IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 12 October 2023, accepted 5 February 2024, date of publication 1 March 2024, date of current version 8 March 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3372516*

RESEARCH ARTICLE

# Machine Learning-Based Problem Space Reduction in Stochastic Programming Models: An Application in Biofuel Supply Chain Network Design

**KOLTON KEITH, KRYSTEL K. CASTILLO-VILLAR, AND ADEL ALAEDDINI**
Mechanical Engineering Department, Texas Sustainable Energy Research Institute, The University of Texas at San Antonio, San Antonio, TX 78249, USA
Corresponding author: Krystel K. Castillo-Villar (krystel.castillo@utsa.edu)

This work was supported by the National Institute of Food and Agriculture (NIFA), U.S. Department of Agriculture (USDA), under the Hispanic Serving Institutions Education Grants Program, under Award 2020-38422-32258.

**ABSTRACT** Biofuels derived from feedstock offer a sustainable source for meeting energy needs. The design of supply chains that deliver these fuels needs to consider quality variability with special attention to shipping costs, because biofuel feedstocks are voluminous. Stochastic programming models that consider all these considerations incur a heavy computational burden. The present work proposes a hybrid strategy that leverages machine learning to reduce the computational complexity of stochastic programming models via problem space reduction. First, numerous randomly generated reduced-space versions of the problem are solved multiple times to generate a set of solution data based on the concept of bootstrapping. Next, a supervised machine learning algorithm is implemented to predict a potentially beneficial mixed integer linear program problem space from which a near-optimal solution can be obtained. Finally, the mixed integer linear program selects the optimal solution from the reduced space generated by the machine learning algorithm. Through extensive numerical experimentation, we determine how much the problem space can be reduced, how many times the reduced space problem needs to be solved and the best performing machine learning techniques for this application. Several supervised learning algorithms, including logistic regression, decision tree, random forest, support vector machine, and k-nearest neighbors, are evaluated. The numerical experiments demonstrate that our proposed solution procedure yields near-optimal outcomes with a considerably reduced computational burden.

**INDEX TERMS** Machine learning, supply chain design, biofuels, stochastic programming.

## I. INTRODUCTION

The design of cost-competitive biofuel supply chains (BSCs) is riddled with considerations that result in robust but computationally complex mathematical models [1]. The high volume of biomass relative to the energy yield makes its transportation strategy of paramount importance. As such, the locations of biomass sources and conversion facilities are often chosen to minimize these costs. In addition, it is often advisable to include depots as preprocessing facilities where the biomass undergoes densification to facilitate shipping to conversion facilities. Optimal depot location in relation to the supplier is critical to overall cost competitiveness, and there exist stochastic optimization models with such objectives [2]. Biomass is also a crop, and thus, there is inherent uncertainty in the yield based on factors such as weather; BSC design models should consider this inherent stochasticity [3]. In addition to yield, biomass also has quality-related variability in relation to conversion to biofuel. Namely, the moisture and ash content should be taken into account in BSC design, as they directly impact conversion costs and yields [4]. All of the above considerations often lead

The associate editor coordinating the review of this manuscript and approving it for publication was Binit Lukose.

to computationally complex stochastic optimization models that seek to optimize facility placement and biomass routing. Based on these challenges, machine learning has gained traction as a tool to reduce the computational burden of complex optimization models [5], [6]. Goettsch et al. utilized machine learning to reduce the number of potential depot locations for biomass cofiring [7].

In this work, we propose a solution procedure that leverages machine learning to solve a stochastic optimization model for optimal BSC design. Inspired by statistical bootstrapping, our proposed solution procedure first randomly builds sets of problems where the space of stage-one decisions, namely, potential biorefinery locations, is reduced. Next, it combines solutions obtained from these reduced space versions of the model into a dataset used for machine learning classification. The classifier then selects the best candidates for the stage-one decision variables at hand and executes the model a final time to obtain the optimal solution. The overall structure of the proposed solution procedure is outlined below in Figure 1.

Our proposed solution procedure is applied to a deterministic version of the BSC design model to assess to what extent the problem space can be reduced when building the dataset for classification, how many runs of the reduced space problem are required to build a quality dataset for classification, and what classification techniques are best suited for this solution procedure. These results are then applied to the stochastic version of the mathematical model.

## II. LITERATURE REVIEW

The design of biofuel supply chains is a vast and complex field covering many different aspects [2], [3], [8], [9]. However, our literature review only explores prior works that most closely influenced our model. Linear programming models are often utilized in the modeling of biomass supply chains. Panichelli and Gnansounou presented a model that makes use of a geographic information system (GIS) approach to determine the optimal location of biomass suppliers in relation to torrefaction plants and gasification facilities [10]. Inter-facility resource competition and variable biomass pricing were the main model considerations. The model's objective function sought to minimize the marginal costs related to the supply of torrefied wood to the gasification plants to determine the plants' locations. Dijkstra's algorithm was used to find the shortest route between forest sites and torrefaction plants to determine the allocation of the existing biomass supply. The method of this paper differs from their approach in that we determine the optimal location of biorefineries and depots rather than focusing on biomass supply. Other implementations of linear programming in relation to biomass supply chains sought to implement the uncertainty in regard to transportation when designing depots [11]. In Cundiff, Dias and Sherali [11], supply uncertainty was explained by four growing season and harvest month weather scenarios. The scenarios are then used in a multistage Liner Program (LP) model to determine the cost and optimal size of monthly inter-facility biomass shipments. This concept of using weather to influence supply chain uncertainty is of vital importance when dealing with cost optimization as it pertains to biomass. The variable growing conditions directly affected the characteristics of the supply and should not be overlooked when seeking to formulate optimization models for biomass supply chains. The complex interactions between biomass logistics and supply chain design and the desire to simultaneously optimize facility locations and biomass supply flow has motivated many authors to utilize mixed integer linear programming [12], [13], [14], [15], [16]. Bowling and El-Halwagi [17] structured their biomass supply chain as a network with nodes containing locations (suppliers, pre-treatment sites and conversion facilities), and the arcs indicate transportation links between them. Within the Mixed Integer Linear Programming (MILP), binary variables were utilized to determine whether facilities are constructed at a particular location, and continuous decision variables are used to describe the biomass flow. The model was optimized based on economic, environmental and energy objectives and considered constraints that included things such as capacities, demands and mass balances. Our proposed model structure is quite similar to the aforementioned model, specifically with regard to network design. Major logistics improvements within targeted strategic areas within biomass supply chains have recently been proposed and offer the potential to increase the economic feasibility of consumer utilization of biofuels. Namely, the moisture and ash content of biomass and their effect on supply chain logistics. To incorporate these quantities into supply chain decisions, researchers often use two-stage stochastic programming models. Castillo et al. [18], made use of trust region cuts and multi-cuts to solve a two-stage stochastic model to determine the optimal or near-optimal location of biorefineries and their associated conversion technologies under the uncertainty that accompanies biomass moisture content. Another innovation in the biomass supply chain network has been the introduction of depots to pre-process and condense biomass prior to transportation to biorefineries. The key advantages of this approach are twofold. First, the preprocessing of biomass offers improved physical and chemical consistency that reduces the variability associated with the conversion of raw biomass into biofuels. Second, the densification and drying that occurs in depots and their proximity to suppliers and mass transportation, i.e., railways, has the potential to reduce total supply chain transportation costs. Aboytes-Ojeda et al. [1] introduced a two-stage stochastic model that used a hub and spoke network including depots that pre-process biomass. In addition to moisture variability, the model also considered the natural variability in biomass ash content when determining the depot locations, biorefinery locations, selection of conversion technologies, and biomass required for bioethanol production. The impact of moisture and ash variability on supply chain decisions was demonstrated by performing a case study in Texas. Numerous examples of using the L-shaped method, among

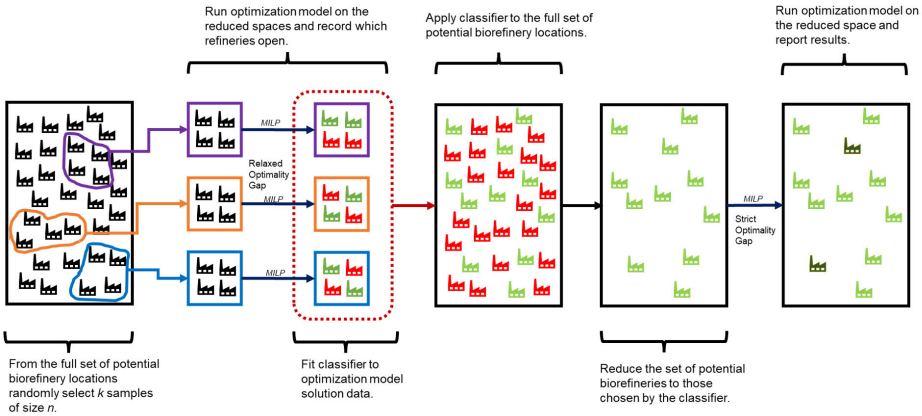


**FIGURE 1.** Outline of the ML-driven problem space reduction solution procedure.

others, to solve large-scale optimization problems can be found in the literature. For example, Marufuzzaman et al. [19] introduced a two-stage stage stochastic model to design and manage a biodiesel supply chain. The solution approach was implemented to solve a case study in the state of Mississippi utilizing a hybrid method between Lagrangian relaxation and L-shaped. The present work is distinguished from a design perspective due to utilizing the L-shaped method for a byproduct (biochar, bioethanol and biodiesel) biomass supply chain to explore more complex interactions. In addition, our model also considers multiple biomass types with unique conversion mass ratios. The conversion technology for biomass in the present work is fast pyrolysis. Fast pyrolysis takes raw biomass and converts it to bio-oil, biochar and syngas. Supply chains incorporating this conversion technology should use more than one of these byproducts rather than focusing on just bio-oil. Casler and Boe [20] demonstrated that pyrolysis was a more effective means of conversion than co-firing for meeting the electricity demands of Taiwan. Additionally, they used bio-oil to meet electricity demands, and biochar was used as a soil additive to increase crop yields in Taiwanese farms. In contrast, the present work seeks to use biochar to meet coal plant demands rather than as a soil additive. Additionally, the present model is distinguished by proposing and testing a novel solution procedure that leverages machine learning to find quality solutions at reduced computational cost. Categorization of the closely related previous works is summarized below in Table 1.

The present work proposes a novel model in the field of bioenergy supply chain network design. The contributions of this work are threefold. From the modeling perspective, we propose a two-stage stochastic programming model that addresses the random nature of the biomass quality-related properties, includes the investment and operational costs, and analyzes the trade-off between bioethanol, biodiesel, and biochar. From the algorithmic development point of view, a novel algorithm is proposed that leverages machine learning to reduce computation time while preserving solution integrity. This solution procedure is aimed at reducing the computational burden of complex two-stage optimization models. From a pragmatic standpoint, we created a realistic case study at the state level to test our solution procedure. Our results show that our algorithm has the potential to generate optimal solutions of high-resolution BSC design problems at the national level.

**TABLE 1.** Present model contributions.

| | Bowling, El-Halwagi, Ponce-Ortega [17] | Castillo-Villar, Eksioglu, Taherkhorsandi, [18] | Aboytes-Ojeda, Castillo-Villar, Eksioglu, [1] | This Work |
|---|---|---|---|---|
| Hub and spoke network | X | X | X | X |
| Considers variability in moisture and ash content | | X | X | X |
| Utilizes depots to reduce transportation costs and processing variability | | | X | X |
| Uses L-shape Method | | X | X | X |
| Considers multiple biomass byproducts | | | | X |
| Considers multiple biomass types | | | | X |
| Utilizes machine learning driven. solution procedure | | | | X |

## III. MATHEMATICAL MODEL

The model is a two-stage stochastic hub-and-spoke problem, where the first-stage decision variables consist of determining which biorefineries and depots to open and the arcs that connect them. The second-stage decision variables determine how much of each type of biomass each supplier should produce as well as the biomass routing through the BSC from supplier to depot to refinery to cities and power plants. The model aims to design a supply chain that converts feedstock to a variety of biofuels via fast pyrolysis to meet known demands. The model considers three types of biomass (corn stover, switchgrass, and miscanthus) that then convert to three fuels (bioethanol, biocoal, and biodiesel). There is

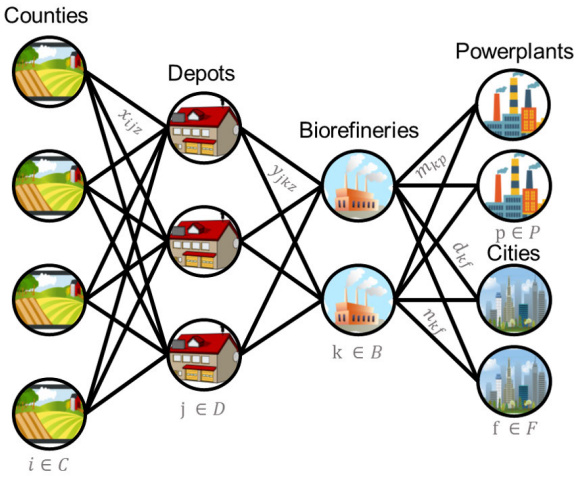


**FIGURE 2.** Hub and spoke network.

stochasticity in the moisture and ash content of the biomass modeled through scenarios. This variability affects supply as well as conversion yields, densification costs and conversion costs. In total, the model consists of 5711 stage-one decision variables and 944,460 stage two decision variables. The mathematical formulation is as follows.

**Sets:**

- $\mathcal{C}$: Set of counties (suppliers) for all $i \in \mathcal{C}$
- $\mathcal{D}$: Set of potential locations for depots for all $j \in \mathcal{D}$
- $\mathcal{B}$: Set of potential locations for biorefineries, $k \in \mathcal{B}$
- $\mathcal{P}$: Set of power plants, $p \in \mathcal{P}$
- $\mathcal{F}$: Set of cities, $f \in \mathcal{F}$
- $\mathcal{T}$: Set of arcs from $\mathcal{C}$ to $\mathcal{D}$, $(i,j) \in \mathcal{T}$
- $\mathcal{R}$: Set of arcs from $\mathcal{D}$ to $\mathcal{B}$, $(j,k) \in \mathcal{R}$
- $\mathcal{U}$: Set of arcs from $\mathcal{B}$ to $\mathcal{P}$, $(k,p) \in \mathcal{U}$
- $\mathcal{V}$: Set of arcs from $\mathcal{B}$ to $\mathcal{F}$, $(k,f) \in \mathcal{V}$
- $\mathcal{Z}$: Set of biomass types, $z \in \mathcal{Z}$
- $\Omega$: Set of scenarios, $o \in \Omega$

**Decision Variables:**

- $x_{ijz}(o)$:Volume of biomass of type $z \in \mathcal{Z}$ along arc $(i,j) \in \mathcal{T}$ under scenario $o \in \Omega$
- $y_{jkz}(o)$: Flow of pre-processed biomass of type $z \in \mathcal{Z}$ along arc $(j,k) \in \mathcal{R}$ under scenario $o \in \Omega$
- $m_{kp}(o)$: Flow of biochar along arc $(k,p) \in \mathcal{U}$ under scenario $o \in \Omega$
- $n_{kf}(o)$: Flow of bioethanol along arc $(k,f) \in \mathcal{V}$ under scenario $o \in \Omega$
- $d_{kf}(o)$: Flow of biodiesel along arc $(k,f) \in \mathcal{V}$ under scenario $o \in \Omega$
- $A_{jk}$: An integer variable denoting the number of unit trains connecting depot $j \in \mathcal{D}$ to biorefinery $k \in \mathcal{B}$
- $\beta_k$: A binary variable that takes the value 1 if $k \in \mathcal{B}$ if the potential location is used as a biorefinery and 0 otherwise
- $W_j$: A binary variable that takes the value 1 if potential location $j \in \mathcal{D}$ is used as a depot and 0 otherwise
- $\pi_1(o)$: Third party coal supply under scenario $o \in \Omega$
- $\pi_2(o)$: Third party bioethanol supply under scenario $o \in \Omega$
- $\pi_3(o)$: Third party biodiesel supply under scenario $o \in \Omega$

**Parameters:**

- $\xi_j$: Investment cost to open a depot at node $j \in \mathcal{D}$
- $e_i(o)$: Moisture level of biomass in County $i \in \mathcal{C}$ under scenario $o \in \Omega$.
- $\varrho_k$: Investment cost to open a biorefinery at location $k \in \mathcal{B}$.
- $\psi_{jk}$: Fixed cost of loading/unloading a unit train along arc $(j,k) \in \mathcal{R}$ every week for a period of one year (52 weeks)
- $p(o)$: Probability of scenario $o \in \Omega$
- $c_{ij}^{\mathcal{T}}$: Unit cost charged per metric ton shipped along arc $(i,j) \in \mathcal{T}$
- $c_{jk}^{\mathcal{R}}$: Unit cost charged per metric ton shipped along arc $(j,k) \in \mathcal{R}$
- $c_{kp}^{\mathcal{U}}$: Unit cost charged per metric ton shipped along arc $(k,p) \in \mathcal{U}$
- $c_{kf}^{\mathcal{V}}$: Unit cost charged per liter shipped along recurrent arc $(k,k) \in \mathcal{B}$
- $\alpha_1$: Represents the penalty cost for demand shortage of coal
- $\alpha_2$: Represents the penalty cost for demand shortage of bioethanol
- $\alpha_3$: Represents the penalty cost for demand shortage of biodiesel
- $s_i(o)$: Available supply in county $i \in \mathcal{C}$ for scenario $o \in \Omega$
- $g_{kf}$: Conversion factor for biomass/bio-oil supplied to city $f \in \mathcal{F}$ applying pyrolysis
- $v_{jk}$: Maximum capacity of a unit train along arc $(j,k) \in \mathcal{R}$
- $u_j$: Represents the preprocessing capacity of depot facility $j \in \mathcal{D}$
- $q_k$: Production capacity of biorefinery $k \in \mathcal{B}$.
- $d_p$: Total demand of biochar for power plant $p \in \mathcal{P}$
- $d_f$: Total demand of bioethanol for city $f \in \mathcal{F}$
- $s_{\mathcal{P}z}$: Production yield of biochar for species $z \in \mathcal{Z}$
- $s_{\mathcal{F}z}$: Production yield of bio-oil for species $z \in \mathcal{Z}$
- $s_{\mathcal{G}z}$: Production yield of syngas for species $z \in \mathcal{Z}$

The stochastic model minimizes the total costs associated with the BCS design, as shown in Eq. (1). These total costs include investment costs to open depots, the operation of biorefineries and their connecting arcs, transportation costs and demand penalty terms.

$$\begin{aligned} Min: \ & \sum_{j\in\mathcal{D}} \xi_j W_j + \sum_{k\in\mathcal{B}} \varrho_k \beta_k + \sum_{j\in\mathcal{D}}\sum_{k\in\mathcal{B}} \psi_{jk} A_{jk} \\ & + \sum_{o\in\Omega} p(o) \left[ \sum_{i\in\mathcal{C}}\sum_{j\in\mathcal{D}}\sum_{z\in\mathcal{Z}} c_{ijz}^{\mathcal{T}}(o) x_{ijz}(o) \right. \\ & + \sum_{j\in\mathcal{D}}\sum_{k\in\mathcal{B}}\sum_{z\in\mathcal{Z}} c_{jk}^{\mathcal{R}}(o) y_{jkz}(o) + \sum_{k\in\mathcal{B}}\sum_{p\in\mathcal{P}} c_{kp}^{\mathcal{U}} m_{kp}(o) \end{aligned}$$

$$+\sum_{k\in\mathcal{B}}\sum_{f\in\mathcal{F}} c^{\mathcal{V}}_{kf}(n_{kf}(o)+d_{kf}(o))$$

$$+\sum_{p\in\mathcal{P}}\alpha_1\pi_1(p,o)+\sum_{f\in\mathcal{F}}(\alpha_2\pi_2(f,o)+\alpha_3\pi_3(f,o))\Bigg] \quad (1)$$

**Subject to** :

$$\sum_{j\in\mathcal{D}} x_{ijz}(o)\le S_{iz}(o) \quad \forall i\in\mathcal{C} \quad o\in\Omega, \quad z\in\mathcal{Z}, \quad (2)$$

$$\sum_{i\in\mathcal{C}}(1-e_i(o))x_{ijz}(o)=\sum_{k\in\mathcal{B}} y_{jkl}(o)\ \forall j\in\mathcal{D},\ z\in\mathcal{Z},\ o\in\Omega, \quad (3)$$

$$\sum_{k\in\mathcal{B}} m_{kp}(o)+\pi_1(p,o)=d_p \quad o\in\Omega, \quad p\in\mathcal{P} \quad (4)$$

$$\sum_{k\in\mathcal{B}} g_{kf}n_{kf}(o)+\pi_2(f,o)=d_f \quad o\in\Omega, \quad f\in\mathcal{F} \quad (5)$$

$$\sum_{z\in\mathcal{Z}} y_{jkz}(o)\le v_{jk}A_{jk} \quad \forall j\in\mathcal{D}, \quad k\in\mathcal{B}, \quad o\in\Omega, \quad (6)$$

$$\sum_{i\in\mathcal{C}}\sum_{z\in\mathcal{Z}} x_{ijz}\le u_jW_j \quad \forall j\in\mathcal{D}, \quad (7)$$

$$\sum_{j\in\mathcal{D}} y_{jkz}(o)\le q_k\beta_k \quad \forall k\in\mathcal{B}, \quad z\in\mathcal{Z}, \quad o\in\Omega, \quad (8)$$

$$\sum_{p\in\mathcal{P}} m_{kp}(o)=\sum_{j\in\mathcal{D}}\sum_{z\in\mathcal{Z}} s_{\mathcal{P}z}y_{jkz}(o),\ \forall k\in\mathcal{B},\ o\in\Omega, \quad (9)$$

$$\sum_{f\in\mathcal{F}} n_{kf}(o)=\sum_{j\in\mathcal{D}}\sum_{z\in\mathcal{Z}} s_{\mathcal{F}z}y_{jkz}(o),\ \forall k\in\mathcal{B},\ o\in\Omega, \quad (10)$$

$$A_{jk}\le 0.5(\beta_k+W_j), \quad \forall k\in\mathcal{B}, \quad j\in\mathcal{D}, \quad (11)$$

$$x_{ijz}\ge 0 \quad \forall i\in\mathcal{C}, \quad j\in\mathcal{D}, \quad z\in\mathcal{Z} \quad o\in\Omega, \quad (12)$$

$$y_{jkz}(o)\ge 0\ \forall j\in\mathcal{D}, \quad k\in\mathcal{B}, \quad z\in\mathcal{Z},\ o\in\Omega, \quad (13)$$

$$m_{kp}(o)\ge 0 \quad \forall k\in\mathcal{B}, \quad p\in\mathcal{P}, \quad o\in\Omega, \quad (14)$$

$$n_{kf}(o)\ge 0 \quad \forall k\in\mathcal{B}, \quad f\in\mathcal{F}, \quad o\in\Omega, \quad (15)$$

$$\pi_1(o)\ge 0 \quad \forall o\in\Omega; \quad \pi_2(o)\ge 0 \quad \forall o\in\Omega, \quad (16)$$

$$W_j\in\{0,1\} \quad \forall j\in\mathcal{D}; \quad \beta_k\in\{0,1\} \quad \forall k\in\mathcal{B}, \quad (17)$$

$$A_{jk}\in\{0,1\} \quad \forall j\in\mathcal{D}, \quad k\in\mathcal{B}. \quad (18)$$

Constraint 2 enforces supply capacity for counties. Constraint 3 ensures mass conservation between wet and dry mass. Constraints 4 and 5 are concerned with demand satisfaction. Constraints 6, 7 and 8 ensure that rail car, depot and biorefinery capacities, respectively, are met. Constraints 9 and 10 enforce mass equality on the conversion of biomass to bio-oil and biochar. Constraint 11 is a connectivity constraint that avoids unrealistic connections. Constraints 12 to 15 are domain constraints ensuring non-negativity in physical quantities as well as enforcing binary status on first-stage decision variables.

## IV. MACHINE LEARNING PROBLEM FORMULATION

We use machine learning to reduce the problem space of the potential biorefinery locations to reduce the computational complexity of the large-scale MILP problem biofuel supply chain network design without significantly sacrificing the quality of the solution. The full list of potential locations is derived from the Bioenergy Atlas site [21]. A suitability analysis conducted on this list identified 167 potential locations. Numerical testing and supply/demand analysis have shown that approximately 10 can be expected to be opened for a given run of the optimization model. Thus, reduction of the potential locations is desirable not only to reduce the number of binary variables in the optimization model associated with the opening of biorefineries ($\beta_k$) but also to reduce the depot connection binaries ($A_{jk}$) as well as continuous biomass inflow ($Y_{jkz}(o)$) and outflow ($m_{kp}(o)$, $n_{kf}(o)$, $d_{kf}(o)$) terms. Specifically, eliminating one potential biorefinery from the problem space results in the elimination of 34 stage 1 variables and 3300 stage 2 variables from the problem space. Unfortunately, the criteria for the elimination of potential biorefineries are complex for a few reasons. Different connecting cities have differing demands, and connecting depots have different biomass amounts, types and quality, to name a few. As such, simple elimination, such as choosing refineries that have the smallest average distance to cities, power plants and depots, is not a sufficient exclusion criterion. Machine learning thus represents a promising tool to capture the complexity of how to select a suitable biorefinery for an optimal solution, and the solution procedure leverages various techniques to exclude refineries from the problem space without heavily compromising optimality in the final solution.

To leverage machine learning to reduce the number of potential biorefineries in the optimization model, it is first necessary to frame the problem in a way that is conducive to generating optimal solutions. A dataset of features and responses must be constructed such that the application of machine learning will yield information on what makes a refinery desirable based on model behavior. In the proposed solution procedure, features are restricted to properties of biorefineries that are present prior to any knowledge gained from running the optimization model. Quantities such as how much biomass a refinery processes and the biomass flow in and out of the refinery are not suitable for features, as the model would have to be run to determine them. As such, since all potential refinery locations have identical capacities and investment costs, the distances from each potential refinery to connecting facilities are the most suitable properties to select as features. Each potential refinery has 33 potential depot connections as well as 17 potential power plant connections and 8 potential city connections for a total of 58 distances. Distances from potential biorefineries to other candidate biorefineries are not included as features in

the present model. For responses, the optimization model will have to be run, from which a binary response is recorded to take the value 1 if the refinery was opened and 0 otherwise. The proposed solution procedure will run the optimization model on a restricted problem space of a randomly selected subset of potential biorefinery locations to generate responses. This process is repeated to yield a large set of data generated over multiple runs of the optimization model that will be suitable for a classification fit. The proposed process has the following inputs: $k$ the number of restricted runs of the MILP model to be conducted, $n$ the number of stage-1 variables (potential biorefinery locations) to be considered in each run, $M$ the mathematical model, $\delta_1$ the optimality gap for the reduced space MILP runs, $\delta_2$ the optimality gap for the final solution MILP run, $C$ the classification algorithm selected, and $X_{(z,f)}$ the features of the stage-1 variable to be reduced (distances). Within the algorithm, $D$ is the dataset used for the classification fit, $\chi$ is the fit classifier and $S$ is the final reported solution to the optimization problem. The data generation process of the proposed machine learning framework, which is based on solving a randomly generated reduced-space version of the problem multiple times, has a close relationship with statistical bootstrapping [22]. Therefore, the resulting dataset enjoys the simplicity, effectiveness, and statistical properties of bootstrapping samples.

**Algorithm 1** Machine Learning-Powered MILP Problem Space Reduction Solution Procedure

**Input:** $k, n, M, \delta_1, \delta_2, C, X_{(z,f)}$

```
Initialize D_(n*k,f+1).
for i = 1, 2, ..., k do
    x_(n,f) ← Randomly select n rows of X_(z,f)
    m ←Reduce M to only consider x_(n,f).
    y_n ← Solve m with optimality gap δ_1.
    for j = 1, 2, ..., n do
        D_((i−1)*n+j) ← [x_(j,f), y_j].
    end for
end for
χ ← Fit C to D.
Y ← Apply χ to X.
r ← 1
for i = 1, 2, ..., z do
    if Y_i = 1 then
        x*_r ← X_i
        r ← r + 1
    end if
end for
m* ←Reduce M to only consider x*.
S ← Solve m* with optimality gap δ_2.
```

**Output:** $S$

The process outlined above introduces key questions that are discussed here. First, how many refineries ($n$) should be selected for a randomly restricted run of the optimization model? There will be a trade-off in information gained from each run of the optimization model and how long each run will take, as a larger problem space per run will result in longer run times. Second, how many times should the optimization model be run ($k$) to build a dataset that gives us confidence in the classifier's predictions? This question has the same consideration as the previous question, with the caveat that more runs will also increase the incidence of multiple responses for the same refinery that could result in complications with the selected classification algorithm. Third, what classification method(s) are best suited for this problem type? This is an interesting question due to the nature of how the responses are generated. Each response is innately coupled with the ones that were generated during the run of the optimization model from which it was generated. In other words, each response does not yield reliable information outright, which will be a hurdle for the selected classifier to overcome. Fourth and ultimately, is the proposed solution procedure suitable for preserving optimality while reducing computational time? A summary of the discussed research questions is shown below.

- What are best practices for the number of potential biorefineries included per run?
- What are best practices for the number of runs of the optimization model required to build a reliable classifier?
- What are the best classification methods and parameters for this application?
- Is the proposed solution procedure suitable for preserving optimality while reducing computational time?

### A. DESIGN OF EXPERIMENTS

The design of experiments (DOE) formulated to answer the above research questions is as follows. To answer question one, 20, 30, 40 and 50 randomly selected refineries per run of the optimization model will be tested. This constitutes keeping 12%, 18%, 24%, and 30% of the potential biorefinery location space. Additionally, this selection will generate training datasets with approximately 50% to 20% positive responses. Since ten refineries are expected to be opened per run, restricting the problem space to 10 would give little to no useful information because most (or all) of the responses would be open. The upper limit of 50 is assigned because any further increase will make it so that the computational time for the generation of the dataset to be used for classification will encroach on that of solving the full problem space optimization model. To determine best practices for question two, datasets will be constructed consisting of 10, 20, 30, 40 and 50 runs of the optimization model. Preliminary testing indicated that values below this range do not yield acceptable classifier performance, while values above this range offer little to no apparent benefit. For question 3, the classification methods to be tested are logistic regression (LR), random forest (RF), K nearest neighbor (KNN), support vector machine (SVM), and decision tree (DT). Neural

networks were also considered for the study, but their high computational cost coupled with unremarkable performance in this application led to their omission. The sections to follow constitute a large DOE that seeks to determine best practices for the number of refineries per run ($n$) and the number of runs of the optimization model used to build the datasets ($k$). In addition, analysis is performed on each of the 5 classification methods to determine best practices in regards to sub-types and hyper-parameter tuning for each of the combinations of $n$ and $k$. To start determining the suitability of and best practices for the proposed solution procedure, it is necessary to generate a large collection of solutions to the randomly reduced optimization model for each of the proposed values of $n$. As a result, the stochastic version of the model is computationally challenged to solve a sufficient number of times to perform the proposed DOE. Thus, the analysis is conducted on the deterministic (1 scenario) version of the model solved to a 1% optimality gap. For each value of $n$, $n$ potential biorefinery locations were randomly selected, and the optimization model was solved. This process is repeated 90 times to yield a total of 90 solutions per value of $n$ for a total of 360 solutions. From these solution sets, $k$ solutions were randomly selected for each value of $k$ considered to build a dataset. To determine the repeatability of the proposed solution procedure, this selection was repeated 10 times for each combination of $n$ and $k$ for a total of 200 datasets.

## V. NUMERICAL EXPERIMENTATION

Computation was carried out on a computer with an Intel(R) Core(TM) i9-7980XE CPU @ 2.60 GHz and 32 GB of RAM. Additionally, IBM ILOG CPLEX 12.8.0 was used to solve the deterministic version of the MILP with an optimality gap of 1% when building the training datasets and 0% when determining the final solutions. MATLAB's Statistics and Machine Learning Toolkit was used to perform the ML fits. For KNN, cityblock distances were used, each dataset was swept over a range of 5-20 nearest neighbors, and the best performer in terms of prediction accuracy was used. For RF, bagging was used with a resampling percentage of 10%, which was determined by preliminary testing. For SVM, polynomial and Gaussian kernels were used based on performance in preliminary testing. Further information on fit parameters and performance can be found in Appendix . For comparison, the full MILP problem was solved and had an objective function value of $4,944,406,330 and computational time of 18,905 seconds. The set of optimal refineries was recorded and used to compute recall. In the following, we analyze our proposed ML-powered solution procedure. First, the computational time to generate the training datasets is discussed, followed by performance in terms of accuracy and recall for the various methods. The section concludes with a comparison of the solution quality and time of selected ML-derived solutions versus the full solution.

**TABLE 2. Comparison of classifier performance.**

| | 10 Runs Per Sample | 20 Runs Per Sample | 30 Runs Per Sample | 40 Runs Per Sample | 50 Runs Per Sample |
|---|---|---|---|---|---|
| 20 Refineries Per Run | LR: 81.97%<br>KNN: 82.62%<br>DT: 80.11%<br>SVM: 82.09%<br>RF: 82.95% | LR: 83.68%<br>KNN: 82.62%<br>DT: 81.72%<br>SVM:83.73%<br>RF: 83.61% | LR: 83.23%<br>KNN:83.52%<br>DT: 82.44%<br>SVM:83.87%<br>RF:84.09% | LR: 83.54%<br>KNN: 83.69%<br>DT: 83.21%<br>SVM:84.14%<br>RF: 84.43% | LR: 83.74%<br>KNN: 83.70%<br>DT: 83.90%<br>SVM:84.37%<br>RF: 84.51% |
| 30 Refineries Per Run | LR: 81.33%<br>KNN: 84.86%<br>DT: 84.90%<br>SVM: 85.35%<br>RF: 85.75% | LR: 81.17%<br>KNN: 84.08%<br>DT: 83.66%<br>SVM: 84.61%<br>RF: 84.95% | LR: 82.42%<br>KNN: 85.62%<br>DT: 84.88%<br>SVM: 85.87%<br>RF:86.10% | LR: 82.89%<br>KNN: 84.44%<br>DT: 85.71%<br>SVM: 86.64%<br>RF:86.73% | LR: 82.17%<br>KNN: 86.10%<br>DT: 86.05%<br>SVM: 86.71%<br>RF: 86.23% |
| 40 Refineries Per Run | LR: 81.18%<br>KNN:86.13%<br>DT: 85.44%<br>SVM:86.30%<br>RF: 86.75% | LR: 82.76%<br>KNN: 87.09%<br>DT: 86.63%<br>SVM:87.58%<br>RF: 87.00% | LR: 83.72%<br>KNN: 86.82%<br>DT: 87.34%<br>SVM:88.15%<br>RF: 87.43% | LR: 83.64%<br>KNN: 87.50%<br>DT: 87.43%<br>SVM:88.10%<br>RF: 87.79% | LR: 84.06%<br>KNN: 88.22%<br>DT: 88.28%<br>SVM: 88.26%<br>RF: 88.26% |
| 50 Refineries Per Run | LR: 84.78%<br>KNN: 86.21%<br>DT: 86.96%<br>SVM: 87.01%<br>RF: 87.81% | LR: 85.19%<br>KNN: 87.86%<br>DT:88.62%<br>SVM:88.88%<br>RF: 88.60% | LR: 85.91%<br>KNN: 88.43%<br>DT: 89.17%<br>SVM:88.80%<br>RF:89.02% | LR: 85.75%<br>KNN: 89.00%<br>DT:89.33%<br>SVM:89.00%<br>RF: 89.34% | LR: 85.70%<br>KNN: 89.02%<br>DT: 88.93%<br>SVM: 88.73%<br>RF: 88.88% |

### A. COMPUTATIONAL TIME TO GENERATE DATASETS

The first step in the implementation of our proposed approach is to generate training sets. This occurs through running the MILP with a relaxed solution gap (%1) multiple times and recording which biorefineries opened in each run. In total, 360 runs of the MILP model were performed, 90 for each case of refineries per run. The runs are randomly selected without replacement to build out datasets for each of the runs per sample considered. Each combination of runs per sample and refineries per run yields 10 distinct datasets which are averaged; the average time to build these datasets is shown in Figure 3. As expected, increasing the runs per sample linearly increases computational time. Beyond the 20 refineries per run case, we observe a considerable jump in computational time when increasing the number of refineries per run. This is logical because increasing the number of refineries per run increases the computational complexity of each run of the MILP.

### B. CLASSIFIER PERFORMANCE

To assess classifier performance, the mean accuracy across the 10 samples (for each combination of refineries per run and runs per sample) is used. Accuracy is determined by performing a classification fit on 80% of the data and applying this fit to the remaining 20%. Table 2 details the classifier performances. The blue text indicates the best performer in that problem set. violet denotes the next best, and red signifies the third best. The results indicate that the RF, KNN, SVM and DT methods generally outperform LR and that the four aforementioned methods perform similarly. From an accuracy consideration, RF slightly beats the other methods in most cases, having the highest accuracy in 11 of 20 cases and never failing to appear in the top 3. We note that due to the nature of the way the data are generated, near 100% accuracy should not be expected. Each run of the MILP model randomly samples refineries with replacement, so each dataset is likely to have multiple repeat variables with differing responses. Figure 4 collects the best performance along each combination of refineries per run

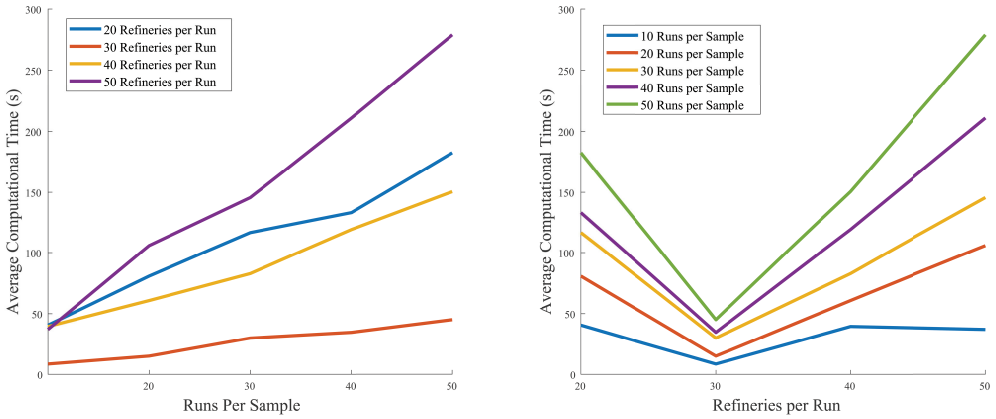


**FIGURE 3.** Average computational time to generate data for classification. Figure.

and runs per sample to determine performance trends. We observe accuracy increases for increasing both refineries per run and runs per sample. Accuracy is more closely related to refineries per run, as this increases the size of potential biorefineries each run of the MILP can use during its computations, thus reducing the instances of "weaker" refineries being selected as optimal. Increasing the number of runs per sample also shows a weak trend in accuracy improvement. However, when considering deviation in accuracy between samples, runs per sample reduces deviation up to 30 runs per sample. From an accuracy perspective 30 runs per sample with 40 biorefineries per run would be we recommendation. An analysis of the classification results considering the full set of potential biorefinery locations indicates that it would be unwieldy to keep all the positive responses. Each ML method loosely follows the trend:

$$BL = \frac{10}{n} * 167 \tag{19}$$

where $n$ is the number of potential biorefineries considered when building the training set and $BL$ is the number of biorefineries given a positive response. In short, the trend shows that the percentage of positive responses per run of the model used to generate the dataset for classification is reflected in the percentage of positive responses yielded when the classifier is applied to the full list of potential biorefinery locations. As such, we propose to narrow responses by only keeping a certain number of the top responses. Figure 5 shows this recall, i.e, the fraction of the optimal potential biorefinery locations are present, when narrowing responses to the top 10, 20 and 30.

Results indicate that once again, increasing the potential biorefineries per run of the MILP model has a larger impact performance when compared to increasing the runs per sample. When keeping 30 of the high values of refineries per run and runs per sample, RF shows the best performance, but for smaller datasets and lower numbers of responses, the KNN shows the best performance. These two methods stand out because the features used for classification are physical distances, and the datasets are generated by combining smaller runs of the model, which is reminiscent of bagging methods. From these results and the accuracy results, we recommend keeping the top 30 responses and considering 40 biorefineries per run for a total of 30 runs. The recommended method is RF for this problem in that region, but if smaller datasets are desirable to further reduce computational time, then KNN is recommended.

### C. FINAL SOLUTIONS

To assess the overall performance of the proposed solution procedure, KNN and RF solutions from the 40 refineries per run and 30 runs per sample case are compared to the full solution and to three heuristics that reduce the space of potential biorefineries. The first heuristic involves generating a score by summing the mean normalized distances from each refinery location to depots, cities and power plants.

$$I_k = \frac{\sum_j d_{jk}}{n_d * \max_j(d_{jk})} + \frac{\sum_f d_{kf}}{n_f * \max_f(d_{kf})} + \frac{\sum_p d_{kp}}{n_p * \max_p(d_{kp})}. \tag{20}$$

where, $d_{jk}$, $d_{kf}$, and $d_{kp}$ denote the distances from potential depot locations to potential biorefinery locations, potential biorefineries to cities and potential biorefineries to power plants, respectively. After each weighed score is obtained, a user-defined cutoff point is used to reduce the space. The second heuristic only considers the distances from refineries to potential depot locations, and the third heuristic considers the distances from each refinery to cities and power plants. These approaches are formulated below in Equations 18 and 19, respectively.

$$I_k = \frac{\sum_j d_{jk}}{n_d * \max_j(d_{jk})}. \tag{21}$$

$$I_k = \frac{\sum_f d_{kf}}{n_f * \max_f(d_{kf})} + \frac{\sum_p d_{kp}}{n_p * \max_p(d_{kp})}. \tag{22}$$

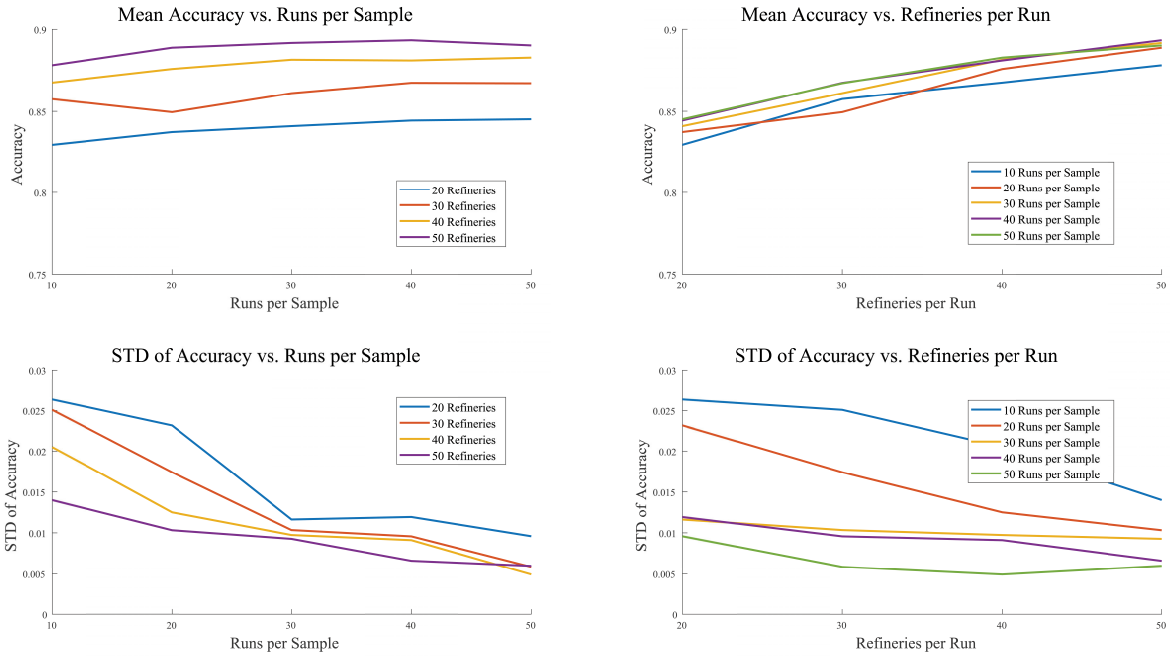


**FIGURE 4.** Best classifier performance. Figure.

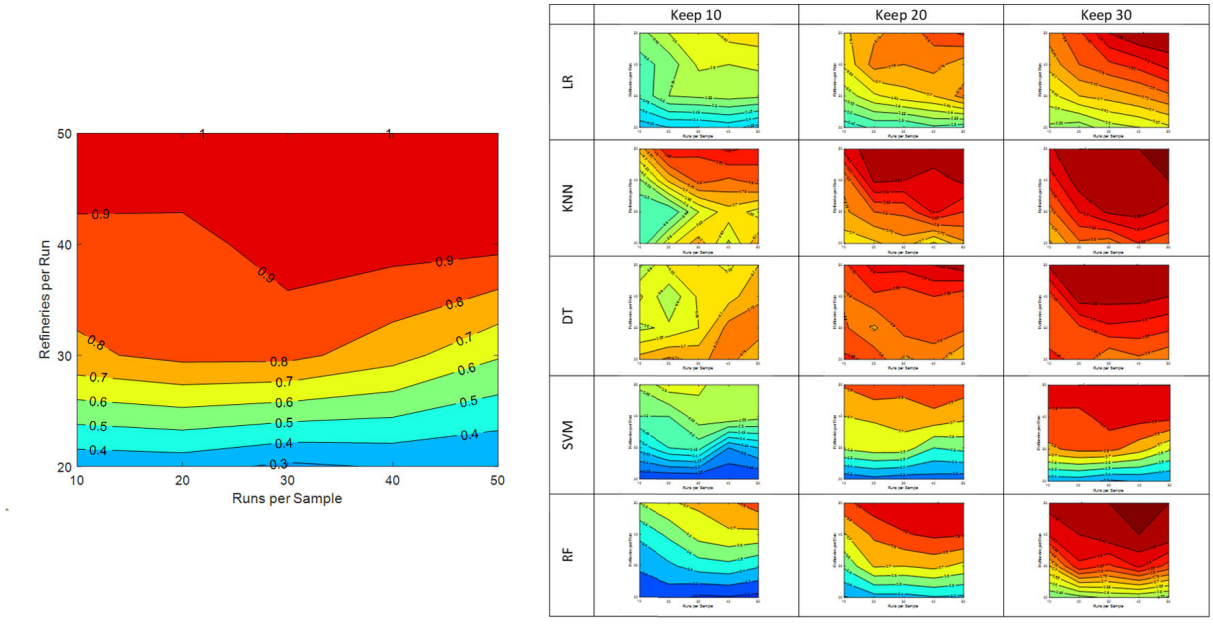


**FIGURE 5.** Recall for ML methods with reference figure.

The cutoff point for each heuristic is 30, which matches the cutoff point for each of the ML-driven solution procedures tested in this section. Each of the 3 heuristics as well as 3 solutions obtained from both the RF and KNN solution procedures are shown below in table 3. Note that the computational times reported include the time to build training sets and fit classifiers. The results indicate that the proposed solution procedure was successful in reducing computation time by approximately 90% with only slight increases in OBJ values. One of the KNN solutions managed to match the optimal solution, but the average RF solution was better across the set of experiments. Both the KNN and RF solutions heavily outperformed the three heuristics tested in regards to optimality. For the three heuristics, the

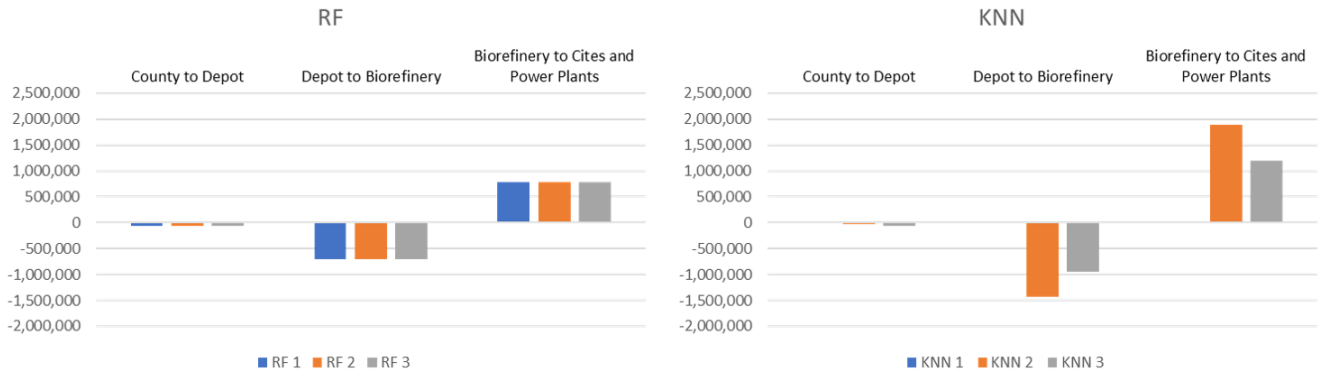


**FIGURE 6.** Cost difference breakdown for KNN and RF solutions.

**TABLE 3.** Comparison of solutions.

| | Average Solution (USD) | Percent Increase | Computational Time (s) | Percent Reduction |
|---|---|---|---|---|
| KNN 1 | 4,944,872,631 | 0.0000% | 1,076 | 94.30% |
| KNN 2 | 4,945,315,690 | 0.0090% | 2,263 | 88.03% |
| KNN 3 | 4,945,043,291 | 0.0035% | 1,644 | 91.30% |
| RF 1 | 4,944,891,670 | 0.0004% | 1,418 | 92.50% |
| RF 2 | 4,944,891,670 | 0.0004% | 1,134 | 94.00% |
| RF 3 | 4,944,891,670 | 0.0004% | 1,545 | 91.83% |
| Heuristic 1 | 4,960,163,615 | 0.3092% | 1,078 | 94.30% |
| Heuristic 2 | 4,984,292,968 | 0.7972% | 3,567 | 81.13% |
| Heuristic 3 | 4,958,548,296 | 0.2766% | 3,010 | 84.08% |
| Full Solution | 4,944,872,631 | N/A | 18,905 | N/A |

| Source | Cost (1000 USD) |
|---|---|
| Shipping from county to depot and densification costs | 237,761.913 |
| Shipping from depot to biorefinery and conversion costs | 428,853.59 |
| Depot investment cost | 114,715.26 |
| Biorefinery investment cost | 2,039,647.97 |
| Unit train contracting cost | 138,005.55 |
| Shipping of byproducts to cities and powerplants | 35,136.85 |
| Demand shortage penalty | 2,009,980.23 |
| **Total** | **5,004,101.36** |

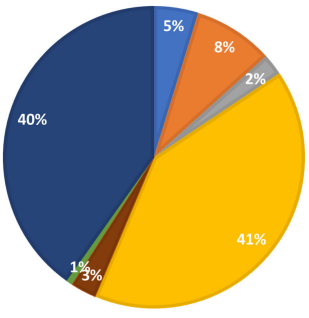


**FIGURE 7.** Cost breakdown for stochastic solution.

solutions indicate that filtering potential biorefineries based on proximity to cities and power plants is more important than depot proximity. This is logically consistent because byproducts are shipped out from refineries via truck and incur a heavier per unit distance cost as opposed to the biomass that is shipped via railroad into each refinery. Further analysis of the KNN and RF solutions revealed that the number of biorefineries, depots and unit trains as well as the third party costs are identical to those of the optimal solution. This result indicates that the cost difference comes from shipping along arcs $\mathcal{U}$, $\mathcal{V}$, $\mathcal{R}$, and $\mathcal{T}$. These differing costs are presented in figure 6 as a comparison to the costs present in the optimal solution.

Interestingly, the ML-derived solutions show reduced costs in depot to biorefinery shipping that are not offset by the increase they experience in byproduct shipping costs. This outcome is attributed to depot distances accounting for 33 of the 58 features used in the classification fits. In conclusion, the results indicate that the ML-driven solution procedure is successful in reducing computational cost while preserving optimality and that both outperform the proposed space-reducing heuristics.

## VI. APPLICATION TO STOCHASTIC MODEL

The best practices discovered in the previous section are applied to the stochastic version of the BSC model here. This model was solved 30 times considering 40 randomly selected potential biorefineries per run. The average run time for each of these was 663 seconds for a total of 13,260 seconds to build the dataset for classification. Each run of the model was solved using CPLEX's built-in Bender's decomposition algorithm with an optimality gap of 5%. From there, RF was selected as the classification algorithm, and the set of potential biorefineries for the final run of the BSC model was restricted to the top 30 responses. Considering only these 30 potential biorefinery locations, the model was run a final time with an optimality gap of 2.5%. The time required to complete this final run of the BSC design model was 128,593 seconds. A breakdown of the solution is shown in Figure 7. The investment and operational costs of biorefineries are the largest contributors to overall network costs. For the shipping arcs, we note that shipments from depots to biorefineries represent the majority of the costs, as the model opts to open depots near suppliers and biorefineries near cities and power plants so that the raw biomass and byproducts that are shipped via truck do not have to travel as far. The bulk of the distance covered is from depots to biorefineries where lower cost railway shipping is utilized.

## VII. CONCLUSION

The development of robust optimization models for the design of biomass supply chains that are cost competitive,

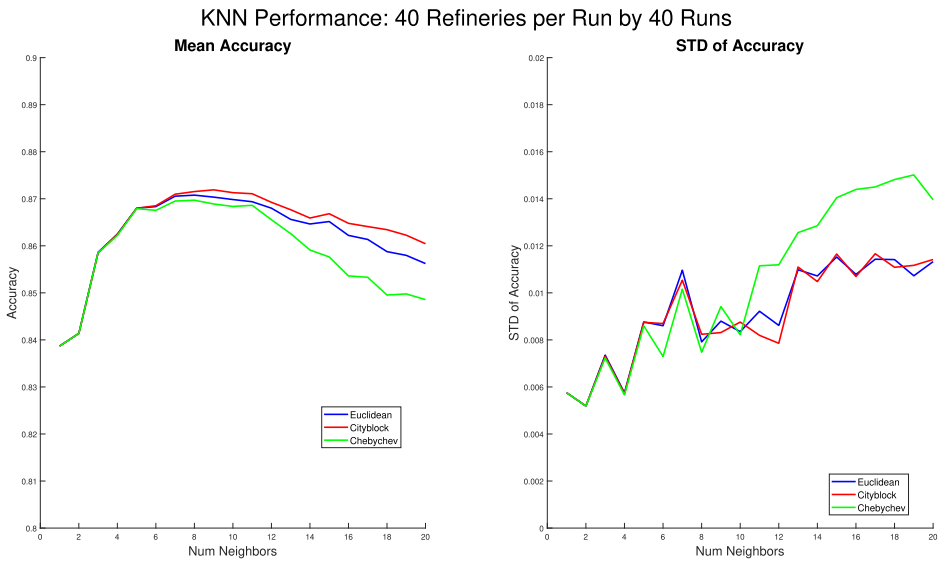


**FIGURE 8.** KNN preliminary performance.

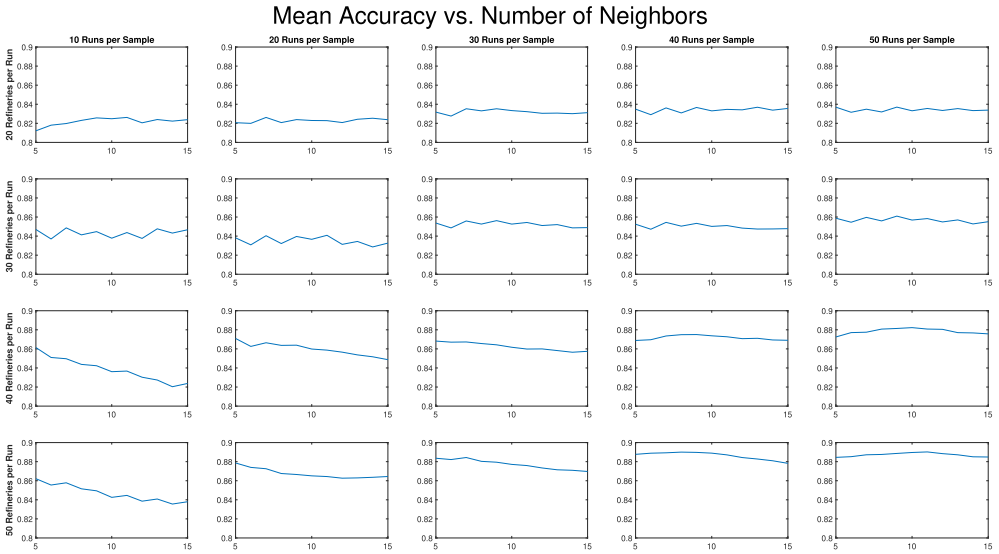


**FIGURE 9.** KNN performance by number of neighbors.

**TABLE 4.** Best performing kernel.

| | 10 Runs Per Sample | 20 Runs Per Sample | 30 Runs Per Sample | 40 Runs Per Sample | 50 Runs Per Sample |
|---|---|---|---|---|---|
| 20 Refineries Per Run | Gaussian | Gaussian | Gaussian | Gaussian | Gaussian |
| 30 Refineries Per Run | Polynomial | Polynomial | Polynomial | Gaussian | Gaussian |
| 40 Refineries Per Run | Polynomial | Polynomial | Polynomial | Polynomial | Polynomial |
| 50 Refineries Per Run | Polynomial | Polynomial | Polynomial | Polynomial | Polynomial |

**TABLE 5.** Best performing resampling percentage.

| | 10 Runs Per Sample | 20 Runs Per Sample | 30 Runs Per Sample | 40 Runs Per Sample | 50 Runs Per Sample |
|---|---|---|---|---|---|
| 20 Refineries Per Run | 25% | 25% | 10% | 10% | 10% |
| 30 Refineries Per Run | 25% | 25% | 25% | 25% | 10% |
| 40 Refineries Per Run | 75% | 25% | 25% | 25% | 25% |
| 50 Refineries Per Run | 50% | 50% | 50% | 50% | 25% |

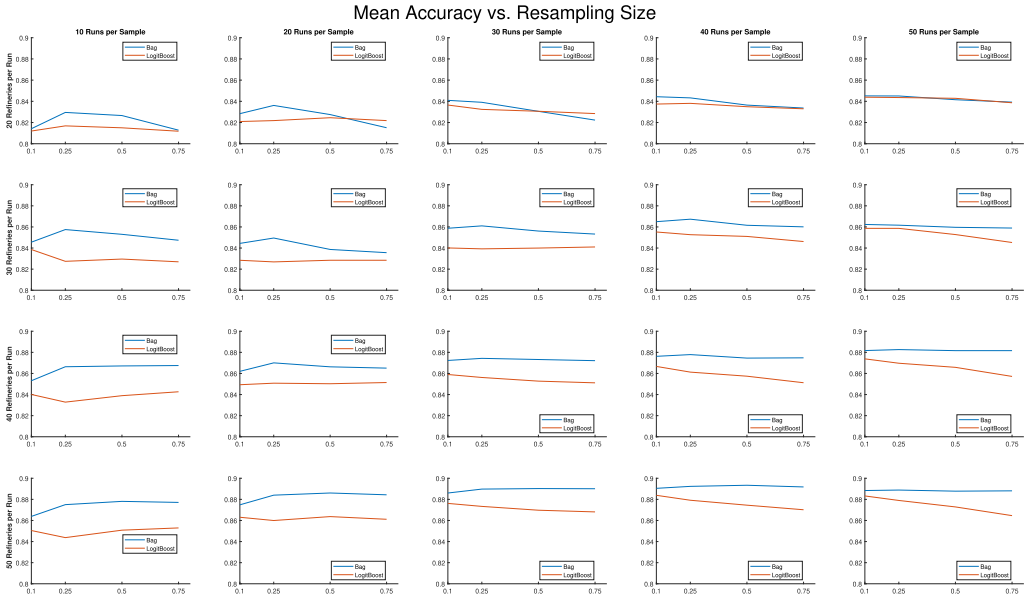


**FIGURE 10.** RF performance by resampling percentage.

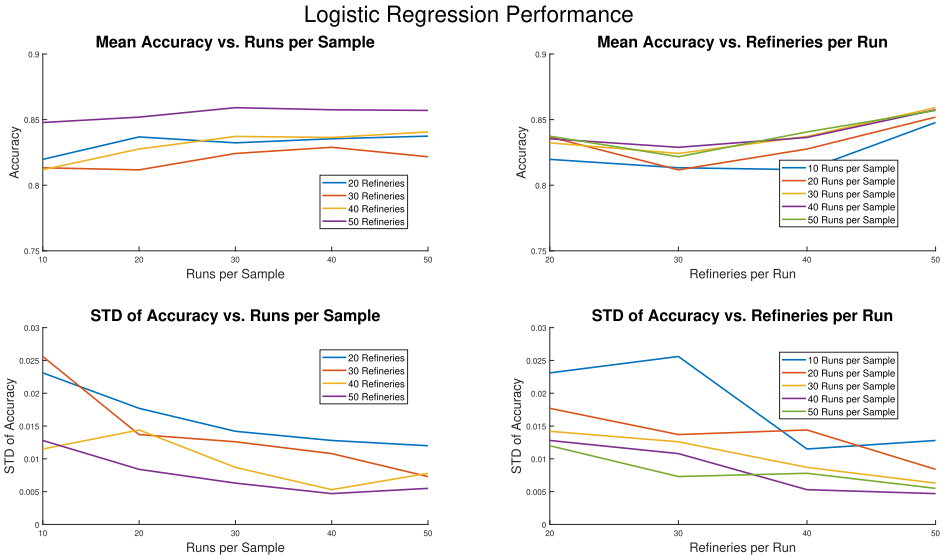


**FIGURE 11.** Logistic regression performance.

satisfy real-world demands and capture the inherent uncertainty in biomass quality leads to heavy computational burdens. Here, we proposed a novel machine learning driven solution procedure to aid in solving stochastic optimization models of this type via stage-one variable space reduction. This solution procedure utilized multiple reduced space runs of the BSC model to build a dataset suitable for classification to derive a reduced space that contains potential biorefinery locations likely to be selected as optimal. Our strategy raised four key questions in modeling this problem: how much can the space be reduced when building the dataset, how many runs are required to build a reliable dataset, what classification techniques are most suitable, and is the proposed approach effective in meeting its goals? Our classification accuracy results indicated that for a consideration of 167 total refineries, selecting 40 refineries per run (which is approximately 25% of the total set) when building the datasets was optimal in terms managing prediction accuracy and computational time. The consideration of fewer refineries lead to reduced accuracy, while the consideration of more refineries generated minimal accuracy gains. Contextualizing 40 refineries requires 4 times the number selected in a given run of the model, meaning that the training data were composed of 25% positive responses (elaborate). Thirty runs per sample was deemed the best number of iterations based on trends observed in the standard deviation of accuracy

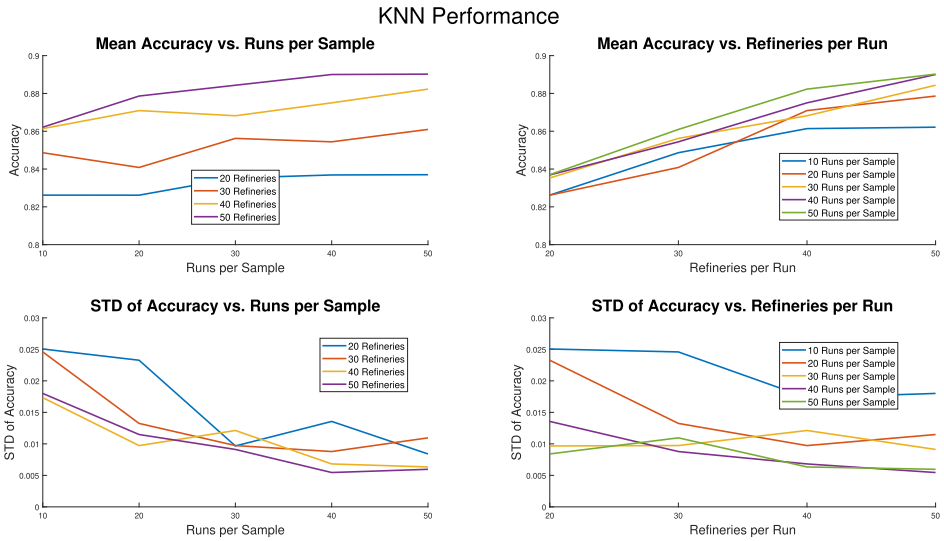


**FIGURE 12.** KNN performance.

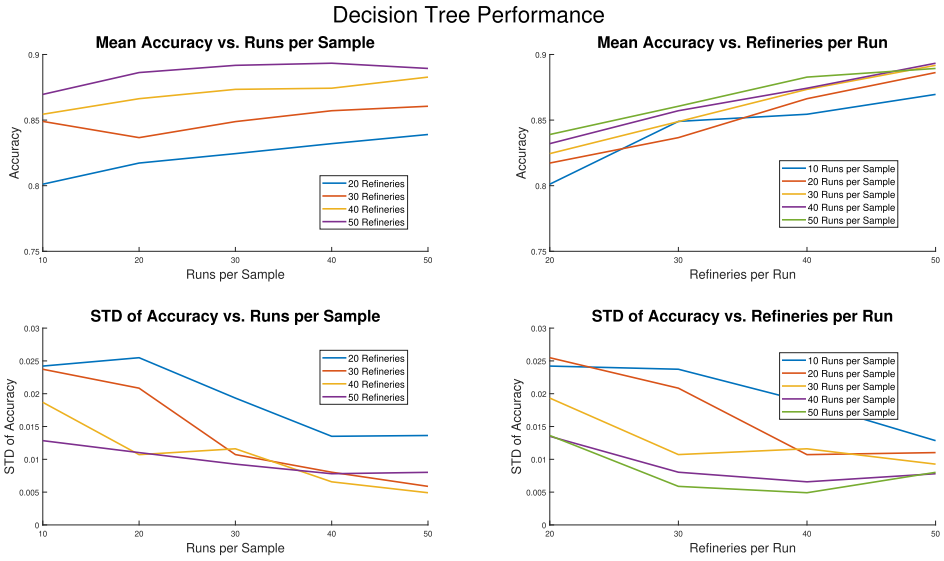


**FIGURE 13.** Decision tree performance.

between samples of the same size. Accuracy and recall considerations demonstrated RF and KNN to be the best of the classification techniques considered. RF achieved the best performance when a higher number of responses were kept for larger datasets. KNN performed well across all trials but stood out for small datasets and situations where only a small number of responses were kept. These outcomes are attributed to the features for classification being physical distances, which is well-suited to KNN-type algorithms. Regarding RF, the solution procedure is reminiscent of bagging methods, as randomly selected reduced space runs of the optimization model are combined to build the training datasets. The solution procedure, when applied to the deterministic version of the model, resulted in an approximately 90% overall computation time reduction with only a small (approximately 0.0004%) increase in the objective value. The solution procedure was then applied to obtain a near-optimal solution of the stochastic version of the BSC model, which otherwise would be computationally exhaustive given the same level of computational resources. For future works, we will consider better modeling of the correlation among the refineries for building the machine learning datasets.

## APPENDIX. RELEVANT TUNING PARAMETERS

Tables 4 and 5, and Fig. 10.

## APPENDIX. VISUALIZATION OF CLASSIFIER PERFORMANCE

See Figs. 11–15.

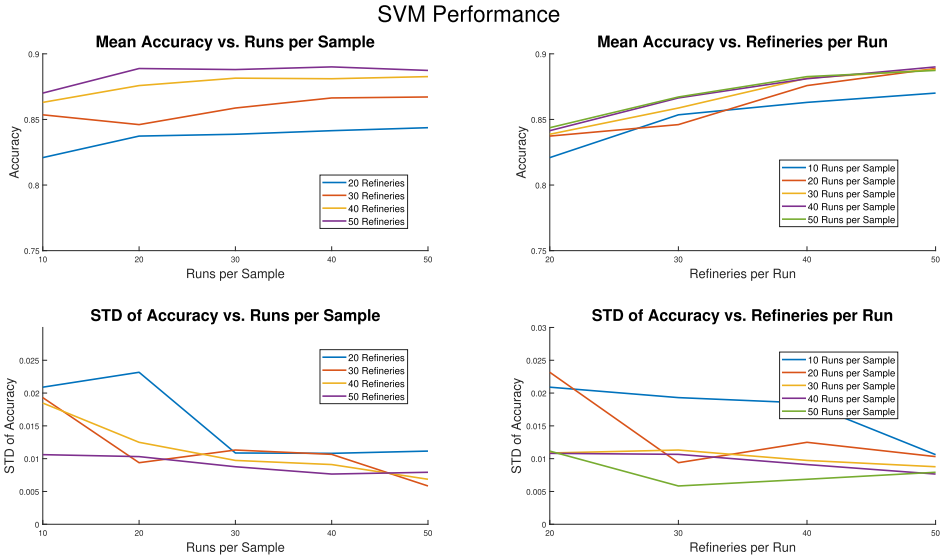


**FIGURE 14.** SVM performance.

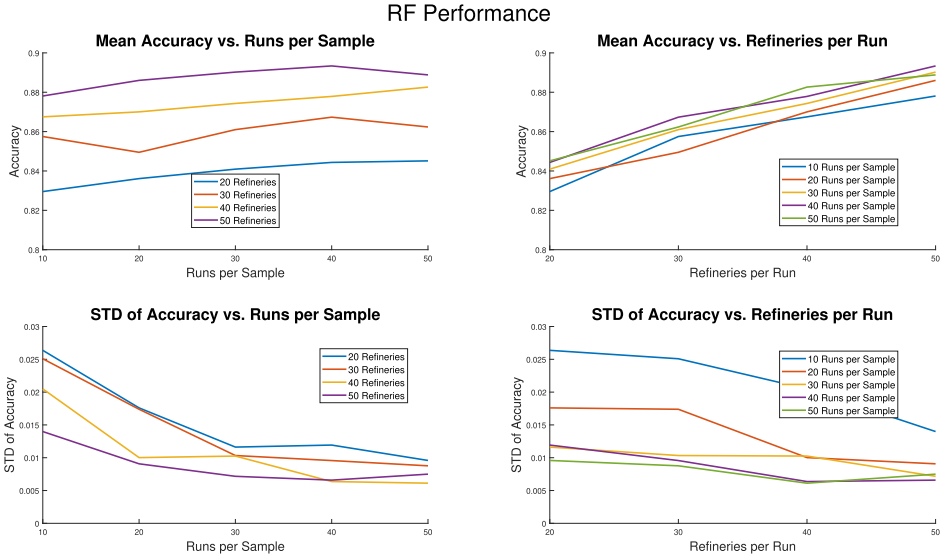


**FIGURE 15.** RF performance.

## REFERENCES


[1] M. Aboytes-Ojeda, K. K. Castillo-Villar, and S. D. Eksioglu, "Modeling and optimization of biomass quality variability for decision support systems in biomass supply chains," *Ann. Oper. Res.*, vol. 314, no. 2, pp. 319–346, Dec. 2019, doi: 10.1007/s10479-019-03477-8.

[2] H. N. Geismar, B. A. McCarl, and S. W. Searcy, "Optimal design and operation of a second-generation biofuels supply chain," *IISE Trans.*, vol. 54, no. 4, pp. 390–404, 2021. [Online]. Available: https://www.tandfonline.com/action/showCitFormats?doi=10.1080%2F24725854.2021.1956022

[3] G. Memişoğlu and H. Üster, "Design of a biofuel supply network under stochastic and price-dependent biomass availability," *IISE Trans.*, vol. 53, no. 8, pp. 869–882, Aug. 2021, doi: 10.1080/24725854.2020.1869870.

[4] F. Nur, M. Aboytes-Ojeda, K. K. Castillo-Villar, and M. Marufuzzaman, "A two-stage stochastic programming model for biofuel supply chain network design with biomass quality implications," *IISE Trans.*, vol. 53, no. 8, pp. 845–868, Aug. 2021, doi: 10.1080/24725854.2020.1751347.

[5] A. Oroojlooyjadid, L. V. Snyder, and M. Takáč, "Applying deep learning to the newsvendor problem," *IISE Trans.*, vol. 52, no. 4, pp. 444–463, Apr. 2020, doi: 10.1080/24725854.2019.1632502.

[6] B. Defourny, D. Ernst, and L. Wehenkel, "Scenario trees and policy selection for multistage stochastic programming using machine learning," *INFORMS J. Comput.*, vol. 25, no. 3, pp. 488–501, Aug. 2013.

[7] D. Goettsch, K. K. Castillo-Villar, and M. Aranguren, "Machine-learning methods to select potential depot locations for the supply chain of biomass co-firing," *Energies*, vol. 13, no. 24, p. 6554, Dec. 2020.

[8] M. M. Aguayo, S. C. Sarin, and J. S. Cundiff, "A branch-and-price approach for a biomass feedstock logistics supply chain design problem," *IISE Trans.*, vol. 51, no. 12, pp. 1348–1364, Dec. 2019.

[9] S. Eksioglu, "Contributions to sustainable bioenergy systems design, planning and operations," *IISE Trans.*, vol. 53, no. 8, pp. 843–844, Aug. 2021.

[10] L. Panichelli and E. Gnansounou, "GIS-based approach for defining bioenergy facilities location: A case study in northern Spain based on marginal delivery costs and resources competition between facilities," *Biomass Bioenergy*, vol. 32, no. 4, pp. 289–300, Apr. 2008.

[11] J. S. Cundiff, N. Dias, and H. D. Sherali, "A linear programming approach for designing a herbaceous biomass delivery system," *Bioresource Technol.*, vol. 59, no. 1, pp. 47–55, Jan. 1997.
[12] O. Akgul, N. Shah, and L. Papageorgiou, "An milp model for the strategic design of the uk bioethanol supply chain," in *Computer Aided Chemical Engineering*, vol. 29. Elsevier, 2011, pp. 1799–1803.
[13] S. Leduc, D. Schwab, E. Dotzauer, E. Schmid, and M. Obersteiner, "Optimal location of wood gasification plants for methanol production with heat recovery," *Int. J. Energy Res.*, vol. 32, no. 12, pp. 1080–1091, Oct. 2008.
[14] J. Ramage and J. Scurlock, "Biomass," in *Renewable Energy: Power for a Sustainable Future*, G. Boyle, Ed. Oxford, U.K.: Oxford Univ. Press, 1996.
[15] I. Awudu and J. Zhang, "Uncertainties and sustainability concepts in biofuel supply chain management: A review," *Renew. Sustain. Energy Rev.*, vol. 16, no. 2, pp. 1359–1368, Feb. 2012.
[16] F. Andersen, F. Iturmendi, S. Espinosa, and M. S. Diaz, "Optimal design and planning of biodiesel supply chain with land competition," *Comput. Chem. Eng.*, vol. 47, pp. 170–182, Dec. 2012.
[17] I. M. Bowling, J. M. Ponce-Ortega, and M. M. El-Halwagi, "Facility location and supply chain optimization for a biorefinery," *Ind. Eng. Chem. Res.*, vol. 50, no. 10, pp. 6276–6286, May 2011.
[18] K. K. Castillo-Villar, S. Eksioglu, and M. Taherkhorsandi, "Integrating biomass quality variability in stochastic supply chain modeling and optimization for large-scale biofuel production," *J. Cleaner Prod.*, vol. 149, pp. 904–918, Apr. 2017.
[19] M. Marufuzzaman, S. D. Eksioglu, and Y. (Eric) Huang, "Two-stage stochastic programming supply chain model for biodiesel production via wastewater treatment," *Comput. Oper. Res.*, vol. 49, pp. 1–17, Sep. 2014.
[20] M. D. Casler and A. R. Boe, "Cultivar × environment interactions in switchgrass," *Crop Sci.*, vol. 43, no. 6, pp. 2226–2233, Nov. 2003.
[21] (2017). *National Renewable Energy Laboratory*. [Online]. Available: https://maps.nrel.gov/bioenergyatlas
[22] D. A. Freedman, "Bootstrapping regression models," *Ann. Statist.*, vol. 9, no. 6, pp. 1218–1228, Nov. 1981.

**KOLTON KEITH** received the Master of Science (M.S.) degree in computational and applied mathematics from Texas A&M University, College Station, TX, USA, in 2017, and the Doctor of Philosophy (Ph.D.) degree in mechanical engineering from The University of Texas at San Antonio, San Antonio, TX, in 2023. His research interests include supply chain optimization, stochastic programming, machine learning, biofuels, and cybersecurity.

**KRYSTEL K. CASTILLO-VILLAR** received the Ph.D. degree from Texas Tech University and the Sc.D. degree in engineering sciences from Monterrey Tech. She is currently the Lutcher Brown Chair of the Department of Mechanical Engineering, the Vice President of Energy Efficiency with the Cybersecurity Manufacturing Innovation Institute (DOE sponsored), and the Director of Texas Sustainable Energy Research Institute, The University of Texas at San Antonio (UTSA). Her research expertise is in mathematical programming and optimization techniques for analyzing large-scale, complex systems under uncertainty, and data analytics. Her research is grounded in relevant applications, such as modeling and optimization of renewable energy systems, cybersecurity manufacturing, and supply chains. She has participated in 44 grants funded by multiple agencies, including the USDA, DOE, NSF, EPA, and Air Force Research Laboratory. For her research contributions to manufacturing, she was inducted to the UTSA Academy of Distinguished Researchers, in 2021. She is a member of the National Academy of Engineering Frontiers of Engineering Alumni and she has participated as a Committee Member for the National Academies Options for a National Smart Manufacturing Plan Study.

**ADEL ALAEDDINI** received the Ph.D. degree in industrial and systems engineering from Wayne State University. He is currently an Associate Professor in mechanical engineering with The University of Texas at San Antonio. He also performed postdoctoral research with the University of Michigan, Ann Arbor. His main research interests include statistical learning in systems modeling and control and data analytics in health care and manufacturing. He has contributed to over 30 peer-reviewed publications in journals, such as *IISE Transactions*, *Production and Operations Management* (POMS), and *Information Sciences*.