Trade Data Harmonization: A Multi-Objective Optimization Approach for Subcategory Alignment and Volume Optimization

Himadri Sikhar Khargharia, Sid Shakya and Dymitr Ruta
EBTIC, Khalifa University, Abu Dhabi, U.A.E.
{*himadri.khargharia, sid.shakya, dymitr.ruta*}*@ku.ac.ae*

Keywords: Trade Data Harmonisation, Non-Dominated Sorting Genetic Algorithm II, Genetic Algorithm, Population-Based Incremental Learning, Distribution Estimation Using MRF and Simulated Annealing.

Abstract: Aligning trade data from disparate sources poses challenges due to volume disparities and category naming variations. This study aims to harmonize subcategories from a secondary dataset with those of a primary dataset, focusing on aligning the number and combined volumes of subcategories. We employ a multi-objective optimization approach using Non-dominated Sorting Genetic Algorithm II (NSGA-II) to facilitate trade-off assessments and decision-making via Pareto fronts. NSGA-II's performance is compared with single-objective optimization techniques, including Genetic Algorithm (GA), Population-based Incremental Learning (PBIL), Distribution Estimation using Markov Random Field (DEUM), and Simulated Annealing (SA). The comparative analysis highlights NSGA-II's efficacy in managing trade data complexities and achieving optimal solutions, demonstrating the effectiveness of meta-heuristic approaches in this context.

1 INTRODUCTION

International trade significantly impacts global economic stability by facilitating the exchange of essential resources such as energy, minerals, metals, and agricultural products (Harding and Harding, 2020). When countries lack key domestic resources, trade shifts from being a strategic option to an economic necessity (Hammoudeh et al., 2009) (Lewrick et al., 2018).

Economists support free trade, highlighting its benefits for growth and welfare (Berg and Lewer, 2015). Accurate and standardized trade data, managed by customs authorities and international bodies, are crucial for quantifying these benefits (Lewrick et al., 2018) (Ferrantino et al., 2012). However, data inconsistencies in volume and category naming across datasets pose challenges, complicating policy analysis and decision-making (Feenstra et al., 1999) (Hansen and Prusa, 1997) (Torres-Espín and Ferguson, 2022) (Khargharia et al., 2023).

In (Khargharia et al., 2023), trade volume disparities were addressed using a subset sum problem framework. Building on this, our research focuses on aligning subcategories across datasets S_1 and S_2 to harmonize traded volumes. Figure 1 illustrates this approach, where rice subcategory volumes from S_1 are matched with those in S_2. This study has two main objectives: (1) aligning subcategory counts and (2) harmonizing combined traded volumes. To achieve this, we employ a multi-objective optimization using the Non-dominated Sorting Genetic Algorithm II (NSGA-II) (Deb et al., 2002), which is widely used for generating Pareto fronts to facilitate trade-off assessments and informed decision-making.

To further assess the effectiveness of NSGA-II, four single-objective optimization techniques are implemented: Genetic Algorithm (GA) (Goldberg, 1989), Population-based Incremental Learning (PBIL) (Baluja, 1994), Distribution Estimation using Markov Random Fields (DEUM) (Shakya and McCall, 2007) (Shakya et al., 2021), and Simulated Annealing (SA) (Kirkpatrick et al., 1983). Scalarization is applied to unify subcategory numbers and volumes into a single optimization criterion. Although direct comparison between single and multi-objective optimization is challenging due to differing cost functions, it is common in EA literature to use the solution closest to the ideal point (refer section 2.2) as a reference for comparison. Additionally, when a Pareto solution outperforms the best single-objective solution across all objectives, the comparison becomes clearer and more meaningful.

For a comprehensive evaluation, we select the so-

Khargharia, H., Shakya, S. and Ruta, D.
Trade Data Harmonization: A Multi-Objective Optimization Approach for Subcategory Alignment and Volume Optimization.
DOI: 10.5220/0013052500003837

In *Proceedings of the 16th International Joint Conference on Computational Intelligence (IJCCI 2024)*, pages 338-345
ISBN: 978-989-758-721-4; ISSN: 2184-3236

lution point in the Pareto front closest to the ideal point (de la Fuente et al., 2018) as one of the reference point for the best solution for NSGA-II. We also scan through the full Pareto set of solutions found by NSGA-II to check if there are any solutions that are better in both objectives in comparison to the best solutions found by single objective algorithms, allowing us to measure the relative performance of NSGA-II against the single-objective optimization techniques.

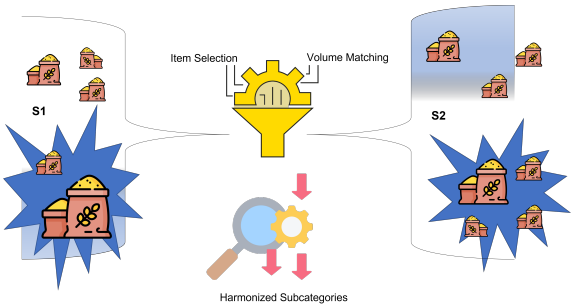


Figure 1: Overview of trade volume harmonization.

The paper is organized as follows: Section 2 discusses the subset sum problem, Pareto fronts, and ideal point calculation. Section 2.5 reviews recent meta-heuristic literature. Section 3 defines the problem, constraints, and objectives with an example. Section 4 covers data preparation. Section 5 presents the methodology and techniques used. Section 6 details the experimental setup and results. Section 7 summarizes findings and suggests future work.

2 BACKGROUND

This section outlines the fundamental concepts for our trade volume harmonization approach, utilizing Pareto optimality, ideal point determination, scalarization, and the subset sum problem for robust alignment of trade data.

2.1 Pareto Optimality

Pareto optimality is essential in multi-objective optimization, where conflicting objectives are minimized. It aligns the number of selected sub-categories and their combined volume with a reference dataset's total import volume. A solution x dominates another solution y if x is less than or equal in all objectives and strictly less in at least one:

$$\text{Pareto Dominance:} \quad \forall i : x_i \leq y_i \text{ and } \exists j : x_j < y_j \quad (x \prec y)$$

$$\text{Pareto Optimality:} \quad x^* \text{ is Pareto optimal} \Leftrightarrow \nexists y \text{ such that } y \prec x^*$$

2.2 Ideal Point Calculation

The ideal point in multi-objective optimization is derived from the final population of solutions, typically the non-dominated set forming the Pareto front. For each objective f_i, z_i^* is determined as the minimum or maximum value across the final population:

$$z_i^* = \min_{x \in \text{Final Population}} f_i(x) \quad \text{or} \quad z_i^* = \max_{x \in \text{Final Population}} f_i(x)$$

The ideal point z^* is then defined as $(z_1^*, z_2^*, \ldots, z_m^*)$. This point, though often theoretical due to conflicting objectives, serves as a reference for evaluating Pareto optimal solutions.

2.3 Scalarization

Scalarization converts multiple objectives into a single scalar function:

$$F(x) = w_1 \cdot f_1(x) + \ldots + w_k \cdot f_k(x).$$

We use this method to balance selected subcategories and their combined volume.

2.4 Subset Sum Problem

The subset sum problem seeks a subset S' of a set S that minimizes:

$$\text{Minimize:} \left| \sum_{v_i \in S'} v_i - T \right|,$$

where T is a target value. It is computationally complex and NP-complete.

2.5 Literature Review

Recent studies (2023-2024) illustrate the versatility of meta-heuristic algorithms in complex optimization. Hosseini et al. (Hosseini et al., 2024) integrated them with deep learning for energy management, while Akter et al. (Akter et al., 2024) applied them to microgrid optimization. Mahmoodi et al. (Mahmoodi et al., 2024) and Abid et al. (Abid et al., 2023) explored their use in financial modeling, improving predictive accuracy. Yahia and Mohammed (Yahia and Mohammed, 2023) optimized UAV path planning.

In network optimization, Priyadarshi (Priyadarshi, 2024) focused on energy-efficient routing in sensor networks, and Ghasemi et al. (Ghasemi et al., 2024) introduced a new engineering optimization method. Khargharia et al. (Khargharia et al., 2023) uniquely applied these techniques to trade data harmonization, balancing trade volumes across subcategories, demonstrating their adaptability in new domains.

3 PROBLEM DESCRIPTION

Khargharia et al. (Khargharia et al., 2023) modeled the alignment of trade volumes between two datasets, S_1 and S_2, for various product categories in a specific country, C_1, as a subset sum problem (see Section 2.4). This section refines this by selecting subcategories from S_2 that match the total trade volume and number of sub-categories in S_1 for each product category. Product categories encompass broader types like rice, edible oils, vegetables etc., while their sub-categories refer to their specific types, such as rice varieties, types of oils like coconut or mustard, or different vegetables.

Let P_{subc} denote the set of sub-categories for a product category Pr_i in S_1:

$$P_{\text{subc}} = \{p_1, p_2, \ldots, p_i\} \tag{1}$$

Similarly, let $\hat{P}_{\text{subc}}$ represent the set of sub-categories for the same product category Pr_i in S_2:

$$\hat{P}_{\text{subc}} = \{\hat{p_1}, \hat{p_2}, \ldots, \hat{p_N}\} \tag{2}$$

3.1 Objective

The goal is to align sub-categories in S_2 with those in S_1 by selecting a subset from S_2 whose combined trade volume approximates that of S_1 while having a similar number of sub-categories. This is expressed mathematically as:

$$\exists \hat{\mathbb{P}}_{\text{subc}} \subset \hat{P}_{\text{subc}} : \sum_{\forall p_i \in P_{\text{subc}}} T_v(p_i) \approx \sum_{\forall \hat{p}_i \in \hat{\mathbb{P}}_{\text{subc}}} T_v(\hat{p}_i) \tag{3}$$

where T_v is the traded volume, $\hat{\mathbb{P}}_{\text{subc}}$ is the selected subset from $\hat{P}_{\text{subc}}$, and $|\hat{\mathbb{P}}_{\text{subc}}| \approx |P_{\text{subc}}|$.

3.2 Constraints

- **Subset Selection:** The solution involves selecting a subset of sub-categories from S_2. This subset must be chosen such that it aligns as closely as possible with both the total trade volume and the number of sub-categories in S_1.

3.3 Example

Consider product rice in S_1 with three sub-categories: Basmati ($T_v(p_1) = 100$), Jasmine ($T_v(p_2) = 200$), and Long-grain ($T_v(p_3) = 300$), giving $P_{\text{subc}} = \{p_1, p_2, p_3\}$ and a total trade volume of 600.

Dataset S_2 has a more detailed breakdown into six sub-categories: White Basmati ($T_v(\hat{p}_1) = 80$), Brown Basmati ($T_v(\hat{p}_2) = 100$), Jasmine ($T_v(\hat{p}_3) = 190$), Organic Long-grain ($T_v(\hat{p}_4) = 310$), Parboiled Long-grain ($T_v(\hat{p}_5) = 270$), and Glutinous rice ($T_v(\hat{p}_6) = 40$), forming $\hat{P}_{\text{subc}} = \{\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4, \hat{p}_5, \hat{p}_6\}$.

The goal is to select $\hat{\mathbb{P}}_{\text{subc}} \subset \hat{P}_{\text{subc}}$ such that the trade volume and number of sub-categories match S_1. For example, $\{\hat{p}_2, \hat{p}_3, \hat{p}_4\}$ gives $100 + 190 + 310 = 600$, matching S_1 in both volume and three sub-categories. Another subset, $\{\hat{p}_2, \hat{p}_3, \hat{p}_5, \hat{p}_6\}$, also sums to 600 but includes four sub-categories, making it non-optimal.

4 DATA PREPARATION AND ANALYSIS

This section describes the preparation and analysis of two datasets, S_1 and S_2, for trade data of country C_1. S_1 is a genuine dataset representing real trade data from reliable sources, while S_2 is a simulated dataset expanding the subcategories in S_1 to increase problem complexity and test robustness.

Table 1: Product categories from S_1 with subcategories and trade volumes.

Category	$\lvert P_{\text{subc}} \rvert$	Trade Vol (KMT)
Pr_1	19	154.103
Pr_2	54	461.450
Pr_3	92	782.301
Pr_4	194	1641.841

Four key product categories are analyzed in S_1: Pr_1, Pr_2, Pr_3, and Pr_4, each with subcategories P_{subc} (see Table 1). To increase complexity, S_2 expands each product's subcategories by approximately ten times, denoted as $\hat{P}_{\text{subc}}$. This ensures controlled scaling for computational efficiency, represented as:

$$N \approx 10 \cdot |P_{\text{subc}}| : \hat{P}_{\text{subc}} = \{\hat{p_1}, \hat{p_2}, \ldots, \hat{p_N}\} \tag{4}$$

where P_{subc} represents the set of subcategories for any product Pr_i from S_1, and $\hat{P}_{\text{subc}}$ denotes the expanded subcategories in S_2.

In S_1, trade volumes $T_v(p_i)$ for subcategories range from 0 to 8 KMT, while $T_v(\hat{p}_i)$ in S_2 ranges from 7 to 10 KMT, increasing complexity and avoiding exact matches. Mathematically, the ranges are:

$$\begin{aligned} p_i \in P_{\text{subc}} : T_v(p_i) &\in \mathbb{R}[0, 8] \\ \hat{p}_i \in \hat{P}_{\text{subc}} : T_v(\hat{p}_i) &\in \mathbb{R}[7, 10] \end{aligned} \tag{5}$$

Table 2 provides details of S_2 with expanded subcategories and corresponding trade volumes.

Table 2: Product categories from S_2 with expanded subcategories and trade volumes.

Category	$\lvert\hat{P}_{\text{subc}}\rvert$	Trade Vol (KMT)
Pr_1	200	762.399
Pr_2	600	2388.052
Pr_3	1000	3955.926
Pr_4	2000	8008.241

5 METHODOLOGY

This section outlines the methodology for modeling the problem from Section 3 as an optimization task, including fitness evaluations for meta-heuristic techniques, solution design, and the specific methods used.

5.1 Solution Representation

To align trade data between datasets S_1 and S_2 (Section 3), we utilize four single-objective binary metaheuristic techniques (Khargharia et al., 2023) and one multi-objective binary metaheuristic technique (Deb et al., 2002). Let $X = \{x_1, ..., x_n\}$, where $n = N = |\hat{P_{\text{subc}}}|$ (from equations 4 and 2). The binary string X, representing solutions, varies in length (SOL) as 200, 600, 1000, or 2000 (Table 2). Each $x_i \in X$ can be 0 or 1.

In a practical example, if dataset S_2 includes eight subcategories (P1, P2, P3, P4, P5, P6, P7, P8) and we aim to select matching subcategories from S_1 such as P2, P3, and P8, the binary solution from any metaheuristic method should resemble the example in Figure 2.

P1	P2	P3	P4	P5	P6	P7	P8
0	1	1	0	0	0	0	1

Figure 2: Illustration of Binary Representation of Solution

5.2 Fitness Evaluation

Fitness evaluation acts as a measure to assess the effectiveness of identified subcategories in aligning trade data, aiming to uncover meaningful patterns for specific commodities.

Let $m = |P_{\text{subc}}|$, representing the number of subcategories associated with any product category Pr_i from dataset S_1 (as defined in equation 1). As mentioned before, T_v is considered as the traded volume. We mathematically model the problem discussed in Section 3 as involving two functions:

$f_1(X)$: Align the number of selected sub-categories.

$f_2(X)$: Align the combined volume of selected sub-categories.

such that:

$$\begin{aligned} f_1(X) &= \left(|m - \sum_{i=1}^{n} x_i| \right) \\ f_2(X) &= \left(|\sum_{i=1}^{m} T_v(p_i) - \sum_{i=1}^{n} x_i \cdot T_v(\hat{p_i})| \right) \end{aligned} \tag{6}$$

where the terms carry their usual meanings as defined in Equation 3 and in Section 5.1.

5.2.1 Using Multi Objective Optimization Techniques

While using a multi-objective optimization technique, $f_1(X)$ and $f_2(X)$ represent the objectives that need to be optimized simultaneously as mathematically represented below.

$$\min_{X=\{x_1,\ldots,x_n\}} f_1(X) \,, \min_{X=\{x_1,\ldots,x_n\}} f_2(X) \tag{7}$$

$f_1(X)$ and $f_2(X)$ are used to create a Pareto front as follows:

Pareto Optimality. The goal is to find solutions X that are not dominated by any other feasible solution in terms of both objectives $f_1(X)$ and $f_2(X)$ (refer section 2.1). A solution X^* is Pareto optimal if there does not exist another feasible solution X' such that:

$$f_1(X') \leq f_1(X^*) \quad \text{and} \quad f_2(X') \leq f_2(X^*) \tag{8}$$

with at least one strict inequality.

Pareto Front. The Pareto front consists of all non-dominated solutions. It represents the trade-offs between $f_1(X)$ and $f_2(X)$ where improving one objective comes at the expense of the other. Points on the Pareto front cannot be improved in one objective without worsening the other.

5.2.2 Using Single Objective Optimization Techniques

To consolidate multiple objectives $f_1(X)$ and $f_2(X)$ into a single-objective form, scalarization is applied using min-max normalization (see section 2.3). Given a set $V = \{v_1, v_2, \ldots, v_n\}$, the normalization function is:

$$\text{Norm}(v_i) = \frac{v_i - \min(V)}{\max(V) - \min(V)} \tag{9}$$

The scalarized fitness function is defined as:

$$\min_X f(X) = w_1 \cdot \text{Norm}\left(f_1(X)\right) + w_2 \cdot \text{Norm}\left(f_2(X)\right) \tag{10}$$

where equal weights $w_1 = w_2 = 1$ are used. The normalization ranges for $f_1(X)$ are 0 and $|m-n|$, and for $f_2(X)$, 0 and $|\sum_{i=1}^{m} T_v(p_i) - \sum_{i=1}^{n} T_v(\hat{p_i})|$.

5.3 Considered Meta-Heuristic Techniques

To address the Trade Data Harmonization problem (Section 3), both single and multi-objective optimization techniques are considered, with NSGA-II (Deb et al., 2002) used for multi-objective optimization.

5.3.1 Non-dominated Sorting Genetic Algorithm II (NSGA-II)

A robust evolutionary algorithm for multi-objective optimization problems, NSGA-II extends Genetic Algorithms by efficiently handling conflicting objectives using two core mechanisms: non-dominated sorting and crowding distance.

- **Non-Dominated Sorting.** Solutions are ranked into non-dominated fronts based on their dominance relationships. A solution X dominates Y if:
$$\forall i,\ f_i(X) \leq f_i(Y), \quad \exists j,\ f_j(X) < f_j(Y)$$
where f_i denotes the objective value for i-th objective. NSGA-II assigns ranks $F_1, F_2, \ldots$ to solutions, with lower ranks indicating better solutions.
- **Crowding Distance.** Maintains diversity in the population. For each front F_n, the distance D_i for each solution i is:
$$D_i = \sum_{j=1}^{m} \frac{f_j(\text{next}_i) - f_j(\text{prev}_i)}{\text{Range}_j}$$
where $f_j(\text{next}_i)$ and $f_j(\text{prev}_i)$ are neighboring objective values, and Range_j is the range of j-th objective in F_n.

NSGA-II evolves populations across generations using selection, crossover, and mutation, balancing convergence and diversity to approach Pareto-optimal solutions. For further details, see (AlShanqiti et al., 2019).

The single objective techniques from (Khargharia et al., 2023) include:

5.3.2 Genetic Algorithm (GA)

GA mimics natural selection on a population of solutions. Using crossover ($cOper$) and mutation ($mOper$) operators with probabilities (cp, mp), it evolves the population to preserve elites (e) and balance exploration (Goldberg, 1989).

5.3.3 Population-Based Incremental Learning (PBIL)

PBIL updates a probability vector based on elite solutions (e) using a learning rate (λ) and selection size (ss), guiding the search in promising regions (Baluja, 1994).

5.3.4 Distribution Estimation Using Markov Random Field (DEUM)

DEUM employs a Markov Random Field model to estimate distributions, adjusting a temperature coefficient (β) for exploration-exploitation balance (Shakya and McCall, 2007).

5.3.5 Simulated Annealing (SA)

SA uses a cooling schedule to control the temperature (τ), shifting from exploration to exploitation as τ decreases (Kirkpatrick et al., 1983).

6 EXPERIMENT SETUP AND ANALYSIS OF RESULTS

Experiments were conducted on a workstation with an `11th Gen Intel Core i7-11800H @ 2.30GHz` processor and `32 GB` RAM, using datasets S_1 and S_2. S_1 has product categories Pr_1 to Pr_4 with subcategory counts of 19, 54, 92, and 194, while S_2 includes sizes 200, 600, 1000, and 2000. Refer to Section 4 for dataset details and Section 3 for experiment descriptions. Each experiment was repeated 15 times with different solution sizes and parameter settings (see Section 6.1).

6.1 Parameter Selection

Optimal parameters were selected empirically through initial trials. Population sizes for GA, PBIL, DEUM, and NSGA-II were set to half the solution length (SOL/2) for SOL values of 200, 600, 1000, and 2000. Maximum generations were set to 10 times the population size, except for SA, where it was set to 10 times $(SOL/2)^2$, due to its single-solution approach.

GA used uniform crossover with 0.79 probability, tournament selection, and 1-bit mutation with 0.034 probability. PBIL's selection size was 0.47 and learning rate 0.16. DEUM had a selection size of 0.06 and a temperature coefficient of 0.86. SA was set with a temperature of 0.023. NSGA-II used 1-point crossover (0.7 probability) and 1-bit mutation (0.0121 probability), with tournament selection.

6.2 Experimental Analysis

In this section, the experimental results for the problem detailed in Section 3 are presented and summarized in Table 3, covering 15 runs for each algorithm-solution size combination. For the single-objective

Table 3: Results of difference in selected Items and volume.

SOL	**Algo**					
2000	GA	PBIL	DEUM	SA	IP (NSGA-II)	B (NSGA-II)
α	55	27	389	216	24	25
β	0.000	0.000	119.871	0.001	0.007	0.000
θ	55.33 ± 0.577	27.0 ± 0.000	394.5 ± 7.778	229.67 ± 12.34	24.11 ± 0.87	25.428 ± 0.494
γ	0.000 ± 0.000	0.000 ± 0.000	123.658 ± 49.197	0.083 ± 0.128	0.026 ± 2.23	0.000 ± 0.000
1000	GA	PBIL	DEUM	SA	IP (NSGA-II)	B (NSGA-II)
α	23	13	167	102	9	9
β	0.002	0.000	74.654	0.337	0.000	0.000
θ	23.0 ± 0.0	13.333 ± 0.577	184.0 ± 15.133	109.667 ± 6.658	9.476±0.67	10.457 ± 1.17
γ	0.086 ± 0.144	0.000 ± 0.000	121.076 ± 51.143	0.425 ± 0.184	1.30±2.43	0.003 ± 0.045
600	GA	PBIL	DEUM	SA	IP (NSGA-II)	B (NSGA-II)
α	50	22	71	62	5	7
β	0.341	0.013	0.496	0.191	7.632	0.000
θ	54.867 ± 2.503	26.0 ± 2.673	73.333 ± 2.082	67.000 ± 4.359	4.78±0.63	7.928±0.593
γ	0.095 ± 0.141	0.034 ± 0.038	0.762 ± 0.258	0.277±0.100	9.23±4.50	0.001±0.003
200	GA	PBIL	DEUM	SA	IP (NSGA-II)	B (NSGA-II)
α	22	9	11	19	2	3
β	0.013	0.022	0.024	0.259	0.888	0.000
θ	23.333 ± 1.155	11.6 ± 1.242	13.0 ± 2.0	24.133 ± 2.386	2.11±0.57	3.268±0.44
γ	0.110 ± 0.089	0.037 ± 0.059	0.035 ± 0.038	0.325 ± 0.116	1.53±2.35	0.004±0.034

optimization algorithms (GA, PBIL, DEUM, and SA), discrepancies in selected sub-categories (*Item Diff*, α) and differences in traded volumes (*VOL Diff*, β) are computed using equation (10) based on the Best Fitness run.

For the multi-objective NSGA-II, two solutions from the Pareto front are reported: *IP (NSGA-II)*, which is the closest point to the ideal (see Section 2.2), and *B (NSGA-II)*, which has superior values for both objectives. The Euclidean distance, normalized to [0,1], ensures equal weight for sub-category differences (*Item Diff*) and volume differences (*VOL Diff*).

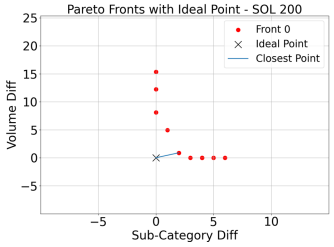
Table 3 also provides the average discrepancies (θ, *Item Diff (Avg ± SD)*) and volume differences (γ, *VOL Diff (Avg ± SD)*) along with standard deviations across 15 runs. For SOLs 200 and 600, all single-objective algorithms achieve near-zero *VOL Diff* with similar sub-categories. For SOLs 1000 and 2000, PBIL achieves the lowest *VOL Diff* with the fewest selected sub-categories. SA and DEUM select more sub-categories for SOLs 1000 and 2000, with SA effectively reducing *VOL Diff*, while DEUM faces challenges in achieving optimal results.

NSGA-II competes effectively with single-objective algorithms, as illustrated in Figures 3 and 4 for solution sizes 200, 600, 1000, and 2000 (based on run 6 out of 15). These figures demonstrate the trade-offs between *Item Diff* and *VOL Diff*, providing insights for decision-makers. Ideal points, plotted as reference points, highlight non-dominated solutions in Front 0, with the closest solutions marked based on Euclidean distance.

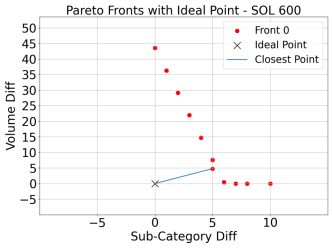
In Table 3, the *B (NSGA-II)* solution consistently outperforms single-objective algorithms across all solution sizes (200, 600, 1000, and 2000), showing superior performance across both objectives compared to the best results from single-objective techniques. The *IP (NSGA-II)* solution prioritizes proximity to the ideal point, yielding results that are generally better or comparable to single-objective approaches. However, for solution size 600, the *IP (NSGA-II)* solution shows a higher *VOL Diff* compared to the best single-objective solution.

6.3 Analysis of Solution Quality Variability

Further analysis of Table 3 is illustrated in Figure 5, showing the distribution of average *Item Diff* (θ) and *VOL Diff* (γ) with their standard deviations across different solution sizes and algorithms. DEUM's outlier performance is excluded to ensure a clear comparison.

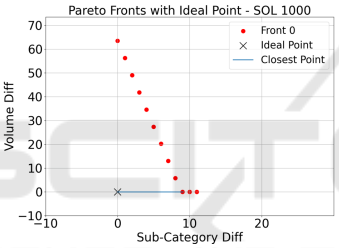


(a) SOL 200

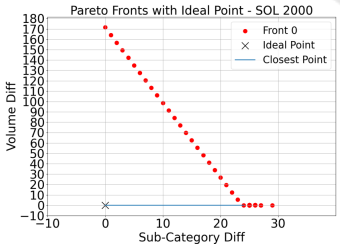


(b) SOL 600

Figure 3: Pareto Front Representation by NSGA-II for SOL.



(a) SOL 1000



(b) SOL 2000

Figure 4: Pareto Front Representation by NSGA-II for SOL.

Results from *B (NSGA-II)* indicate that NSGA-II consistently achieves superior outcomes with negligible variation for both *Item Diff* and *VOL Diff* across all solution sizes and algorithms. Among the single-objective algorithms, PBIL provides the best results with minimal variability, while DEUM is less likely to yield optimal results.

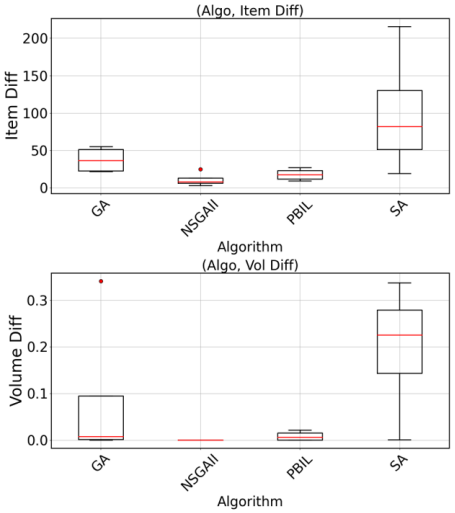


Figure 5: Spread of average Item Diff and Volume Diff across all Solution Size.

7 CONCLUSION

This paper tackled a trade data harmonization problem with multiple optimization objectives, evaluating NSGA-II's performance against single-objective techniques. A Pareto front was generated to help decision-makers balance trade-offs between sub-category numbers and combined volumes. Scalarization and normalization converted multiple objectives into a single scalar form for fair comparison with single-objective algorithms. Results showed that NSGA-II consistently outperformed single-objective methods, finding better solutions for both objectives. Among the single-objective techniques, PBIL often performed best, while DEUM had the lowest performance.

In conclusion, this study demonstrates the effectiveness of multi-objective techniques, particularly NSGA-II, in trade data harmonization. These methods handle multiple objectives directly, avoiding the need for normalization or scalarization, and produce a Pareto set of solutions, giving users more flexibility in selecting the optimal solution. Future work could refine these methods, explore other multi-objective algorithms, and apply them to real-world case studies to address trade data harmonization challenges further.

REFERENCES

Abid, M., El Kafhali, S., Amzil, A., and Hanini, M. (2023). An efficient meta-heuristic methods for travelling salesman problem. In *The International Confer-*

ence on Advanced Intelligent Systems and Informatics. Springer.

Akter, A., Zafir, E., Dana, N., Joysoyal, R., and Sarker, S. (2024). A review on microgrid optimization with meta-heuristic techniques: Scopes, trends and recommendation. *Energy Strategy Reviews*.

AlShanqiti, K., Poon, K., Shakya, S., Sleptchenko, A., and Ouali, A. (2019). A multi-objective design of in-building distributed antenna system using evolutionary algorithms. In *Artificial Intelligence XXXVI: 39th SGAI International Conference on Artificial Intelligence, AI 2019, Cambridge, UK, December 17–19, 2019, Proceedings 39*, pages 253–266. Springer.

Baluja, S. (1994). Population-based incremental learning. a method for integrating genetic search based function optimization and competitive learning. Technical report, Carnegie-Mellon Univ Pittsburgh Pa Dept Of Computer Science.

Berg, H. V. d. and Lewer, J. J. (2015). International trade and economic growth.

de la Fuente, D., Vega-Rodríguez, M. A., and Pérez, C. J. (2018). Automatic selection of a single solution from the pareto front to identify key players in social networks. *Knowledge-Based Systems*, 160:228–236.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197.

Feenstra, R. C., Hai, W., Woo, W. T., and Yao, S. (1999). Discrepancies in international data: an application to china–hong kong entrepôt trade. *American Economic Review*, 89(2):338–343.

Ferrantino, M. J., Liu, X., and Wang, Z. (2012). Evasion behaviors of exporters and importers: Evidence from the us–china trade data discrepancy. *Journal of international Economics*, 86(1):141–157.

Ghasemi, M., Golalipour, K., Zare, M., and Mirjalili, S. (2024). Flood algorithm (fla): an efficient inspired meta-heuristic for engineering optimization. *The Journal of Supercomputing*.

Goldberg, D. (1989). Genetic algorithms in search. *Optimization, and Machine Learning, Addison Wesley*.

Hammoudeh, S., Sari, R., and Ewing, B. T. (2009). Relationships among strategic commodities and with financial variables: A new look. *Contemporary Economic Policy*, 27(2):251–264.

Hansen, W. L. and Prusa, T. J. (1997). The economics and politics of trade policy: An empirical analysis of itc decision making. *Review of Int. Economics*, 5(2):230–245.

Harding, R. and Harding, J. (2020). *Strategic Trade as a Means to Global Influence*, chapter 6, pages 143–172. John Wiley & Sons, Ltd.

Hosseini, E., Al-Ghaili, A., and Kadir, D. (2024). Metaheuristics and deep learning for energy applications: Review and open research challenges (2018–2023). *Energy Strategy Reviews*.

Khargharia, H., Shakya, S., and Ruta, D. (2023). Comparative analysis of metaheuristics techniques for trade data harmonization. In *Proceedings of the 15th International Joint Conference on Computational Intelligence - Volume 1: ECTA*, pages 206–213. INSTICC, SciTePress.

Kirkpatrick, S., Gelatt Jr, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *science*, 220(4598):671–680.

Lewrick, U., Mohler, L., and Weder, R. (2018). Productivity growth from an international trade perspective. *Review of International Economics*, 26(2):339–356.

Mahmoodi, A., Hashemi, L., and Mahmoodi, A. (2024). Novel comparative methodology of hybrid support vector machine with meta-heuristic algorithms to develop an integrated candlestick technical analysis model. *Journal of Capital Markets Studies*.

Priyadarshi, R. (2024). Energy-efficient routing in wireless sensor networks: a meta-heuristic and artificial intelligence-based approach: a comprehensive review. *Archives of Computational Methods in Engineering*.

Shakya, S. and McCall, J. (2007). Optimization by estimation of distribution with deum framework based on markov random fields. *International Journal of Automation and Computing*, 4:262–272.

Shakya, S., Poon, K., AlShanqiti, K., Ouali, A., and Sleptchenko, A. (2021). Investigating binary eas for passive in-building distributed antenna systems. In *2021 IEEE Congress on Evolutionary Computation (CEC)*, pages 2101–2108. IEEE.

Torres-Espín, A. and Ferguson, A. R. (2022). Harmonization-information trade-offs for sharing individual participant data in biomedicine.

Yahia, H. and Mohammed, A. (2023). Path planning optimization in unmanned aerial vehicles using meta-heuristic algorithms: A systematic review. *Environmental Monitoring and Assessment*.