A multi-scale analysis of learning dynamics in data-driven MCDA: Evidence from the INCOME method

Bartłomiej Kizielewicz[1,2][0000−0001−5736−4014]

1 National Institute of Telecommunications, Szachowa 1, 04-894 Warsaw, Poland
b.kizielewicz@il-pib.pl
2 Research Team on Intelligent Decision Support Systems, Department of Artificial Intelligence and Applied Mathematics, Faculty of Computer Science and Information Technology, West Pomeranian University of Technology in Szczecin, ul. Żołnierska 49, 71-210 Szczecin, Poland

Abstract. The Intelligent Characteristic Objects Method (INCOME) replaces the expert-driven evaluation stage of the COMET method with a data-driven artificial expert based on k-Nearest Neighbours regression. Although previous studies demonstrate high agreement between INCOME and expert-based COMET, the data requirements and small-sample behaviour of the method remain systematically unexplored. This paper presents an empirical learning curve analysis of INCOME, examining how training set size affects ranking accuracy and stability as measured by the WS rank similarity coefficient. The results reveal a concave learning trajectory with rapid gains in the small-sample regime, followed by a practical saturation region beyond which additional data collection yields negligible improvement relative to its cost. A simple logarithmic approximation, adopted as a descriptive model motivated by classical learning curve literature, captures the overall trend and identifies an early diminishing-return point marking the transition from steep initial improvement to progressively decelerating gains. Furthermore, a hyperparameter sensitivity analysis demonstrates that non-monotonic variance patterns observed in the small-sample regime are attributable to suboptimal configuration rather than structural limitations of the method, and can be resolved through stability-constrained hyperparameter selection. These findings provide preliminary empirical evidence on INCOME's data requirements for the studied problem class and offer initial guidance for data collection planning and deployment readiness assessment in data-driven multi-criteria decision analysis.

Keywords: Multi-criteria decision analysis · Artificial expert · Learning curve analysis · Hyperparameter sensitivity · Ranking stability

1 Introduction

Expert judgment remains the basis of multi-criteria decision analysis (MCDA), but its scalability limitations are becoming increasingly apparent [1]. Methods

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_28

such as Analytic Hierarchy Process (AHP) require decision-makers to perform pairwise comparisons between alternatives and criteria, a process that grows rapidly intractable as problem dimensionality increases [9]. These comparisons are conducted manually by domain experts, making the process both time-intensive and costly. One method that inherently relies on such pairwise evaluations is the Characteristic Objects METhod (COMET) [5]. The method constructs $t = c^n$ characteristic objects from the Cartesian product of c characteristic values across n criteria. Evaluating these objects requires an expert to perform $p = \frac{c^n(c^n-1)}{2}$ pairwise comparisons to construct the Matrix of Expert Judgment (MEJ), yielding a complexity of $\mathcal{O}(c^{2n})$ with respect to the number of criteria. Such exponential growth in the number of required comparisons renders exhaustive expert evaluation impractical as problem complexity increases, severely constraining COMET's real-world applicability despite its theoretical advantages.

To solve this problem, recent studies have proposed replacing human experts with artificial experts based on machine learning, trained on historical data on decisions made. The Intelligent Characteristic Objects MEthod (INCOME) constructs such an artificial expert using k-Nearest Neighbors regression, approximating expert preference functions directly from historical decision records [4]. Instead of engaging a domain expert to perform pairwise comparisons, this method predicts characteristic object evaluations from the k most similar historical cases, weighted by their distance in the criteria space. Initial validation demonstrates that INCOME achieves high rank correlation with expert-based COMET, reducing the expert evaluation burden from $\mathcal{O}(c^{2n})$ pairwise comparisons to a one-time computational training procedure.

While INCOME eliminates the expert evaluation bottleneck, it introduces a data dependency whose implications remain systematically unexplored in the literature. Existing studies show that INCOME produces rankings consistent with the expert-based COMET, but none of them provide a systematic characterization of the minimum training set size required for reliable implementation, the marginal utility of additional samples, or the stability of predictions in independent replications. This absence of empirical guidance creates a symmetric risk for practitioners: insufficient training data yields unstable and unreliable recommendations, while excessive data collection imposes unnecessary resource costs without commensurate improvement in model quality. Absent such guidance, deployment decisions are made without empirical grounding. This undermines reproducibility and, more broadly, the case for data-driven MCDA as a credible alternative to expert elicitation.

The study of how model performance scales with training set size has a well-established tradition in machine learning, typically framed as *learning curve analysis* [6, 2, 11]. Classical results characterise learning curves in terms of power-law or logarithmic decay of generalization error, and relate their shape to the bias-variance tradeoff of the underlying estimator. However, this literature focuses almost exclusively on predictive accuracy in supervised learning settings and does not address the specific requirements of MCDA, where the output of

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_28

interest is a ranking rather than a point prediction, and where top-rank fidelity matters disproportionately. The present study adapts the learning curve framework to the MCDA context by using the WS rank similarity coefficient as the evaluation metric, thereby extending the classical analysis to a setting where performance is defined by ordinal agreement rather than numerical error.

This guidance gap becomes particularly pronounced in small-sample regimes, where the relationship between training set size and model behaviour exhibits unexpected characteristics. Specifically, empirical observations indicate a non-monotonic relationship between training set size and prediction stability, where incremental increases in training data can paradoxically yield higher variance in ranking outcomes, contradicting the standard statistical assumption that larger samples produce more stable estimates. Whether this instability reflects a fundamental limitation of the method or an artifact of suboptimal hyperparameter configuration, specifically the choice of k in the underlying k-Nearest Neighbors model, remains an open question with direct practical consequences. If such anomalies are attributable to misconfiguration rather than inherent method constraints, they can be mitigated through adaptive hyperparameter selection, potentially reducing data requirements without sacrificing ranking accuracy. Resolving this distinction is therefore essential both for establishing principled deployment criteria and for avoiding premature rejection of a viable methodology.

To address these gaps, this paper presents a systematic empirical investigation of INCOME's data requirements and small-sample behaviour, structured around three interconnected research questions:

(RQ1): How does training set size affect the accuracy and stability of INCOME rankings relative to a reference ranking, and at what point do marginal gains become negligible in practice?
(RQ2): How strong is the relationship between training set size and ranking accuracy in INCOME, and does the learning curve exhibit a saturation effect?
(RQ3): How does the choice of hyperparameter k affect the accuracy and stability of INCOME in small-sample regimes, and can adaptive k selection improve the consistency of results?

Together, these contributions establish an empirical framework for assessing data collection requirements in data-driven MCDA, characterising how training set size shapes ranking accuracy and stability across the full sample spectrum (RQ1), quantifying the strength and saturation behaviour of the learning curve to inform cost-effective data collection strategies (RQ2), and determining whether apparent small-sample instabilities reflect fundamental constraints of the method or correctable hyperparameter configuration choices (RQ3).

The remainder of this paper is organised as follows. Section 2 introduces the INCOME method and the WS ranking similarity coefficient used throughout the study. Section 3 describes the dataset, experimental protocol, evaluation metrics, and hyperparameter sensitivity design. Section 4 presents the empirical results structured around the three research questions. Section 5 interprets the findings,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_28

discusses practical implications, and identifies limitations. Section 6 summarises the contributions and outlines future research directions.

2 Methodology

2.1 The INCOME Method

The Characteristic Objects Method (COMET) [3] constructs a continuous preference model through five steps: defining characteristic values for each criterion, generating characteristic objects (COs) from their Cartesian product, obtaining a preference ordering of COs through exhaustive pairwise comparison by a domain expert, building a fuzzy rule base, and interpolating preferences for arbitrary alternatives via Mamdani fuzzy inference. A key structural property of COMET is its complete resistance to the rank reversal paradox, as each alternative is evaluated independently against a fixed rule base [5]. However, the method requires an expert to perform $p = t(t-1)/2$ pairwise comparisons over $t = \prod_{i=1}^{n} c_i$ characteristic objects, where c_i is the number of characteristic values for criterion i, yielding an overall complexity of $O(c^{2n})$ with respect to the number of criteria. For the configuration used in this study ($n = 4$, $c_i \in \{5, 5, 6, 5\}$, $t = 750$), this amounts to 281,625 pairwise comparisons, a volume that renders manual expert evaluation impractical.

The INtelligent Characteristic Objects METhod (INCOME) [4] addresses this bottleneck by replacing the human expert with a k-Nearest Neighbours (kNN) regression model trained on historically evaluated decision records. Given a training set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, where $\mathbf{x}_i \in \mathbb{R}^n$ is a vector of criterion values and $y_i \in \mathbb{R}$ is the associated decision outcome, the kNN model predicts the value of any query point $\mathbf{x}'$ as the distance-weighted average of its k nearest neighbours in the criteria space:

$$\hat{y}(\mathbf{x}') = \frac{1}{k} \sum_{(\mathbf{x}_i, y_i) \in D_{\mathbf{x}'}} w_i \cdot y_i, \tag{1}$$

where $D_{\mathbf{x}'}$ denotes the set of k training instances closest to $\mathbf{x}'$ under the Chebyshev distance, and $w_i = 1 - d_i / \sum_{j=1}^{k} d_j$ is the normalized inverse-distance weight for the i-th neighbour at distance d_i. The Chebyshev distance was adopted from the original INCOME formulation [4], where it was established as the default metric for the kNN artificial expert.

To construct the Matrix of Expert Judgment (MEJ) in INCOME, each pair of characteristic objects CO_i and CO_j is evaluated through the trained kNN model. The comparison outcome is determined as:

$$\alpha_{ij} = \begin{cases} 0.0, & \text{if } \hat{y}(CO_i) < \hat{y}(CO_j), \\ 0.5, & \text{if } \hat{y}(CO_i) = \hat{y}(CO_j), \\ 1.0, & \text{if } \hat{y}(CO_i) > \hat{y}(CO_j). \end{cases} \tag{2}$$

The summed judgments $SJ_i = \sum_{j=1}^{t} \alpha_{ij}$ yield preference values for each CO, and the resulting rule base supports Mamdani inference for ranking arbitrary alternatives, exactly as in the standard COMET procedure. The critical advantage

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_28

is that the $O(c^{2n})$ expert comparisons are replaced by a one-time computational procedure whose quality depends on the size and representativeness of the training set D rather than on expert availability. This dependency, however, raises the question of how much training data is necessary to produce reliable rankings, which constitutes the central focus of this paper.

2.2 Ranking Similarity Evaluation

Ranking accuracy is measured using the WS rank similarity coefficient [7], selected for its top-rank sensitivity: discrepancies at the top of a ranking receive exponentially higher penalties than those at the bottom, which aligns with the practical importance of correctly identifying the best-performing alternatives in MCDA applications. For two rankings $\mathbf{x}$ and $\mathbf{y}$ of N alternatives, the WS coefficient is defined as:

$$WS(\mathbf{x}, \mathbf{y}) = 1 - \sum_{i=1}^{N} \left(2^{-x_i} \cdot \frac{|x_i - y_i|}{\max(|x_i - 1|,\ |x_i - N|)} \right), \tag{3}$$

where x_i and y_i denote the ranks of the i-th alternative in the reference and predicted rankings, respectively. The coefficient is asymmetric by design: the first argument serves as the reference ranking, with the exponential weighting scheme ensuring that top-rank errors carry disproportionate penalties. A detailed formal analysis of this asymmetry and its decision-making motivation is provided in [7, 8]. WS takes values in $(0, 1]$, where $WS = 1$ indicates identical rankings.

3 Experimental Setup

3.1 Dataset and Decision Problem

The empirical investigation uses the *Combined Cycle Power Plant* dataset [10], publicly available from the UCI Machine Learning Repository. The dataset comprises N = 9,568 operational records of a combined cycle power plant, each describing a full-load state of the plant. It is continuous, real-valued, and moderately sized, with naturally occurring measurement noise, making it suitable for analysing ranking stability under subsampling.

Four physical measurements serve as decision criteria: ambient temperature C_1 [°C], relative humidity C_2 [%], ambient pressure C_3 [mbar], and exhaust vacuum C_4 [cm Hg], with their characteristic values listed in Table 1. The target variable is net hourly electrical output P [MW]. Higher output corresponds to a more favourable plant state, so the reference ranking is constructed in descending order of P, with rank 1 assigned to the highest-output alternative. This ranking serves as the ground truth against which all INCOME predictions are evaluated. The characteristic values yield t = 750 Characteristic Objects in the COMET structure.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_28

Table 1. Decision criteria and characteristic values used in the COMET structure.

C_i **Name**	**Unit**	**Characteristic values**
C_1 Ambient temperature	[°C]	$\{1, 14, 18, 25, 40\}$
C_2 Relative humidity	[%]	$\{25, 40, 54, 68, 82\}$
C_3 Ambient pressure	[mbar]	$\{992, 1003, 1011, 1019, 1026, 1035\}$
C_4 Exhaust vacuum	[cm Hg]	$\{25, 60, 80, 95, 101\}$

3.2 Experimental Protocol

The dataset is partitioned once into a fixed training pool and a fixed test set using an 80/20 random split (random seed 42), yielding 7,654 training candidates and 1,914 test instances. A fixed test set is used to isolate the effect of training subset variability from data partition variability, as repeated splits would conflate two distinct sources of instability. No feature normalization was applied: as all criteria represent physically meaningful measurements with comparable operational relevance, no artificial rescaling was introduced, and the raw criterion ranges reflect the natural operating envelope of the power plant, preserving the physical interpretation of neighbourhood structure under the Chebyshev distance metric.

For each training set size $m \in \mathcal{M} = \{50, 100, 200, 500, 1000, 2000, 5000\}$, we perform $n = 30$ independent replications. In each replication r, a random subset of m instances is drawn without replacement from the training pool. An INCOME model is then fitted on this subset using the kNN artificial expert with Chebyshev distance and $k = 15$ neighbors, following the configuration established in the original INCOME study [4]. The fitted model is evaluated on the fixed test set, producing a predicted ranking $\mathbf{r}_r(m)$. The value $n = 30$ was selected as a standard compromise between statistical reliability and computational feasibility.

3.3 Evaluation Metrics

Formally, for n replications (with $n = 30$ in this study) at training set size m, the mean accuracy is:

$$\overline{WS}(m) = \frac{1}{n} \sum_{r=1}^{n} WS_r(m) \tag{4}$$

and the inter-replication standard deviation is:

$$\sigma(m) = \sqrt{\frac{1}{n-1} \sum_{r=1}^{n} \Big(WS_r(m) - \overline{WS}(m) \Big)^2} \tag{5}$$

The standard deviation $\sigma(m)$ specifically measures sensitivity to training subset selection: a large value indicates that model quality varies substantially depending on which m instances are drawn, constituting a practical reliability risk independent of mean accuracy.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_28

Internal consistency is measured as the mean pairwise WS coefficient across all $\binom{n}{2} = 435$ replication pairs:

$$WS_{\text{internal}}(m) = \frac{2}{n(n-1)} \sum_{r<s} WS\big(\mathbf{r}_r(m),\, \mathbf{r}_s(m)\big) \tag{6}$$

This metric captures mutual agreement among independent model instances trained on different subsets of the same size, independently of the reference ranking.

3.4 Hyperparameter Sensitivity Protocol

The main study addresses RQ1 and RQ2 under fixed $k = 15$. Preliminary analysis revealed a non-monotonic variance pattern between $m = 50$ and $m = 100$, contradicting the standard expectation that larger samples yield more stable estimates. RQ3 tests whether this instability is a structural limitation of INCOME or an artefact of suboptimal hyperparameter configuration.

To this end, a grid search over $k \in \{3, 5, 10, 15, 20, 30, 40\}$ is conducted for the small-sample regime $m \in \{50, 100, 200\}$, as hyperparameter sensitivity is theoretically most pronounced where training data are scarce: with fixed k, the ratio k/m varies dramatically across sample sizes, creating fundamentally different local neighbourhood structures. For larger sample sizes ($m \geq 500$), the marginal influence of k diminishes as the neighbourhood becomes increasingly representative of the underlying decision function.

Each of the $7 \times 3 = 21$ configurations is evaluated over $n = 30$ independent replications. The optimal k for each m is defined as the value minimising $\sigma(m, k)$, subject to the constraint that $\overline{WS}(m, k)$ does not decrease below the baseline mean WS established at $k = 15$ under identical replication settings. This constraint prevents trivial stability gains obtained by choosing overly smooth neighbourhoods that reduce variance at the expense of ranking accuracy.

4 Experiments and Results

To demonstrate how training set size shapes the behaviour of INCOME, this section presents the empirical results structured around the three research questions. We begin by examining the learning curve and identifying the practical saturation boundary (Section 4.1), then quantify the global effect size and derive an early diminishing-return threshold (Section 4.2), and finally investigate whether the instabilities observed under fixed hyperparameters can be attributed to correctable configuration choices (Section 4.3).

4.1 Effect of training set size

We first examine how ranking accuracy and stability evolve as the training set grows from $m = 50$ to $m = 5{,}000$. Figure 1 presents both the mean WS coefficient and the inter-replication standard deviation as a function of m. As expected, ranking accuracy improves rapidly in the small-sample regime: $\overline{WS}$ rises

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_28

from 0.738 at $m = 50$ to 0.958 at $m = 500$. Beyond this point, the curve enters a plateau where further gains become incremental, reaching 0.970 at $m = 1{,}000$ and 0.989 at $m = 5{,}000$. Stability, measured by σ, follows a broadly complementary trajectory, decreasing from 0.077 at $m = 50$ to 0.002 at $m = 5{,}000$.

An unexpected pattern emerges at $m = 100$, where the standard deviation ($\sigma = 0.081$) actually exceeds the value observed at $m = 50$, contradicting the intuitive expectation that more data should yield more stable predictions. This anomaly persists across all 30 replications and is investigated in detail in Section 4.3, where we show that it is attributable to suboptimal hyperparameter configuration rather than a structural limitation of the method. It is worth noting that internal consistency ($\overline{WS}_{\text{int}}$, Eq. 6) remains at or above 0.979 for all m, indicating that independent model instances agree well with each other regardless of their absolute accuracy.

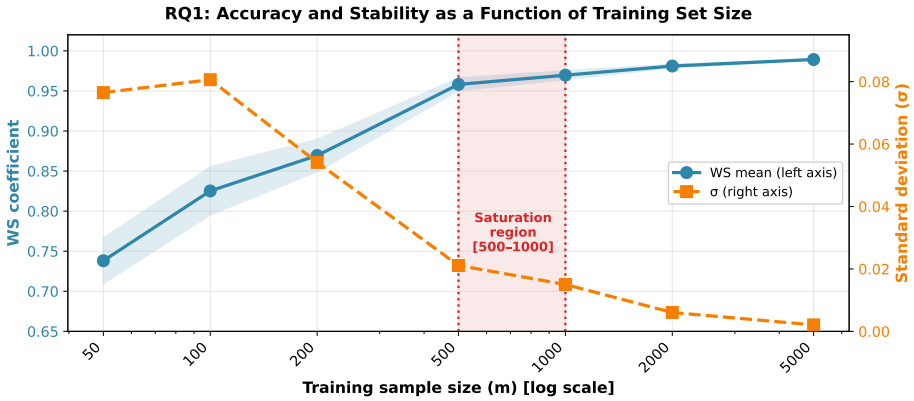

Fig. 1. Ranking accuracy and stability as a function of training set size ($k = 15$, $n = 30$ replications).

To better understand where additional data collection ceases to be cost-effective, Table 2 quantifies the per-transition accuracy gain and the efficiency-normalized return on investment. The highest per-sample return occurs in the first transition ($m = 50 \rightarrow 100$, $\text{ROI}_{WS} = 0.174$), and returns remain above the negligibility threshold $\tau = 0.01$ through $m = 200 \rightarrow 500$ ($\text{ROI}_{WS} = 0.030$). However, at the transition from $m = 500$ to $m = 1{,}000$, the ROI drops sharply to 0.002, falling well below τ. Stability ROI converges to the same region: all transitions beyond $m = 500$ yield $\text{ROI}_{\sigma} < \tau$. The coincidence of both indicators identifies the saturation region $m \in [500, 1{,}000]$ as the practical boundary beyond which additional data collection produces negligible improvement relative to its cost.

To obtain a continuous approximation of the observed trend, we fit a simple logarithmic model of the form $\overline{WS}(m) = \alpha + \beta \ln(m)$ to the seven empirical

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_28

Table 2. Marginal accuracy and stability changes with efficiency-normalized ROI per 100 additional samples.

Transition	$\Delta\overline{WS}$	ROI_{WS}	$\Delta\sigma$	ROI_{σ}
$50 \to 100$	+0.0871	+0.1742	+0.0041	−0.0082
$100 \to 200$	+0.0442	+0.0442	−0.0265	+0.0265
$200 \to 500$	+0.0887	+0.0296	−0.0330	+0.0110
$500 \to 1{,}000$	+0.0115	+0.0023	−0.0061	+0.0012
$1{,}000 \to 2{,}000$	+0.0114	+0.0011	−0.0090	+0.0009
$2{,}000 \to 5{,}000$	+0.0081	+0.0003	−0.0040	+0.0001

means. This functional form is motivated by classical learning curve literature; for kNN regression, Stone's theorem implies an error decay of $O\left(m^{-\frac{2}{d+2}}\right)$, which in the present four-dimensional setting yields an exponent of approximately $-\frac{1}{3}$, qualitatively consistent with the slow concave improvement the logarithm captures. The fit yields $\alpha = 0.571$, $\beta = 0.054$, and $R^2 = 0.870$. While this value indicates a reasonable approximation, it also suggests that the logarithmic form does not fully capture all features of the data. The residual variance can be attributed to two factors: the σ anomaly at $m = 100$ and a slight departure from log-linearity in the high-sample regime ($m \geq 2{,}000$), where empirical values lie marginally above the prediction.

4.2 Global effect size and early diminishing returns

The previous subsection characterised the learning curve through pairwise transitions. Here, we complement that analysis by quantifying the overall strength of the training-size effect and deriving an analytical characterisation of when learning gains begin to decelerate.

A one-way ANOVA across the seven training set sizes confirms that the effect of m on ranking accuracy is statistically significant, with $F(6, 203) = 117.07$ ($p = 3.95 \times 10^{-63}$). The associated effect size $\eta^2 = 0.776$ (Table 3) indicates that approximately 78% of the total variance in WS scores is attributable to differences in training set size, constituting a very large effect by conventional benchmarks. The remaining variance reflects the stochastic variability introduced by random subset selection within each size condition.

Table 3. Quantitative summary of the global effect size, model fit, and early diminishing-return point.

	η^2	R^2	m^*_{early}
Value	0.7758	0.8704	≈ 51

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_28

The logarithmic approximation introduced in Section 4.1 ($R^2 = 0.870$, Figure 2) provides a simple continuous description of the learning curve:

$$\overline{WS}(m) = \alpha + \beta \ln(m),$$

with $\beta = 0.054$. The instantaneous marginal gain implied by this model is given by

$$\frac{d\overline{WS}}{dm} = \frac{\beta}{m}.$$

Rather than imposing an arbitrary fixed threshold, we define an *early diminishing-return point* as the smallest training size m for which the marginal gain falls below a small fraction of the empirically observed inter-replication variability. Specifically, saturation is defined by

$$\frac{\beta}{m} < \alpha_\sigma \cdot \text{median}(\sigma(m)),$$

where $\sigma(m)$ denotes the inter-replication standard deviation and $\alpha_\sigma = 0.05$. Using the empirical median $\text{median}(\sigma(m)) = 0.021$, this criterion yields $m^*_{\text{early}} \approx 51$. At this point, the instantaneous learning gain becomes smaller than 5% of typical stochastic variability induced by random training subset selection, marking the end of the steep initial learning regime, as illustrated in Figure 2.

Importantly, this derivative-based threshold captures the transition from rapid improvement to gradually diminishing returns, but should not be interpreted as a practical deployment boundary. In contrast, the ROI-based saturation region $m \in [500, 1{,}000]$ identified in RQ1 reflects a cost-sensitive notion of practical saturation, where additional data collection yields negligible improvement relative to its cost. Together, these perspectives provide a multi-scale characterisation of INCOME's learning dynamics: an early reduction in marginal gains (around $m \approx 50$) followed by a later practical stabilisation region (around $m \geq 500$).

4.3 Hyperparameter sensitivity in small-sample regimes

The preceding analyses revealed two noteworthy patterns under fixed $k = 15$: the non-monotonic σ anomaly at $m = 100$ and substantial variance at small sample sizes more generally. A natural question arises: are these instabilities inherent to the INCOME method, or can they be mitigated through better hyperparameter selection? To answer this, we conduct a grid search over $k \in \{3, 5, 10, 15, 20, 30, 40\}$ for each small-sample size.

Figure 3 reveals that the sensitivity profiles differ qualitatively across the three regimes, which is particularly informative. For $m = 200$, σ is minimised at small k values ($k = 5$) and increases gradually with neighbourhood size, as one might expect. For $m = 50$, the curve is comparatively flat in the range $k \in [5, 15]$ but rises sharply for $k \geq 30$, reflecting oversmoothing when the neighbourhood encompasses a large fraction of the training set. The most striking pattern emerges at $m = 100$: here, σ attains its minimum at $k = 3$, rises to a peak

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_28

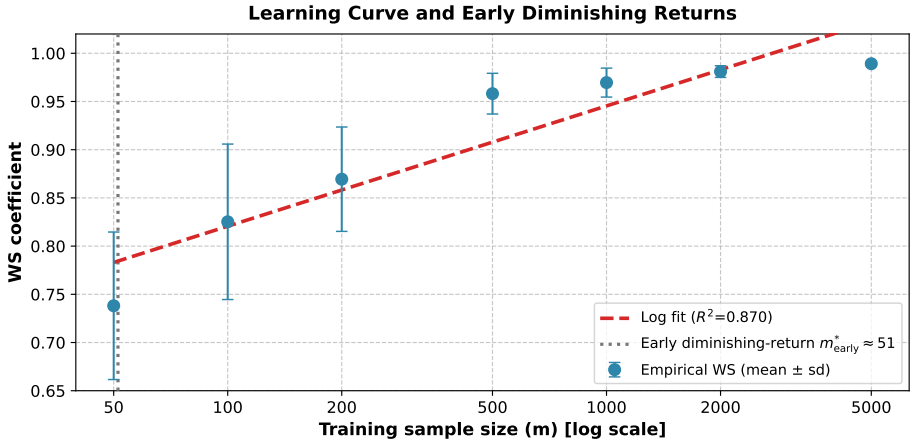

Fig. 2. Logarithmic learning curve fit with derivative-based early diminishing-return point $m^*_{\text{early}} \approx 51$.

at $k = 10$, and decreases monotonically thereafter, yet remains substantially above the optimum across the entire evaluated range $k \geq 10$. Consequently, the baseline configuration $k = 15$ falls well above the stability optimum for this specific sample size, explaining the anomaly observed in Section 4.1. These divergent profiles confirm that a fixed k cannot be uniformly optimal across training set sizes and that the interaction between k and m is non-trivial.

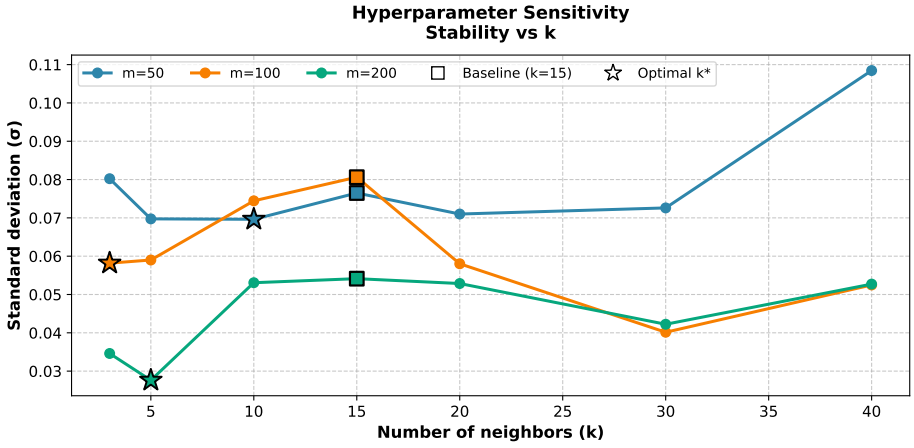

Fig. 3. Inter-replication standard deviation as a function of k for the small-sample regime.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_28

Having established that the sensitivity profiles are qualitatively different, we now identify the optimal k for each sample size. Table 4 compares the baseline ($k = 15$) with the stability-constrained optimal k^*. The optimal values ($k^* = 10$ for $m = 50$, $k^* = 3$ for $m = 100$, and $k^* = 5$ for $m = 200$) correspond to k^*/m ratios of 20%, 3%, and 2.5%, respectively. In all three cases, selecting k^* simultaneously improves both accuracy and stability relative to the baseline, as illustrated in Figures 4 and 5.

Table 4. Baseline ($k = 15$) versus stability-constrained optimal k^* in the small-sample regime.

m	k_{fixed}	k^*	$\overline{WS}_{\text{fixed}}$	$\overline{WS}_{\text{opt}}$	σ_{fixed}	σ_{opt}
50	15	10	0.7381	0.7787	0.0765	0.0696
100	15	3	0.8252	0.9190	0.0806	0.0582
200	15	5	0.8694	0.9434	0.0541	0.0276

The case of $m = 100$ is especially interesting: accuracy increases by +0.094 (from 0.825 to 0.919) and σ decreases by 28% (from 0.081 to 0.058) simply by changing k from 15 to 3, without collecting any additional training data. At $m = 200$, the stability improvement is even more pronounced, reaching 49% (σ drops from 0.054 to 0.028). At $m = 50$, the improvements are more modest (+0.041 in accuracy, 9% in σ), which is consistent with the fundamental information constraint at very small sample sizes, where there is simply too little data for any configuration to perform reliably.

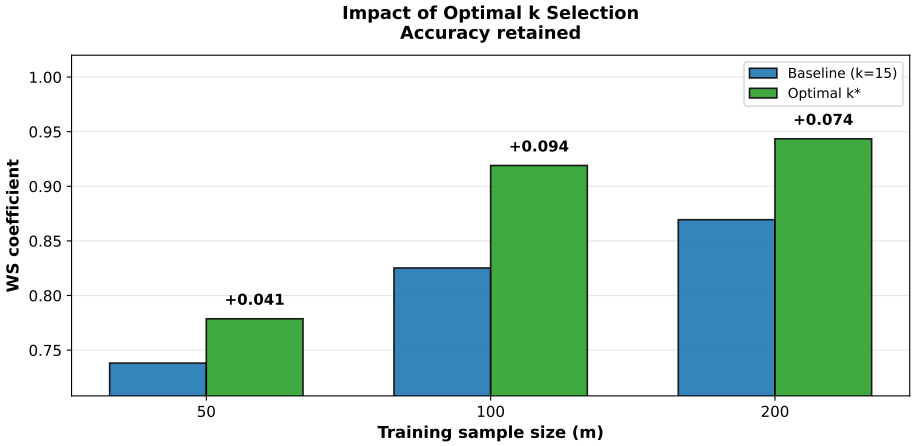

Fig. 4. Accuracy improvement under optimal k^* selection relative to the fixed baseline.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_28

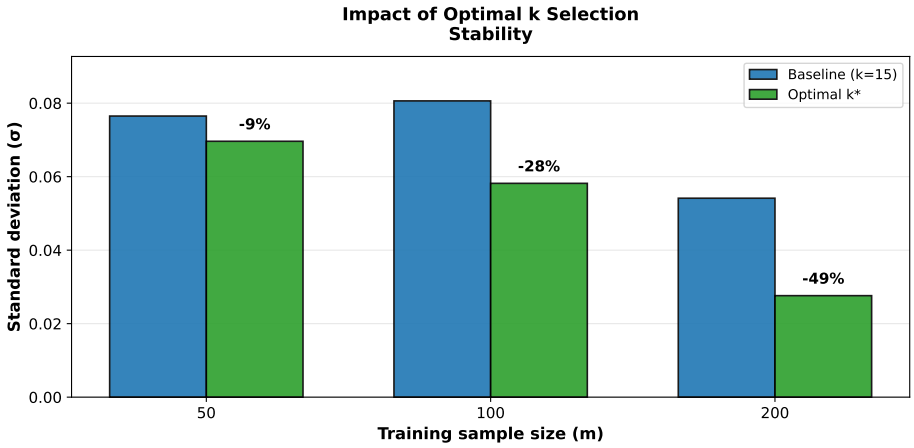

Fig. 5. Stability improvement under optimal k^* selection relative to the fixed baseline.

Crucially, the non-monotonic variance pattern from Section 4.1, where $\sigma(m{=}100) > \sigma(m{=}50)$ under fixed $k = 15$, is fully resolved under optimal k^* selection. With $k^* = 3$ at $m = 100$, the standard deviation drops to 0.058, which is now below $\sigma_{\text{opt}} = 0.070$ at $m = 50$, restoring the expected monotonic relationship between sample size and prediction stability.

This finding has an important practical implication: the instabilities observed in Section 4.1 are not inherent to the INCOME method but rather an artefact of hyperparameter misconfiguration. When combined with the derivative-based early diminishing-return point from Section 4.2 ($m^*_{\text{early}} \approx 50$), the results paint a coherent picture: the steep initial learning phase is both hyperparameter-sensitive and structurally transient. Beyond this regime, model behaviour becomes progressively more stable, ultimately reaching the practical saturation region identified in Section 4.1 ($m \geq 500$).

5 Discussion

For practitioners, these findings suggest a concrete data collection strategy: 200 records suffice to exit the hyperparameter-sensitive regime, while approximately 500 records yield deployment-ready accuracy ($WS > 0.95$). Below this threshold, adaptive k-selection can partially compensate for data scarcity, and at $m = 100$ switching from $k = 15$ to $k = 3$ improved WS by 0.094 without collecting any additional data. Although the present analysis relies on the WS coefficient, the qualitative learning curve shape is expected to be robust to metric choice; rank-uniform measures such as Spearman's ρ may shift precise thresholds but should preserve the concave trajectory.

The observed concave learning trajectory is qualitatively consistent with classical characterisations of learning curves for non-parametric estimators [6,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_28

11]. The logarithmic model adopted here provides a reasonable descriptive fit ($R^2 = 0.870$), although other functional forms (such as power-law or exponential saturation models) could potentially yield a closer approximation. The present setting also differs from standard supervised learning benchmarks in that performance is defined by ordinal agreement rather than numerical error, and the top-rank sensitivity of the WS coefficient amplifies the practical consequences of small improvements in the high-sample regime. This motivates the distinction between the derivative-based early saturation point and the later ROI-based saturation region.

Several limitations should be noted. The study relies on a single dataset with four continuous criteria and a monotonic target function. The generalizability of the identified thresholds and saturation behaviour to problems with discrete or mixed criteria, higher dimensionality, or non-monotonic preference structures remains to be established. The reference ranking derives from the target variable rather than expert judgments, which may understate the difficulty of real-world ranking tasks. Sensitivity to metric choice (e.g., Euclidean, Manhattan) remains an open empirical question, and the operational thresholds ($\tau = 0.01$, $\alpha\sigma = 0.05$) are empirically motivated. Finally, the use of a single regressor (kNN) leaves open whether the observed learning curve characteristics generalize to other artificial expert implementations. Although different threshold values would shift exact numerical boundaries, the qualitative pattern of early structural deceleration followed by later practical stabilisation remains robust.

6 Conclusion

This paper presented a systematic empirical investigation of the data requirements and small-sample behaviour of the INCOME method. The learning curve analysis revealed that ranking accuracy follows a concave trajectory, well approximated by a simple logarithmic model, with training set size explaining approximately 78% of the variance in ranking quality. Two complementary saturation thresholds were identified: a derivative-based early diminishing-return point at approximately $m^*_{\text{early}} \approx 51$, marking the end of the steep initial learning regime, and a cost-sensitive practical saturation region at $m \in [500, 1{,}000]$, beyond which marginal gains in both accuracy and stability become negligible. For the specific dataset and configuration studied, INCOME can produce reliable rankings ($\overline{WS} > 0.95$) with roughly 500 training instances.

The hyperparameter sensitivity analysis further demonstrated that the non-monotonic variance anomaly observed in the small-sample regime is an artefact of fixed-k configuration rather than an inherent methodological limitation. Stability-constrained k selection simultaneously improved accuracy and reduced variance without requiring additional training data, confirming that adaptive hyperparameter tuning is essential in data-scarce settings. The extent to which these findings, particularly the identified saturation thresholds and the effectiveness of adaptive configuration, generalize to higher-dimensional problems,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_28

alternative distance metrics, and expert-derived reference rankings remains an important direction for future research.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ferretti, V., Montibeller, G.: Key challenges and meta-choices in designing and applying multi-criteria spatial decision support systems. Decision Support Systems **84**, 41–52 (2016)
2. Figueroa, R.L., Zeng-Treitler, Q., Kandula, S., Ngo, L.H.: Predicting sample size required for classification performance. BMC medical informatics and decision making **12**(1), 8 (2012)
3. Habeeb, R., Hussain, I., Al-Ansari, N., Sammen, S.S.: A proposed comparative algorithm for regional crop yield assessment: An application of characteristic objects method. Mathematical Problems in Engineering **2022**(1), 8224953 (2022)
4. Kizielewicz, B., Shekhovtsov, A., Więckowski, J., Wątróbski, J., Sałabun, W.: Intelligent characteristic objects method (income): a data knowledge-based multi-criteria decision analysis. Artificial Intelligence Review **57**(10), 266 (2024)
5. Paradowski, B., Olender, P., Sałabun, W.: A comparative study on the efficiency of the modified comet in decision-making. Procedia Computer Science **246**, 103–112 (2024)
6. Perlich, C., Provost, F., Simonoff, J.S.: Tree induction vs. logistic regression: A learning-curve analysis. Journal of Machine Learning Research **4**(Jun), 211–255 (2003)
7. Sałabun, W., Urbaniak, K.: A new coefficient of rankings similarity in decision-making problems. In: International conference on computational science. pp. 632–645. Springer (2020)
8. Shekhovtsov, A.: How strongly do rank similarity coefficients differ used in decision making problems? Procedia Computer Science **192**, 4570–4577 (2021)
9. Tavana, M., Soltanifar, M., Santos-Arteaga, F.J.: Analytical hierarchy process: Revolution and evolution. Annals of operations research **326**(2), 879–907 (2023)
10. Tüfekci, P.: Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. International Journal of Electrical Power & Energy Systems **60**, 126–140 (2014)
11. Viering, T., Loog, M.: The shape of learning curves: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(6), 7799–7819 (2022)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_28