Linking Sanitation and Water Access to Socio-Economic Outcomes in Indonesia: Insights from Canonical Correlation Analysis

Amsal Anugrah
Statistics Dept.
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11530
amsal.anugrah@binus.ac.id

Ilham Hadi Shahputra
Statistics Dept.
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11530
ilham.shahputra001@binus.ac.id

Kayla Nur Khosyatillah Sudza
Statistics Dept.
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11530
kayla.sudza002@binus.ac.id

***Abstract*—This study investigates the relationship between household sanitation, water access, and community welfare indicators across Indonesian provinces using canonical correlation analysis. Data from the Badan Pusat Statistika Indonesia for 2023, encompassing variables such as educational attainment, life expectancy, poverty rates, handwashing facilities, open defecation practices, and access to safe drinking water, were analyzed. The first canonical correlation was found to be significant (0.870), indicating a strong relationship between improved sanitation, water access, and enhanced community welfare outcomes. Specifically, higher educational attainment and lower poverty rates were positively associated with better sanitation practices and access to safe drinking water. These findings underscore the critical importance of integrated public health policies aimed at improving sanitation and water infrastructure to promote overall community welfare in Indonesia. The study provides actionable insights for policymakers to design targeted interventions that can lead to substantial improvements in health, education, and socio-economic conditions. Future research should explore longitudinal impacts and incorporate additional socio-economic and environmental variables to further elucidate these relationships.**

Keywords—Canonical Correlation Analysis, Community Welfare, Sanitation, Water Access, Indonesia, Public Health Policy

I. INTRODUCTION

Access to clean water and proper sanitation facilities remains a critical challenge across many developing regions, significantly impacting community welfare and socioeconomic development. In Indonesia, disparities in household sanitation and water access can have profound implications on educational outcomes, health, and poverty levels. This study seeks to explore these dynamics using canonical correlation analysis to understand the relationship between household sanitation and water access indicators and community welfare metrics across all provinces in Indonesia.

Recent studies have highlighted the significant impact of environmental health factors on community development. According to the World Health Organization (WHO) 2022 report, improving sanitation and water access in underdeveloped areas can lead to substantial improvements in poverty reduction and educational attainment. The report underscores that achieving universal access to safe and sustainably managed water, sanitation, and hygiene (WASH) is essential to prevent the devastating impacts on health and socio-economic well-being, particularly in vulnerable communities [1].

Furthermore, the UN-Water Global Analysis and Assessment of Sanitation and Drinking-Water (GLAAS) 2022 report reveals that many countries, including Indonesia, need to accelerate their efforts to meet the Sustainable Development Goal (SDG) 6 targets for water and sanitation by 2030. The report highlights the urgent need for increased investment and stronger governance in WASH services to address the gaps in access and ensure the resilience of these systems against climate-related disruptions [2].

This paper aims to extend these findings by examining the specific context of Indonesia, a nation with diverse environmental and socio-economic landscapes. The problem addressed by this research is multifaceted: firstly, how household sanitation and water access indicators correlate with community welfare across Indonesian provinces in 2023; secondly, the extent to which these sanitation and water access factors influence education completion rates, life expectancy, and poverty rates; and thirdly, the presence of significant correlations between sanitation and clean water conditions and welfare indicators.

The objectives of this research are to analyze the relationships between "Household Sanitation and Water Access" indicators and "Community Welfare" outcomes, identify the influences of sanitation and water access on educational, health, and economic metrics, and determine significant correlations to inform policy recommendations aimed at improving the quality of life in Indonesian communities.

By bridging the gap between environmental health factors and socio-economic outcomes, this study contributes to a nuanced understanding of the underlying drivers of inequality in Indonesia, offering evidence-based insights for policymakers and stakeholders involved in regional development and public health.

II. RELATED WORKS

In recent years, numerous studies have explored the intricate relationship between sanitation, water access, and community welfare, highlighting their critical impact on public health and socio-economic development. This section reviews relevant research to contextualize the current study.

A. Sanitation and Quality of life

A significant study conducted by researchers across Ethiopia, Malawi, Mozambique, and Zambia introduced the Sanitation-related Quality of Life index (SanQoL-5), which assessed the impact of various sanitation facilities on the quality of life. The study found that improved sanitation facilities, such as those with solid walls and water seals,

significantly enhanced individuals' perceived quality of life and well-being. This research underscores the importance of robust sanitation infrastructure in improving health outcomes and overall life satisfaction in both urban and rural settings [3].

B. Sanitation and Education

Another study by the Asian Development Bank examined the effects of improved sanitation on child health and educational outcomes. The findings indicated that better sanitation facilities not only reduced child mortality rates but also increased school enrollment and attendance, particularly among girls. This highlights the broader social benefits of sanitation improvements beyond immediate health impacts [3].

C. Water, Sanitation, and Child Nutrition

Research in Southern Punjab, Pakistan, explored the relationship between water, sanitation, hygiene (WASH) practices, and stunting among children under five. The study found a significant association between poor WASH conditions and higher rates of stunting, emphasizing the role of clean water and sanitation in preventing malnutrition and supporting child development. The hierarchical regression analysis showed that factors like water source quality were critical in influencing stunting outcomes [4].

D. Impact on Socio-economic Development

Lastly, studies have consistently shown that access to clean water and improved sanitation is strongly linked to socio-economic development. Improved WASH services contribute to higher productivity, reduced healthcare costs, and enhanced educational outcomes, creating a positive feedback loop that promotes sustainable community welfare [5].

These studies collectively underscore the multifaceted benefits of improved sanitation and water access, reinforcing the relevance of examining their relationship with community welfare indicators in Indonesia. By understanding these dynamics, the current study aims to provide actionable insights for policymakers to enhance public health and socio-economic outcomes through targeted WASH interventions.

III. MATERIALS AND METHOD

A. Dataset

This study utilizes data from Badan Pusat Statistika Indonesia, focusing on recorded observations from 34 provinces in Indonesia in year 2023. The data is categorized into two primary groups: " Community Welfare" and "Household Sanitation and Water Access," each group having $p = 3$ variables. Hereby, the data will be denoted by a variable $X_p^{(i)}$ where p and i are the variables' within-group index and group index.

The categorization of variables into "Household Sanitation and Water Access" and "Community Welfare" is supported by substantial evidence from various studies and reports.

Research has consistently shown that educational attainment, life expectancy, and poverty rate are vital indicators of community welfare. Higher educational levels lead to better job prospects and economic growth, while increased life expectancy reflects overall health improvements. Reducing poverty is essential for achieving sustainable socio-economic development. These indicators collectively provide a comprehensive view of the well-being and socio-economic status of a population [3] [4].

A study by the World Health Organization emphasized the importance of sanitation and clean water in preventing diseases and improving public health outcomes. It highlighted that access to handwashing facilities, safe drinking water, and reduced open defecation are fundamental components of effective WASH (Water, Sanitation, and Hygiene) programs, which are critical for enhancing community health [1] [5].

TABLE I. DATA DESCRIPTION

Group	Variables	Description
Household Sanitation and Water Access (Group 1)	$X_1^{(1)}$	Percentage of Households with Handwashing Facilities with Soap and Water
	$X_2^{(1)}$	Percentage of Households Still Practicing Open Defecation
	$X_3^{(1)}$	Percentage of Households with Safe Drinking Water Sources
Community Welfare (Group 2)	$X_1^{(2)}$	Percentage of High-School Diploma Attainment
	$X_2^{(2)}$	Life Expectancy
	$X_3^{(2)}$	Percentage of Population Living Below Poverty Line

B. Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a multivariate statistical method used to explore the relationships between two sets of variables. This technique identifies pairs of canonical variables linear combinations of the original variables, that maximize the correlation between the two sets [6].

1) **Theory**: Canonical Correlation Analysis was first introduced by Harold Hotelling in 1936. The primary goal of CCA is to find the linear combinations of variables in two datasets that are maximally correlated. These combinations are referred to as canonical variates. CCA extends beyond simple correlation by examining the multivariate relationship, thereby allowing for a more comprehensive analysis of the data structure [7].

2) Formula for Matrix X and Y: The canonical correlation between two sets of variables *X* and *Y* is found by solving the following eigenvalue problem:

$$\boldsymbol{A} = \boldsymbol{P}_{11}^{-\frac{1}{2}} \boldsymbol{P}_{12} \boldsymbol{P}_{22}^{-1} \boldsymbol{P}_{21} \boldsymbol{P}_{11}^{-\frac{1}{2}}$$

$$\boldsymbol{B} = \boldsymbol{P}_{22}^{-\frac{1}{2}} \boldsymbol{P}_{21} \boldsymbol{P}_{11}^{-1} \boldsymbol{P}_{12} \boldsymbol{P}_{22}^{-\frac{1}{2}}$$

Where,

- P_{11} is the covariance matrix of the set $X^{(1)}$.
- P_{22} is the covariance matrix of the set $X^{(2)}$.
- P_{12} is the covariance matrix between $X^{(1)}$ and $X^{(2)}$.
- P_{21} is the transpose of P_{12}, which is the cross-covariance matrix between $X^{(2)}$ and $X^{(1)}$.

3) Canonical Variables (U and V): Canonical variables U are linear combinations of two sets of variables X^1 and X^2 that maximize the correlation between these sets. The first pair U_1 and V_1 has the highest possible correlation, and each subsequent pair is orthogonal to the previous pairs, ensuring uncorrelated linear combinations. These variables are determined using eigenvalue and eigenvector analysis of the covariance matrices [7].

Formula :

$$U_k = e_k' \Sigma_{11}^{-\frac{1}{2}} X^{(1)} = a_k' X^{(1)}$$
$$V_k = f_k' \Sigma_{22}^{-\frac{1}{2}} X^{(2)} = b_k' X^{(2)}$$

4) Assumption

Canonical Correlation Analysis relies on several assumptions:

- Linearity: The relationships between the variables are assumed to be linear.
- Multivariate normality: Both sets of variables are assumed to be normally distributed.
- Homoscedasticity: The variances within each set of variables are assumed to be equal.
- Absence of multicollinearity: The variables within each set should not be highly collinear [8].

C. EigenValue

Eigenvalues in CCA represent the squared canonical correlations. These values provide insight into the amount of shared variance between the canonical variates. Higher eigenvalues indicate stronger relationships between the variable sets. They are found by solving the characteristic equation:

$$\det(A - \lambda I) = 0$$

Where:

- A is a square matrix
- λ is the eigenvalue
- and I is the identity matrix

Eigenvalues are crucial in determining the properties of a matrix and are used in various applications, including stability analysis and vibration analysis

D. Canonical Correlation Test

1) Simultaneous Test

For independence of Two Set Variables

Hypothesis

$$H_0: \Sigma_{xy} = 0$$
$$H_1: \Sigma_{xy} \neq 0$$

Formula

$$\Lambda_1 = \frac{|S|}{|S_{yy}||S_{xx}|} = \frac{|R|}{|R_{yy}||R_{xx}|}$$

Critical Values

$$\Lambda_\alpha = \Lambda_{p,q,n-1-q}$$

where p ≤ q; p, q are number of variables in each group

2) Partial Test

Partial test in canonical correlation analysis refers to the statistical assessment of the significance of a subset of canonical variable pairs while controlling for the influence of other variables. Canonical Correlation Analysis (CCA) is a statistical technique used to explore the relationship between two sets of variables [9].

Hypothesis

$$H_0: \rho_1 = \rho_2 = \rho_3 = 0$$
$$H_1: at\ least\ \rho_1 \neq 0$$

$$H_0: \rho_2 = \rho_3 = 0$$
$$H_1: at\ least\ \rho_2 \neq 0$$

$$H_0: \rho_3 = 0$$
$$H_1: at\ least\ \rho_3 \neq 0$$

Test Statistics

$$\Lambda_k = \prod_{i=k}^{s} (1 - r_i^2)$$

s is the canonical correlation count or minimum variable in group.

E. Canonical Weights

Canonical weights are the coefficients of the linear combinations of the original variables that form the canonical variates. These weights are derived from the eigenvectors of the matrices involved in the eigenvalue problem of CCA [10].

F. Canonical Loadings

Canonical loadings, also known as canonical structure coefficients, measure the correlation between the original variables and the canonical variates. These loadings help interpret the canonical variates by showing the contribution of each original variable to the canonical variates [11].

IV. EXPERIMENT

A. Environment

The experiment is conducted within the RStudio environment using R version 4.2.2. A set of packages beyond the standard base packages were used in facilitating statistical analysis and modeling. The packages are: 'car' (3.1-2), 'CCA' (1.2.2), 'expm' (0.999), 'MVN' (5.9), and 'MASS' (7.3-60).

B. Analysis

The initial phase of our analysis involves the exploration of the data first to check the data' descriptive statistics to aid our understanding of the data itself. Next step, we check whether the data met the assumptions of canonical correlation analysis; if assumptions are not met, attempt to give treatments are made based on which assumptions get violated. The canonical correlation analysis comes after.

The computations in our analysis are done using R following the formula stated in methodology—but the end results get cross-checked by '`cc()`' function from 'CCA' package. This ensures the validity of the analysis.

V. RESULTS AND DISCUSSION

We first examine the descriptive statistics of the data for each variable within the two groups: Community Welfare and Household Sanitation and Water Access. This provides an overview of the central tendency and variability in the data. The descriptive statistics in Table II and III provide an overview of the data distribution for each variable.

Community Welfare Indicators

For $X_1^{(2)}$, on average, 65.81% of households have access to handwashing facilities, with a standard deviation of 77.58%, with a standard deviation of 11.68%. This indicates considerable variation in educational attainment among the provinces. For $X_2^{(2)}$, the average percentage of households practicing open defecation is 5.18 % with a standard deviation of 4.53%, showing variability in health outcomes across different regions. For $X_3^{(2)}$, the mean percentage of households with access to safe drinking water is 88.19%, with a standard deviation of 7.75%, highlighting variability in water access.

Household Sanitation and Water Access Indicators

For $X_1^{(1)}$, the average percentage of individuals who have completed high school across the provinces is 10.69%, indicating differences in hygiene practices. For $X_2^{(1)}$, the mean life expectancy is 70.75 years, with a standard deviation of 2.42 years. For $X_3^{(1)}$, the mean percentage of the population living below the poverty line is 10.09%, with a standard deviation of 5.18%, reflecting significant socio-economic disparities.

The correlation matrix in Table III presents the Pearson correlation coefficients between the variables within the Community Welfare and Household Sanitation and Water Access groups. This matrix is instrumental in understanding the relationships between these variables and identifying potential multicollinearity issues.

The significant positive correlations between household sanitation indicators and community welfare indicators underscore the interconnectedness of these variables. For instance, the percentage of high-school diploma attainment ($X_1^{(2)}$) is positively correlated with the percentage of households with handwashing facilities ($X_1^{(1)}$) and negatively correlated with the percentage of households practicing open defecation ($X_2^{(1)}$). These relationships highlight the broader impacts of sanitation on educational and health outcomes.

TABLE II. DESCRIPTIVE STATISTICS

	Mean	**Standard Deviation**
$\boldsymbol{X_1^{(1)}}$	77.58	11.68
$\boldsymbol{X_2^{(1)}}$	5.18	4.53
$\boldsymbol{X_3^{(1)}}$	88.19	7.75
$\boldsymbol{X_1^{(2)}}$	65.81	10.69
$\boldsymbol{X_2^{(2)}}$	70.75	2.42
$\boldsymbol{X_3^{(2)}}$	10.09	5.18

Before proceeding with the canonical correlation analysis, it is essential to verify that the data meet the assumptions required for this analysis: linearity, normality, and multicollinearity. These assumptions are crucial because violating them can significantly affect the validity and interpretability of the results. If the assumptions are not satisfied, the results may be biased, the statistical tests may lack power, and the overall conclusions drawn from the analysis could be invalid.

Linearity refers to the relationship between the variables within each set being linear. This assumption is necessary because canonical correlation analysis aims to find the linear combinations of variables that maximize the correlation between the sets. If the relationships are non-linear, the canonical variates derived from the analysis may not capture the true nature of the data, leading to misleading conclusions [11]. To check this assumption, we refer to the correlation matrix presented in Table III. The correlation matrix provides Pearson correlation coefficients between the variables within each group. The significant correlations observed in the matrix indicate that the relationships between the majority of variable pairs are linear. Specifically, the significant positive and negative correlations suggest that as one variable increases, the other variable tends to increase or decrease in a predictable linear manner.

Next, we examine the normality assumption, which requires that the variables and their linear combinations follow a multivariate normal distribution. This assumption is crucial because it affects the validity of the statistical tests used in canonical correlation analysis. To assess normality, we employed Mardia's test for skewness and kurtosis, with the results summarized in Table IV. The test results for the Community Welfare group indicate significant deviant from normality, particularly for $X_1^{(1)}$ and $X_2^{(1)}$. The Mardia skewness and kurtosis values for these variables were significantly different from zero, suggesting that the data does not meet the normality assumption.

TABLE III. INTERCORRELATION MATRIX (PEARSON)

	$X_2^{(1)}$	$X_3^{(1)}$	$X_1^{(2)}$	$X_2^{(2)}$	$X_3^{(2)}$
$X_1^{(1)}$	-0.60 ***	0.45 **	0.56 ***	0.51 **	-0.74 ***
$X_2^{(1)}$		-0.39 *	-0.59 ***	-0.64 ***	0.68 ***
$X_3^{(1)}$			0.49 **	0.45 **	-0.28
$X_1^{(2)}$				0.55 ***	-0.57 ***
$X_2^{(2)}$					-0.61 ***

[a.] *, **, ***: the correlation is statistically significant (two-tailed) within 0.05, 0.01, 0.001 level of confidence.

To address this issue, we applied the Box-Cox transformation to $X_1^{(1)}$ and $X_2^{(1)}$. The Box-Cox transformation is a commonly used technique to stabilize variance and make the data more normally distributed. After applying this transformation, the Mardia's test was repeated, and the results in Table V show the transformed variables now follow a multivariate normal distribution.

TABLE IV. MULTIVARIATE NORMALITY RESULT (BEFORE)

Test	Statistic	p-value
	$X^{(1)}$	
Mardia Skewness	64.46	< 0.001
Mardia Kurtosis	6.03	< 0.001
	$X^{(2)}$	
Mardia Skewness	16.60	0.08
Mardia Kurtosis	1.10	0.27

TABLE V. MULTIVARIATE NORMALITY RESULT (AFTER)

Test	Statistic	p-value
	$X^{(1)}$	
Mardia Skewness	11.37	0.33
Mardia Kurtosis	0.27	0.79

The final assumption to check before proceeding with the canonical correlation analysis is multicollinearity. Multicollinearity occurs when there are high inter-correlations among the predictor variables within each set. This can inflate the variance of the estimated coefficients and make it difficult to assess the individual effect of each variable. To assess multicollinearity, we use the Variance Inflation Factor (VIF), which quantifies how much the variance of a regression coefficient is inflated due to collinearity with other predictors.

Table VI presents the VIF values for each variable in our study. VIF value greater than 10 typically indicates significant multicollinearity. However, the VIF values for our variables are well below this threshold, suggesting that multicollinearity is not a concern in this dataset. The VIF values indicate that none of the variables exhibit problematic multicollinearity.

TABLE VI. VARIANCE INFLATION FACTOR (VIF) RESULT

Variable	VIF
$X_1^{(1)}$	1.29
$X_2^{(1)}$	1.20
$X_3^{(1)}$	1.24
$X_1^{(2)}$	1.63
$X_2^{(2)}$	1.75
$X_3^{(2)}$	1.82

Since all assumptions are met, we can continue the analysis with confidence that the results will be reliable and interpretable. Canonical correlation analysis was performed to explore the relationships between $X^{(1)}$ and $X^{(2)}$.

The canonical correlation analysis identified pairs of canonical variates (U and V) that maximize the correlation between the two sets of variables. Table VII summarizes the canonical correlations and their significance testing results. The first canonical correlation (U_1V_1) was found to be 0.870, indicating a strong relationship between the canonical variates of the two variable sets. This high correlation suggests that there is a substantial amount of shared variance between the Community Welfare indicators and the Household Sanitation and Water Access indicators.

To determine the statistical significance of the canonical correlations, we employed the Wilks' Lambda test. This test assesses whether the canonical correlations are significantly different from zero, indicating meaningful relationships between the variable sets.

Table VII presents the Wilks' Lambda statistic and the approximated F-value for each canonical pair. The results showed that the first pair of canonical variates was statistically significant at the 0.05 confidence level (Wilks' $\Lambda = 0.20$, $F = 7.13$, $p < 0.001$). This indicates that the relationship captured by the first pair of canonical variates is highly significant and explains a substantial portion of the shared variance between the two sets of indicators.

TABLE VII. CANONICAL CORRELATION AND SIGNIFICANCE

Canonical Pair	Canonical Correlation	Squared Canonical Correlation	Wilks' Λ-statistic	F-statistic (Approximation)	F-critical
U_1V_1	0.87	0.750	0.20	7.13	2.02
U_2V_2	0.45	0.200	0.79	1.77	2.53
U_3V_3	0.06	0.004	1.00	0.11	4.17

Subsequent canonical pairs were not statistically significant, with F-statistics failing to exceed the critical values at the 0.05 confidence level. This suggests that the additional pairs of canonical variates do not provide significant explanatory power beyond what is captured by the first pair.

The canonical coefficients, which are used to form the canonical variates, are presented in Table VIII. These coefficients represent the weights applied to each original variable to construct the canonical variates. The first pair of canonical variates are of particular interest due to its statistical significance and high canonical correlation.

TABLE VIII. CANONICAL COEFFICIENTS

	$X^{(1)}$		
a_1	0.45	−0.65	0.17
a_2	−0.92	−0.25	0.88
a_3	0.49	0.85	0.67
	$X^{(2)}$		
b_1	−0.35	−0.43	0.41
b_2	0.71	0.61	1.28
b_3	1.00	−1.09	−0.15

These coefficients indicate the contribution of each variable to the respective canonical variates. For instance,

- In the Household Sanitation and Water Access group, the percentage of households with handwashing facilities and the percentage of households with safe drinking water sources have positive coefficients, suggesting that these variables contribute positively to the canonical variate U_1. Conversely, open defecation has a negative coefficient, indicating an inverse relationship within the canonical variate.
- in the Community Welfare group, the positive coefficient for the percentage of people living below poverty line suggest that these variables contribute positively to the canonical variate V_1. The negative coefficient for the high-school diploma attainment and life expectancy indicating an inverse relationship within the canonical variate.

The canonical loadings, also known as structure coefficients, measure the correlation between the original variables and the canonical variates. These loadings provide insight into the contribution of each variable to the canonical variates and help interpret the results of the canonical correlation analysis. Table IX presents the canonical loadings for each variable in relation to the canonical variates.

TABLE IX. CANONICAL LOADINGS

	$X^{(1)}$		
$X^{(1)}U_1$	1.00	−0.37	0.40
$X_2^{(1)}U_2$	−1.00	0.37	−0.40
$X_3^{(1)}U_3$	1.00	−0.37	0.41
	$X^{(2)}$		
$X_1^{(2)}V_1$	−0.92	−0.73	0.82
$X_2^{(2)}V_2$	0.68	0.23	0.20
$X_3^{(2)}V_3$	0.98	0.37	−0.53

These loadings reveal the strength and direction of the relationship between each variable and the canonical variates. For instance,

- in the Household Sanitation and Water Access group, the high positive loadings for the percentage of households with handwashing facilities and percentage of households with safe drinking water sources indicate that these variables are strongly related to the canonical variate U_1.
- Similarly, in the Community Welfare group, the strong positive loading for the percentage of people living below the poverty line indicates a significant relationship with the canonical variate V_1.

Overall, the canonical loadings highlight the key variables that drive the relationships captured by the canonical variates. The significant correlations between these variables and the canonical variates underscore the importance of educational attainment, life expectancy, and poverty levels in the Community Welfare group, and handwashing facilities, open defecation practices, and access to safe drinking water in the Household Sanitation and Water Access group. These findings emphasize the interconnectedness of sanitation, water access, and community welfare, providing valuable insights for policy interventions aimed at improving these critical areas in Indonesian provinces.

VI. CONCLUSION

The primary objective of this study was to explore the relationship between Community Welfare indicators and Household Sanitation and Water Access indicators across Indonesian provinces using canonical correlation analysis. Specifically, the study aimed to understand how variables such as high-school diploma attainment, life expectancy, and poverty rate are related to household sanitation practices and water access.

A. Key Findings

The canonical correlation analysis revealed a strong relationship between the two sets of indicators. The first canonical correlation (U_1V_1) was significant, with a high correlation coefficient of 0.870, indicating a substantial shared variance between Community Welfare indicators and Household Sanitation and Water Access indicators. Specifically, higher educational attainment and lower poverty rates were positively associated with better sanitation practices and improved access to safe drinking water. Conversely, higher rates of open defecation were negatively associated with these welfare metrics.

Subsequent canonical pairs (U_2V_2 and U_3V_3) were not statistically significant, suggesting that additional pairs of canonical variates did not provide significant explanatory power beyond what was captured by the first pair.

The findings highlight the critical impact of household sanitation and water access on community welfare. Improved sanitation practices and access to clean water are strongly associated with better educational outcomes, higher life expectancy, and lower poverty rates. These results underscore the importance of integrated public health and socio-economic policies that focus on enhancing sanitation infrastructure and water accessibility to improve overall community welfare.

B. Policy Recommendations

Based on the study's findings, policymakers should prioritize interventions aimed at improving sanitation and water access. Key recommendations include increasing the availability of handwashing facilities, implementing measures to reduce open defecation, and ensuring the provision of safe drinking water across all provinces. Such targeted interventions can lead to significant improvements in educational attainment, health outcomes, and poverty reduction.

C. Limitations

While this study provides valuable insights, it is limited by the availability and granularity of the data. The analysis is based on cross-sectional data from 2023, which may not capture temporal dynamics or causality. Additionally, potential biases in data reporting and collection could affect the findings. These limitations suggest that the results should be interpreted with caution and complemented with further research.

Future research should consider longitudinal studies to examine the long-term impacts of sanitation and water access improvements on community welfare. Exploring additional variables such as healthcare access, economic policies, and environmental factors could provide a more comprehensive understanding of the dynamics between these indicators. Employing different methodological approaches, such as structural equation modeling, could also enhance the robustness of the findings.

REFERENCES

[1] "Global water, sanitation and hygiene: Annual report 2022," WHO, 2022. [Online]. Available: https://www.who.int/publications/i/item/9789240076297

[2] "Global Analysis and Assessment of Sanitation and Drinking-Water (GLAAS) 2022 report," WHO, 2022. [Online]. Available: https://glaas.who.int/glaas/un-water-global-analysis-and-assessment-of-sanitation-and-drinking-water-(glaas)-2022-report

[3] "The Sanitation-related Quality of Life index (SanQoL-5)," Research Square, 2023

[4] "Impacts of Sanitation on Child Mortality and School Enrollment: A Policy Brief," Asian Development Bank, 2021.

[5] "Relationship of stunting with water, sanitation, and hygiene (WASH) practices among children under the age of five," BMC Public Health, 2023.

[6] Canonical Correlation Analysis | STATA Data Analysis Examples. (n.d.). https://stats.oarc.ucla.edu/stata/dae/canonical-correlation-analysis/

[7] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Cengage Learning.

[8] Iweka, F., & Magnus-Arewa, A. (2018). Canonical Correlation Analysis, A Sine Qua Non for Multivariant Analysis in Educational Research. International Journal of Humanities Social Sciences and Education (IJHSSE), 5(7), 116-126.

[9] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Cengage Learning.

[10] Wang, W., & Zhou, Y.-H. (2021). Eigenvector-based sparse canonical correlation analysis: Fast computation for estimation of multiple canonical vectors. Journal of Statistical Planning and Inference, 211, 1-15. https://doi.org/10.1016/j.jspi.2020.10.002

[11] Mihalik, A., Chapman, J., Adams, R. A., Winter, N. R., Ferreira, F. S., Shawe-Taylor, J., & Mourão-Miranda, J. (2022). Canonical Correlation Analysis and Partial Least Squares for Identifying Brain–Behavior Associations: A Tutorial and a Comparative Study. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 7(8), 785-797. https://doi.org/10.1016/j.bpsc.2022.03.009

Akses Air Bersih dan Sanitasi	Variabel	Link (BPS)	Definisi	Reference Link
Proporsi Air Bersih	Y1	Proporsi Rumah Tangga Yang Memiliki Fasilitas Cuci Tangan Dengan Sabun Dan Air Menurut Provinsi - Tabel Statistik - Badan Pusat Statistik Indonesia (bps.go.id)	Proporsi Rumah Tangga Yang Memiliki Fasilitas Cuci Tangan dengan Sabun dan Air	https://www.sciencedirect.com/science/article/pii/S1438463920305307
Persentase Tempat Buang Air Besar	Y2	https://www.bps.go.id/id/statistics-table/2/MjE3NiMy/persentase-rumah-tangga--yang-masih-mempraktikkan-buang-air-besar-sembarangan--babs--di-tempat-terbuka-menurut-provinsi-dan-tipe-daerah--persen-.html	Persentase Rumah Tangga yang Mempraktekan Defekasi Terbuka	https://ec.europa.eu/echo/files/evaluation/watsan2005/annex_files/USAID/USAID1%20-%20Water%20and%20sanitation%20indicators%20measurement.pdf
Persentase Air Minum	Y3	Persentase Rumah Tangga menurut Provinsi dan Sumber Air Minum Layak - Tabel Statistik - Badan Pusat Statistik Indonesia (bps.go.id)	Persentase Rumah Tangga dengan sumber Air Minum Layak	article.php (kemdikbud.go.id)

Kesejahteraan Masyarakat	Variabel	Link (BPS)	Definisi	Reference Link
Tingkat Penyelesaian Pendidikan	X1	https://www.bps.go.id/id/statistics-table/2/MTk4MCMy/tingkat-penyelesaian-pendidikan-menurut-jenjang-pendidikan-dan-provinsi.html	Tingginya suatu tingkat penyelesaian pendidikan pada suatu daerah.	https://dspace.uii.ac.id/bitstream/handle/123456789/48840/19313164.pdf?sequence=1&isAllowed=y

Angka Harapan Hidup	X2	https://www.bps.go.id/id/statistics-table/2/NTAxIzI=/angka-harapan-hidup--ahh--menurut-provinsi-dan-jenis-kelamin--tahun-.html	Angka harapan hidup pada tiap-tiap provinsi menurut jenis kelamin (rata-rata).	https://jurnal.dharmawangsa.ac.id/index.php/bisnet/article/viewFile/3884/2597
Persentase Penduduk Miskin	X3	https://www.bps.go.id/id/statistics-table/2/MTkyIzI=/persentase-penduduk-miskin--p0--menurut-provinsi-dan-daerah--persen-.html	Persentase penduduk miskin.	https://journal.unespadang.ac.id/JIEE/article/view/93/92

38 Provinsi	Y1	Y2	Y3	X1	X2	X3
ACEH	72,59	9	89,74	74,46	70,385	14,45
SUMATERA UTARA	73,92	4,78	92,19	74,43	70,03	8,15
SUMATERA BARAT	88,77	9,31	85,59	68,64	70,24	5,95
RIAU	72,29	2,64	90,47	67,79	72,285	6,68
JAMBI	73,41	5,45	80,02	66,62	71,82	7,58
SUMATERA SELATAN	75,24	5,38	87,19	64,81	70,71	11,78
BENGKULU	81,41	5,12	73,08	63,41	69,965	14,04
LAMPUNG	79,39	1,64	82,78	64,54	71,295	11,11
KEP. BANGKA BELITUNG	88,17	1,85	81,64	68,96	71,28	4,52
KEP. RIAU	85,97	0,46	92,1	78,97	70,965	5,69
DKI JAKARTA	76,83	0,13	99,42	88,1	73,745	4,44
JAWA BARAT	82,03	2,52	93,86	66,47	74,1	7,62
JAWA TENGAH	86,15	2,9	93,76	58,35	74,785	10,77
DI YOGYAKARTA	84,29	0,59	96,69	89,69	75,215	11,04
JAWA TIMUR	83,37	5,3	96,01	68,65	72,16	10,35
BANTEN	82,79	4,95	92,95	70,07	70,82	6,17
BALI	90,54	2,61	98,31	76,51	73,03	4,25
NUSA TENGGARA BARAT	75,48	8,49	96,03	63,66	67,52	13,85
NUSA TENGGARA TIMUR	43,5	5,84	88,35	43,46	67,81	19,96

KALIMANTAN BARAT	77,59	6,08	82,08	55,58	71,365	6,71
KALIMANTAN TENGAH	75,67	2,53	77,72	63,93	70,325	5,11
KALIMANTAN SELATAN	83,31	1,93	76,29	68,35	69,47	4,29
KALIMANTAN TIMUR	78,87	1,06	87,9	73,63	74,79	6,11
KALIMANTAN UTARA	76,24	1,22	90,19	59,5	72,735	6,45
SULAWESI UTARA	86,48	5,67	94,37	67,57	72,45	7,38
SULAWESI TENGAH	78,81	10,4	86,85	55,69	69,22	12,41
SULAWESI SELATAN	85,52	2,26	92,12	67,41	71,26	8,7
SULAWESI TENGGARA	84,74	4,34	94,8	68,28	71,525	11,43
GORONTALO	82,44	9,42	96	46,19	68,885	15,15
SULAWESI BARAT	77,18	8,88	79,86	54,79	66,055	11,49
MALUKU	74,11	10,41	92,98	75,01	66,835	16,42
MALUKU UTARA	83,62	5,37	89,01	64,61	69,155	6,46
PAPUA BARAT	65,38	3,38	81,57	59,99	66,845	20,49
PAPUA	31,78	24,3	66,49	39,5	66,49	26,03