)/\sigma_v(x)$ is not constant. However, our empirical implementation requires selecting a *single* quantile level $k \in (0, 1)$ for estimation. This raises the question: how do we reconcile the pointwise optimality of $\tau(x)$ with the need for a global quantile estimator?

Since $\tau(x)$ varies with x whenever the signal-to-noise ratio $\sigma_u(x)/\sigma_v(x)$ is not constant, the frontier cannot be represented by a unique quantile level in a strict structural sense. In empirical applications, however, a global frontier curve is required. We therefore seek the quantile level k^* whose implied Jradi value $\hat{\tau}(k)$ is most consistent with the observed distribution of residuals. This global k^* can be interpreted as a distribution-weighted average of the pointwise optimal quantiles, providing a coherent and stable empirical frontier.

We address this issue by searching over a grid of candidate quantiles $k \in \mathcal{K} = \{k_1, \ldots, k_K\}$ (e.g., $\mathcal{K} = \{0.60, 0.61, \ldots, 0.98\}$ in our implementation) and selecting the value k^* that minimizes the self-consistency criterion

$$k^* = \arg\min_{k \in \mathcal{K}} \left| k - \hat{\tau}(k) \right|, \tag{A.8}$$

where

$$\hat{\tau}(k) = 0.5 + \frac{1}{\pi} \arcsin\left(\frac{\hat{m}_1(k)}{\hat{m}_A(k)} \right), \tag{A.9}$$

and the sample moments are computed from the residuals $Y_i - \hat{Q}_k(Y_i|X_i)$:

$$\hat{m}_1(k) = -\frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \hat{Q}_k(Y_i|X_i) \big), \quad \hat{m}_A(k) = \frac{1}{n} \sum_{i=1}^{n} \big| Y_i - \hat{Q}_k(Y_i|X_i) \big|. \tag{A.10}$$

Here, $\hat{Q}_k(Y_i|X_i)$ is the k-th conditional quantile estimated from the GAMLSS model. Specifically, given the fitted GAMLSS location $\hat{\mu}(X_i)$ and scale $\hat{\sigma}(X_i)$, we compute

$$\hat{Q}_k(Y_i|X_i) = \hat{\mu}(X_i) + \hat{\sigma}(X_i) \cdot \Phi^{-1}(k), \tag{A.11}$$

where Φ^{-1} is the quantile function of the standard normal distribution.

The procedure in equations (A.8)–(A.11) can be interpreted as follows:

1. For each candidate quantile k, we compute the conditional quantile curve $\hat{Q}_k(Y_i|X_i)$ and treat it as a provisional estimate of the frontier.

2. We compute the residuals $Y_i - \hat{Q}_k(Y_i|X_i)$ and evaluate their sample moments $\hat{m}_1(k)$ and $\hat{m}_A(k)$.

3. We apply the Jradi formula (A.9) to these moments, which yields an *implied* frontier quantile level $\hat{\tau}(k)$. This is the quantile level that the data "suggest" as optimal, given the provisional frontier $\hat{Q}_k(Y_i|X_i)$.

4. We select k^* as the quantile that is *self-consistent*, i.e., for which the implied level $\hat{\tau}(k)$ is closest to k itself. This ensures that the chosen quantile k^* is consistent with the underlying Normal–Half–Normal structure.

The selected quantile k^* can be interpreted as a data-driven approximation to a population-weighted average of the pointwise optimal levels $\tau(x)$. Formally, if we define

$$\bar{\tau} = \int \tau(x)\, dF_X(x),$$

where F_X is the marginal distribution of X_i, then k^* provides a consistent estimate of $\bar{\tau}$ under suitable regularity conditions. In settings where $\lambda(x)$ is approximately constant—or varies slowly with x—we expect $\tau(x) \approx \tau$ for all x, and k^* provides a consistent estimate of the global frontier quantile. When $\lambda(x)$ exhibits substantial variation, k^* captures an "average" frontier level, and the GAMLSS framework ensures that the estimated conditional quantile $\hat{Q}_{k^*}(Y_i|X_i)$ adapts locally to heteroskedasticity through the estimated scale function $\hat{\sigma}(X_i)$.

Remark 4 (Computational implementation). *In practice, the optimization in* (A.8) *is straightforward because $\mathcal{K}$ is a finite grid. For each $k \in \mathcal{K}$, we:*

1. *Compute $\hat{Q}_k(Y_i|X_i)$ from the fitted GAMLSS model using equation* (A.11)*.*

2. *Compute the sample moments $\hat{m}_1(k)$ and $\hat{m}_A(k)$ from the residuals.*

3. *Evaluate $\hat{\tau}(k)$ using equation* (A.9)*.*

4. *Store the value $|k - \hat{\tau}(k)|$.*

We then select $k^* = \arg\min_{k\in\mathcal{K}} |k - \hat{\tau}(k)|$. The final frontier estimate is $\hat{f}(X_i) = \hat{Q}_{k^*}(Y_i|X_i)$.

The grid $\mathcal{K}$ is typically chosen to cover a range of quantiles in the upper half of the distribution (e.g., $[0.60, 0.98]$ for cost frontiers, where higher quantiles correspond to lower costs after accounting for the sign convention). The grid spacing (e.g., 0.01) is chosen to balance computational cost and precision. In our simulations and empirical application, a spacing of 0.01 proves sufficient.

Appendix B. Details of the Fan-type correction

For completeness, we summarise the full ex post correction applied to the GAMLSS mean fit. Let $\widehat{\mu}_i$ denote the estimated conditional mean from the baseline GAMLSS, and define residuals

$$\widehat{\varepsilon}_i = y_i - \widehat{\mu}_i.$$

Assuming a Normal–Half–Normal structure for the composed error $\varepsilon_i = v_i - u_i$, we profile the shape parameter

$$\lambda = \frac{\sigma_u}{\sigma_v}$$

by maximising the log-likelihood function derived in Fan (1996) over a compact interval.

Given $\widehat{\lambda}$ and the empirical variance of the residuals, the variances of v_i and u_i are recovered using the moment expressions of the Normal–Half–Normal model. Mean inefficiency is then

$$\widehat{\mu}_u = E[u_i] = \sqrt{\frac{2}{\pi}} \frac{\widehat{\sigma}_u^2}{\sqrt{\widehat{\sigma}_u^2 + \widehat{\sigma}_v^2}}.$$

The Fan-corrected frontier is finally obtained as

$$\widehat{y}_i^{\text{FRONT, Fan}} = \widehat{\mu}_i + \widehat{\mu}_u.$$