統計数理(2026)
第 74 巻 第 1 号 5–19

特集「統計エキスパート人材育成プロジェクト」
[原著論文]

木構造に基づく確率分布推定を用いた DEA 効率性評価

趙 宇†

(受付 2025 年 5 月 29 日；改訂 8 月 26 日；採択 9 月 2 日)

要　　旨

標準的な Data Envelopment Analysis(DEA)モデルは決定論的であり，観測データに測定誤差や外生的ノイズが存在しないことを前提としている．しかし，実データには確率的変動が不可避であり，これを無視すると効率性推定に深刻な誤差が生じる可能性がある．既存のブートストラップ法や回帰型アプローチは統計的推論を部分的に導入しているものの，データの不確実性を直接的に扱う枠組みは確立されていない．本研究では，DEA における効率性および信頼区間を推定するため，密度木と密度フォレストに基づく新たなサンプリング手法を提案する．提案手法は，観測データをクラスタに自動分類し，各クラスタ内でガウス型エントロピー関数に基づく情報利得を最大化することで局所的な確率構造をモデル化する．さらに，アンサンブル学習により推定の安定性と頑健性を確保し，観測データのばらつきやノイズを反映した擬似データを生成する．これにより，DEA における統計的不確実性を柔軟かつ実証的に評価できる．電力企業データによる実証分析では，観測データに基づく DEA 効率値と，提案手法による推定効率値との間で Kolmogorov–Smirnov 検定を行った結果，統計的に有意な分布差は認められなかった．この結果は，提案手法が観測データの分布的特性を適切に再現し，測定誤差やノイズを考慮した効率性評価を可能にする有効なアプローチであることを示唆している．

キーワード：密度木，密度フォレスト，サンプリング，信頼区間推定，不確実性．

1. はじめに

Data Envelopment Analysis(DEA)は，効率性を評価するためのノンパラメトリック手法であり，経済，医療，教育，エネルギーなど多様な分野でパフォーマンス評価や政策立案に広く用いられてきた．DEA における効率性とは，一般に，多入力・多出力を有する意思決定単位(Decision Making Unit；DMU)において，最小限の入力で最大限の出力を得るという観点から，入力が出力へと変換される生産過程の効率を指す．しかし，従来の DEA モデルは入出力データに誤差が存在しないことを前提としており，実データに不可避な測定誤差や外生的ノイズなどの確率的変動を考慮していない．これらは効率性推定の精度に大きな影響を及ぼすため，近年では DEA に統計的推論を導入し，不確実性を考慮した効率性評価の研究が進展している．

Simar and Wilson (1998)は，DEA におけるブートストラップ手法を初めて提案し，効率性の信頼区間推定や仮説検定を可能にした．その後，Kneip et al. (2008)が漸近理論を統合して統計的基盤を強化し，Moradi-Motlagh and Emrouznejad (2022)は過去 20 年間のブートストラップ

† 東京理科大学 経営学部：〒 102–0071 東京都千代田区富士見 1–11–2; yu.zhao@rs.tus.ac.jp

手法の理論的発展を整理するとともに，ロバスト確率的手法（Aragon et al., 2005; Cazals et al., 2002）や条件付き効率性推定（Daraio and Simar, 2014; Daraio et al., 2020）の重要性を指摘した．また，Dia et al.（2022），Kerstens et al.（2022），Walheer（2022），Ngo and Tsui（2022），Michali et al.（2023），Lin and Lu（2024），Kang et al.（2024）なども近年の主要な発展に位置づけられる．これらの中で，Simar and Wilson（1998）は最も広く用いられているが，本稿の 2.2 節で示すとおり，その手法はデータに内在する誤差構造や異質性を十分に捉えられないという限界がある．このため，観測された入出力ベクトルを確率変数として扱い，その確率構造に基づいた柔軟なサンプリングおよび信頼区間推定手法の開発が求められている．

別の流れとして，Kuosmanen and Kortelainen（2012）による Stochastic Nonparametric Envelopment of Data（StoNED）がある．StoNED は DEA と（Aigner et al., 1977, SFA）を統合し，非効率性とノイズを合成した誤差項をセミノンパラメトリックに推定可能とした．ここでいう非効率性は，入力の過剰や出力の不足といった改善余地を示す尺度であり，効率性と表裏一体の概念である．これにより確率的変動を考慮した効率性推定が可能となったが，Mergoni et al.（2025）が指摘するように，計算負荷や次元の呪いといった課題が残る．

さらに近年，機械学習の進展も DEA 研究に影響を与えている．Zhu et al.（2021），Aparicio et al.（2023），Boubaker et al.（2025），Esteve et al.（2023），Shi and Zhao（2024）は，DEA と機械学習を統合した効率性測定・予測手法を提案しており，また Valero-Carreras et al.（2021），Aparicio and Esteve（2023），Guillen et al.（2023），Mergoni et al.（2025）はフロンティア推定の精度向上を目指している．しかし，これらの研究はいずれも推定精度やフロンティア構造の改善に主眼を置いており，効率性の信頼区間推定に関する議論はほとんど見られない．

以上を踏まえ，本研究では確率的変動を適切に考慮した DEA 効率性評価手法の開発を目的とし，木構造（密度木および密度フォレスト）に基づく新たなサンプリングアルゴリズムを提案する．本手法は，観測データから高密度領域を自動抽出し，ガウス型エントロピー関数（Criminisi and Shotton, 2013）による情報利得最大化で局所的確率構造をモデル化する．さらに，アンサンブル学習により推定の安定性と頑健性を高め，入出力データの確率的特性を反映した疑似データ生成を実現する．

本論文の構成は以下のとおりである．第 2 章で DEA の概要と不確実環境下での効率性評価の課題を述べ，第 3 章で密度木に基づくサンプリングアルゴリズムを提案する．第 4 章で効率性の信頼区間構築手法を示し，第 5 章で電力企業データによる実証分析を行い，第 6 章で結論と今後の課題を示す．

2. DEA の概要

2.1 決定論的アプローチとしての DEA

少ない入力から多くの出力を得ようとする多入力・多出力を有する DMU を対象とする．入力および出力のベクトルを $\boldsymbol{v} \in \mathbb{R}^d_+ := (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^m_+ \times \mathbb{R}^s_+$ と定義する．ただし，$\boldsymbol{x}$ は入力，$\boldsymbol{y}$ は出力を表す．すべての $\boldsymbol{v}$ によって構成される集合 P は，実現可能な入出力の組み合わせを表し，生産可能集合と呼ばれる．この集合 P の境界は生産フロンティアと呼ばれ，その上に位置する DMU は効率的であるとされる．一般に，効率的でない DMU を効率的にするためには，その入出力水準を生産フロンティア上の水準まで改善する必要がある．改善の方向には複数の選択肢が存在するが，本稿では実務で広く用いられている入力指向型モデルに基づく枠組みを採用する[1)]．以下に，その定式化を示す．

生産可能集合 P に対して，入力指向型の効率性を δ としたとき，効率値 $\hat{\delta}$ は次の線形計画問題を解くことで得られる（Banker et al., 1984）：

$$\hat{\delta} := \min\{\delta > 0 \mid (\delta \boldsymbol{x}, \boldsymbol{y}) \in P\}. \tag{2.1}$$

ここで δ は，所与の出力 $\boldsymbol{y}$ を維持したまま，入力 $\boldsymbol{x}$ をどれだけ縮小できるかを表す尺度であり，効率的な DMU では $\hat{\delta} = 1$，非効率的な DMU では $\hat{\delta} < 1$ となる．

今，n 個の DMU が観測されたとする．DEA では，複数の観測された DMU の入出力データに基づき，それらを包含する経験的生産可能集合 P^{DEA} を構築する：

$$P^{DEA} := \left\{ (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}_+^{m+s} \;\middle|\; \sum_{h=1}^{n} \lambda_h \boldsymbol{x}_h \leq \boldsymbol{x}, \quad \sum_{h=1}^{n} \lambda_h \boldsymbol{y}_h \geq \boldsymbol{y}, \quad \sum_{h=1}^{n} \lambda_h = 1, \quad \boldsymbol{\lambda} \geq \boldsymbol{0} \right\}. \tag{2.2}$$

ここで，$\boldsymbol{x}_h \in \mathbb{R}_+^m$ および $\boldsymbol{y}_h \in \mathbb{R}_+^s$ は，第 h 番目の DMU における入力および出力ベクトルを表し，$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^\top$ は n 次元の非負ベクトルである．制約 $\sum_{h=1}^{n} \lambda_h = 1$ を課すことで，P^{DEA} は観測された DMU の凸包およびその凸包より大きい入力・小さい出力をもつ点の集合となり，このとき規模に関する収穫は可変(Variable Returns to Scale, VRS)となる（Banker, 1984）[2)]．

評価対象となる $\mathrm{DMU}_o, o \in \{1, 2, \ldots, n\}$ について，P^{DEA} に基づき，DEA における入力指向型の効率性は，次の式によって定式化できる：

$$\hat{\delta}_o = \min\{\delta > 0 \mid (\delta \boldsymbol{x}_o, \boldsymbol{y}_o) \in P^{DEA}\}. \tag{2.3}$$

この問題は BCC(Banker, Charnes and Cooper)モデル（Banker, 1984）と呼ばれ，各 DMU ごとに個別に解く必要がある．以降，DEA における記法の慣習に従い，混乱の生じるおそれがない場合には，DMU の番号を表す添字を省略して記述する．

2.2 不確実環境下における DEA 効率性評価の課題

従来，DEA に基づく効率性に対して統計的推論を導入する手法として，Simar and Wilson (1998)によるブートストラップ法(以下，Simar–Wilson 法)が広く採用されてきた．Simar–Wilson 法では，観測された入力および出力に測定誤差や外生的ノイズが含まれる可能性は考慮されておらず，データのばらつきはすべて各 DMU の効率性の違いに起因すると仮定されている．この仮定のもとでは，経験的生産可能集合 P^{DEA} は真の生産可能集合 P の部分集合 ($P^{DEA} \subset P$)となるため，効率値 $\hat{\delta}_o$ は系統的に上方バイアスされることが知られている（Simar and Wilson, 2004）．すなわち，有限サンプルに基づく DEA フロンティアは真のフロンティアより内側に位置し，結果として効率値が過大推定される傾向が生じる．Simar–Wilson 法はブートストラップによるバイアス補正を試みるが，データ生成過程に内在する測定誤差や外生的ノイズを直接モデル化していないため，これらの影響を十分に補正できるわけではない．さらに，Simar–Wilson 法はすべての DMU が同一の効率性分布から独立に抽出されているとする同質性の仮定を前提としている．この仮定のもとでは，効率性のばらつきは主として確率的な変動に起因するものとみなされるが，実際には技術水準や経営環境の異なる集団が混在している場合が多く，一律の効率性分布を仮定することは現実的ではない．Olesen and Petersen（2016）はこの点を批判し，Simar–Wilson 法では効率性のばらつきを全体的な傾向として扱うため，異常値や特異な DMU の構造的特徴が適切に捉えられず，不公平な評価を招く可能性があると指摘している．

Simar–Wilson 法は，DEA に統計的推論を導入する重要な突破口を提供した一方で，観測データに内在する誤差構造や異質性を十分に捉えられないという限界を有する．これらの課題を克服するためには，観測された入出力ベクトルを確率変数として扱い，その確率構造に基づいたより柔軟なサンプリングおよび信頼区間の推定手法の開発が必要である．

3. 木構造に基づくサンプリング

3.1 準備：密度木(Density Tree)

入出力ベクトル $\boldsymbol{v} \in \mathbb{R}_+^d$ を考える．ここで，$\mathcal{S}_0 := \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ は n 個の DMU(データ点)の集合である．本研究で用いる密度木 (Criminisi and Shotton, 2013)は，頂点(ノード)とそれらを結ぶ辺から構成される階層的な二分木であり，各ノード $j \in \{0, 1, \ldots, J\}$ にはデータ集合 $\mathcal{S}_j \subseteq \mathcal{S}_0$ が対応する．

各ノード j では，分割関数 $h_j(\boldsymbol{v}, \boldsymbol{\theta}) \in \{0, 1\}$ に基づき，データ $\mathcal{S}_j$ を左右に分割する．ここで分割パラメータ $\boldsymbol{\theta} := \{d, t\}$ は，$d \in \{1, \ldots, m, m+1, \ldots, m+s\}$(分割方向)および閾値 $t \in \mathbb{R}$ から構成され，定義域は $\mathcal{T} = \{(d, t) \mid d \in \{1, \ldots, m, m+1, \ldots, m+s\}, t \in \mathbb{R}\}$ である．

ノード j に到達したデータ $\mathcal{S}_j$ は次のように二分割される：

$$\mathcal{S}_j^L = \{\boldsymbol{v} \in \mathcal{S}_j \mid h_j(\boldsymbol{v}, \boldsymbol{\theta}) = 0\}, \tag{3.1}$$

$$\mathcal{S}_j^R = \{\boldsymbol{v} \in \mathcal{S}_j \mid h_j(\boldsymbol{v}, \boldsymbol{\theta}) = 1\}. \tag{3.2}$$

分割パラメータ $\boldsymbol{\theta}$ は，エントロピー関数に基づく情報利得を最大化するように逐次的に決定される．情報利得 $I(\mathcal{S}_j, \boldsymbol{\theta})$ は以下で定義される：

$$I(\mathcal{S}_j, \boldsymbol{\theta}) = H(\mathcal{S}_j) - \sum_{i \in \{L, R\}} \frac{|\mathcal{S}_j^i|}{|\mathcal{S}_j|} H(\mathcal{S}_j^i), \tag{3.3}$$

ここで，$|\cdot|$ は集合の濃度を表し，$H(\cdot)$ はクラスタリング目的に用いるエントロピー関数である．

最適な分割パラメータ $\boldsymbol{\theta}^*$ は次式で与えられる：

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \mathcal{T}} I(\mathcal{S}_j, \boldsymbol{\theta}). \tag{3.4}$$

二分木による分割構造を図 1 に示す．情報利得 $I(\mathcal{S}_j, \boldsymbol{\theta})$ が大きいほど，分割により親ノードの不確実性 $H(\mathcal{S}_j)$ が大幅に減少したことを意味し，その分割がデータの内部構造を効果的に捉えていると解釈できる．したがって，密度木は情報利得を最大化する分割を逐次的に適用することで，階層的かつ確率的に意味のあるクラスタ構造を形成できる．

なお，本研究では Criminisi and Shotton (2013) に倣い，各クラスタ(ノード)をガウス分布で近似できると仮定し，エントロピー $H(\cdot)$ にはガウス分布に基づくエントロピーを用いる．すなわち，$\mathcal{S}_j$ に属する各データ $\boldsymbol{v}$ を多変量正規分布に従う確率変数 $\boldsymbol{V}$ の実現値とみなす：

$$\boldsymbol{V} \sim \mathcal{N}_d(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j). \tag{3.5}$$

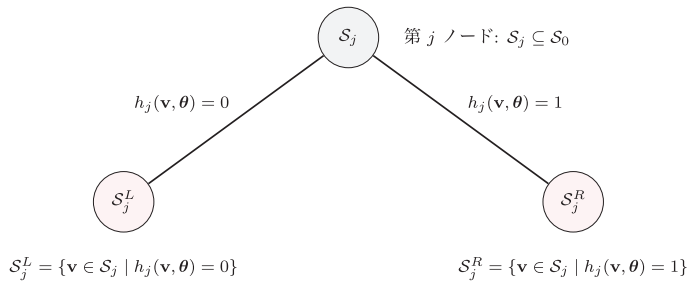

図 1. 二分木による分割構造：ノード $\mathcal{S}_j$ は分割関数 $h_j(\boldsymbol{v}, \boldsymbol{\theta})$ に基づき左右に分割される．

ここで，$\boldsymbol{\mu}_j \in \mathbb{R}_+^d$ は平均ベクトル，$\boldsymbol{\Sigma}_j$ は正定値対称共分散行列である．対応する確率密度関数は以下の通りである：

$$p(\boldsymbol{v}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \frac{1}{(2\pi)^{d/2} \det(\boldsymbol{\Sigma}_j)^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{v} - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{v} - \boldsymbol{\mu}_j)\right). \tag{3.6}$$

このときのエントロピーは次式により与えられる（Cover and Thomas, 2012, p.256）：

$$H(\mathcal{S}_j) = \frac{1}{2}\log((2\pi e)^d \det(\boldsymbol{\Sigma}_j)). \tag{3.7}$$

従って，式(3.3)は次のように書き換えられる：

$$I(\mathcal{S}_j, \boldsymbol{\theta}) = \log(\det(\boldsymbol{\Sigma}_j)) - \sum_{i\in\{L,R\}} \frac{|\mathcal{S}_j^i|}{|\mathcal{S}_j|} \log(\det(\boldsymbol{\Sigma}_j^i)). \tag{3.8}$$

この情報利得を最大化する分割パラメータを逐次的に選定することで，$\mathcal{S}_0$ を複数のクラスタに分割できる．各クラスタは密度木の葉ノードに対応し，クラスタサイズによる重み付けにより，単点クラスタの生成を抑制する設計となっている．

葉ノードのインデックスを $l(\boldsymbol{v}) : \mathbb{R}_+^d \to \mathbb{N}$ とすると，密度木全体における混合分布は次式で表される（Criminisi and Shotton, 2013）：

$$p(\boldsymbol{v}) = \sum_{l(\boldsymbol{v})} \frac{|\mathcal{S}_{l(\boldsymbol{v})}|}{|\mathcal{S}_0|} p(\boldsymbol{v}|\boldsymbol{\mu}_{l(\boldsymbol{v})}, \boldsymbol{\Sigma}_{l(\boldsymbol{v})}). \tag{3.9}$$

ここで $\sum_{l(\boldsymbol{v})} \frac{|\mathcal{S}_{l(\boldsymbol{v})}|}{|\mathcal{S}_0|} = 1$ を満たす．これはガウス混合モデルとして解釈できる．

図 2 は，密度木を用いた数値実験の結果を示している．ここでは，$X \sim \mathrm{Uniform}(0, 1)$ および $U \sim \mathrm{Exp}(1/\mu), \mu = 1/5$ に従う確率変数に基づき，関数 $Y = X^{0.5}\exp(-U)$ により生成された 200 個の二次元データを対象としている．深さ 0 から 5 まで密度木による分割を適用した結果，深さ 0 では全データを 2 つのクラスタに分割し，各クラスタにおける情報利得は 0.78 で

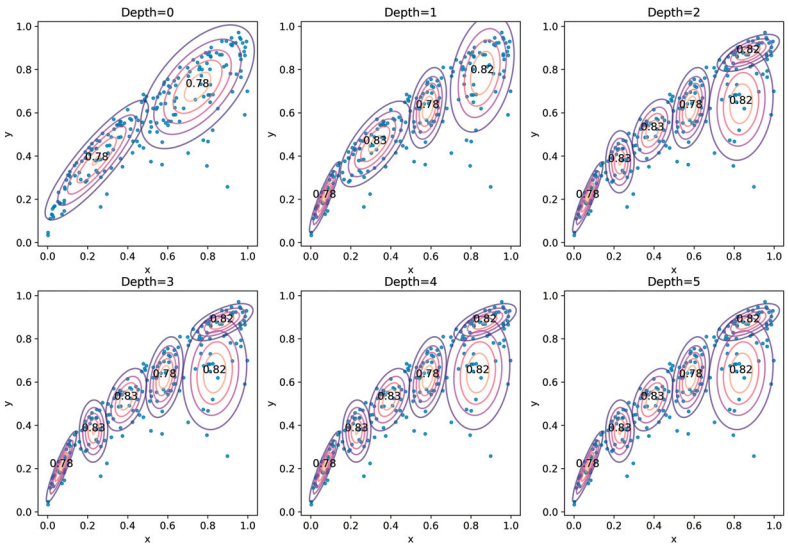

図 2．密度木によるクラスタリング．

あった．深さが増すごとにクラスタ数は増加し，情報利得の更新が観察されるが，深さ 2 以降は分割構造が安定し，情報利得の増加も停滞する．この結果より，密度木は高密度領域に中心を持つクラスタを形成する傾向があることが確認できる．

なお，初期仮定として $\boldsymbol{V} \sim \mathcal{N}_d(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ を置いているが，図 2 の例に見られるように，全体分布がガウス分布から逸脱していても密度木は有効に機能する．これは分割基準の設計がガウス分布を前提としており，各クラスタが局所的にガウス近似されることに起因する．実際，この生成モデルにおいては，$X = x$ の条件下で Y はスケールされたベータ分布に従い，$Y \mid X = x \sim x^{0.5}Z$, $Z \sim \mathrm{Beta}(5, 1)$ が成り立つことは明らかである．

3.2 木構造に基づくサンプリングアルゴリズム

本節では，Simar–Wilson 法のように各 DMU に同一の効率性分布を仮定することなく，観測データに基づき入出力ベクトルの擬似データを生成するための木構造ベースのサンプリング手法を提案する．本手法は，密度木およびそのアンサンブルである密度フォレストに基づき，局所的な確率分布を推定し，これを利用して擬似データを生成するものである．生成されたデータは，後述する効率性の統計的性質（推定値の分布や信頼区間など）の評価に用いられる．

ここで，真の生産フロンティアは一意に存在すると仮定し，入力ベクトル $\boldsymbol{x}$ を確率変数 $\boldsymbol{X}$ の実現値として，また効率値 $\hat{\delta}$ を確率変数 D の実現値として扱う．このとき，以下を仮定する：

$$E(D) = \mu > 0, \quad \mathrm{Var}(D) < \infty. \tag{3.10}$$

この枠組みの下で，所与の出力 $\boldsymbol{y}$ に対する入力 $\boldsymbol{X}$ の不確実性や効率性分布の統計的性質を推定するため，以下のサンプリングアルゴリズムを導入する．

Algorithm 1 密度木に基づくサンプリングアルゴリズム（ガウス混合分布）

Require: 観測された入出力ベクトル集合 $\{(\boldsymbol{x}_h, \boldsymbol{y}_h)\}_{h=1}^n$
Ensure: 生成された擬似入力ベクトル集合 $\{\boldsymbol{x}_h^b\}_{h=1}^n, \quad b = 1, \ldots, B$

密度木の構築：
for 各ノード $j = 0, 1, \ldots, J$ **do**
　情報利得を最大化する分割パラメータを決定：

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \mathcal{T}} I(\mathcal{S}_j, \boldsymbol{\theta})$$

　分割関数 $h_j(\cdot, \boldsymbol{\theta}^*)$ に基づき，データを左右に分割：

$$\mathcal{S}_j^L = \{\boldsymbol{v} \in \mathcal{S}_j \mid h_j(\boldsymbol{v}, \boldsymbol{\theta}^*) = 0\},$$
$$\mathcal{S}_j^R = \{\boldsymbol{v} \in \mathcal{S}_j \mid h_j(\boldsymbol{v}, \boldsymbol{\theta}^*) = 1\}$$

end for
葉ノードにおける局所分布推定：
for 各葉ノード $\mathcal{S}_l$ **do**
　平均ベクトル $\boldsymbol{\mu}_l$ および共分散行列 $\boldsymbol{\Sigma}_l$ を推定
　葉ノードの事前確率を計算：

$$p(L = l) = \frac{|\mathcal{S}_l|}{|\mathcal{S}_0|}$$

end for
条件付きサンプリング：
for 各データ点 $(\boldsymbol{x}_h, \boldsymbol{y}_h)$ **do**

13: ベイズの定理により各クラスタに属する事後確率を計算：

$$p(L=l \mid \boldsymbol{Y}=\boldsymbol{y}_h)=\frac{p(\boldsymbol{y}_h \mid L=l)\,p(L=l)}{\sum_{l'} p(\boldsymbol{y}_h \mid L=l')\,p(L=l')}$$

14: 事後確率 $p(L=l \mid \boldsymbol{Y}=\boldsymbol{y}_h)$ に基づきクラスタ l^* をサンプリング

15: クラスタ l^* における条件付き平均および分散を計算：

$$\boldsymbol{\mu}_{x|y}^{(l^*)}=\boldsymbol{\mu}_x^{(l^*)}+\boldsymbol{\Sigma}_{xy}^{(l^*)}(\boldsymbol{\Sigma}_{yy}^{(l^*)})^{-1}(\boldsymbol{y}_h-\boldsymbol{\mu}_y^{(l^*)}),$$
$$\boldsymbol{\Sigma}_{x|y}^{(l^*)}=\boldsymbol{\Sigma}_{xx}^{(l^*)}-\boldsymbol{\Sigma}_{xy}^{(l^*)}(\boldsymbol{\Sigma}_{yy}^{(l^*)})^{-1}\boldsymbol{\Sigma}_{yx}^{(l^*)}$$

16: 条件付き分布 $\mathcal{N}_m(\boldsymbol{\mu}_{x|y}^{(l^*)},\boldsymbol{\Sigma}_{x|y}^{(l^*)})$ から $\boldsymbol{x}_h^b$ をサンプリング

17: **end for**

アルゴリズム 1 では，情報利得に基づき再帰的に分割を行うことにより，密度木を構築する．その後，各葉ノードに局所的な多変量正規分布を推定し，これらを事前確率に基づいて重ね合わせることで，全体としてガウス混合モデルを構成する．さらに，観測された出力 $\boldsymbol{y}_h$ に対してベイズの定理を用いてクラスタの事後確率を計算し，擬似入力 $\boldsymbol{x}_h^b$ をサンプリングする．

このとき，ガウス混合モデルにおける $\boldsymbol{Y}=\boldsymbol{y}$ に対する条件付き分布は次式で表される：

$$p(\boldsymbol{X} \mid \boldsymbol{Y}=\boldsymbol{y})=\sum_l \frac{|\mathcal{S}_l|}{|\mathcal{S}_0|}\,p(\boldsymbol{X} \mid \boldsymbol{Y}=\boldsymbol{y};\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l). \tag{3.11}$$

しかし，式(3.11)に従って単純にサンプリングを行う場合，クラスタ選択は事前確率 $\frac{|\mathcal{S}_l|}{|\mathcal{S}_0|}$ のみに依存する．このため，$\boldsymbol{Y}=\boldsymbol{y}$ に対する尤度 $p(\boldsymbol{y} \mid L=l)$ が非常に低いクラスタからも $\boldsymbol{X}$ が生成される恐れがある．特に，データ点がクラスタ境界付近に位置する場合，実際には複数の局所分布の影響を受けるはずであるにもかかわらず，単一クラスタに基づく条件付き分布のみを用いると，隣接クラスタの情報が無視される結果，不自然なサンプルが生成される可能性がある．この問題を回避するため，本研究ではベイズの定理に基づいて事後確率 $p(L=l \mid \boldsymbol{Y}=\boldsymbol{y})$ を計算し，$\boldsymbol{y}$ に適合するクラスタを優先的に選択する．これにより，局所的なデータ構造を反映しつつ，多様性と一貫性を兼ね備えた自然なサンプリングが可能となる．

観測データから構築された単一の密度木は，データのばらつきや偶然的な構成に強く依存する．このため，観測されていない領域に対して新規データ点を生成する場合，単一木の分割構造に由来する推定誤差やバイアスが顕著に現れ，推定される分布が不安定となる可能性がある．特に，データ密度が低い領域や，外挿を要する領域では，密度木の局所的な偏りが不自然なサンプル生成を引き起こしやすい．この問題を回避するため，本研究では複数のランダム化された密度木を用いたアンサンブル学習により，分布推定の安定性を向上させる．以下に示すアルゴリズム 2 は，複数の密度木からなるフォレスト(密度フォレスト)に基づくサンプリング手法である．

Algorithm 2 密度フォレストに基づくサンプリングアルゴリズム

Require: 観測された入出力ベクトル集合 $\{(\boldsymbol{x}_h,\boldsymbol{y}_h)\}_{h=1}^n$，木の数 T
Ensure: 生成された擬似入力ベクトル集合 $\{\boldsymbol{x}_h^b\}_{h=1}^n,\quad b=1,\ldots,B$

1: 密度フォレストの構築：
2: **for** $t=1$ **to** T **do**
3: ブートストラップ標本 $\mathcal{S}_0^{(t)}$ を作成
4: ランダムな特徴量選択に基づき密度木を構築

5: 各葉ノード $\mathcal{S}_l^{(t)}$ に対し，平均 $\boldsymbol{\mu}_l^{(t)}$ および共分散 $\boldsymbol{\Sigma}_l^{(t)}$ を推定
6: **end for**
7: **各木に対する条件付き分布推定：**
8: 各木に対して，アルゴリズム 1 を適用
9: **フォレスト全体の密度推定：**
10: 各木から得られた条件付き密度を平均化し，フォレスト全体の密度関数を構築：

$$p_{\text{forest}}(\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{1}{T}\sum_{t=1}^{T} p^{(t)}(\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y})$$

11: **条件付きサンプリング：**
12: **for** 各データ点 $(\boldsymbol{x}_h, \boldsymbol{y}_h)$ **do**
13: 平均化された密度 $p_{\text{forest}}(\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y}_h)$ に基づき，$\boldsymbol{x}_h^b$ をサンプリング
14: **end for**

なお，アルゴリズム 2 におけるランダム特徴量選択は，Breiman (2001) において提案された Random Forest の構築原理に基づいている．すなわち，各ノードにおいて，全特徴量集合 $\mathcal{F}$ の中からサイズ k の部分集合 $\mathcal{F}_j \subset \mathcal{F}$ を一様ランダムに抽出し，その中で最も情報利得 $I(\mathcal{S}_j, \boldsymbol{\theta})$ を最大化する分割パラメータ $\boldsymbol{\theta}^*$ を選択する：

$$\mathcal{F}_j \sim \text{Uniform}(\mathcal{P}_k(\mathcal{F})), \tag{3.12}$$

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \mathcal{T}_j(\mathcal{F}_j)} I(\mathcal{S}_j, \boldsymbol{\theta}), \tag{3.13}$$

ここで，$\mathcal{P}_k(\mathcal{F})$ は $\mathcal{F}$ から k 個を選択する部分集合の全体を表し，$\mathcal{T}_j(\mathcal{F}_j)$ は $\mathcal{F}_j$ に対応する分割パラメータ空間である．k の選択については，例えば $k = \lfloor\sqrt{|\mathcal{F}|}\rfloor$ のように，特徴量数の平方根に比例させる設定が標準的である (Breiman, 2001)．このようなランダム特徴量選択により，各密度木の分岐構造に多様性を導入でき，特定の特徴量や局所的なデータ構成に依存するリスクを低減できる．これにより，単一の密度木に起因する推定誤差やバイアスを緩和し，アンサンブル全体の分布推定の安定性と汎化性能を向上させる効果が期待される．

4. 木構造に基づくサンプリングによる DEA 効率性の信頼区間推定

本節では，密度木または密度フォレストに基づき生成された擬似入力データを用いて，DEA 効率性の推定および信頼区間を構築する方法について述べる．

まず，入力指向型 DEA モデル (2.1) は，サンプリングされた入力ベクトル $\{\boldsymbol{x}_h^b\}_{h=1}^n$ を用いて以下のように表される：

$$\hat{\delta}^b = \min\left\{\delta^b > 0 \;\middle|\; \sum_{h=1}^{n} \lambda_h \boldsymbol{x}_h^b \le \delta^b \boldsymbol{x}^b, \quad \sum_{h=1}^{n} \lambda_h \boldsymbol{y}_h \ge \boldsymbol{y}, \quad \sum_{h=1}^{n} \lambda_h = 1, \quad \boldsymbol{\lambda} \ge \boldsymbol{0}\right\}. \tag{4.1}$$

ここで，$b = 1, \ldots, B$ はサンプリングインデックスを表し，$\{\boldsymbol{x}_h^b\}_{h=1}^n$ はアルゴリズム 1 またはアルゴリズム 2 によって生成された擬似入力データである．

それぞれのサンプルに対して得られる B 個の効率性推定値 $\hat{\delta}^b$ について，その平均を以下により定義する：

$$\bar{\delta} = \frac{1}{B}\sum_{b=1}^{B} \hat{\delta}^b. \tag{4.2}$$

前節で述べた通り，入力指向型 DEA モデル(2.1)に関して，入力ベクトル $\boldsymbol{X}$ を確率変数，対応する DEA 効率値を確率変数 D とみなし，$E(D) = \mu > 0$，$\mathrm{Var}(D) < \infty$ が成立すると仮定する．この仮定のもとで，大数の法則により，サンプル数 $B \to \infty$ の極限において，$\bar{\delta}$ は母平均 $E(D)$ に一致し，すなわち真の効率値に収束することが保証される．この性質を利用することで，$\{\hat{\delta}^b\}_{b=1}^B$ の経験分布に基づき，推定された DEA 効率値に対する信頼区間を構築することが可能となる．

本研究では，$\bar{\delta}$ の信頼区間を構築するために，ノンパラメトリック・ブートストラップ法を適用する．具体的には，効率値の集合 $\{\hat{\delta}^b\}_{b=1}^B$ から，復元抽出により K 個のブートストラップ標本 $\{\tilde{\delta}^{b,k}\}_{b=1}^B,\ k = 1, \ldots, K$ を生成する．各標本に対して平均値

$$\bar{\delta}^{(k)} = \frac{1}{B} \sum_{b=1}^{B} \tilde{\delta}^{b,k} \tag{4.3}$$

を計算し，その経験分布に基づいて信頼区間を構築する．信頼区間の構築には，ブートストラップ平均のパーセンタイル法を用いる．すなわち，$\{\bar{\delta}^{(k)}\}_{k=1}^K$ を昇順に並べ，その 2.5 パーセンタイル値および 97.5 パーセンタイル値をそれぞれ信頼区間の下限および上限として採用する．

これにより，DEA 効率性推定に伴う不確実性をノンパラメトリックな枠組みで定量的に評価することが可能となる．本研究と既存手法である Simar–Wilson 法との本質的な相違は以下の点にある．2.2 節で述べたように，Simar–Wilson 法は観測データに含まれる測定誤差や外生的ノイズを明示的にモデル化せず，データ生成過程に内在する誤差構造を十分に補正するものではない．これに対し，本研究では観測データ $\boldsymbol{x}_h$ に測定誤差や外生的ノイズが含まれる可能性を明示的に考慮する．この場合，P^{DEA} と P の包含関係($P^{DEA} \subset P$)は必ずしも成立せず，従来のバイアス補正とは異なるアプローチが必要となる．具体的には，観測データのばらつきやノイズを反映した確率的分布モデルを学習し，このモデルに基づいて多数の擬似データセットを生成する．各擬似データセットに DEA を適用して効率値を再推定することで，効率性推定値に対する経験的な分布を構築し，信頼区間を推定する．このアプローチにより，観測誤差を考慮したうえで，DEA 推定に伴う統計的不確実性を柔軟かつ実証的に評価することが可能となる．

5. シミュレーション実験

本節では，文献 Fare et al. (1989)で用いられた電力会社データ(表 1)を用いて，提案手法の有効性を検証する．このデータセットは，19 の電力会社(DMU)に対し，3 つの入力変数(Labor，Fuel，Capital)と 1 つの出力変数(Output)を含む．

アルゴリズム 1 により構築した密度木と，そこから得られたクラスタごとの DMU の割り当て結果を，それぞれ図 3 および表 2 に示す．表 2 より，クラスタ 1 は入出力変数の水準が低い小規模 DMU，クラスタ 2 は中程度の水準を示す中規模 DMU，クラスタ 3 は Fuel や Capital および Output が顕著に大きい大規模 DMU で構成されていることが分かる．さらに，図 3 に示すように，各クラスタにおける平均ベクトルおよび共分散行列には明確な差異が認められ，密度木による分類がデータの潜在構造を的確に捉えていることが確認できる．

次に，表 3 に，観測データに基づく DEA 効率値($\hat{\delta}$)，アルゴリズム 1 および 2 による推定効率値(それぞれ $\bar{\delta}^1$ および $\bar{\delta}^2$)，ならびにブートストラップ($B = 2000$)による 95% 信頼区間を示す．なお，密度フォレストはブートストラップ標本から 10 本の密度木を生成して構築した．

観測データに基づく DEA 効率値($\hat{\delta}$)と，密度木により生成した擬似データに基づく推定効

表 1. 入出力データ.

DMU の番号	Labor	Fuel	Capital	Output
1	179.0	3969.1	1005.5	3531.4
2	96.0	896.32	194.6	841.3
3	94.0	803.17	200.0	690.5
4	141.0	1629.9	564.1	1498.1
5	147.0	3264.4	617.4	3347.7
6	261.0	1427.5	347.1	1263.2
7	448.0	6365.3	1680.4	5734.0
8	223.0	4911.0	1319.4	4107.7
9	279.0	7779.3	1785.6	6964.2
10	214.0	2307.5	690.0	1961.0
11	367.0	4334.7	1268.9	4207.2
12	97.0	1870.7	416.5	1752.7
13	143.0	2730.6	779.8	2667.6
14	95.0	540.61	305.0	457.2
15	241.0	9509.2	1892.1	9539.9
16	147.0	1561.1	718.5	1360.6
17	306.0	1699.0	306.3	1689.7
18	182.0	881.16	182.3	750.3
19	500.0	2548.9	650.1	2439.5

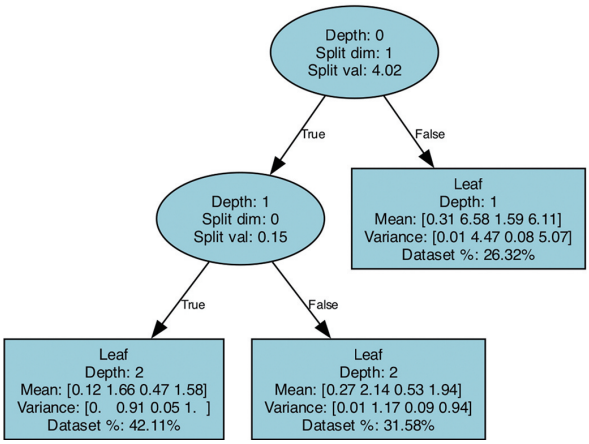

図 3. アルゴリズム 1 で構築した密度木.

表 2. クラスタの割り当て.

クラスタ	DMU の番号
1	2, 3, 4, 5, 12, 13, 14, 16
2	1, 6, 10, 17, 18, 19
3	7, 8, 9, 11, 15

表 3. アルゴリズム 1 およびアルゴリズム 2 による効率性推定結果.

DMU		アルゴリズム 1(密度木)				アルゴリズム 2(密度フォレスト)			
番号	$\hat{\delta}$	$\bar{\delta}^1$	SD($\bar{\delta}^1$)	2.5%	97.5%	$\bar{\delta}^2$	SD($\bar{\delta}^2$)	2.5%	97.5%
1	0.8691	0.9151	0.0760	0.9119	0.9183	0.9180	0.0674	0.9149	0.9209
2	1.0000	0.9505	0.0777	0.9470	0.9537	0.9869	0.0339	0.9853	0.9883
3	1.0000	0.9635	0.0715	0.9603	0.9667	0.9914	0.0261	0.9903	0.9925
4	0.9307	0.9425	0.0615	0.9398	0.9453	0.9228	0.0510	0.9206	0.9249
5	1.0000	0.9193	0.0752	0.9160	0.9226	0.9519	0.0606	0.9493	0.9546
6	0.9071	0.9410	0.0663	0.9380	0.9438	0.9206	0.0505	0.9185	0.9228
7	0.8909	0.9321	0.0544	0.9298	0.9345	0.9080	0.0558	0.9056	0.9104
8	0.8208	0.8971	0.0711	0.8940	0.9001	0.9224	0.0654	0.9195	0.9252
9	0.8885	0.9629	0.0421	0.9611	0.9646	0.9185	0.0435	0.9167	0.9205
10	0.8469	0.9377	0.0658	0.9347	0.9405	0.9287	0.0549	0.9263	0.9311
11	0.9531	0.8972	0.0696	0.8942	0.9003	0.9258	0.0643	0.9230	0.9286
12	1.0000	0.9373	0.0636	0.9346	0.9402	0.9424	0.0437	0.9406	0.9442
13	0.9602	0.9346	0.0664	0.9316	0.9374	0.9362	0.0642	0.9334	0.9390
14	1.0000	0.9953	0.0443	0.9931	0.9970	1.0000	0.0006	1.0000	1.0000
15	1.0000	1.0000	0.0000	1.0000	1.0000	1.0000	0.0000	1.0000	1.0000
16	0.8885	0.9416	0.0628	0.9388	0.9444	0.9149	0.0569	0.9124	0.9173
17	1.0000	0.9403	0.0626	0.9375	0.9430	0.9275	0.0508	0.9251	0.9298
18	1.0000	0.9566	0.0762	0.9530	0.9599	0.9859	0.0368	0.9843	0.9874
19	0.9441	0.9346	0.0652	0.9316	0.9375	0.9487	0.0502	0.9466	0.9508
平均	0.9421	0.9420	0.0617	0.9393	0.9447	0.9448	0.0461	0.9428	0.9468

率値($\bar{\delta}^1$)との分布を比較するため，Kolmogorov–Smirnov 二標本検定を適用した．Kolmogorov–Smirnov 検定は，二つの独立標本が同一分布に由来するかを検定するノンパラメトリック手法であり，DEA 効率値の分布形状に仮定を置かない点で本分析に適している．検定の結果，統計量 $D = 0.3684$，$p = 0.1532$ が得られ，有意水準 5% では帰無仮説(両分布が同一である)を棄却できなかった．このことから，密度木による擬似データは観測データに基づく効率性分布を統計的に再現しているといえる．

さらに，密度フォレストについても同様に検定を行った結果，$D = 0.3684$，$p = 0.1532$ となり，同様に帰無仮説は棄却されなかった．すなわち，密度フォレストによる擬似データも観測データと統計的に同等の効率性分布を保持していることが確認された．特に，密度フォレストは特徴量選択にランダム性を導入するにもかかわらず，効率性分布を保持した上で推定の不確実性を反映した擬似データ生成が可能であることが示された．

最後に，密度木と密度フォレストを比較すると，効率性推定値の標準偏差の平均は密度木で 0.0617，密度フォレストで 0.0461 となり，アンサンブル化による分散低減効果が確認された．また，95% 信頼区間の幅の平均も密度木($|0.9447 - 0.9393| = 0.0054$)に比べ密度フォレスト($|0.9468 - 0.9428| = 0.004$)で狭まり，より安定した推定が得られることが分かった．以上より，密度フォレストを用いることで，単一密度木に依存した場合に生じる推定の不安定性を回避し，より頑健で信頼性の高い DEA 効率性推定が可能であることが実証された．

6. 結論

本研究では，DEA における効率性推定および信頼区間推定に向けて，密度木および密度フォ

レストに基づくサンプリングアルゴリズムを提案した．提案手法は，観測データに内在するばらつきやノイズを明示的に取り込み，局所的な確率構造に基づいてデータ生成を行うことで，DEA 推定に伴う統計的不確実性を柔軟かつ実証的に評価可能とするものである．

本研究の主な貢献は，以下の三点に整理できる．

(a) 従来の Simar–Wilson 法とは異なり，観測された入力ベクトルを確率変数として扱い，測定誤差や外生的ノイズを内在化した擬似データ生成を可能とした点である．これにより，観測データの確率的変動を考慮できると同時に，効率性も確率変数として扱うことが可能となり，バイアス補正を要せずブートストラップ法により信頼区間の推定が可能となった．さらに，観測された出力ベクトルを確率変数とすることで，提案手法は出力指向型 DEA モデルにも容易に拡張できる．

(b) 単一の密度木に依存する際に生じる推定の不安定性を回避するため，複数のランダム化された密度木からなる密度フォレストを導入した点である．これにより，標準偏差のばらつきや信頼区間幅のばらつきが抑制され，アンサンブル化による分散低減効果と推定安定性の向上が実証的に確認された．

(c) 実データを用いた検証の結果，Kolmogorov–Smirnov 検定において，観測データに基づく DEA 効率値と提案手法による推定効率値との分布に有意差は確認されなかった．このことは，提案手法が観測データに内在する効率性構造を統計的に保持している可能性を示唆するものである．

さらに，密度木の構築においては，各葉ノード単位で局所的にガウス近似を仮定しているものの，全体分布がガウス分布であることを仮定していない点に注意を要する．実際，図 2 に示したように，全体として非ガウス的な分布（たとえばベータ型分布）であっても，局所的なガウス近似に基づいて効果的なモデリングとサンプリングが可能であることが示された．したがって，提案手法は，全体分布がガウス性から逸脱している場合でも有効に機能する柔軟性を有している．

また，本研究で直接比較対象としなかった StoNED や SFA といった回帰型効率性推定手法に対しても，以下の点で提案手法は優位性を有すると考えられる．StoNED や SFA では，非効率性と誤差項を合成した条件付き分布に基づき効率性を点推定するため，追加的なパラメトリック仮定が必要となり，効率性推定値の大きさ自体に実務的意味を持たせることは難しい．一方で，提案手法は効率性そのものの確率分布に着目し，古典的な統計手法を通じた直感的な信頼区間推定を可能としており，実務上の解釈性に優れる．

もっとも，本研究には限界も存在する．提案手法と Simar–Wilson 法に基づく信頼区間推定法との比較は実データに基づいて実施しておらず，その理由は，両手法が前提とするデータ生成過程が本質的に異なること，および実データでは真の投入量・産出量が不明であることに起因する．また，現実の生産可能集合を厳密に定義するための情報が不十分であることも，比較を困難にしている．この点を補完するためには，人工データを用いた多様な実験設計に基づく包括的検証が今後必要である．

今後の研究課題としては，多様な DEA モデルへの拡張，ガウス分布以外の柔軟な局所分布モデルの導入，単一密度木に基づくサンプリングにおける外れ値影響の定量的評価，ならびに密度木構築・サンプリング手順の頑健性と計算効率のさらなる向上が挙げられる．

注．

[1] 指向性（改善の方向）によってさまざまなモデルが拡張可能である（森田浩, 2024）が，本稿

の提案手法は指向性に依存せず適用可能であるため，ここでは簡便のため入力指向を前提に議論を進める．

[2] 一方，$\sum_{h=1}^{n} \lambda_h = 1$ を除去すると，規模に関する収穫一定(Constant Returns to Scale, CRS)下の経験的生産可能集合が得られ，この場合は入力を k 倍すると出力も k 倍となる．なお，規模に関する収穫には VRS や CRS のほか，Increase Return to Scale(IRS)，Decrease Return to Scale(DRS)，および General Return to Scale(GRS)などがある（森田浩, 2024）．本稿の提案手法は，規模に関する収穫の仮定に依存せず適用可能であるため，以下では簡便のため VRS を前提に議論を進める．

参 考 文 献

Aigner, D., Lovell, C. K. and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models, *Journal of Econometrics*, **6**(1), 21–37.

Aparicio, J. and Esteve, M. (2023). How to peel a data envelopment analysis frontier: A cross-validation-based approach, *Journal of the Operational Research Society*, **74**(12), 2558–2572.

Aparicio, J., Kapelko, M. and Ortiz, L. (2023). Enhancing the measurement of firm inefficiency accounting for corporate social responsibility: A dynamic data envelopment analysis fuzzy approach, *European Journal of Operational Research*, **306**(2), 986–997.

Aragon, Y., Daouia, A. and Thomas-Agnan, C. (2005). Nonparametric frontier estimation: A conditional quantile-based approach, *Econometric Theory*, **21**(2), 358–389.

Banker, R. D. (1984). Estimating most productive scale size using data envelopment analysis, *European Journal of Operational Research*, **17**(1), 35–44.

Banker, R. D., Charnes, A. and Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis, *Management Science*, **30**(9), 1078–1092.

Boubaker, S., Le, T. D., Ngo, T. and Manita, R. (2025). Predicting the performance of MSMEs: A hybrid DEA-machine learning approach, *Annals of Operations Research*, **350**(2), 555–577.

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

Cazals, C., Florens, J.-P. and Simar, L. (2002). Nonparametric frontier estimation: A robust approach, *Journal of Econometrics*, **106**(1), 1–25.

Cover, T. M. and Thomas, J. A. (2012). *Elements of Information Theory*, John Wiley & Sons, New York.

Criminisi, A. and Shotton, J. (2013). *Decision Forests for Computer Vision and Medical Image Analysis*, Springer Science & Business Media, United Kingdom.

Daraio, C. and Simar, L. (2014). Directional distances and their robust versions: Computational and testing issues, *European Journal of Operational Research*, **237**(1), 358–369.

Daraio, C., Simar, L. and Wilson, P. W. (2020). Fast and efficient computation of directional distance estimators, *Annals of Operations Research*, **288**(2), 805–835.

Dia, M., Takouda, P. M. and Golmohammadi, A. (2022). Assessing the performance of Canadian credit unions using a three-stage network bootstrap DEA, *Annals of Operations Research*, **311**(2), 641–673.

Esteve, M., Aparicio, J., Rodriguez-Sala, J. J. and Zhu, J. (2023). Random Forests and the measurement of super-efficiency in the context of Free Disposal Hull, *European Journal of Operational Research*, **304**(2), 729–744.

Fare, R., Grosskopf, S. and Kokkelenberg, E. C. (1989). Measuring plant capacity, utilization and technical change: A nonparametric approach, *International Economic Review*, **30**(3), 655–666.

Guillen, M. D., Aparicio, J. and Esteve, M. (2023). Gradient tree boosting and the estimation of production frontiers, *Expert Systems with Applications*, **214**, https://doi.org/10.1016/j.eswa.2022.119134.

Kang, H. J., Kim, C. and Choi, K. (2024). Combining bootstrap data envelopment analysis with social networks for rank discrimination and suitable potential benchmarks, *European Journal of Operational Research*, **312**(1), 283–297.

Kerstens, K., Sadeghi, J., Van de Woestyne, I. and Zhang, L. (2022). Malmquist productivity indices and plant capacity utilisation: New proposals and empirical application, *Annals of Operations Research*, **315**(1), 221–250.

Kneip, A., Simar, L. and Wilson, P. W. (2008). Asymptotics and consistent bootstraps for DEA estimators in nonparametric frontier models, *Econometric Theory*, **24**(6), 1663–1697.

Kuosmanen, T. and Kortelainen, M. (2012). Stochastic non-smooth envelopment of data: Semiparametric frontier estimation subject to shape constraints, *Journal of Productivity Analysis*, **38**, 11–28.

Lin, S.-W. and Lu, W.-M. (2024). Efficiency assessment of public sector management and culture-led urban regeneration using the enhanced Russell-based directional distance function with stochastic data, *Journal of the Operational Research Society*, **75**(8), 1624–1642.

Mergoni, A., Emrouznejad, A. and De Witte, K. (2025). Fifty years of data envelopment analysis, *European Journal of Operational Research*, **326**(3), 389–412.

Michali, M., Emrouznejad, A., Dehnokhalaji, A. and Clegg, B. (2023). Subsampling bootstrap in network DEA, *European Journal of Operational Research*, **305**(2), 766–780.

Moradi-Motlagh, A. and Emrouznejad, A. (2022). The origins and development of statistical approaches in non-parametric frontier models: A survey of the first two decades of scholarly literature (1998–2020), *Annals of Operations Research*, **318**(1), 713–741.

森田浩 (2024). DEA の例解, オペレーションズ・リサーチ = Communications of the Operations Research Society of Japan: 経営の科学, **69**(1), 12–19.

Ngo, T. and Tsui, K. W. H. (2022). Estimating the confidence intervals for DEA efficiency scores of Asia-Pacific airlines, *Operational Research*, **22**(4), 3411–3434.

Olesen, O. B. and Petersen, N. C. (2016). Stochastic data envelopment analysis—A review, *European Journal of Operational Research*, **251**(1), 2–21.

Shi, Y. and Zhao, W. (2024). An Integrated machine learning and DEA-predefined performance outcome prediction framework with high-dimensional imbalanced data, *INFOR*: *Information Systems and Operational Research*, **62**(1), 100–129.

Simar, L. and Wilson, P. W. (1998). Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models, *Management Science*, **44**(1), 49–61.

Simar, L. and Wilson, P. W. (2004). Performance of the bootstrap for DEA estimators and iterating the principle, *Handbook on Data Envelopment Analysis* (eds. W. W. Cooper, L. M. Seiford and J. Zhu), 265–298, Springer US, Boston, Massachusetts.

Valero-Carreras, D., Aparicio, J. and Guerrero, N. M. (2021). Support vector frontiers: A new approach for estimating production functions through support vector machines, *Omega*, **104**, https://doi.org/10.1016/j.omega.2021.102490.

Walheer, B. (2022). Global Malmquist and cost Malmquist indexes for group comparison, *Journal of Productivity Analysis*, **58**(1), 75–93.

Zhu, N., Zhu, C. and Emrouznejad, A. (2021). A combined machine learning algorithms and DEA method for measuring and predicting the efficiency of Chinese manufacturing listed companies, *Journal of Management Science and Engineering*, **6**(4), 435–448.

Tree-based Distribution Estimation for DEA Efficiency Evaluation

Yu Zhao

School of Management, Tokyo University of Science

Standard Data Envelopment Analysis (DEA) models are deterministic, assuming that the observed data are free from measurement errors and exogenous noise. However, in real-world data, stochastic variations are inevitable, and ignoring them can lead to substantial errors in efficiency estimation. Although bootstrap-based and regression-based approaches have been proposed to partially introduce statistical inference into DEA, a framework that directly addresses data uncertainty has not yet been established. In this study, we propose a novel sampling method based on density trees and density forests to estimate efficiency and confidence intervals in DEA. The proposed method automatically clusters the observed data and models local probabilistic structures by maximizing information gain based on a Gaussian entropy function within each cluster. Furthermore, by employing an ensemble learning framework, it ensures estimation stability and robustness while generating pseudo-datasets that reflect the variability and noise inherent in the observed data. This enables a flexible and empirical assessment of statistical uncertainty in DEA. An empirical analysis using data from electric utility companies demonstrated the effectiveness of the proposed method. Specifically, a Kolmogorov-Smirnov test comparing the efficiency scores derived from observed data with those estimated by the proposed method revealed no statistically significant differences in their distributions. This result suggests that the proposed approach can appropriately reproduce the distributional characteristics of the observed data and provide an effective means of efficiency evaluation that accounts for measurement errors and noise.

Key words: Density tree, density forest, sampling, confidence interval estimation, uncertainty.