ISSN 2424-5321

2025, 29(1), 35–40. Received 8 December 2025. Accepted 30 December 2025.

QUANTITATIVE EVALUATION OF EMPLOYEE SUBSTITUTABILITY IN MANUFACTURING USING COSINE SIMILARITY AND PRINCIPAL COMPONENT ANALYSIS

Donatas Dervinis

Šiaulių valstybinė kolegija / Higher Education Institution

ORCID id: orcid.org/0000-0002-3517-8859

Annotation

Employee substitutability in manufacturing plays a key role in ensuring production efficiency and seamless workflow transitions. This study explores algorithms and computational methods to quantify substitutability using operational performance data. In work are used employ Cosine Similarity and Principal Component Analysis (PCA) to evaluate how closely employees' work patterns align. Cosine Similarity measures the resemblance between employee task performance, while PCA reduces dimensionality to highlight key differences in skill sets. A dataset containing employee operation counts and execution times was analyzed using these methods. The results indicate that employees with high Cosine Similarity scores exhibit comparable performance levels, while PCA effectively identifies clusters of employees with similar efficiency patterns. Combining both techniques provides a comprehensive methodology for employee substitutability, workforce optimization and facilitating task reallocation.

Keywords: employee substitutability, principal component analysis (PCA), cosine similarity.

Introduction

Employee substitutability in manufacturing is a crucial factor in maintaining production efficiency and ensuring smooth workflow transitions [1]. The ability to replace one worker with another without significant performance loss depends on multiple factors, including task complexity, skill levels, and operational dependencies. Objective evaluation of substitutability requires analyzing key production data such as product type, operations performed, output quantity, and standard production times. By leveraging this data, companies can develop metrics to quantify how seamlessly a new employee can take over a given task [2].

One method to measure substitutability is by assessing the variance in production efficiency when different employees perform the same operation. This can be done by comparing the actual production time of a new worker against the standard time and the previous worker's performance [3]. Additionally, machine learning techniques can be applied to predict potential efficiency losses based on historical data [4]. Such analyzes help manufacturers optimize workforce allocation and develop targeted training programs to improve flexibility within the production line [5]. The substitutability of employees is important not only for evaluating production efficiency but also for understanding employee psychology [6].

Task: Develop a quantitative framework for assessing employee substitutability in manufacturing using Cosine Similarity and Principal Component Analysis (PCA), based operational performance data.

Research goals:

1. To explore possible computational methods for evaluating employee substitutability.
2. To develop employee substitutability models based on Cosine Similarity and PCA methods.
3. To realize and validate the models using real data.

Methodology: literature review, computer modeling, data analysis.

ISSN 2424-5321

2025, 29(1), 35–40. Received 8 December 2025. Accepted 30 December 2025.

1. Review of possible methods for evaluating employee substitutability

Employee substitutability can be assessed using mathematical, statistical, big data, machine learning, and AI techniques, which enable a comprehensive evaluation of how effectively one employee can replace another. The following are the most used methods in various practices:

One of first possible – Cosine Similarity measures the similarity between two vectors representing the tasks performed by employees. Cosine similarity calculates the cosine of the angle between two vectors, quantifying how similar their performances are [7]. A higher cosine similarity score indicates that the employees perform similar tasks at comparable levels, suggesting a higher potential for substitutability.

Principal Component Analysis (*PCA*) [8;9] is another effective technique that reduces the dimensionality of the employee data while retaining the most significant variance. By analyzing the skill sets of employees in a reduced space of principal components, PCA helps to identify the most important features that differentiate employees.

The Euclidean Distance method provides a straightforward way to compare employees by treating their skill sets or task performance metrics as vectors in a multi–dimensional space. By calculating the straight – line distance between these vectors, organizations can determine how similar or dissimilar two employees are in terms of their capabilities. A smaller Euclidean distance indicates that the employees are closer in skill, which implies a higher likelihood that one can replace the other with minimal performance degradation [10].

In addition to these traditional methods, Clustering Algorithms such as K – means or hierarchical clustering can group Employees with similar skill sets and task performances. These algorithms partition Employees into clusters where each cluster contains workers whose characteristics and performance metrics closely align. Employees within the same cluster are more likely to substitute for each other seamlessly, as they share similar competencies. Clustering provides a visual and statistical approach to identifying substitution groups within a workforce [11].

More advanced methods involve the use of Machine Learning (*ML*) [12] and Artificial Intelligence (*AI*), which can provide insights into Employee substitutability. For instance, Reinforcement Learning (*RL*) can optimize workforce allocation by continuously learning the most efficient ways to assign tasks based on Employee performance data. Over time, the system adapts and identifies which Employees can effectively replace others, minimizing performance degradation. Similarly, Support Vector Machines (*SVM*) and other machine learning models can be trained on historical data to classify Employees based on their substitutability.

2. Evaluation of Employee substitutability using Cosine Similarity and PCA

Considering the available dataset, which includes Employee ID, Operation ID, the daily count of completed operations, and the time required for each operation, along with the methodological application specifics, this Employee substitutability method was designed those two evaluation approaches: Cosine Similarity and Principal Component Analysis (PCA). The data came from an electronic device designed in 2022 as part of the project “INOSTART” implementation [14].

Cosine Similarity mathematical model

Cosine similarity is a metric used to evaluate the similarity between two vectors [13;14], which in this case represents the performance of Employees based on the number of operations they complete per day and the time required to perform each operation. A

ISSN 2424-5321

2025, 29(1), 35–40. Received 8 December 2025. Accepted 30 December 2025.

higher cosine similarity score indicates that the Employees will perform similar tasks at comparable levels, suggesting a higher potential for substitutability. Employees who exhibit similar characteristics in the principal components are considered more substitutable. This method allows us to assess how closely the work patterns of two Employees are aligned.

Data preparation:

Our dataset comprises Employee ID, Operation ID, the number of operations performed daily, and the time required to complete each operation. Also, can calculate the total time spent by each Employee on each day.

Initially, for each Employee, a vector is constructed, where each dimension represents the total execution count of a specific operation specific Employee and operation times.

For example, if an Employee performs operations 1, 2, and 3, their vector may look like:

$$Vector = [operation1\_count \times operation1\_time, operation2\_count \times operation2\_time, \ldots] \quad (1)$$

Default the *Cosine Similarity* for two arguments *A* and *B* formula:

$$CosineSimilarity(A, B) = \frac{A \cdot B}{\|A\| \cdot \|B\|}; \quad (2)$$

where:

$$A \cdot B = \sum_i A_i \times B_i, \quad (3)$$

$$\|A\| = \sqrt{\sum_i A_i^2}\ , \|B\| = \sqrt{\sum_i B_i^2}\ . \quad (4)$$

3. Principal Component Analysis (PCA) algorithm

Principal Component Analysis (PCA) is a widely used dimensionality reduction technique that transforms high – dimensional data into a lower – dimensional space while retaining as much variance as possible. It achieves this by identifying new orthogonal axes, called principal components, that maximize data variance.

PCA sequence algorithm [9; 15]

1. Data Standardization: The input dataset is standardized by centering each feature. This is achieved by subtracting the meaning of each feature from the corresponding data points and scaling by the standard deviation, ensuring that each feature has a meaning of zero and a variance of one. This step ensures that all features contribute equally to the analysis, especially when the features have different units or scales.

2. Covariance Matrix Calculation: The covariance matrix is computed to capture the relationships between the features of the standardized data. It quantifies the variance of each feature and the covariance between pairs of features, indicating how features vary together. The covariance matrix is symmetric, and its diagonal elements represent variances, while the off – diagonal elements represent covariances.

3. Eigen Decomposition: The covariance matrix undergoes eigen decomposition to identify the principal components of the data. Eigenvectors represent the directions (principal components), and the corresponding eigenvalues indicate the amount of

ISSN 2424-5321

2025, 29(1), 35–40. Received 8 December 2025. Accepted 30 December 2025.

variance explained by each component. Eigenvalues are sorted in descending order to prioritize the components that explain the most variance.

4. Data Projection: The standardized data is projected onto the top *k* eigenvectors (principal components). This reduces the dataset's dimensionality while retaining the most significant variance. The new data representation is obtained by multiplying the standardized data matrix by the matrix of selected eigenvectors, resulting in a lower – dimensional representation of the original data.

5. Recast the Data Along the Principal Components Axes: The recasting of data along the principal components axes refers to the process of reorienting the data in the new coordinate system defined by the principal components. This transformation is accomplished by projecting the original, standardized data onto the principal component vectors, thereby re – representing the data in a new basis where the axes correspond to the directions of maximum variance. This step allows for visualization and analysis in a reduced – dimensional space while preserving the most important features of the original data.

4. Practical implementation

All the methods *Cosine Similarity* and *PCA* were executed in Python using libraries: *numpy, pandas, and sklearn*. Using 5 Employees and with the following 10 – 15 operation each performance testing data from Table 1.

Table 1. Employees work primary data (part)

Rec. No.	Employee ID	Operation ID	Operation count	Operation time (min)	Total time (min)
1	4	10	43	2,60	111,80
2	5	8	50	1,73	86,50
3	3	10	30	2,58	77,40
4	1	10	38	2,70	102,60
5	5	8	36	2,05	73,80
…	…	…	…	…	…
187	2	8	21	1,87	39,27

Using the derived formulas (1 – 4), a Cosine Similarity matrix was computed to assess the similarity between all Employees. The results are presented in Table 2.

From Table 2, it can be observed that Employees 3 and 5 exhibit a high Cosine Similarity (0.836), suggesting they perform similar tasks with comparable efficiency. Employees 2 and 4 show a moderate similarity (0.559), indicating that they share somewhat related work patterns. Employees 1 and 2 (0.371) have low similarity. Employees 1 – 4 or 1– 5 of 2 – 3 or 4 – 5 have very low cosine similarity values (less 0.1), implying that they are engaged in entirely different tasks and can't correctly compare or their performance result very different.

Table 2. Employees Cosine Similarity matrix

Employee ID	1	2	3	4	5
1	–	0.371	0.062	0.083	0.181
2	0.371	–	0.045	0.559	0.365
3	0.062	0.045	–	0.199	0.836
4	0.083	0.559	0.199	–	0.103
5	0.181	0.365	0.836	0.103	–

ISSN 2424-5321

2025, 29(1), 35–40. Received 8 December 2025. Accepted 30 December 2025.

The PCA analysis (Fig.1) reveals insights into the work patterns and efficiency of the Employees. The cosine similarity results suggest that Employees 3 and 5 exhibit the most similar work behavior, as indicated by their high similarity score and different work patterns compared to the other Employees. This implies they may be performing tasks that are closely related in terms of both type and efficiency. On the other hand, Employees 1, 3, and 4 have negligible similarities, meaning their tasks or performance are quite distinct from each other.

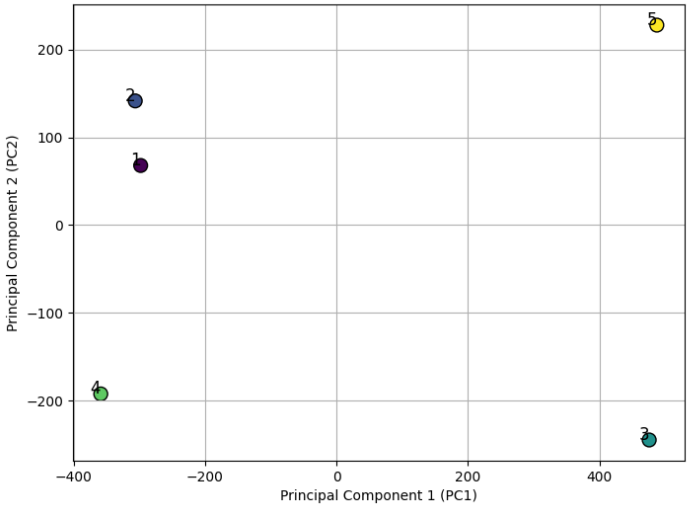

Fig. 1 PCA Analysis of Employee Performance

In conclusion, Employees 3 and 5 appear to be the most distinct in terms of their task performance and efficiency, while Employees 1, 2, and 4 show more aligned work patterns.

This analysis could help in identifying teams' groups with complementary work patterns or in optimizing task assignments to enhance productivity.

The results from the Cosine Similarity and PCA analysis are similar. Both methods show that Employees 3 and 5 are most alike in their tasks and efficiency, with high PC1 similarity scores indicating that they work on related tasks. Employees 2 – 4 or 1 – 2 or 2 – 5 have moderate points, *PC1* or *PC2* has similarity – their work patterns are somewhat related. The very low similarity values for pairs 1 – 4, 1 – 5 and 2 – 3 on both methods indicate distinct task assignments or significant performance.

Conclusion

1. This paper examines various methods for assessing employee substitutability, including Cosine Similarity, Principal Component Analysis, Euclidean Distance, Clustering Algorithms, Machine Learning, and Artificial Intelligence.

2. After evaluating the available data and the suitability of each method, Cosine Similarity and Principal Component Analysis were selected for further analysis. Mathematical models for these methods were developed and a Python – based software tool was created to compute the relevant indicators.

3. The calculations performed using both methods on real data revealed that the two approaches yield similar results but at the same time show different information. Cosine Similarity method – better shows the similarity of employee efficiency, PCA allows you to see groups of employees based on work efficiency and job specifics. When combined

2025, 29(1), 35–40. Received 8 December 2025. Accepted 30 December 2025.

and using both, these results facilitate the classification of employees into groups with similar productivity levels, allowing for the identification of comparable work styles and efficiency. This enables substitutability for employee substitution without decreasing work efficiency and lets the formation of teams or work groups based on aligned work behaviors.

Reference

1. L. Argote, P. Ingram. Knowledge transfer: A basis for competitive advantage in firms. Organizational Behavior and Human Decision Processes, 82(1), p. 150–169, 2000. https://doi.org/10.1006/obhd.2000.2893

2. M. Mehta et al. Achieving World Class Manufacturing Excellence: Integrating Human Factors and Technological Innovation. Sustainability (24), 2024. https://doi.org/10.3390/su162411175

3. Weckenborg et al. Flexibility in manufacturing system design: A review of recent approaches from Operations Research. European Journal of Operational Research, Science Direct, vol. 315, p. 413 – 441, 2024. https://doi.org/10.1016/j.ejor.2023.08.050

4. L. Myeongju et al. Machine Learning – Based Causality Analysis of Human Resource Practices on Firm Performance. Administrative sciences. 14(4), 2024. https://doi.org/10.3390/admsci14040075

5. J. Ingels, J., Maenhout, B. Employee substitutability as a tool to improve the robustness in personnel scheduling. Spectrum (39), Springer, p. 623–658, 2017. https://doi.org/10.1007/s00291-017-0476-0

6. H. Hijry et al. Real Time Worker Stress Prediction in a Smart Factory Assembly Line. *IEEE Access*, vol. 12, p. 116238–116249, 2024, https://doi.org/10.1109/ACCESS.2024.3446875

7. M. Kirişci. New cosine similarity and distance measures for Fermatean fuzzy sets and TOPSIS approach. Knowl. Inf. Syst. 65, Springer, p. 855–868, 2023. https://doi.org/10.1007/s10115-022-01776-4

8. Principal Component Analysis for Special Types of Data. Springer Series in Statistics. Springer, New York, NY, 2002. https://doi.org/10.1007/0-387-22440-8_13

9. M. Greenacre et al. Principal component analysis. Reviews Methods Primers. Nature, vol. 2 (100), 2023. https://doi.org/10.1038/s43586-022-00184-w.

10. C. Khalaf, et al. How to Assess the Transferability of Worker Skills: A Hybrid Clustering Approach. Journal of Regional Analysis and Policy, vol. 51, p. 67–78, 2021. https://doi.org/10.22004/ag.econ.339948

11. N. L. Ratniasihm, R. Diaz. Comparison Of Clustering Algorithm in Employee Training Management Recommendations. *3rd International Conference on Cybernetics and Intelligent System (ICORIS)*, IEEE Xplore, p. 1–4, 2021. https://doi.org/10.1109/ICORIS52787.2021.9649503

12. Z. Mourad, A. Noura and C. Mohamed. Towards a new method for classifying employee performance using machine learning algorithms. *International Conference on Intelligent Systems and Computer Vision (ISCV)*, Fez, Morocco, 2022, p. 1–5, https://doi.org/10.1109/ISCV54655.2022.9806118

13. L. Zhou, et al. Intuitionistic fuzzy ordered weighted cosine similarity measure. Group Decis Negot, vol. 23, p. 879–900, 2014. https://doi.org/10.1007/s10726-013-9359-1

14. J. Ye. Cosine similarity measures for intuitionistic fuzzy sets and their applications. Mathematical and Computer Modelling, ScienceDirect. Vol.53, Iss. 1–2, p. 91–97, 2011. https://doi.org/10.1016/j.mcm.2010.07.022

15. Z. Jaadi. Principal Component Analysis (PCA): A Step–by–Step Explanation Links: https://builtin.com/data-science/step-step-explanation-principal-component-analysis, 2024.

16. R. Plačinskaitė, D. Dervinis. Gamybos duomenų surinkimo, monitoringo ir valdymo sistemos projektavimas, Verslas, naujos technologijos ir sumani visuomenė. Šiaulių valstybinė kolegija, ISSN 2783–6932, 2023.