Journal of Industrial and Engineering System (JIES)
e-ISSN: 2722-7979
Vol. 7 No. 1, Hal 10-21

Small-Sample Capability Estimation for Recruitment Speed: A DMAIC Case Study on an Indonesian Outsourcing Platform

Filscha Nurprihatin*[1], Enrico Halim[1], Williem[2], Glisina Dwinoor Rembulan[3], Ali Vaezi[4]
[1]Department of Industrial Engineering, Sampoerna University, L'Avenue Building, Jl. Raya Pasar Minggu No. Kav. 16, Pancoran, South Jakarta 12780, Indonesia
[2]Department of Informatics, President University, Jababeka Education Park, Jl. Ki Hajar Dewantara, Cikarang Baru, Bekasi Regency, West Java 17530, Indonesia
[3]Department of Management, Universitas Bunda Mulia, Jl. Lodan Raya No. 2, Pademangan, North Jakarta 14430, Indonesia
[4]Department of Logistics and Supply Chain Management, Trent University, 55 Thornton Road South, Oshawa, ON L1J 5Y1, Canada
e-mail: *[1]filscha.nurprihatin@sampoernauniversity.ac.id, [1]enrico.halim@my.sampoernauniversity.ac.id, [2]williem@president.ac.id [3]grembulan@bundamulia.ac.id, [4]alivaezi@trentu.ca

*Corresponding: filscha.nurprihatin@sampoernauniversity.ac.id

ABSTRACT

Workforce sourcing, the process by which outsourcing and gig-economy intermediaries recruit job applicants for client firms, lacks standardized capability metrics that are comparable to those used in manufacturing quality control. This technical case study reports the design, implementation, and critical evaluation of a Define-Measure-Analyze-Improve-Control (DMAIC) procedure used to establish provisional sourcing-speed benchmarks (pessimistic, average, optimistic) for an Indonesian human-resource outsourcing platform, stratified by province and job category. Because most province–job-type strata contained fewer than ten observations (range 1–19), a Chebyshev-inequality-based trimming rule was adopted instead of parametric three-sigma control limits, automated through Excel VBA for scalability. Results show that benchmark reliability is highly stratum-dependent: only the highest-volume stratum (n = 19) yielded values interpretable with reasonable confidence, while low-volume strata produced control bands too wide for operational use. This paper discusses why a fixed 2-standard-deviation trim is a weaker foundation than sample-size-adjusted alternatives recently proposed in the outlier-detection literature, and outlines the validation and data-aggregation work still required before such benchmarks can be deployed as genuine capability limits.

Keywords: *Sourcing Speed, DMAIC, Chebyshev's Inequality, Process Capability, Small-sample Statistics, Workforce Outsourcing, Gig Economy*

INTRODUCTION

Over the past five years, digital labor-market intermediaries, including outsourcing firms and gig-economy platforms, have rapidly expanded in Southeast Asia by connecting blue-collar workers to short-term and contract job opportunities. Recent systematic reviews show that these platforms reduce entry barriers and improve job matching. However, they also focus on operational risks stemming from the speed and reliability of the applicant-sourcing pipeline. When the sourcing process is slow or unpredictable, both the platform's revenue and the client firm's staffing continuity are at risk (Alauddin et al., 2025). In Indonesia specifically, technological support and platform-mediated coordination have been shown to materially affect worker productivity and engagement outcomes at the sector level (Alfarizi et al., 2025), which places renewed emphasis on the internal operational metrics platforms use to plan and staff their own sourcing capacity.

The operational significance of sourcing speed is often underestimated. In this context, sourcing speed refers to the rate at which a platform accumulates qualified applicants for a specific project within a defined timeframe. Despite its importance, this metric is infrequently measured with the same rigor applied to metrics such as manufacturing throughput or defect rates. Six Sigma and its Define-Measure-Analyze-Improve-Control (DMAIC) roadmap are the dominant frameworks for establishing such rigor in production contexts. A recent literature review confirms that DMAIC implementation remains measurably less consistent in-service industries (60%) than in manufacturing (72%), with service-sector variation more often attributed to human behavior and weak systems than to the Man, Machine, Material, Method, Environment (4M1E) factors typical of factories (Trimarjoko et al., 2020). This asymmetry is directly relevant to recruitment operations, which are service processes almost by definition.

A small but growing body of work has begun applying DMAIC specifically to recruitment. Lean Six Sigma tools were utilized within a DMAIC framework to achieve a reduction in recruitment cycle time by 25.2% at a U.S. firm (Al-Rifai, 2025). Similar gains in processing speed and data accessibility were reported in a comparable case study in supply-chain logistics recruitment (Sundram et al., 2023). Strategic HR literature further argues that Six Sigma deployment in HR functions improves proactive service delivery and internal customer satisfaction (Singh & Srivastava, 2023). However, The studies in question focus primarily on the reduction of cycle time, which refers to the duration required for a single process step to be completed. This emphasis on cycle time contrasts with the objective of capability estimation, which involves establishing a statistically defensible range within which sourcing performance is expected to fall.

Outside recruitment specifically, the DMAIC framework has been extended to encompass a diverse array of operational metrics. These include supply-chain disruption risk (Andry et al., 2022), labor-driven product defects (Nurprihatin, Ayu, et al., 2023), machine downtime assessed through Overall Equipment Effectiveness (Nurprihatin, Rembulan, et al., 2023), customer satisfaction measured by Net Promoter Score (Nurprihatin et al., 2022), and waste in food manufacturing (Widiwati et al., 2025). This expansion demonstrates the adaptability of the DMAIC framework across various metric types. However, it is noteworthy that none of these studies establishes a small-sample, distribution-free capability limit. Consequently, this gap remains unaddressed not only within the context of the recruitment subset but also within the broader DMAIC literature. This distinction matters because capability limits, unlike cycle-time averages, are meant to function as ongoing control boundaries against which future performance is judged, and their credibility depends heavily on the statistical procedure used to derive them.

The empirical setting for this study is an Indonesian human-resource outsourcing company that manages both manpower outsourcing (candidate placement only) and business-power outsourcing (placement plus ongoing HR administration) for client firms. The company sources applicants through several digital and community channels and, prior to this study, set sourcing-speed targets arbitrarily from historical recollection rather than from a documented statistical procedure. The key operational question driving this work was whether a defensible and stratified sourcing-speed benchmark, categorized by province and job type, could be developed from the company's historical records to replace the current ad hoc target-setting practices.

The central methodological complication addressed in this paper lies in the violation of a standard assumption inherent in classical Six Sigma capability analysis. Specifically, this pertains to the issue of inadequate sample size within each stratum. This paper's primary contribution is the examination of this discrepancy and its implications for operational data analysis. Most manufacturing DMAIC applications draw on hundreds or thousands of units per batch; here, most province–job-type strata contained between one and nineteen completed projects over a five-month observation window. Under such conditions, neither a fitted-distribution approach nor a conventional three-sigma rule can be applied with confidence, since neither the shape nor the spread of the underlying distribution can be reliably estimated. This context underscores the importance of employing a distribution-free bound, specifically Chebyshev's inequality. Recently, this approach has garnered renewed interest in the outlier-detection literature. Its appeal lies in the applicability of Chebyshev's inequality to datasets that either lack a known distribution or have sample sizes that are insufficient for traditional parametric assumptions to be valid(Alves et al., 2024; Aydın, 2023).

This context highlights the importance of using a distribution-free bound, specifically Chebyshev's inequality. Recently, this approach has gained renewed interest in the literature on outlier detection. The appeal of Chebyshev's inequality lies in its applicability to datasets that either do not have a known distribution or have sample sizes that are small. Section 2 situates the work within recent literature on DMAIC in recruitment, small-sample outlier detection, and the gig-economy workforce context. Section 3 details the data and analytical procedure. Section 4 presents the stratified results and a sensitivity analysis of the trimming rule and discusses the procedure's practical and statistical limitations. Section 5 concludes with recommendations for the data-collection and validation work required before these estimates could be adopted as genuine control limits.

LITERATURE REVIEW

Table 1 exhibits the sources cover several primary domains, sucah as DMAIC applications, advanced statistical methods for outlier detection such as stratified benchmarking and Chebyshev-inequality.

DMAIC in Recruitment and HR Operations

DMAIC's extension beyond manufacturing into HR and recruitment processes is a recent but active research thread. A study applied Lean Six Sigma tools within DMAIC to a U.S. recruitment pipeline, evaluating each hiring stage and tracking cycle times before and after intervention, reporting a 25.2% overall reduction and a 50.0% reduction in approval times specifically (Al-Rifai, 2025). Another study demonstrated a parallel application in supply-chain logistics recruitment, showing that DMAIC's structure improved process efficiency and data accessibility for new-hire recruiting (Sundram et al., 2023). Another paper argued more broadly that Six Sigma deployment in HR shifts practices toward proactive, data-driven service delivery, improving both internal customer satisfaction and employee motivation (Singh & Srivastava, 2023). Adjacent case studies in warehousing (A. Adeodu et al., 2023) and manufacturing rejection-rate reduction (Mittal et al., 2023) confirm DMAIC's general effectiveness as a structured improvement roadmap. However, all of these studies measure cycle time and defect rate, but do not address the capability limits for a continuously monitored operational metric. This study fills that gap, including those specific to recruitment.

Six Sigma Capability Estimation Under Small-Sample Conditions

A persistent limitation noted across the DMAIC literature is inconsistency in how the Control phase is implemented once improvements are made. A study found that only 60% of service-sector case studies consistently implemented the full DMAIC roadmap, compared to 72% in manufacturing, attributing the gap partly to human-behavior-driven variability and weaker systematization in services (Trimarjoko et al., 2020). Recruitment operations are marked by inconsistent project volumes, diverse job categories, and a heavy reliance on human judgment at various stages. This aligns with a context of high variability and low consistency, which reflects the challenges of sample size and standardization faced in the current situation.

Distribution-Free Outlier Detection and Chebyshev's Inequality

Because classical three-sigma control limits assume approximate normality and adequate sample size, several recent studies have revisited Chebyshev's inequality as a distribution-free alternative suited to small or unknown-distribution datasets. A study proposed a boundary-aware, density-based outlier-detection method that uses Chebyshev's inequality to draw neighborhood boundaries without assuming a specific data distribution, reporting improved performance over prior methods particularly on datasets with few points (Aydın, 2023). A systematic review of

102 studies confirms Chebyshev's inequality as one of only three outlier-removal techniques (alongside boxplot methods and principal component analysis) in consistent use across the literature, underscoring its continued relevance despite being a comparatively old result (Alves et al., 2024). However, a recent study showed that fixed-fence rules, like Tukey's boxplot and Chauvenet's criterion, have a common issue, such as using a uniform fence that does not consider sample size (Lin et al., 2026). This issue is similar to the fixed 2-standard-deviation Chebyshev trim (Lin et al., 2026). To address this weakness, the study proposed a sample-size-adjusted boxplot criterion specifically designed to improve upon these fixed-fence methods (Lin et al., 2026). Another study further showed that Bienaymé–Chebyshev-based rules can be extended with Bayesian bootstrap methods to quantify the uncertainty attached to each outlier classification, rather than treating the threshold as fixed and certain (Sartore et al., 2024). These findings directly bear on the present study's methodological choice.

Gig-Economy and Outsourced Workforce Context in Indonesia

The context of this study involves a technology-driven outsourcing intermediary that connects blue-collar applicants with employers. This fits within the broader literature on platform-mediated gig work. A systematic review of 18 studies published between 2019 and 2023 found that digital platforms lower entry barriers and facilitate task allocation but simultaneously expose workers to algorithmic control and downward wage pressure, with effects that are highly context dependent (Alauddin et al., 2025). In the Indonesian setting specifically, another study found that technological support and hybrid work coordination significantly improve gig-worker productivity and engagement, with engagement contributing more than productivity to broader employment-quality outcomes (Alfarizi et al., 2025). Neither strand of literature adequately addresses the internal operational metrics, such as sourcing speed, that influence the reliability of these platforms in staffing the specific positions that are the focus of this evaluation. This gap in understanding is the area that this case study aims to explore.

Table 1. Related Works

Authors	Methods			Objectives
	Lean Six Sigma (DMAIC)	Stratified Benchmarking	Chebyshev-inequality	
(A. O. Adeodu et al., 2020)	✓			To develop an improvement framework for warehouse processes in a 3PL company.
(Mittal et al., 2023)	✓			To reduce the rejection rate of rubber weather strips in a manufacturing context.
(Sundram et al., 2023)	✓			To demonstrate LSS adaptability in improving recruitment processes for supply chain logistics.
(Singh & Srivastava, 2023)	✓	✓		To explore the deployment of Six Sigma in strategic Human Resource Management (HRM).
(Aydın, 2023)			✓	To propose a fast algorithm for outlier detection in high-dimensional datasets.
(Sartore et al., 2024)			✓	To study uncertainty in identifying cellwise outliers in multivariate survey data.
This Paper	✓	✓	✓	To establish sourcing-speed benchmarks for an Indonesian

outsourcing platform using small-sample data

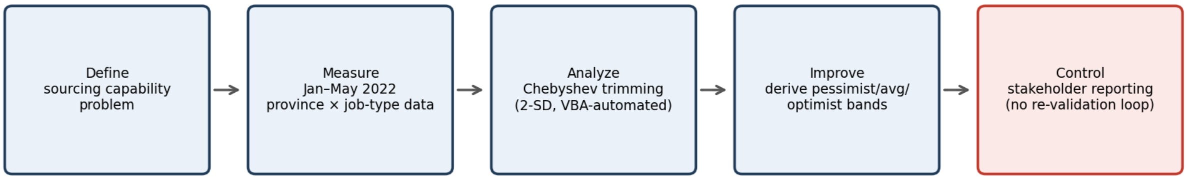

Figure 1. DMAIC Procedure

METHODS

This study employs the DMAIC roadmap, as illustrated in Figure 1. It focuses on a specific operational metric known as sourcing speed. Sourcing speed is defined as the number of qualified applicants gathered per calendar day for a given project at a human-resource outsourcing company based in Indonesia. The unit of analysis is a stratum defined by the combination of province and job-type code; each completed sourcing project contributes exactly one observation (a single sourcing-speed value) to its corresponding stratum. Data were extracted from the company's internal project database for the period January–May 2022, a five-month window chosen at the stakeholders' discretion rather than through a formal power or sample-size calculation.

In the Define phase, the operational problem was specified as the absence of a documented, repeatable procedure for setting sourcing-speed targets; targets had previously been set from managers' recollection of past projects rather than from a defined dataset or statistical rule. In the Measure phase, data were extracted for each completed project, focusing on three key fields, including province, job-type code, and realized sourcing speed. Subsequently, strata were established by grouping projects that shared the same province and job type. This process resulted in the identification of eleven distinct strata, with project counts (n) ranging from 1 to 19, as presented in Table 2.

Table 2. Stratified Sourcing-Speed Estimates (Applicants/Day) by Province and Job Type, January–May 2022, After Chebyshev-Based ($k = 2$) Trimming

Province	Job type	Project count (n)	Pessimist	Average	Optimist
A	D	1	1.38	1.38	1.38
A	K	2	1.50	2.94	4.38
A	S	8	0.00	0.47	1.67
A	SP	1	0.40	0.40	0.40
B	A	10	0.83	2.61	9.33
B	D	2	5.00	6.83	8.67
B	K	1	8.17	8.17	8.17
B	L	4	0.21	0.61	1.14
B	PK	2	0.71	1.43	2.14
B	S	19	0.06	0.56	1.20
B	SP	2	1.67	2.21	2.75

Note: Project count (n) is presented alongside each estimate because the reliability of the band heavily depends on n and should not be inferred solely from the point estimates.

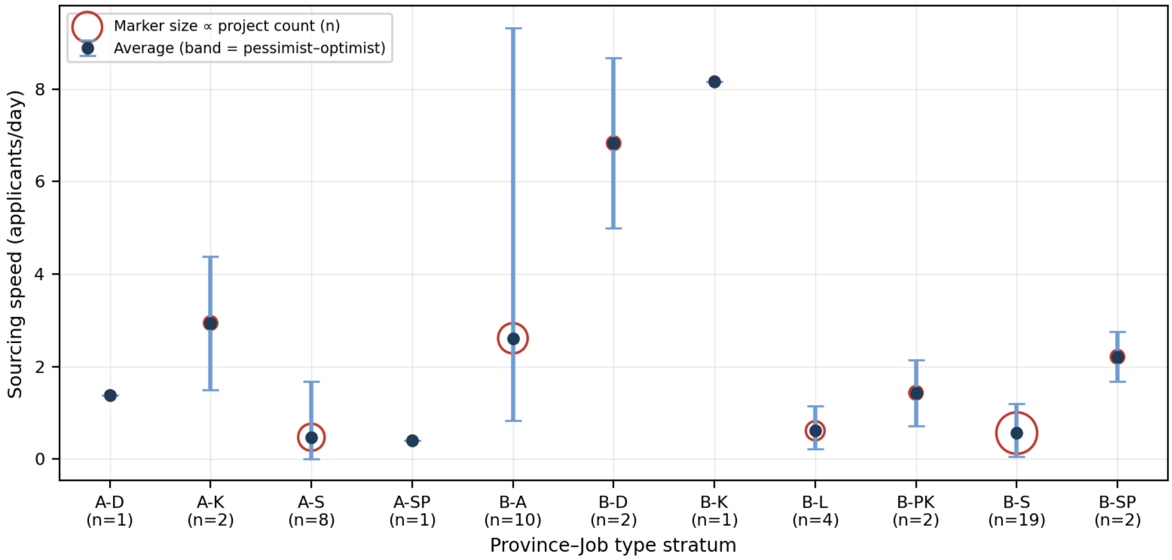

Figure 2. Sourcing-speed Control Bands by Stratum

Note: Marker size is proportional to project count (n); band width does not decrease monotonically with n but is visibly narrowest only for the single stratum (B-S) with n = 19.

The Analyze phase required selecting an outlier-trimming rule appropriate to the data's structure. Because most strata contained fewer than thirty observations, distributional shape could not be identified with confidence, ruling out a standard three-sigma rule calibrated to normality. Chebyshev's inequality was therefore adopted, which guarantees that at least $1 - 1/k^2$ of any dataset's mass lies within k standard deviations of the mean regardless of distribution (Alves et al., 2024; Aydın, 2023). Setting $k = 2$ implies that at least 75% of observations in each stratum should lie within two standard deviations of that stratum's mean; observations outside this range were treated as outliers. Because per-stratum standard deviations were themselves unreliable at low n, the rule was operationalized empirically: observations were sorted in ascending order and the outer 12.5% of ranks at each tail were removed directly, approximating the 75%-retention criterion without requiring a stable standard-deviation estimate. This procedure was automated in Excel VBA to apply consistently across all eleven strata, given the impracticality of manual, stratum-by-stratum recalculation.

In the Improve phase, three summary values were computed for each stratum after trimming: the arithmetic mean ('average'), the maximum retained value ('optimist'), and the minimum retained value ('pessimist'). The proposed parameters were presented to stakeholders as a working analogue to Six Sigma control limits. Specifically, the pessimist value serves as a lower bound, indicating that sourcing is considered underperforming when below this threshold. Conversely, the optimist value acts as an upper bound, signifying that sourcing is deemed ahead of target when above this limit. Furthermore, the project count (n) was explicitly retained alongside the results for each stratum. This approach enables stakeholders to visually assess the level of confidence that a given band warrants, as illustrated in Table 2 and Figure 2.

The Control phase, as implemented, consisted of presenting these stratified bands to company stakeholders for immediate operational use in setting future project targets. It is important to state plainly that this Control phase was a one-time report-out rather than a closed monitoring loop with a defined re-measurement cadence, a distinction that is elaborated as a limitation and illustrated in Figure 1. The original data collection did not involve a holdout period, backtesting, or comparisons with alternative trimming methods. Where it is feasible to conduct such comparisons using only the published summary statistics, the analysis is presented analytically

Submitted: **14/06/2026**; Revised: **20/06/2026**; Accepted: **27/06/2026**; Published: **28/06/2026**

rather than being claimed as an empirical validation.

Two supporting artifacts are presented alongside the numerical results. Table 2 reproduces the stratum-level pessimist/average/optimist values together with project counts, consistent with the source operational dataset. Figure 2 visualizes the same data as a scatterplot with asymmetric error bars spanning the pessimist–optimist band, with marker size scaled to project count, to make the relationship between sample size and band reliability visually inspectable rather than left implicit in a table of numbers. Figure 1 summarizes the DMAIC sequence as implemented, explicitly marking the Control phase as a single reporting event.

RESULTS AND DISCUSSION

Stratum-Level Capability Estimates

Table 2 reports the pessimist, average, and optimist sourcing-speed values for all eleven observed province–job-type strata, together with the project count underlying each estimate. The project exhibits a significant variation, ranging from a single observation in strata A-D, A-SP, and B-K, to as many as nineteen observations in stratum B-S. This represents a nineteen-fold difference in evidentiary weight across the strata. However, this disparity is not effectively communicated in the format of the output received by stakeholders, as all eleven strata were reported using the same three-number summary, regardless of the sample size (n). Average sourcing speed across strata ranges from 0.40 to 8.17 applicants per day, and the width of the pessimist–optimist band ranges from 0.00 (strata with $n = 1$, where all three statistics coincide trivially) to 8.50 applicants per day (stratum B-A, $n = 10$), illustrating that band width does not decrease monotonically with n and is also sensitive to the underlying heterogeneity of job type and province.

Figure 2 vividly illustrates this pattern. Notably, only stratum B-S ($n = 19$) meets the project count threshold that is generally recommended for even non-parametric interval estimation. This stratum yields a control band of 0.06 to 1.20 applicants per day, which is sufficiently narrow to serve as a plausible operational control range. Strata with $n \leq 2$ produce bands that are either degenerate (identical pessimist, average, and optimist values when $n = 1$) or implausibly wide relative to the average (e.g., stratum B-D, $n = 2$, ranges from 5.00 to 8.67 around a mean of 6.83). Presenting these seven low-n strata with the same apparent precision as stratum B-S risks conveys false confidence to stakeholders who were not shown the underlying project counts in the original deployment. This issue is addressed by consistently co-displaying n.

Sensitivity of the Fixed 2-Standard-Deviation Trimming Rule

A central claim tested here is whether the fixed $k = 2$ Chebyshev trim is robust to the small samples it is applied to. Chebyshev's inequality provides a conservative estimate because it is derived as a worst-case bound applicable to all possible distributions. When $k = 2$, the inequality guarantees that at least 75% of the distribution's mass lies within the specified range. This is a much weaker assurance compared to the approximately 95% that a normal distribution would provide for the same value of k. When applying the concept of 'trimming 12.5% from each tail' to strata with $n = 1$ or $n = 2$, this practice becomes operationally meaningless. Specifically, there are insufficient data points available for removal. Consequently, the perspectives of both the pessimist and the optimist converge to the singular or dual observations that are present. The resultant 'band' therefore merely reflects sampling noise, rather than representing a genuine performance range. This is not a flaw specific to this company's implementation; it is a structural limitation of applying any tail-trimming rule to samples below roughly $n = 10$–15, and it explains why a previous study explicitly designed a sample-size-adjusted fence rather than retaining a fixed-percentage cut for all sample sizes (Lin et al., 2026).

To illustrate the practical consequence, consider stratum B-A ($n = 10$): removing or retaining a single additional extreme observation at this sample size shifts the empirical 12.5% trim boundary by roughly one full data point out of ten, i.e., a 10 percentage-point swing in what counts as 'trimmed.' At $n = 19$ (stratum B-S), the same single-observation change represents only a 5 percentage-point swing, which is consistent with the visibly tighter band observed for that stratum in Figure 2. This back-of-envelope comparison is computed directly from the reported project counts, rather than relying on re-simulated data. The findings demonstrate that the stability of the method scales with sample size (n) in the manner that sample-size-adjusted alternatives are intended to formalize (Aydın, 2023; Lin et al., 2026). Furthermore, these results support the adoption of such an alternative when larger per-stratum samples become available. This is preferable to the continued application of a fixed percentage cut uniformly across the data.

Comparison with Sample-Size-Adjusted Alternatives

The outlier-detection literature offers at least two directions for improving on the fixed-trim approach used here, neither of which was applied retroactively to the present dataset but both of which are directly actionable for future data collection cycles. First, the density-based method utilizes Chebyshev's inequality to define local neighborhood boundaries rather than a single global fence (Aydın, 2023). This approach could, in principle, be adapted to weight each stratum's trim by its own local density of observations, rather than applying an identical 12.5% cut to every stratum regardless of n (Aydın, 2023). Second, the fence width Chauvenet-type boxplot is explicitly parameterized as a function of sample size (Lin et al., 2026). A third option involves the application of a Bayesian bootstrap extension to Chebyshev-based rules (Sartore et al., 2024). This approach enables stakeholders to visualize a confidence interval surrounding each estimate provided by pessimists and optimists, rather than relying on a single deterministic figure (Sartore et al., 2024). By doing so, it directly addresses the concern of false precision in estimations (Sartore et al., 2024). Adopting any of these would require re-implementing the Analyze phase, not merely re-labeling the existing output, and is recommended as a concrete next step rather than a rhetorical caveat.

Managerial and Operational Implications

Despite its statistical limitations, the procedure delivered practical value that a purely arbitrary, memory-based target-setting practice did not: it made the evidentiary basis for each target explicit, attached a project count to every claim, and automated recomputation via VBA so that the exercise could in principle be repeated as new data accrue. For the one stratum with adequate volume (B-S, n = 19), the resulting band is plausibly usable as an interim operational control range, consistent with how the company reported using it. For the remaining ten strata, the appropriate managerial use of the output is not as a control limit but as a diagnostic flag identifying which province–job-type combinations require further data accumulation, or aggregation into coarser categories (e.g., pooling job types within a province), before any target should be treated as binding. The reframing of the concept of 'benchmark' to that of a 'diagnostic, contingent on n' represents a significant managerial contribution. While this shift is defensible and potentially beneficial, it nonetheless does not achieve the status of a validated capability model.

Limitations

Several limitations constrain the generalizability and current operational readiness of this work. First, the dataset is drawn from a single organization over a single five-month window chosen without a formal sample-size justification, limiting external validity to other outsourcing contexts or seasons. Second, the study did not conduct a holdout or backtesting procedure to assess whether the bands derived from data collected between January and May 2022 would accurately classify the performance of subsequent months as either in-control or out-of-control. Furthermore, the application of these bands by stakeholders indicates a level of satisfaction based on face validity, rather than on rigorous statistical validation. Third, the underlying data were reported in censored form to preserve commercial confidentiality, which is disclosed here for transparency but means the present paper cannot be independently re-analyzed from raw records. Fourth, no comparison against an alternative outlier-detection method was empirically executed.

The discussion is analytical and literature-grounded rather than a head-to-head empirical benchmark. The next phase of this research requires addressing four key points. These include the establishment of a longer observation window, the implementation of a formal validation split, the arrangement for data-sharing or the use of synthetic data, and a comparison of various implemented methods. Each of these components is essential for advancing the study.

CONCLUSIONS

This paper reported a DMAIC-based procedure for estimating sourcing-speed capability at an Indonesian human-resource outsourcing platform, using a Chebyshev-inequality-based trimming rule to accommodate the small, uneven sample sizes characteristic of stratified operational HR data. The central empirical finding highlights that the reliability of the resulting pessimist/average/optimist bands is strongly and unevenly dependent on the sample size at the stratum level. Notably, only the stratum containing nineteen observations yielded a band that was narrow enough to potentially serve as an operational control range. In contrast, the remaining ten strata, which had sample sizes ranging from one to ten, produced either degenerate bands or excessively wide bands. The apparent precision of these broader bands could mislead stakeholders who are not privy to the underlying project counts.

The sensitivity analysis reveals, based on the study's own reported data rather than simulations, that a fixed-percentage trimming rule exhibits structural instability at low sample sizes (n). In the lowest-volume strata, the addition or removal of a single observation can result in a shift of the effective trim boundary by ten percentage points or more. In contrast, in the highest-volume stratum, this shift is approximately five percentage points. This issue is not unique to the company under investigation. Instead, it arises from a foreseeable consequence linked to the implementation of fixed-fence outlier detection rules. Such rules include those based on Chebyshev's theorem, the three-sigma rule, and standard Tukey boxplots. These methodologies tend to exhibit limitations when utilized on samples containing fewer than approximately ten to fifteen observations. This inherent weakness highlights the need for recent sample-size-adjusted alternatives in the outlier detection literature, which are specifically designed to address these concerns.

Accordingly, the principal contribution of this paper is not the introduction of a validated new benchmarking method. Instead, it presents a documented and transparent account that identifies where a plausible-looking, already-deployed benchmarking procedure is statistically defensible and where it is not. Furthermore, the paper includes a concrete roadmap, grounded in existing literature, aimed at addressing and closing the identified gaps. This paper recommend three specific next steps for this and comparable organizations. First, extend the observation window well beyond five months, or aggregate job-type categories within a province, so that a larger share of strata reach the $n \approx 15–20$ threshold at which distribution-free bounds begin to behave stably. Second, replace the fixed 2-standard-deviation, fixed-percentage trim with a sample-size-adjusted fence, such as the Chauvenet-type boxplot, or a Bayesian-bootstrap extension of Chebyshev's inequality that reports an explicit confidence interval around each pessimist/optimist estimate rather than a single deterministic value (Lin et al., 2026; Sartore et al., 2024). Third, before implementing any revised procedures operationally, it is essential to conduct a formal backtest. This involves deriving bands from one time window and subsequently evaluating a held-out window. The goal is to assess whether the observed sourcing speeds that fall outside these bands correspond to periods identified independently by the company as either under-performing or over-performing.

This case underscores a significant gap in the existing DMAIC literature. While DMAIC has been utilized effectively to reduce recruitment cycle times (Al-Rifai, 2025; Sundram et al., 2023), its application in capability estimation is less developed. Capability estimation involves establishing a defensible statistical range for ongoing monitoring, rather than relying solely on a one-time before-and-after comparison. This underdevelopment is particularly evident in service and HR contexts. Furthermore, it aligns with the documented lower consistency in the full implementation of the DMAIC roadmap within service industries compared to manufacturing sectors (Trimarjoko et al., 2020). Extending DMAIC-based capability analysis to small and unevenly distributed operational datasets is a promising avenue for future research in industrial engineering. This is particularly relevant for contexts such as human resources, recruitment, and other service operations, which often differ from the large production batches for which Six Sigma was originally developed. Furthermore, the sample-size-adjusted outlier-detection methods reviewed in this context are directly applicable to these smaller datasets.

Finally, this paper serves as a note for assessing the practical maturity of the work presented. It is important to note that this study originated as an applied industry case study rather than from a controlled experimental framework. At this stage, the principal value of the research lies in its diagnostic and methodological contributions. Specifically, it aims to clarify which outputs can currently be trusted, and to what extent, rather than providing prescriptive recommendations. This paper consider this an appropriate and honest scope for a first published account of the problem, to be superseded by the validated, sample-size-adjusted follow-up study outlined above.

REFERENCES

Adeodu, A., Maladzhi, R., Kana-Kana Katumba, M. G., & Daniyan, I. (2023). Development of an improvement framework for warehouse processes using lean six sigma (DMAIC) approach. A case of third party logistics (3PL) services. *Heliyon*, *9*(4), 1–19. https://doi.org/10.1016/j.heliyon.2023.e14915

Adeodu, A. O., Kanakana-Katumba, M. G., & Maladzhi, R. (2020). Implementation of lean six sigma (LSS) methodology, through DMAIC approach to resolve down time process; A case of a paper manufacturing company. *Proceedings of the International Conference on Industrial Engineering and Operations Management*, *59*, 37–47.

Alauddin, F. D. A., Aman, A., Ghazali, M. F., & Daud, S. (2025). The influence of digital platforms on gig workers: A systematic literature review. *Heliyon*, *11*(1), 1–14. https://doi.org/10.1016/j.heliyon.2024.e41491

Alfarizi, M., Noer, L. R., & Noer, B. A. (2025). Technological support, hybrid work, and national employment policies: Catalysts for worker productivity and SDG 8 achievement in Indonesia's gig economy. *Jurnal Ketenagakerjaan*, *20*(1), 1–21. https://doi.org/10.47198/jnaker.v20i1.440

Al-Rifai, M. (2025). Application of the DMAIC methodology and lean six sigma (LSS) tools to improve recruitment cycle time. *Measuring Business Excellence*, *29*(3), 586–608. https://doi.org/10.1108/MBE-09-2024-0142

Alves, F., de Souza, E. G., Sobjak, R., Bazzi, C. L., Hachisuca, A. M. M., & Mercante, E. (2024). Data processing to remove outliers and inliers: A systematic literature study. *Revista Brasileira de Engenharia Agricola e Ambiental*, *28*(9), 1–12.

Andry, J. F., Nurprihatin, F., & Liliana, L. (2022). Supply chain disruptions mitigation plan using six sigma method for sustainable technology infrastructure. *Management and Production Engineering Review*, *13*(4), 88–97. https://doi.org/10.24425/mper.2022.142397

Aydın, F. (2023). Boundary-aware local density-based outlier detection. *Information Sciences*, *647*(July), 1–17. https://doi.org/10.1016/j.ins.2023.119520

Lin, H., Zhang, R., & Tong, T. (2026). When Tukey meets Chauvenet: A new boxplot criterion for outlier detection. *Journal of Computational and Graphical Statistics*, *35*(1), 198–211.

https://doi.org/10.1080/10618600.2025.2520577

Mittal, A., Gupta, P., Kumar, V., Al Owad, A., Mahlawat, S., & Singh, S. (2023). The performance improvement analysis using Six Sigma DMAIC methodology: A case study on Indian manufacturing company. *Heliyon*, *9*(3), 1–11. https://doi.org/10.1016/j.heliyon.2023.e14625

Nurprihatin, F., Ayu, Y. N., Rembulan, G. D., Andry, J. F., & Lestari, T. E. (2023). Minimizing product defects based on labor performance using linear regression and six sigma approach. In *Management and Production Engineering Review* (Vol. 14, Number 2, pp. 88–98). Polska Akademia Nauk. https://doi.org/10.24425/mper.2023.146026

Nurprihatin, F., Mardhatillah, N. R., Young, M. N., Liman, S. D., Redi, A. A. N. P., & Prasetyo, Y. T. (2022). Integration of Net Promoter Score and DMAIC approach to measure customer satisfaction in packaging industry. *International Conference on Computational Modelling, Simulation and Optimization*, 220–225. https://doi.org/10.1109/ICCMSO58359.2022.00052

Nurprihatin, F., Rembulan, G. D., Andry, J. F., Lubis, M., Widiwati, I. T. B., & Vaezi, A. (2023). Integration of overall equipment effectiveness and six sigma approach to minimize product defect and machine downtime. *Management and Production Engineering Review*, *14*(4), 71–91. https://doi.org/10.24425/mper.2023.147205

Sartore, L., Chen, L., & Bejleri, V. (2024). Empirical inferences under Bayesian framework to identify cellwise outliers. *Stats*, *7*(4), 1244–1258. https://doi.org/10.3390/stats7040073

Singh, P., & Srivastava, A. (2023). Strategic human resource management: enhancing performance with six sigma approach. *International Journal of Innovative Research and Growth*, *12*(3), 50–55. https://doi.org/10.26671/ijirg.2023.3.12.105

Sundram, V. P. K., Ghapar, F., Lian, C. L., & Muhammad, A. (2023). Engaging lean six sigma approach using DMAIC methodology for supply chain logistics recruitment improvement. *Information Management and Business Review*, *15*(1), 46–53.

Trimarjoko, A., Purba, H. H., & Nindiani, A. (2020). Consistency of DMAIC phases implementation on six sigma method in manufacturing and service industry: A literature review. *Management and Production Engineering Review*, *11*(4), 34–45. https://doi.org/10.24425/mper.2020.136118

Widiwati, I. T. B., Liman, S. D., & Nurprihatin, F. (2025). The implementation of Lean Six Sigma approach to minimize waste at a food manufacturing industry. *Journal of Engineering Research*, *13*(2), 611–626. https://doi.org/10.1016/j.jer.2024.01.022