Optimizing Operator-Task Assignments via Linear Programming

Kevin P. Rodriguez

December 2025

Abstract

We consider the problem of optimizing operator-task assignments in a data collection operation supporting autonomous system development. In this setting, human operators collect training data across multiple task types, and the quality of that data directly impacts downstream model performance. We introduce a bespoke quantitative recipe for maximizing data volume and data quality simultaneously, accounting for individual operator aptitude (*i.e.* QA pass rate and data volume output), data collection site count, site supervisor count, autonomous unit availability, and task types. The methods used here are based on state-of-the-art operations research (OR) techniques including the simplex method, decision analysis, linear programming (LP), mixed-integer programming, and more. Current implementation is a static, fixed-shift analysis. Further research will account dynamically for real time changes in autonomous unit and operator availability.

1 Introduction

Modern autonomous systems, including but not limited to: self-driving vehicles, "embodied AI" robots for industrial and consumer use, *etc.*, depend on large volumes of high-quality training data collected by human operators. In such operations, maximizing data volume output per unit time is essential for accelerating development, while simultaneously maximizing data quality is critical for producing models that generalize reliably. This is a classic operations research (OR) problem subject to several constraints, including:

- number of data collection sites,
- number of task types (denoted T1, T2, ..., T_T) to be completed according to current research objectives,
- number of operators on shift,
- individual operator volume production, and
- individual operator QA pass/fail rates (*task-specific*).

Historically, data volume (lump sum hours per shift) and data quality (individual operator pass/fail rates) have been addressed and prioritized as separate but related entities. The methods that follow will show that under proper optimization, there is a **unique solution** of operator + task + site assignments that produces the maximal volume of data while simultaneously maintaining the highest mathematically possible data quality, on average.

2 LP Model Definitions

In this section we quantitatively define our input parameters, decision variables, and objective function in the context of data volume and data quality maximization.

2.1 Quality-Scaled Operator Data Collection Rate

There are three subtleties in defining individual operator collection rates:

1. Some operators are better at certain tasks than others.

2. An operator that produces lots of volume is not necessarily producing quality data.

3. Autonomous units go down regularly, artificially altering operator volume production on paper.

To address these problems, rather than using raw operator volume as an input parameter, we make the definition:

$$r_{ij} = \langle v_{ij} \rangle \times p_{ij}, \tag{1}$$

where

$$\begin{aligned} \langle v_{ij} \rangle &= \text{average volume per hour of the } i\text{th operator on task } j \\ p_{ij} &= \text{QA pass/fail rate of the } i\text{th operator on task } j. \end{aligned}$$

We will refer to r_{ij} as the **quality-scaled volume rate** of operator i on task j, or simply "**scaled rate**" for short.[1] See footnote for details regarding the units of v_{ij}.

Example 1: Scaled Rate Calculation

If Operator Bob collects 1.5 Hours of data during an 8 hour shift on task T1 with a QA pass/fail rate of 80% (the minimum acceptable threshold), we would have the definitions:

$$\begin{aligned} \langle v_{\text{Bob, T1}} \rangle &= \left(\frac{1.5\,\text{Hr}}{8\,\text{hr}} \right) \\ p_{\text{Bob, T1}} &= 0.80. \end{aligned}$$

Plugging these values into Equation 1, we see that Bob's scaled rate for T1 would be

$$r_{\text{Bob, T1}} = \left(\frac{1.5\,\text{Hr}}{8\,\text{hr}} \right) \times 0.80 = \boxed{0.15\,\text{Hr/hr}}$$

The units, while awkward looking, are informative: Bob produces 0.15 Hours (or 9 minutes) of usable data volume (quality = `Acceptable` or `Exceptional`) per shift hour of actual work, on average. In practice, both $\langle v_{ij} \rangle$ and r_{ij} would be computed from running averages of his existing work.

Example 2: Volume vs. Quality

Consider another operator Alice, who collects 1.35 Hours of data during an 8 hour shift on T1, but with an exceptional QA pass/fail rate of 98%.

Operator	Volume (Hr/8hr)	Pass Rate	Scaled Rate (Hr/hr)
Bob	1.50	80%	0.150
Alice	1.35	98%	0.165

Despite lower raw volume, Alice produces ∼10% more usable data per hour due to her higher pass rate. This illustrates why raw volume alone is a misleading metric.

[1] Note that the units of v_{ij} – being measured in data-volume-hours per actual-shift-hours – are peculiar, and are technically dimensionless. To avoid confusion, units of hours of data (volume) collected will be denoted as "Hours" (Hr), while actual elapsed hours as measured by clock will remain simply "hours" (hr). Thus, v_{ij} has units of Hr/hr.

Example 3: Task-Dependent Optimization

Consider two operators with different strengths across tasks:

Operator Charlie:

Task	Volume (Hr/8hr)	Pass Rate	Scaled Rate (Hr/hr)
T1	2.4	95%	0.285
T2	1.2	85%	0.128

Operator Dana:

Task	Volume (Hr/8hr)	Pass Rate	Scaled Rate (Hr/hr)
T1	1.6	80%	0.160
T2	2.0	95%	0.238

A naive assignment might place both on T1 (assuming it is the "priority" task). However, the optimal assignment is:

$$\begin{aligned} \text{Charlie} \to \text{T1} \quad (r = 0.285) \\ \text{Dana} \to \text{T2} \quad (r = 0.238) \end{aligned}$$

yielding a combined output of $0.285 + 0.238 = 0.523$ Hr/hr. The reverse assignment yields only $0.128 + 0.160 = 0.288$ Hr/hr—**45% less usable data.**

While the correct choice is obvious in this simple case, optimizing decisions for 100+ operators and tasks at half a dozen sites is highly nontrivial, and mathematically intractable on paper. This is the motivation for this Model developed in the following sections, which provides the foundation for a computer-based optimization that resolves in seconds.

2.2 Input Parameters

Before defining our decision variables, let us enumerate the input parameters (known quantities) necessary for us to compute the optimal assignment solution[2]:

N = total number of available autonomous units
M = total number of data collection sites (= total site supervisor count)
R = total number of operators on shift
t_j = minimum number of operators required to be on task j
s_k = maximum number of operators to be stationed at the kth collection site
T = total number of task targets
r_{ij} = quality-scaled operator collection rate

In this Model, each site k is assumed to have only one associated site supervisor, such that they are mathematically one and the same. The Model will additionally allow preferential assignments of particular operators to particular sites (supervisors) when desired.

2.3 Decision Variables

We are now in a position to define our set of decision variables (unknowns) which will represent the assignments of operators to particular tasks and locations. We define the binary decision variable:

$$x_{ijk} \in \{0, 1\} \tag{2}$$

where $x_{ijk} = 1$ if operator i is assigned to task j at site k, and $x_{ijk} = 0$ otherwise. Keep in mind that the subscripts i, j, and k will range from 1 to R, T, and M, respectively. Thus – neglecting other constraints – there will be a total of $R \times T \times M$ decision variables of the form x_{ijk} to determine.

[2]Recall that knowledge of r_{ij} requires both $\langle v_{ij} \rangle$ and p_{ij} as defined in the previous section.

2.4 Objective Function

Now with our definitions of r_{ij} as the quality-scaled operator collection rate and x_{ijk} as the assignment decision variable, we can sum over all feasible combinations and define the **total projected quality-scaled volume** as

$$V = \sum_{i=1}^{R} \sum_{j=1}^{T} \sum_{k=1}^{M} r_{ij} \cdot x_{ijk} \tag{3}$$

This is our objective function, which we seek to maximize under the constraints in the following section.

2.5 Constraints

Below we define the linear constraint equations relevant to our problem at hand:

$$\sum_{j=1}^{T} \sum_{k=1}^{M} x_{ijk} = 1, \quad \forall\, i \qquad \text{(each operator assigned exactly once)} \tag{4}$$

$$\sum_{i=1}^{R} \sum_{k=1}^{M} x_{ijk} \geq t_j, \quad \forall\, j \qquad \text{(minimum task staffing)} \tag{5}$$

$$\sum_{i=1}^{R} \sum_{j=1}^{T} x_{ijk} \leq s_k, \quad \forall\, k \qquad \text{(maximum site staffing)} \tag{6}$$

$$\sum_{i=1}^{R} \sum_{j=1}^{T} \sum_{k=1}^{M} x_{ijk} \leq N \qquad \text{(maximum unit availability)} \tag{7}$$

In words:

- Equation (4) says that each operator i may only be assigned to a single task j and at a single site k at a time.
- Equation (5) says that each task j must have at least t_j operators working on it per shift. In practice, priority tasks will be assigned the desired number of operators ($t_j > 0$), while lower priority tasks may be set as $t_j = 0$. The Model will automatically assign operators to the *optimal* task j outside of priority in this case.
- Equation (6) says that each site k may only have a maximum of s_k operators present for a given shift.
- Finally, Equation (7) demands that there may never be more operators than autonomous units in a given shift.[3]

3 Worked Model Example: Single-Site Task Assignments

In this section, we will use a small set of mock data as input parameters, along with the decision variables, objective function, and constraints as defined in the previous section to demonstrate the utility of our Model.

Consider a small shift with the following parameters:

$$\begin{aligned} R &= 5 \text{ operators (Alice, Bob, Charlie, Dana, Eve)} \\ T &= 3 \text{ tasks (T1, T2, T3)} \\ M &= 1 \text{ site} \\ N &= 4 \text{ available autonomous units} \\ s_1 &= 5 \text{ (max site capacity)} \\ t_{\mathrm{T1}} = 1, \quad t_{\mathrm{T2}} &= 1, \quad t_{\mathrm{T3}} = 1 \text{ (task minimums)} \end{aligned}$$

[3] In practice, operators may be without units if this condition is not met – the Model will then preferentially assign existing units to operators which will maximize V. Alternatively, this condition may be relaxed if excess operators are allowed to shadow operators who do have unit priority.

Suppose that the quality-scaled data volume rates (r_{ij}) of each operator for the three tasks are calculated from their previous work history as:

Operator	T1	T2	T3
Alice	0.18	0.12	0.15
Bob	0.15	0.20	0.10
Charlie	0.14	0.22	0.19
Dana	0.16	0.14	0.24
Eve	0.11	0.13	0.12

Table 1: Quality-scaled rates r_{ij} (Hr/hr) for each operator-task pair.

3.1 Decision Variables (example)

Because there are $R = 5$ operators, $T = 3$ tasks, and $M = 1$ sites, we have a total of $5 \times 3 \times 1 = 15$ binary decision variables x_{ijk}. These are our unknowns, which – when solved for – will assign operators to the optimal task.

Additionally, since $M = 1$, we may simplify the notation of our decision variables for this example, defining:

$$x_{ij} \equiv x_{ij1}.$$

3.2 Objective Function (example)

With only one site, the objective function collapses to a double sum over operators (i) and tasks (j):

$$V = \sum_{i=1}^{5} \sum_{j=1}^{3} r_{ij} \cdot x_{ij}.$$

Expanding the sum with the data in Table 1 gives

$$\begin{aligned} V = {} & 0.18x_{A,1} + 0.12x_{A,2} + 0.15x_{A,3} \\ & + 0.15x_{B,1} + 0.20x_{B,2} + 0.10x_{B,3} \\ & + 0.14x_{C,1} + 0.22x_{C,2} + 0.19x_{C,3} \\ & + 0.16x_{D,1} + 0.14x_{D,2} + 0.24x_{D,3} \\ & + 0.11x_{E,1} + 0.13x_{E,2} + 0.12x_{E,3} \end{aligned}$$

3.3 Constraints (example)

Each operator is assigned at most once (with unit constraint, some may be unassigned):

$$\sum_{j=1}^{3} x_{ij} \leq 1, \quad \forall i$$

Task minimums requirement:

$$\sum_{i=1}^{5} x_{ij} \geq t_j, \quad \forall j$$

Unit availability requirement:

$$\sum_{i=1}^{5} \sum_{j=1}^{3} x_{ij} \leq 4$$

3.4 Solution (example)

By inspection of the r_{ij} matrix in Table 1, each task's best performer:

- T1: Alice (0.18)
- T2: Charlie (0.22)
- T3: Dana (0.24)

With one unit remaining, assign the best remaining operator-task pair. Bob on T2 (0.20) beats all of Eve's options.

Optimal assignment:

$$\begin{aligned} \text{Alice} &\to \text{T1} \quad (r = 0.18) \\ \text{Bob} &\to \text{T2} \quad (r = 0.20) \\ \text{Charlie} &\to \text{T2} \quad (r = 0.22) \\ \text{Dana} &\to \text{T3} \quad (r = 0.24) \\ \text{Eve} &\to \text{unassigned (standby/shadow)} \end{aligned}$$

Plugging back into the objective function, we find that the expected maximized quality-scaled volume per hour for this group is: $V^* = 0.18 + 0.20 + 0.22 + 0.24 = 0.84$ Hr/hr.

In other words, we would expect approximately 0.84 Hr/hr $\times$ 8 hr = $\boxed{6.72\,\text{Hr}}$ of **usable quality** (`Acceptable` / `Exceptional`) data from this particular arrangement of operators **per shift**. For comparison, here are the outputs for a several different arrangements:

T1	T2	T3	Unassigned	V (Hr/hr)	$V \times 8$ (Hr)
Alice	Bob, Charlie	Dana	Eve	**0.84**	**6.72**
Alice	Bob	Charlie, Dana	Eve	0.81	6.48
Alice, Bob	Charlie	Dana	Eve	0.79	6.32
Bob	Alice, Charlie	Dana	Eve	0.73	5.84
Alice	Charlie	Bob	Dana	0.62	4.96

Table 2: Comparison of selected feasible assignments.

3.5 Discussion

This toy example, while contrived from a small set of fictitious data, demonstrates the power and utility of our LP Model. The true utility of the Model shines when there are many more variables than can be accounted for by hand as we did here. For example, a realistic shift would have perhaps 30 operators across 6 sites with 40 different tasks. This would create a system of $30 \times 40 \times 6 = 7200$ variables. In this case, and in the following section, we make use of the Python package `PuLP` to aid in calculation.

4 Comparison of LP Model to Random and Greedy Assignments

In this section we directly compare the numerical results between our LP model and four different naive candidate task assignment algorithms ("Methods"): random, volume greedy, pass-rate greedy, and quality-scaled volume greedy. Briefly, the descriptions of the different assignment algorithms are as follows:

1. **Random**: Operators are assigned randomly to different tasks, regardless of their existing QA pass rates or volume rates per task.
2. **Volume Greedy**: Operators are randomly ordered and then assigned to tasks based exclusively on their individual $\langle v_{ij} \rangle$, *i.e.*, average volume rate for operator i on task j.
3. **Pass-Rate Greedy** Operators are randomly ordered and then assigned to tasks based exclusively on their individual p_{ij}, *i.e.*, QA pass percentage for operator i on task j.

4. **Quality-Scaled Greedy** Operators are randomly ordered and then assigned to tasks based on their task–specific quality-scaled data volume rates r_{ij} as introduced in Equation (1).

5. **Optimized LP Model**: Operators are assigned to tasks based upon the unique configuration which maximizes the objective scaled-volume function in Equation (3), subject to task, site, and unit availability constraints as defined in Equations (4) - (7).

Current operator assignments are based on a combination of Methods 1-4, with some tasks being assigned based on quality considerations (especially those operators handpicked by the research team for particular tasks), some assigned with intent to maximize volume (often by supervisors who wish to increase operator volume regardless of quality), and the remaining essentially random: up to the discretion of individual supervisors and/or operators themselves, within allowed limits of task/site/unit staffing and research goals.

4.1 Monte Carlo Simulations

Methods 1-4 involve large degrees of freedom and inherent randomness. It is in our interest to investigate **all** possible outcomes in such scenarios – a single set of operator data may lead to innumerable outcomes for these algorithms, especially when the number of autonomous units is not equal to the number of operators. To better understand and visualize the total range and likelihood of possible outcomes, we use the Monte Carlo method to simulate $N = 10,000$ assignments for each Method 1-4, thus generating a probability mass function $f_n(V)$ for $n \in \{1, 2, 3, 4\}$, where V is the total quality-scaled volume over an 8 hour shift. The functions $f_n(V)$ are normalized such that

$$1 = \int_0^{\max V_n} f_n(V)\,\mathrm{d}V.$$

For example, $f_2(35\,\mathrm{Hr})\mathrm{d}V$ would represent the probability of producing 35 quality-Hours of data in a single shift with under Method 2: volume greedy assignments. The corresponding mean μ_n of each probability density function can then be calculated in general as

$$\mu_n \equiv \langle V_n \rangle = \int_0^{\max V_n} V\, f_n(V)\,\mathrm{d}V$$

to give the expected value of the total quality-scaled data volume for the nth Method.

Note in particular that Method 5 (our optimized LP model) will not yield a probability density function, the reason being that LP optimization produces only a *single* configuration to maximize the objective function, by definition. For notational consistency, however, we will refer to the corresponding expected quality-scaled volume of this Method as μ_5.

4.2 Input Data

The input data for this analysis has been synthetically generated to replicate known characteristics of real operator performance data, based on the author's domain knowledge:

- **Pass rate distribution:** Task-averaged operator pass rates are approximately normally distributed, bounded below at 50% (the minimum acceptable threshold), with population mean $\mu_p \approx 75\%$. Formally, $\langle p_i \rangle \in [0.50, 1.00]$ with $\mathbb{E}[\langle p_i \rangle] \approx 0.75$.

- **Volume rate distribution:** A small subset of operators produces disproportionately high volume, resulting in a right-skewed distribution. Volume rates span $\langle v_i \rangle \in (0.0, 0.4)$ Hr/hr, with most operators clustered toward the lower end.

- **Skill consistency correlation:** Higher-performing operators tend to exhibit more uniform performance across tasks, while lower-performing operators show greater task-to-task variability. Mathematically, the per-operator standard deviation of pass rates decreases with as average performance increases: $\sigma_{p_i} \to 0$ as $\langle p_i \rangle \to 1$.

These characteristics are illustrated in Figures 1 and 2 for a synthetic batch of 32 operators.

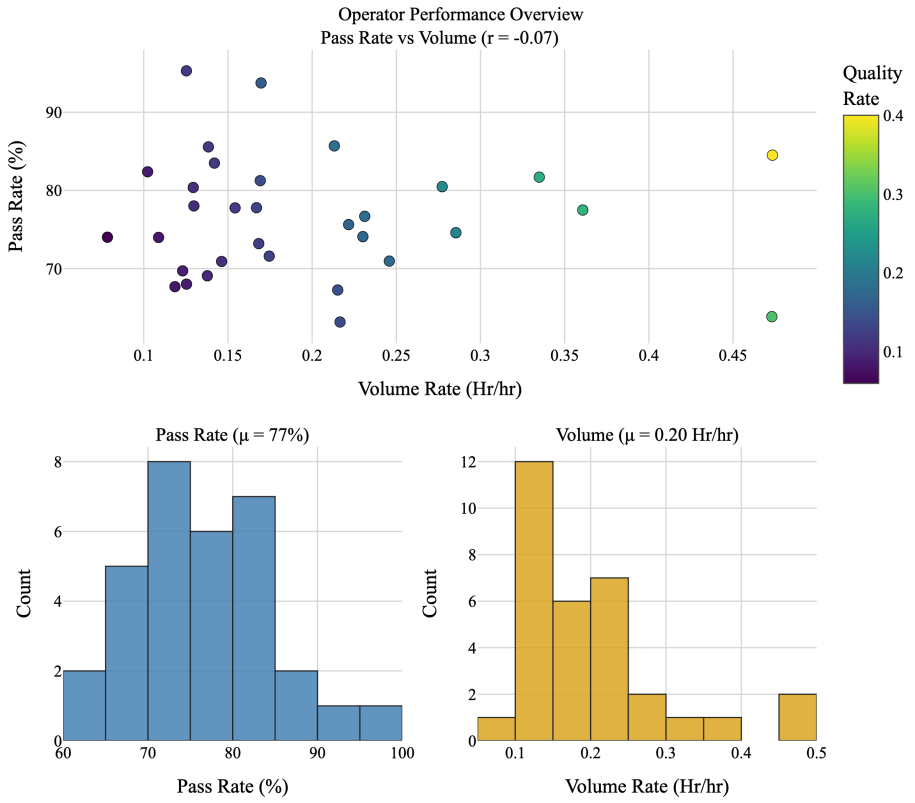

Figure 1: Overview of synthetic operator performance data. **Top:** Scatter plot of mean pass rate versus mean volume rate for each operator, colored by quality-scaled rate $\langle r_i \rangle = \langle v_i \rangle \times \langle p_i \rangle$. The weak correlation ($r = -0.06$) indicates that high-volume operators are not necessarily high-quality. **Bottom left:** Distribution of task-averaged pass rates across operators. **Bottom right:** Distribution of task-averaged volume rates, showing characteristic right skew.

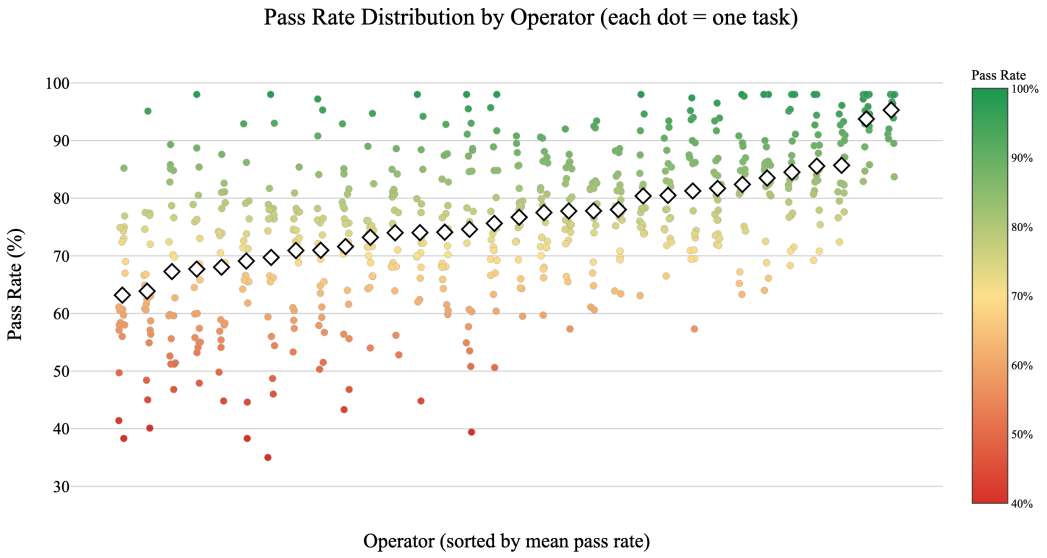

Figure 2: Task-level pass rate distribution by operator, sorted by mean performance. Each column represents one operator; individual dots are pass rates for specific tasks; white diamonds indicate the operator's mean across all tasks. The decreasing vertical spread from left to right illustrates the skill consistency correlation: higher-performing operators exhibit less task-to-task variability ($\sigma_{p_i} \to 0$ as $\langle p_i \rangle \to 1$).

4.3 Results

Figure 3 shows the outcome distributions $f_n(V)$ for each assignment method across $N = 10{,}000$ Monte Carlo trials. The simulation uses the synthetic operator data from Section 4.2 with shift parameters $N = 30$ effective autonomous units, $R = 32$ operators, $T = 21$ task types, and $M = 6$ sites. Task staffing minimums range from $t_j \in [0, 2]$ and site capacities from $s_k \in [6, 15]$.

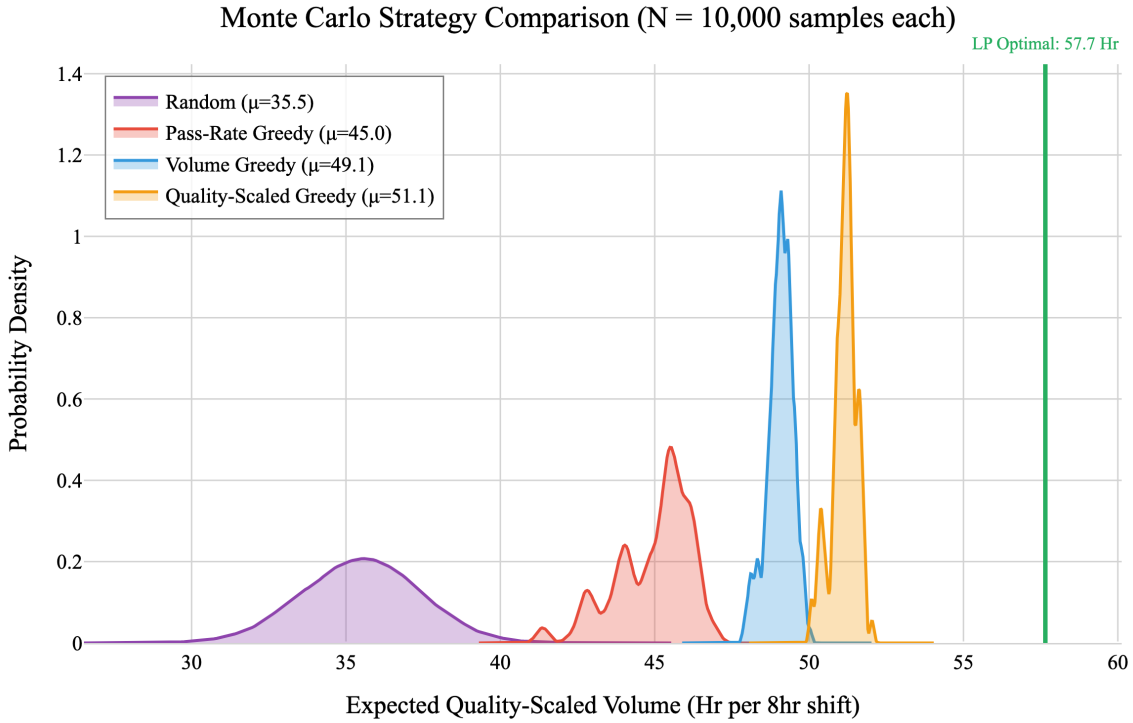

Figure 3: Monte Carlo comparison of assignment strategies. Each curve represents the probability density of total quality-scaled volume for $N = 10{,}000$ simulated shifts. The green vertical line indicates the LP optimal solution with value $\mu_5 = 57.7$ Hr.

Table 3 below shows the percent improvement for each method over the random baseline. Most notably, the LP optimal solution produces 62.5% more quality-scaled volume than random assignment—an absolute gain of 22.2 Hr per shift. For reference, the corresponding raw volume (ignoring pass rate) under LP assignment is 66.1 Hr, yielding an efficiency ratio of $57.7/66.1 = 87.3\%$. This exceeds the average operator pass rate in the input data (77%), indicating that the optimizer preferentially assigns high-quality operators to their strongest tasks, while simultaneously respecting task and site staffing constraints. To visualize how the LP model achieves this improvement, Figure 4 shows the full operator-task assignment matrix.

Method	Algorithm	μ_n (Hr)	Improvement over Random
1	Random	35.5	—
2	Pass-Rate Greedy	45.0	26.8%
3	Volume Greedy	49.1	38.3%
4	Quality-Scaled Greedy	51.1	43.9%
5	LP Optimal	57.7	62.5%

Table 3: Mean quality-scaled volume per 8-hour shift for each assignment method. The LP optimal solution improves upon the best greedy heuristic by 6.6 Hr (12.9%).

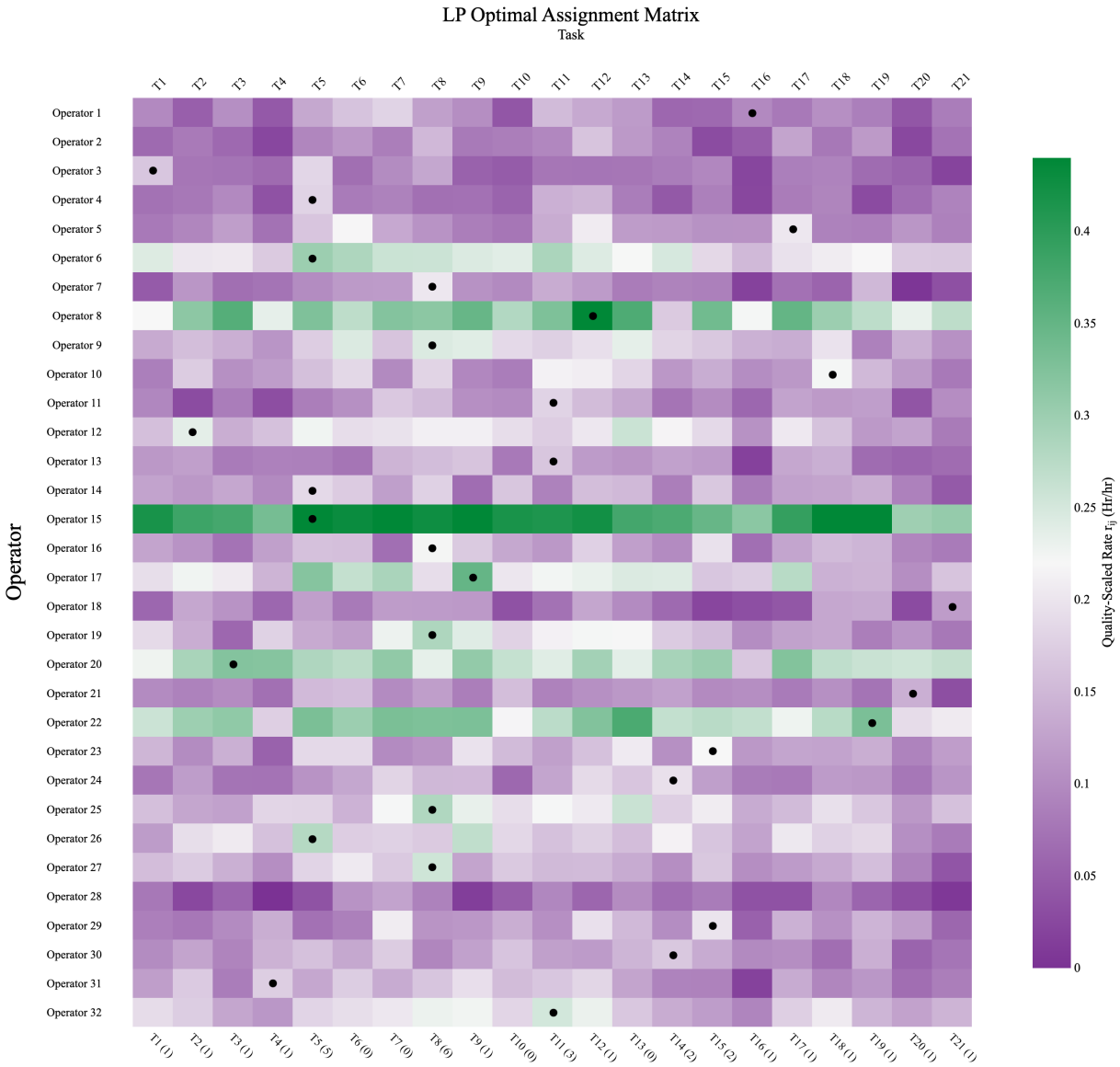

Figure 4: LP optimal assignment matrix. Cell color indicates the quality-scaled rate r_{ij} for each operator-task pair, with black markers denoting the optimal assignments. The optimizer preferentially assigns operators to their highest-rate tasks subject to staffing and capacity constraints. Note that Operators 2 and 28 are unassigned due to unit availability constraints in this instance.

5 Future Extensions

The model presented here assigns each operator to a single task per shift. In practice, operators typically complete 1–4 assigned tasks (approximately 1–4 shift hours of work) before transitioning to discretionary task selection appropriate within research and volume goals. A natural extension generalizes the decision variable from binary to integer-valued:

$$x_{ijk} \in \{0, 1, 2, \ldots, B\} \tag{8}$$

where B represents the maximum number of task blocks per operator per shift, and x_{ijk} now indicates the number of blocks operator i spends on task j at site k. The objective function retains its form, with the interpretation that higher x_{ijk} values contribute proportionally more to total quality-scaled volume. This formulation constitutes a Mixed-Integer Program (MIP), which remains tractable for standard LP solvers such as `PuLP`, which was used to produce the results of Section 4.

A more realistic variant would weight assignments by expected (averaged or target) task duration d_j rather than block count, constraining each operator's total assigned work to a target window:

$$D_{\min} \leq \sum_{j=1}^{T} \sum_{k=1}^{M} d_j \cdot x_{ijk} \leq D_{\max}, \quad \forall\, i \tag{9}$$

For example, $D_{\min} = 60$ and $D_{\max} = 120$ minutes would ensure each operator receives 1–2 hours of optimized assignments before discretionary work begins. Implementation is planned once task duration data becomes available. A further extension supporting mid-shift reassignment—re-optimizing based on updated unit availability or task completion—is left for future work.