Modeling Air Quality Determinants in Indonesia Using Generalized Linear Models for Sustainable Development

Restu Arisanti[1]*[ID], Aisya Putri Syarnurli[2][ID], Dianda Destin[2][ID], Maharani Rizki Febrianti[2][ID], Yuyun Hidayat[1][ID], Irlandia Ginanjar[1][ID], Titi Purwandari[1][ID]

[1] Department of Statistics, Padjadjaran University, 45363 Jatinangor, Indonesia
[2] Bachelor Degree Program of Statistics Department, Padjadjaran University, 45363 Jatinangor, Indonesia

* Correspondence: Restu Arisanti (r.arisanti@unpad.ac.id)

Abstract: The Sustainable Development Goals (SDGs), particularly Goal 11 (Sustainable Cities and Communities) and Goal 13 (Climate Action), underscore the interconnectedness between air quality and climate change. Escalating levels of air pollution in both urban and rural regions of Indonesia necessitate a deeper understanding of the factors contributing to air quality degradation. This study employs a generalized linear modeling approach, specifically focusing on ordinal logistic regression, to explore the determinants influencing the Air Quality Index (AQI) across 34 provinces in Indonesia. Key predictors, including motor vehicle density, population density, Greenhouse Gas (GHG) emissions, and forest cover, are analyzed to assess their impact on air quality levels. The findings indicate that the number of motor vehicles and the extent of forest cover are significant predictors of air quality. Elevated motor vehicle density is shown to deteriorate the AQI, while larger forest cover areas are associated with improvements in air quality. These results emphasize the importance of targeted environmental interventions, particularly those aimed at reducing vehicle emissions and preserving forest ecosystems. The study highlights the need for the development and enforcement of policies that promote sustainable urban mobility and forest conservation to mitigate air pollution. By providing a comprehensive statistical framework through ordinal logistic regression, this research offers actionable insights for policymakers. The findings can guide the formulation of effective environmental management strategies, supporting efforts to achieve sustainable development objectives. Moreover, this study demonstrates the relevance of adopting rigorous statistical models to address complex environmental challenges, contributing to the broader discourse on sustainability and climate action.

Keywords: Air Quality Index (AQI); Ordinal logistic regression; Generalized linear models; Environmental determinants; Sustainable development; Air pollution; Greenhouse Gas (GHG) emissions

1. Introduction

Air is a mixture of gases that covers the Earth's surface. Its composition consists of 78% nitrogen, 20% oxygen, 0.93% argon, and 0.30% carbon dioxide (CO_2), along with trace amounts of other gases (Mehrpooya et al., 2020). Air is essential for sustaining life across all living organisms on Earth. The primary components of air are nitrogen at 78.09% and oxygen at 20.94% (Vamsi, 2020). Although these proportions remain relatively constant, they are influenced by environmental factors such as air temperature and humidity. Under certain conditions, the composition of air can change due to the introduction of foreign substances, which alter its physical and chemical properties. This phenomenon is referred to as air pollution. When pollutants persist in the atmosphere at significant concentrations for extended periods, they disrupt environmental sustainability, posing risks to the survival of plants, animals, and humans.

Air quality is a parameter used to quantify the condition of the air in a specific region or location. Multiple gases, including carbon monoxide (CO), CO_2, sulfur dioxide (SO_2), automobile emissions, industrial emissions, and domestic air waste, can contribute to a decline in air quality. Deviation from optimal air quality has detrimental

effects on human health. Outdoor air quality, sometimes known as ambient air, is of primary concern to many individuals. Human exposure to atmospheric air pollutants directly affects both human health and environmental parameters. The main atmospheric pollutants consist of CO, dust particles, nitrogen oxides, and lead (Pb), principally generated by motor vehicle emissions (Zamorategui-Molina et al., 2021).

Based on Government Regulation of the Republic of Indonesia Number 41 of 1999, air as a natural resource that affects human life and other living things must be maintained and preserved for the maintenance of human health and welfare and protection for other living things. Recently, environmental issues in Indonesia have become a serious concern of the Indonesian people, and one of these environmental problems that has recently been highlighted is air pollution. The degradation of air quality in densely populated and economically disadvantaged regions highlights the broader need for sustainable policies that address both environmental and social challenges.

Over the past decade, Indonesia has experienced an increase in particulate pollution. Currently, more than 93% of Indonesia's 262 million population lives in areas where annual average PM2.5 levels exceed the WHO guideline threshold, and average life expectancy would be reduced by 2.5 years relative to what it would be if the WHO guideline threshold was 10 g/m^3. Some areas in Indonesia have air pollution exceeding the national air quality average. In Jakarta, the average resident would live 5.5 years shorter if PM2.5 levels remained at current levels. In total, Indonesia's current population will lose about 643 million years of life due to particulate pollution (Lee & Greenstone, 2021).

IQ Air data in October 2015 revealed that air pollution in Indonesia was triggered by nearly 5,000 cases of simultaneous forest and peatland fires. In one day, about 80 million metric tons of CO_2 were released. Besides forest fires, pollution is also caused by the transportation sector and energy production. In Jakarta, much of the air pollution comes from rapidly increasing emissions from coal-fired power plants. This condition is exacerbated by emissions from transportation, households, the construction industry, road dust, and uncontrolled burning of forests and agricultural land. All of these factors occur daily and impact the lives of its 25 million inhabitants. In 2017, air quality in Indonesia was recorded with a PM2.5 reading of 29.7 $\mu g/m^3$, indicating "moderate" air quality. In 2018, this figure increased to 45.3 $\mu g/m^3$, which means "unhealthy for sensitive groups."

In 2019, Indonesia was ranked as the 6th most polluted country out of 98 countries worldwide. The average PM2.5 concentration in Indonesia in 2019 reached 49.4 µg/m³. In South Tangerang, for 10 months of the year, air quality is classified as "Unhealthy" with concentrations between 55.5 and 150.4 µg/m³, while the remaining 2 months fall into the "Unhealthy for Sensitive Groups" category with concentrations between 35.5 and 55.4 µg/m³. In Pekanbaru City, the air quality in September 2019 was recorded as "Very Unhealthy" with a concentration of 214.9 µg/m³. Based on data from weather stations in Central Jakarta during the first 6 months of 2019, there was a slight improvement in measured air quality, with PM2.5 concentrations reaching 28.57 µg/m³. In 2020, the PM2.5 level in Jakarta was 24.33 µg/m³, lower than the same period in 2019 which was recorded at 28.57 µg/m³ (Kumar et al., 2022). In 2023, according to the IQAir World Air Quality Report released in March 2023, Indonesia ranked 14th as the country with the highest air pollution levels in the world and noted that Indonesia topped the list as the country with the highest pollution levels in the Southeast Asia region (IQAir, 2023). This increase in air pollution is of great concern as it can adversely affect human respiratory health. Improvement of air quality has a direct impact on the quality of life in urban environments, as does the focus of research conducted to evaluate urban livability. In Indonesia, air pollution is influenced by a range of socio-economic and environmental determinants. Furthermore, better air quality can significantly improve people's health and well-being, which is in line with broader sustainability goals (Poveda, 2023). Maintaining high air quality is essential not only for the environment but also for achieving a sustainable future that is safe and just for all living beings (Orr & Kish, 2022).
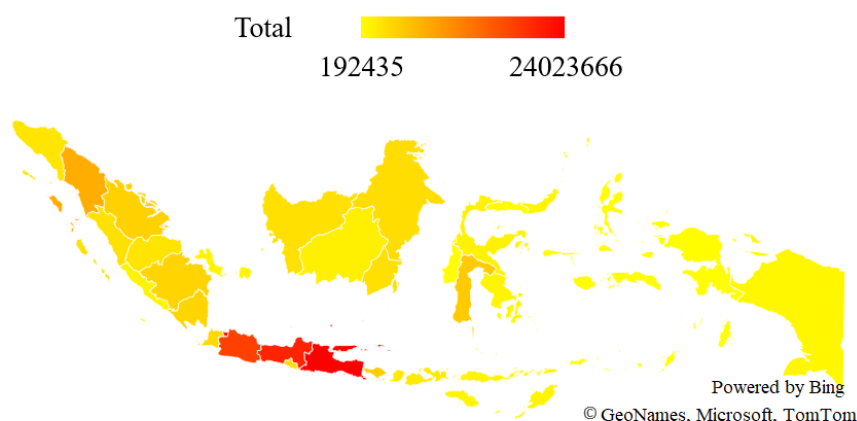
Figure 1. Maps of the number of motor vehicles in Indonesia in 2022

This research highlights the worsening air quality that is influenced by several factors. The factor of population density and the increase in the number of motor vehicles is one of the reasons that air quality is getting worse because it causes the number of emissions produced to increase and increase. The Korlantas Polri (Indonesian Police) noted that the total population of motor vehicles in Indonesia that were active until February 9, 2023 reached 153,400,392 units. This figure includes motorcycles by 87% (Shidqi & Supriyatna, 2023). It can be seen in Figure 1 that the island of Java has a darker color because of the largest number of motor vehicles in Indonesia.

The increasing number of motor vehicles in urban areas will result in a decrease in clean air quality due to emissions from fuel combustion and become emissions that have a large contribution to the concentration of nitrous oxide (N_2O) and CO in the air, which amounts to more than 50% (Wang et al., 2022). and based on the results of previous literature, examining the effect of population growth on air quality gives negative results, which means that if population growth increases, the air index will decrease or get lower (Vohra et al., 2022).

In addition, GHG emissions are also a serious problem affecting air quality in Indonesia. Indonesia is one of the countries with the highest deforestation rate in the world, and the forest and land fires that occur every year also produce large GHG emissions. According to data from the Central Bureau of Statistics (BPS), the area of forest land in Indonesia decreased by 6.02 million hectares between 2009 and 2019. This contributes to an increase in GHG emissions, especially from the land use/forest sector, which accounts for around 35.0% of total global GHG emissions in 2020 (Grassi et al., 2022).

Based on the complexity of the factors affecting air quality and the need to understand their relative influence as previously described, the selection of appropriate analytical methods is very important. In this study, the development of an ordinal logistic regression model is the main modelling technique to analyze the determinants of air quality in Indonesia. Ordinal logistic regression is appropriate when the dependent variable is ordinal, meaning that it has a natural order but the distance between levels is not the same. Furthermore, in the context of air quality, each category in the ordinal data reflects a different severity of pollution.

This model is specifically designed to handle dependent variables that are ordinal in nature. Therefore, the ordinal logistic regression model can be used to explore the relationship between determinants and ordered air quality categories. Furthermore, the advantage of ordinal logistic regression lies in its ability to analyze and quantify the intrinsic properties of ordinal data, which cannot be well captured by other linear regression models. In addition, this model also has a clear interpretation of the odds ratio, allowing for a deeper understanding of the influence of the independent variables on changes in air quality categories.

This study aims to explore the factors that influence air quality in Indonesia using ordinal logistic regression analysis. The main research question is to identify whether factors such as percentage of motorized vehicles, population density, GHG emissions, and area of forest cover significantly affect the IKU (abbreviation AQI in Indonesia). In addition to aforementioned factors, this study also aims to determine the probability and probability ratio of the influence of each factor on air quality in Indonesia. The hypothesis of this study states that the higher the number of motorized vehicles, increased population density, greater GHG emissions, and reduced forest cover will lead to worse air quality.

2. Methodology

2.1 Data Source

The data used in this study are secondary data related to the environment, climate change, and AQI in 2022. The research variables used are quantitative variables for independent variables (X) and qualitative variables for dependent variable (Y).

2.1.1 Dependent variable (Y)

The dependent variable of this study is air quality in 2023. This air quality data comes from the performance report data of the Directorate of Air Pollution Control, Directorate General of Pollution Control and Environmental Damage in 2022 (Direktorat Pengendalian Pencemaran Udara, 2022). The index is grouped into five groups based on the publication of the Ministry of Environment and Forestry entitled Indonesia Environmental Quality Index 2019 (KLHK, 2019). Table 1 shows the AQI (IKU) categories, which are used to measure the level of air cleanliness and its impact on human health. Table 1 shows the categories that classify air quality into levels based on the concentration of pollutants in the air.

Table 1. AQI categories

IKU Group	Indicator	Description
1	IKU > 91	Excellent
2	81 < IKU ≤ 91	Good
3	71 < IKU ≤ 81	Fair
4	61 < IKU ≤ 71	Poor
5	51 < IKU ≤ 61	Very poor

2.1.2 Independent variables (X)

The independent variables used are the number of vehicles, population density, GHG emissions, and forest cover area. All independent variables are based on data from 2023.

(a) Percentage of Motor Vehicles (X_1)

This variable is secondary data obtained from the website of BPS-Statistics Indonesia based on the year 2022 (BPS-Statistics Indonesia, 2022a).

(b) Percentage of GHG Emissions (X_2)

This variable is secondary data obtained from the website of the Ministry of Environment and Forestry (KLHK) covering data from all provinces in Indonesia (Pelaporan Dan Verifikasi, 2022).
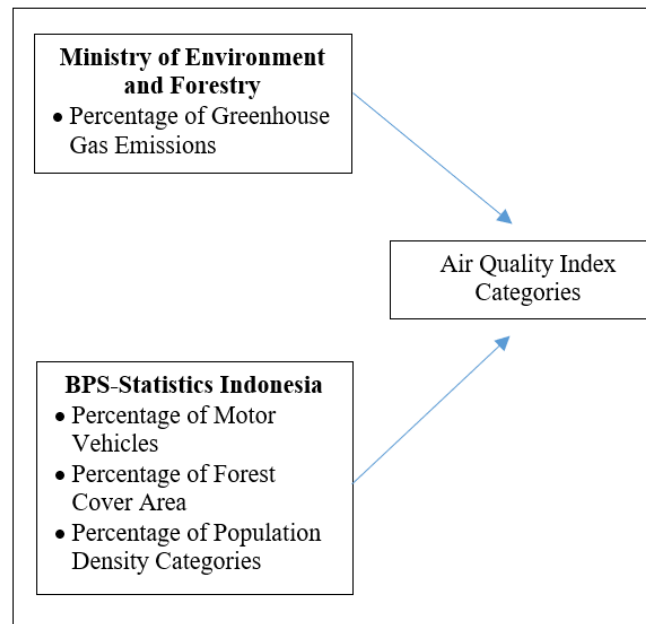
(c) Percentage of Forest Cover Area (X_3)

This variable is secondary data obtained from the website of BPS-Statistics Indonesia for each province in 2022 (BPS-Statistics Indonesia, 2022b).

(d) Percentage of Population Density (X_4)

This variable is secondary data obtained from the website of BPS-Statistics Indonesia based on provinces in 2022 (BPS-Statistics Indonesia, 2022c).

2.2 Data Processing

The data processing stage is using RStudio software, which attaches syntax and text mining. Figure 2 below illustrates the relationship between the various factors that influence the AQI (IKU) in Indonesia.

Figure 2. Influencing factors of AQI in Indonesia

As shown in Figure 2, the percentage of GHG emissions, percentage of motor vehicles, forest cover area, and population density play a direct role in determining the air quality category. The arrows pointing to these IKU categories indicate that changes in each of these factors can impact air quality, which is then categorized based on the level of health risk it poses.

2.3 Data Analysis

2.3.1 Generalized linear model

GLM is an extension of the linear regression model, which allows the distribution of the response variable not to be normally distributed. The distribution of the response variable in GLM belongs to the exponential family. This model uses a link function to relate the expectation of the response variable to a linear combination of the predictors. GLM does not require classical regression assumptions (Arisanti et al., 2024).

A special case of GLM where the response variable has levels of categories but the distances between the categories are not necessarily equal is called Ordinal Logistic Regression. With this approach, we can model the cumulative probability of an ordinal response based on the predictor factors, making it very useful in situations where the response outcomes are in a certain order.

2.3.2 Ordinal logistic regression

Ordinal logistic regression is a statistical method for analysing ordinal data scale dependent variables consisting of three or more categories and categorical or continuous data scale independent variables consisting of two or more variables (Hosmer et al., 2013). The ordinal logistic regression model is a cumulative logit model in which the ordinal nature of the response Y will be used as a cumulative probability. If the independent variable Y has G ordinal scale categories and x_i expresses the vector of the p i-th observation variables, then $x_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]^T$ with i = 1,2,3,.., then the logit model will be stated as follows:

$$\text{Logit}\left[P\left(Y_i \leq g \mid x_i\right)\right] = \alpha_g + x_i^T \beta, g = 1, 2, 3, \ldots, G-1 \tag{1}$$

where, $P(Y_i\ g|x_i)$ is the cumulative probability that is less than or equal to the gth category if x_i is known. g is an intercept parameter, with $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{G-1}$ and $\beta = [\beta_1, \beta_2, \ldots, \beta_p]^T$ is the vector of regression coefficients corresponding to $x_1, x_2, \ldots, x_p$.

Then the ordinal logistic regression model can be stated as follows:

$$\text{Logit}\left[P\left(Y_i \leq g \mid x_i\right)\right] = \ln\left[\frac{P\left(Y_i \leq g \mid x_i\right)}{1 - P\left(Y_i \leq g \mid x_i\right)}\right], g = 1, 2, 3, \ldots, G-1 \tag{2}$$

So, the ordinal logistic regression model can be expressed as follows:

$$\text{Logit}\left[P\left(Y_i \leq g \mid x_i\right)\right] = \ln\left[\frac{P\left(Y_i \leq g \mid x_i\right)}{1 - P\left(Y_i \leq g \mid x_i\right)}\right], = \alpha_g + x_i^T \beta, g = 1, 2, 3, \ldots, G-1 \tag{3}$$

The probability for each response category can be expressed as follows:

$$\pi_g\left(x_i\right) = \frac{\exp\left(\alpha_g + x_i^T \beta\right)}{1 + \left(\alpha_g + x_i^T \beta\right)} - \frac{\exp\left(\alpha_{g-1} + x_i^T \beta\right)}{1 + \left(\alpha_{g-1} + x_i^T \beta\right)}, g = 1, 2, \ldots, G \tag{4}$$

2.3.3 Parameter estimation

Parameter estimation in ordinal logistic regression can be done using the Maximum Likelihood Estimation (MLE) method. If n sample vectors of random variables $Y_1, Y_2, \ldots, Y_n$, with $Y_i = [Y_{i1}, Y_{i2}, \ldots, Y_{ip}]^T$, are distributed multinomially with the probability of the g-th category outcome being $\pi_g(x_i)$.

Then the likelihood function is:

$$\ell(\theta) = \prod_{i=1}^{n}\prod_{g=1}^{G}\left(\pi_g\left(x_i\right)\right)^{y_{ig}} \prod_{i=1}^{n}\prod_{g=1}^{G}\left[\frac{\exp\left(\alpha_g + x_i^T \beta\right)}{1 + \exp\left(\alpha_g + x_i^T \beta\right)} - \frac{\exp\left(\alpha_{g-1} + x_i^T \beta\right)}{1 + \exp\left(\alpha_{g-1} + x_i^T \beta\right)}\right] \tag{5}$$

Then, a log transformation is applied to the likelihood function, resulting in the log-likelihood function as follows:

$$L(\theta) = \ln[\ell(\theta)] = \sum_{i=1}^{n}\sum_{g=1}^{G}\ln\left[\frac{\exp\left(\alpha_g + x_i^T \beta\right)}{1 + \left(\alpha_g + x_i^T \beta\right)} - \frac{\exp\left(\alpha_{g-1} + x_i^T \beta\right)}{1 + \left(\alpha_{g-1} + x_i^T \beta\right)}\right] \tag{6}$$

There are 4 categories of the response variable Y (G = 4), so the log-likelihood function becomes:

$$
\begin{aligned}
L(\theta) = \ & Y_{i1}\ ln\left[\frac{exp\left(\alpha_g + x_i^T \beta\right)}{1 + \left(\alpha_g + x_i^T \beta\right)} - \frac{exp\left(\alpha_{g-1} + x_i^T \beta\right)}{1 + \left(\alpha_{g-1} + x_i^T \beta\right)}\right] \\
& + Y_{i2}\ ln\left[\frac{exp\left(\alpha_g + x_i^T \beta\right)}{1 + \left(\alpha_g + x_i^T \beta\right)} - \frac{exp\left(\alpha_{g-1} + x_i^T \beta\right)}{1 + \left(\alpha_{g-1} + x_i^T \beta\right)}\right] \\
& + Y_{i3}\ ln\left[\frac{exp\left(\alpha_g + x_i^T \beta\right)}{1 + \left(\alpha_g + x_i^T \beta\right)} - \frac{exp\left(\alpha_{g-1} + x_i^T \beta\right)}{1 + \left(\alpha_{g-1} + x_i^T \beta\right)}\right] \\
& + Y_{i4}\ ln\left[\frac{exp\left(\alpha_g + x_i^T \beta\right)}{1 + \left(\alpha_g + x_i^T \beta\right)} - \frac{exp\left(\alpha_{g-1} + x_i^T \beta\right)}{1 + \left(\alpha_{g-1} + x_i^T \beta\right)}\right]
\end{aligned} \tag{7}
$$

However, estimating parameters from the nonlinear regression equation is not easy using MLE calculations, thus an iterative method is needed (Jung & Lee, 2021). The iterative method used is the Fisher-scoring algorithm.

2.3.4 Best model selection

The best model can be determined by looking at the RMSE or Deviance values. The Root Mean Squared Error (RMSE) is a method to measure the average magnitude of errors made by the model when predicting observation outcomes. By using RMSE, it can be identified which model provides high accuracy in making predictions. The lower the RMSE or Deviance value, the better the model and the more accurate the results (Hodson, 2022).

Systematically, the formula is written as follows:

$$RMSE = \sqrt{\left(\frac{\sum(y_i - \hat{y}_l)}{n}\right)} \tag{8}$$

Or

$$Devians = -2\log\left(L_C - L_F\right) \tag{9}$$

2.3.5 Simultaneous test

In ordinal logistic regression, the simultaneous test (or often called the overall model test) is used to evaluate whether the model used is significantly better than the null model, which has no independent variables at all. This test uses the likelihood ratio test (LRT) (McCullagh & dan Nelder, 1989). This test aims to determine whether there is a significant relationship between the ordinal dependent variable and a set of independent variables. If the results of the simultaneous test show a significant p-value, this means that there is at least one independent variable that has a significant relationship with the dependent variable, and the model used can be considered better in explaining the data compared to the model that does not contain independent variables.
Hypothesis:
$H_0: \beta_1, \beta_2, \beta_3, \beta_4 = 0$ (No simultaneous effect of factors)
$H_1: at\ least\ one\ \beta_k \neq 0$ (There is a simultaneous effect of factors)
Significance level: 5%
Test statistic: G^2 test or LRT

$$G^2 = -2\left(L_0 - L_1\right) \sim X_k^2 \tag{10}$$

where:
L_0: log-likelihood of the null model.
L_1: log-likelihood of the maximum model.
The G^2 statistic, theoretically follows a X^2 distribution with k degrees of freedom. When the test statistic G^2 is $G^2 > X^2$, H_0 is rejected.

2.3.6 Partial test

Partial test is used to test the significance of each independent variable in the ordinal logistic regression model. This test aims to determine whether there is a significant contribution of each independent variable individually to the dependent variable, of course, after considering the influence of other variables in the model. Partial tests are conducted as evaluating the relative contribution of each predictor variable individually in relation to the response using the following test statistics (Pisică et al., 2022):
Hypothesis:
$H_0: \beta_i = 0$ (The i-th independent variable has no significant effect on the independent variable)
$H_1: \beta_i \neq 0$ (The i-th independent variable has a significant effect on the independent variable)
Significance level: 5%
Test statistics: Wald or W test

$$W = \frac{\beta_l}{SE\left(\beta_l\right)} \sim Z \tag{11}$$

In theory, this W statistic follows the standard normal distribution if H_0 is true. If the statistical value $|W| > Z_{\alpha/2}$ or p-value < then H_0 is rejected.

2.3.7 Goodness of fit

The goodness of fit test is carried out to determine the suitability of a model. Testing the suitability of the model can be done by comparing the observed value for a subject with the predicted value for that subject (Chen et al., 2020). This test statistic used is the LRT statistic, with the following hypothesis:

Hypothesis:
H_0: the model is appropriate
H_1: the model is not appropriate
Significance level: 5%
Test statistics: LRT test

$$LRT = -2(L_0 - L_1) \sim X^2_{G-1} \qquad (12)$$

Description:
L_0: log-likelihood of the tested model
L_1: log-likelihood of the model with indicator variables
G: Number of groups

The null hypothesis (model fit) is rejected when the LR value is greater than the value of the X^2_{G-1} or the p-value is smaller than the alpha value.

2.3.8 Model interpretation

Model interpretation is a form of defining the unit of change in the independent variable caused by the independent variable and determining the functional relationship between the independent variable and the independent variable. To make it easier to interpret the model, the odds ratio value was used (Arisanti et al., 2024). The odds ratio value used for the interpretation of the ordinal logistic regression coefficient is a value that shows the comparison of the level of tendency of two or more categories in one independent variable with one of the categories used as a comparison (Williams & Quiroz, 2020).

$$
\begin{aligned}
\text{Odds ratio } &= \frac{P(Y \geq j \mid x=a)/(P(Y < j \mid x=a)}{P(Y \leq j \mid x=a)/(P(Y > j \mid x=a)} \\
&= \frac{\left[\dfrac{\exp(a_j + b_k)}{1+\exp(a_j+b_k)}\right] \Big/ \left[\dfrac{1}{1+\exp(a_j)}\right]}{\left[\dfrac{\exp(a_j)}{1+\exp(a_j)}\right] \Big/ \left[\dfrac{1}{1+\exp(a_j)}\right]} = \exp(\beta_k)
\end{aligned}
\qquad (13)
$$

3. Results

3.1 Statistics Descriptive
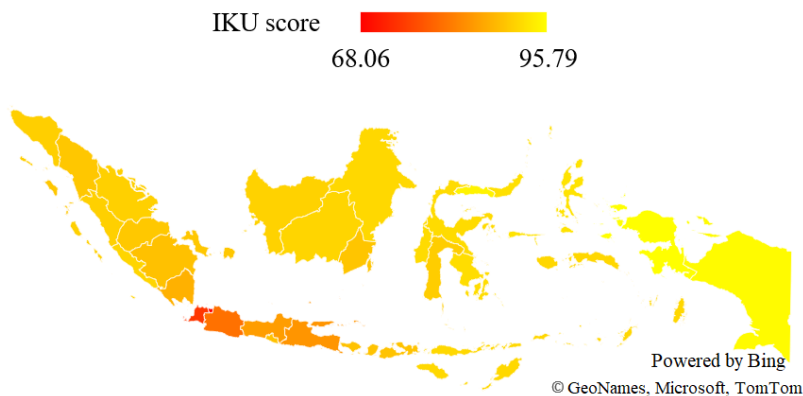
Figure 3. Maps of AQI in Indonesia

Descriptive statistical analysis aims to identify data characteristics by displaying minimum, maximum, mean, median, first to third quartile values, as well as measuring correlations between variables. Figure 3 shows a map of the AQI in Indonesia, with each province colored according to its air pollution level. The level of air pollution is indicated by a gradation of red, where a deeper red color indicates a higher level of pollution.

From Figure 3, it can be seen that DKI Jakarta province has the most intense red color, indicating that it experiences the highest levels of air pollution compared to other provinces. Meanwhile, the surrounding provinces have a lighter shade of red, indicating lower pollution levels.
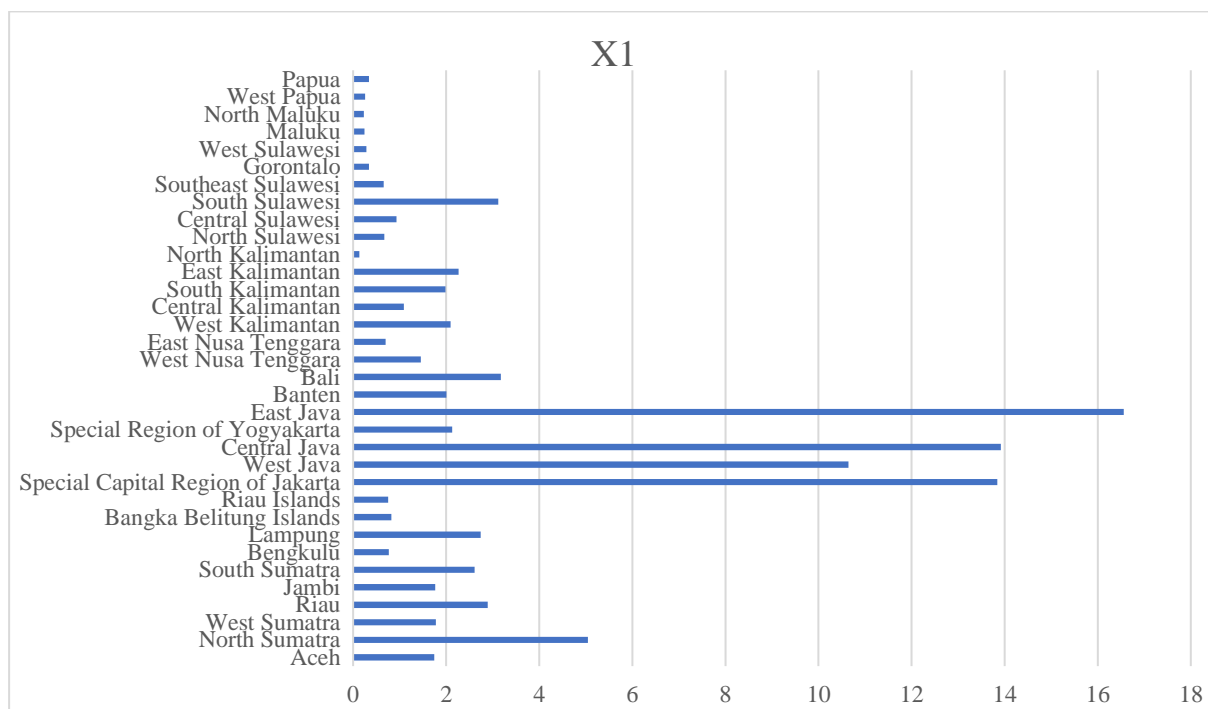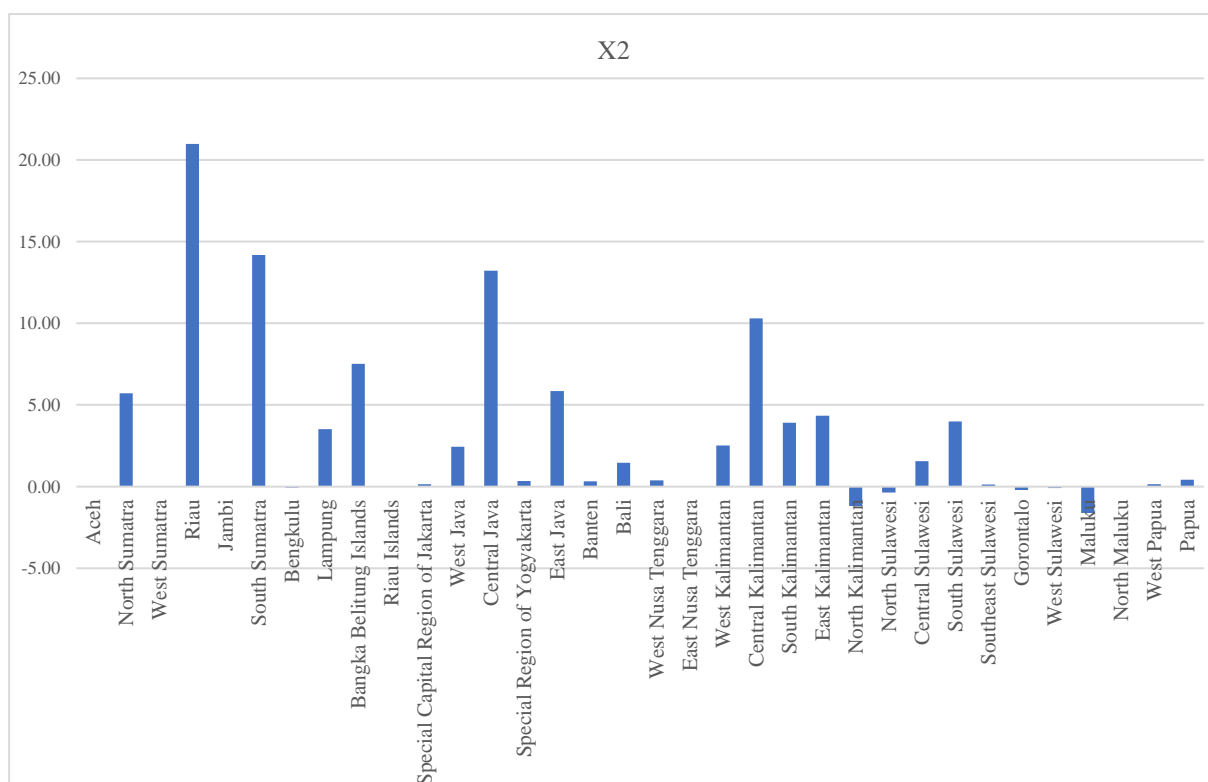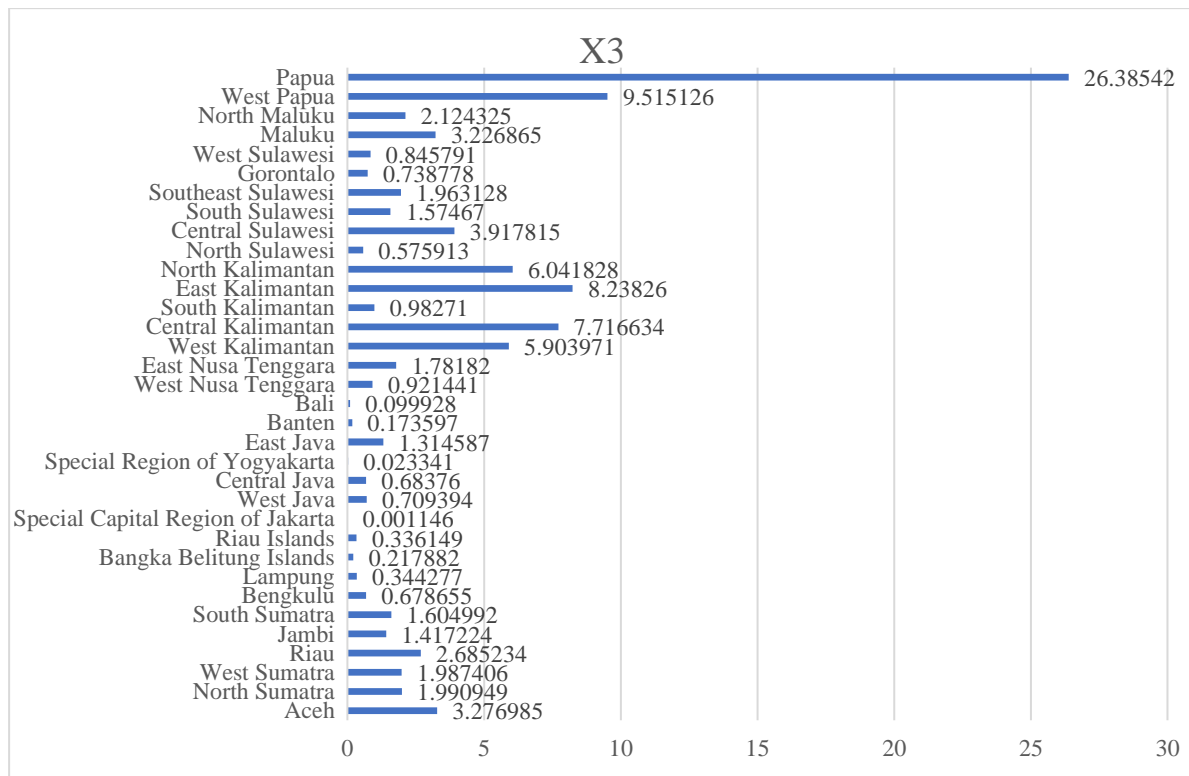

Figure 4. Percentage of motor vehicles (X_1)

Figure 5. Percentage of GHG emissions (X_2)

Figure 6. Percentage of forest cover area (X_3)

Data on the percentage of the number of motor vehicles in Indonesia refers to the ratio between the number of motor vehicles registered or operating in a particular area and the total number of motor vehicles in Indonesia, expressed as a percentage (%). Through Figure 4, it can be seen on the bar chart that East Java has the highest percentage of motor vehicles, followed by Central Java and DKI Jakarta. Meanwhile, North Kalimantan is the province with the lowest percentage of motor vehicles.

GHG emission percentage refers to the proportion or share of total GHG emissions produced by a particular source or sector, compared to the total emissions produced by all relevant sources or sectors, expressed as a percentage (%). These GHGs include CO_2, methane (CH_4), N_2O, and other gases that contribute to global warming and climate change. Based on the interpretation of Figure 5, the largest percentage of GHG emissions in 2022 is Riau Province, while Maluku and North Kalimantan provinces experience a significant minus percentage of GHG additions, this means that these provinces do not contribute GHG but add clean air.

Forest cover area refers to the state or condition where a certain area of land is covered by forest vegetation, including trees, shrubs, and other plants. Forest coverage area describes how much of an area is still covered by natural vegetation or trees that make up a forest. From Figure 6, it can be seen that the highest area of forest cover in 2022 is in Papua Province, while the lowest forest land cover is in DKI Jakarta Province.

The population percentage of each province in Indonesia is the ratio of the total population in a province to the total population in Indonesia, expressed as a percentage (%). In Figure 7, it can be seen that West Java is the province with the highest percentage of population in Indonesia, followed by East Java and Central Java. Meanwhile, North Kalimantan is the province with the lowest percentage of population in Indonesia.

The descriptive statistics table can be given in Table 2.

Descriptive statistics for the independent variables in this study are presented in Table 2. The table displays the average number of motor vehicles (X_1), GHG emissions (X_2), forest cover area (X_3), and population density (X_4) across Indonesia. Notably, provinces such as Jakarta and West Java exhibit high population density and motor vehicle usage, while provinces like Kalimantan and Papua show higher forest cover and lower population density.

Table 2. Descriptive statistics

Statistics	X_1	X_2	X_3	X_4
Min	0.131	-1.60	0.001	0.260
1st Qu.	0.680	0.010	0.679	0.833
Median	1.753	0.365	37.207	1.560
Mean	2.941	2.940	39.607	2.941
3rd Qu.	2.708	3.975	54.727	2.955
Max	16.56	20.990	93.143	17.920
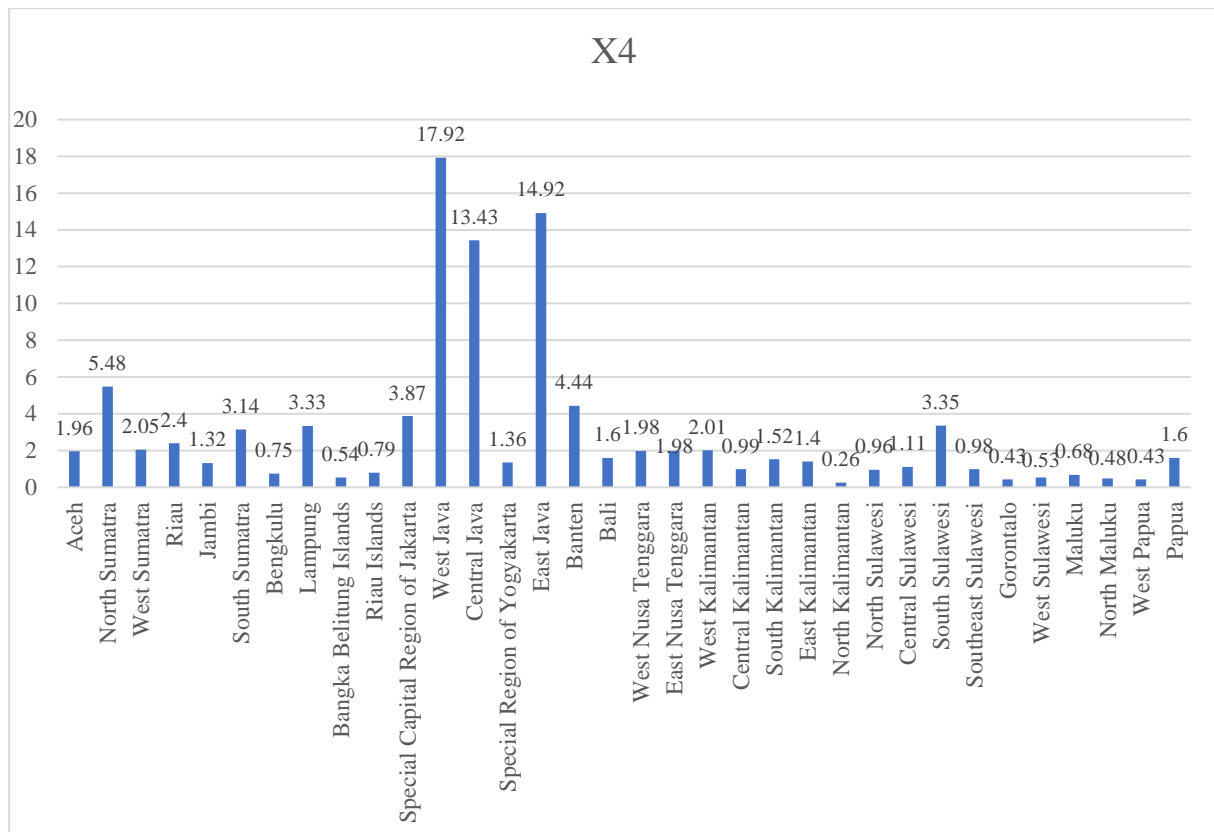
Figure 7. Percentage of population density (X_4)

From Table 3, the classification of AQI (IKU) across Indonesian provinces reveals that the majority, comprising 17 provinces, fall into the first category, indicating good air quality. Two provinces are classified under the second category, suggesting moderate air quality, which may pose risks to sensitive groups. One province is categorized under the third tier, indicating poorer air quality that could be harmful to vulnerable populations and the general public. Additionally, 14 provinces fall into the fourth category, where the air quality is deemed unhealthy for all individuals, highlighting the need for targeted interventions to improve these conditions.

Table 3. Provinces per category

Y	Total Provinces
1	17 provinces
2	2 provinces
3	1 province
4	14 provinces

Table 4 shows the correlation between variables. From the table there is a significant relationship between variables, where X_1 has a strong positive correlation with X_4 (0.853) and a moderate positive correlation with Y (0.608), while X_3 has a strong negative correlation with Y (-0.727). This indicates that an increase in the value of X_1 tends to be followed by an increase in the value of X_4 and Y, while an increase in the value of X_3 tends to be followed by a decrease in the value of Y. The identified correlations indicate a close relationship between several variables that can significantly affect the results of the study.

Table 4. Correlation between variables

	X_1	X_2	X_3	X_4	Y
X_1	1	0.302	-0.497	0.853	0.608
X_2	0.302	1	-0.338	0.294	0.160
X_3	-0.497	-0.338	1	-0.397	-0.727
X_4	0.853	0.294	-0.397	1	0.475
Y	0.608	0.160	-0.727	0.475	1

3.2 Ordinal Logistic Regression

3.2.1 Multicollinearity
Table 5 shows the multicollinearity of the predictor variables. Multicollinearity in the predictor variables will be detected using the Variance Inflation Factor (VIF) value. The predictor variable will have a multicollinearity problem if the VIF value > 5 (Kyriazos & Poga, 2023). Based on the calculation, there is no VIF value > 5 so there is no multicollinearity problem.

Table 5. VIF value

X_1	X_2	X_3	X_4
3.791	1.107	1.055	3.687

3.2.2 Parameter estimation
The MLE method was used to obtain estimates of logistic regression parameters, while the Cumulative Logistic Model (CLM) was used for model building.

Table 6. CLM model estimation value

Statistics	Estimate	Std. Error	Z Value	Pr(>|z|)
X_1	0.338	0.193	1.758	0.079
X_2	0.041	0.083	0.492	0.622
X_3	-0.486	0.239	-2.037	0.042 *
X_4	-0.016	0.164	-0.098	0.922
Y = 1|Y = 2	-0.712	0.679	-1.049	
Y = 2|Y = 3	3.756	1.387	2.708	
Y = 3|Y = 4	5.287	1.702	3.106	

The results of the CLM estimation are shown in Table 6, which shows that among the variables studied, only the percentage of forest cover area (X_3) has a statistically significant impact on the model, with a p-value of 0.0416. The negative coefficient for X_3 (-0.486) suggests that as the percentage of forest cover increases, the likelihood of being in a higher category of the dependent variable Y decreases. This implies that higher forest cover is associated with lower categories of Y, potentially indicating better outcomes or conditions related to the dependent variable. In contrast, motor vehicle use (X_1), GHG emissions (X_2), and population density (X_4) do not show statistically significant associations with air pollution levels in this model. However, these variables remain critical to understanding broader environmental dynamics, particularly in urban areas.

$$\text{Logit}[P(Y \le 1| X)] = -0.712 + 0.338X_1 + 0.041X_2 - 0.486X_3 - 0.016X_4$$
Or
$$P(Y \le 1| X) = \frac{\exp(-0.712 + 0.338X_1 + 0.041X_2 - 0.486X_3 - 0.016X_4)}{1 + \exp(-0.712 + 0.338X_1 + 0.041X_2 - 0.486X_3 - 0.016X_4)}$$

$$\text{Logit}[P(Y \le 2| X)] = 3.756 + 0.338X_1 + 0.041X_2 - 0.486X_3 - 0.016X_4$$
Or
$$P(Y \le 2| X) = \frac{\exp(\text{Logit}[P(Y \le 2| X)] = 3.756 + 0.338X_1 + 0.041X_2 - 0.486X_3 - 0.016X_4)}{1 + \exp(3.756 + 0.338X_1 + 0.041X_2 - 0.486X_3 - 0.016X_4)}$$

$$\text{Logit}[P(Y \le 3| X)] = 5.287 + 0.338X_1 + 0.041X_2 - 0.486X_3 - 0.016X_4$$
Or
$$P(Y \le 3| X) = \frac{\exp(\text{Logit}[P(Y \le 3| X) = 5.288 + 0.338X_1 + 0.041X_2 - 0.486X_4}{1 + \exp(5.288 + 0.338X_1 + 0.041X_2 - 0.486X_3 - 0.016X_4)}$$

The model equations provided reflect the estimated probabilities for each category of Y, demonstrating the cumulative nature of the logistic regression applied in this analysis. The following model is generated based on the parameter estimation.

3.2.3 Best model selection
The selection of the best model is based on the small Devians and RMSE values so that the results are displayed

in Table 7.

Based on Table 7, with the smallest devians value and the smallest RMSE, the best model is model number 15, but when the LRT is carried out, there is no significance when using all variables compared to only X_1 and X_3 (see point 3.2.6). So the best model is model number 6, with the second smallest devians and the smallest RMSE. Model number 6 is the model with the variable with the percentage of motor vehicles (X_1) and the variable percentage of forest land cover (X_3). The best model obtained from the variable selection results was carried out ordinal logistic regression analysis. The results of the analysis are shown in Table 8.

Table 7. Devians and RMSE value

No.	Model	Devians	RMSE
1	X_1	53.171	1.431
2	X_2	64.878	1.484
3	X_3	55.776	1.445
4	X_4	57.488	1.451
5	$X_1 + X_2$	53.152	1.432
6	$X_1 + X_3$	45.772	1.392
7	$X_1 + X_4$	53.149	1.429
8	$X_2 + X_3$	53.414	1.434
9	$X_2 + X_4$	57.096	1.447
10	$X_3 + X_4$	49.163	1.410
11	$X_1 + X_2 + X_3$	45.522	1.393
12	$X_1 + X_3 + X_4$	45.765	1.391
13	$X_1 + X_2 + X_4$	53.128	1.430
14	$X_2 + X_3 + X_4$	48.338	1.405
15	$X_1 + X_2 + X_3 + X_4$	45.512	1.392

Table 8. Best model of logistic ordinal regression

Statistics	Estimate	Std. Error	Z Value	Pr(>|z|)
X_1	0.344	0.1318	2.613	0.009 **
X_3	-0.469	0.2340	-2.005	0.045 *
Y = 1|Y = 2	-0.746	0.6741	-1.107	
Y = 2|Y = 3	3.765	1.4103	2.670	
Y = 3|Y = 4	5.350	1.7442	3.067	

Based on Table 8, we obtained three logit equations from the ordinal logistic regression model as follows:

$$\text{Logit}[P(Y \leq 1 | X)] = -0.746 + 0.344X_1 - 0.469X_3$$
Or
$$P(Y \leq 1 | X) = \frac{\exp(-0.746 + 0.344X_1 - 0.469X_3)}{1 + \exp(-0.746 + 0.344X_1 - 0.469X_3)}$$

$$\text{Logit}[P(Y \leq 2 | X)] = 3.765 + 0.344X_1 - 0.469X_3$$
Or
$$P(Y \leq 2 | X) = \frac{\exp(3.765 + 0.344X_1 - 0.469X_3)}{1 + \exp(3.765 + 0.344X_1 - 0.469X_3)}$$

$$\text{Logit}[P(Y \leq 3 | X)] = 5.350 + 0.344X_1 - 0.469X_3$$
Or
$$P(Y \leq 3 | X) = \frac{\exp(5.350 + 0.344X_1 - 0.469X_3)}{1 + \exp(5.350 + 0.344X_1 - 0.469X_3)}$$

The refined logistic ordinal regression model highlights those two variables, X_1 and X_3, are statistically significant predictors of the dependent variable Y. Specifically, X_1 (with a positive estimate of 0.344 and a p-value of 0.009) significantly increases the likelihood of Y falling into a higher category, suggesting that as X_1 increases, so does the probability of being in a more severe outcome category. On the other hand, X_3 (with a negative estimate

of -0.469 and a p-value of 0.045) significantly reduces the likelihood of Y being in a higher category, indicating that an increase in X_3 is associated with a shift towards lower categories of Y, which could imply better outcomes. The model equations provide the logit transformations and corresponding probabilities, allowing for the prediction of Y's categorical outcomes based on the values of X_1 and X_3. This model, focusing on the most significant predictors, enhances the accuracy of forecasting the dependent variable's distribution.

3.2.4 Simultaneous test

The ratio of motor vehicles to the percentage of land use was the subject of a simultaneous test that was carried out in order to ascertain the degree to which the AQI (IKU) is considerably impacted by the combined effects of two particular variables. In the absence of any concurrent impact exerted by the components, the null hypothesis (H_0) asserts that all of the coefficients (β_1, β_3) are equal to zero. As opposed to this, the alternative hypothesis (H_1) suggests that there is at least one coefficient that is not equal to zero, which demonstrates that there is a simultaneous effect.

Table 9. P-value LRT

Chi Square Table	G^2	p-Value
21.66428	34.675	$2.954e^{-08}$ ***

In order to determine whether or not the model had a good fit, the LRT was utilized, and the significance level was established at 5%. Based on Table 9, the test gave a G^2 value of 34.675, which exceeded the necessary Chi-Square value from the table (X^2 table = 21.66428). Furthermore, the p-value that was obtained was $2.954e^{-08}$, which is a number that is much lower than the probability level of 0.05.

Because the G^2 statistic is more than the crucial value and the p-value is significantly lower than the significance threshold, we are able to unequivocally reject the null hypothesis. This is because both factors are significantly lower than the threshold. With this new information, there is sufficient evidence to show that the proportion of land cover area and the number of motor vehicles both have a simultaneous and considerable impact on the IKU.

3.2.5 Partial test

A partial test was conducted to analyze the separate effects of two carefully chosen independent variables, namely the proportion of motor vehicles (X_1) and the fraction of land cover area (X_3), on the dependent variable, the AQI (IKU).

The null hypothesis (H_0) for the IKU states that there is no statistically significant influence of any of the independent variables (β_i). Conversely, the alternative hypothesis (H_1) posited that each independent variable indeed exerts a substantial impact on the solution.

Table 10. P-value wald test

Variable	p-Value
X_1	0.001005 **
X_3	0.0453 *

The statistical significance of each independent variable using the Wald test (commonly referred to as the W test) is displayed in Table 10. Hypotheses X_1 and X_3 yield estimated p-values of 0.001 and 0.0453, respectively. Any test statistic |W| that exceeds the crucial value $Z_{\alpha/2}$ or a p-value below the significance threshold α should lead to the rejection of the null hypothesis H_0. This conclusion is drawn with the significance threshold always maintained at 5%.

The p-values for both X_1 and X_3 are below the significance threshold of 0.05, indicating that we can definitively reject the null hypothesis for both variables. Therefore, it can be inferred with a 95% degree of confidence that the proportion of land cover area (X_3) and the fraction of motor vehicles (X_1) exert a substantial impact on the IKU. By emphasizing the importance of these elements in relation to their influence on air quality, these results provide evidence to justify the incorporation of these components in future research on environmental policy and environmental related matters.

3.2.6 Model fit test

The goal of the model fit test was to ascertain whether or not it would be beneficial to add additional factors to the model, namely the percentage of GHG emissions (X_2) and the percentage of population density (X_4), in order to enhance its ability to forecast the AQI (IKU). It was hypothesized—known as the null hypothesis, or H_0—that adding these extra elements would not produce a statistically significant improvement. This suggests that $\varepsilon_1, \varepsilon_3, ..., \varepsilon_n$, the other parameters, are all equivalent to zero. The hypothesis that there is a difference between zero and at

least one of the new parameters, known as the alternative hypothesis (H_1), stated that the statistical model's ability to match the data was improved.

Table 11. P-value LRT

Chi Square Table	G^2	p-Value
23.952	0.203	0.904

To assess the model, the LRT was employed at a significance level of 5%. Based on Table 11, this test yields a G^2 value of 0.203, significantly below the necessary Chi-Square value calculated from the table (X^2 table = 23.952). In addition, the calculation yielded a p-value of 0.904, which exceeds the significance level of 0.05.

In accordance with the test requirements, we fail to reject the null hypothesis that G^2 value is below the critical value and the p-value exceeds the significance level. Based on a 95% confidence level, we definitely determine that including the variables percentage of GHG emissions (X_2) and percentage of population density (X_4) does not have a substantial impact on enhancing the model's predictive accuracy for the IKU. Thus, the most simplified model that has a substantial impact on the IKU comprises solely the proportion of motor vehicles (X_1) and the proportion of land cover area (X_3). These findings indicate that these two factors are the main determinants of air quality in the specific setting of this investigation.

3.2.7 Model interpretation

The interpretation of the odds ratios provides insight into how different predictors—such as the number of motor vehicles, forest land cover, and population density—affect the likelihood of AQI (IKU) outcomes across different categories. For instance, the odds ratio $Y_{1|2} = \exp(-0.7460) = 0.474$ indicates that a province's probability of having a very good AQI (group 1) or lower is less than half (0.474 times) compared to the likelihood of falling into a lower air quality category (groups 2, 3, or 4). This suggests that provinces are more likely to experience declining air quality unless significant measures are taken to control factors like motor vehicle emissions and forest preservation. As it stands, the odds are against maintaining excellent air quality, highlighting the urgent need for policies that curb vehicle emissions and protect forests.

Similarly, the odds ratio $Y_{2|3} = exp(3.765) = 43.162$ shows that provinces with better control over motor vehicle use and forest management have a much higher probability—43 times greater—of having air quality in the good (group 2) or very good (group 1) range compared to fair (group 3) or poor (group 4). This large odds ratio implies that areas with relatively better environmental management (including motor vehicle use control and forest land preservation) have a greater chance of maintaining good air quality. Policy implications include the need for strict emission controls and urban planning policies that encourage green open spaces, which help maintain good air quality.

When examining the odds ratio $Y_{3|4} = exp(5.340) = 210.497$, it becomes even clearer that proactive environmental management dramatically reduces the risk of poor air quality. Provinces with better forest cover and fewer vehicles are over 210 times more likely to avoid poor air quality (group 4) and maintain a higher level of air cleanliness (groups 1, 2, or 3). This powerful statistic reinforces the importance of prioritizing reforestation, sustainable urban planning, and the reduction of motor vehicle emissions as vital strategies for preventing severe air quality deterioration.

Furthermore, every unit increase in the number of motor vehicles significantly impacts air quality, as indicated by an odds ratio $exp(0.344) = 1.411$. This means that provinces with more motor vehicles are 1.4 times more likely to see their air quality decline from excellent to poor. The implications here are clear: managing the growth of motor vehicles through policies like the promotion of public transportation and electric vehicles, as well as stricter emissions standards, is essential for improving air quality, particularly in urban areas where motor vehicle density is higher.

In contrast, the odds ratio X_3, $exp(-0.469) = 0.626$ means that every 1 unit increase in forest cover area causes the ratio of AQI from excellent to good, good to fair, or fair to poor to decrease by 0.626 times. In other words, provinces with 1 unit of additional forest cover have a lower chance of having an AQI of "less" than provinces with less forest cover. This implies that provinces with greater forest cover are less likely to experience poor air quality, as forests act as a buffer against environmental pollution. Forests play a crucial role in absorbing pollutants and mitigating the effects of emissions. Policymakers should focus on reforestation, deforestation prevention, and the expansion of protected forest areas to safeguard air quality and promote long-term environmental sustainability.

4. Discussion

This research discusses environmental and climate change that focuses on air quality in Indonesia, where air quality can be influenced by several factors. These factors are modeled to see which ones have a significant effect

on air quality. Research from the Center for Research on Energy and Clean Air (CREA) indicates that, while seasonal patterns significantly influence the spread of pollution, there is strong evidence of persistent transboundary pollution issues. The latest World Air Quality 2023 report by IQ Air ranked Indonesia as the country with the 14th highest average PM2.5 concentration in the world in 2023, weighted by population. Jakarta ranked 7th among world capitals in the country, region, and area categories in descending order. The report also notes a 20% increase in PM2.5 levels compared to 2023, putting Indonesia back as the country with the highest air pollution levels in Southeast Asia.

This study employs Ordinal Logistic Regression Analysis to investigate the factors influencing the AQI. The analyzed factors include the number of vehicles, population density, GHG emissions, and forest cover area. These four factors are assessed for their impact on air quality. Assumption checks are performed, including tests for multicollinearity, partial and simultaneous significance, and parameter estimation. The process also involves selecting the best-fitting model, evaluating goodness-of-fit, and interpreting the results. From this research, the best model is the 6th model because it has the smallest RMSE or Devian's value and the model is X_1, namely the number of motor vehicles and X_3 Forest cover area.

If associated with the data, factor X_1 or the number of motor vehicles is known as one of the main sources of air pollution due to harmful gas emissions such as CO, N_2O and particulates (PM2.5). The results of the ordinal logistic regression model show that the percentage of vehicles has a significant positive influence on the AQI. Each unit increase in motor vehicles increases the chance of poor air quality by 1.411 times. This identifies that areas with high vehicle density tend to experience poorer air quality. Previous studies have also shown that transportation is a major source of air pollution. Air pollution and the transportation sector account for an average of 70% of total air pollution (Anenberg et al., 2019). The WHO report (1992) states that at least 90% of CO in urban air comes from motor vehicle emissions.

Forest cover plays an important role in purifying the air by absorbing CO_2 and filtering pollutants. The results show that the percentage of forest cover area has a significant negative influence on air pollution. Each unit increase in the forest cover area increases the chance of better air quality by 0.625 times. This indicates that areas with more forest cover tend to have cleaner air quality. Forests serve as natural air filters that capture particulates and absorb GHG, thus improving overall air quality. This significant negative relationship emphasizes the importance of reforestation and conservation efforts in reducing air pollution, especially in Indonesia, which faces a serious deforestation problem.

This analysis shows that among the four factors studied, the percentage of motor vehicles and the area of forest cover are the most significant in determining air quality in Indonesia. Motor vehicles significantly contribute to air pollution, while forest cover helps reduce its negative impact by filtering pollutants. Population density and GHG emissions were not shown to have a significant influence in this model, although their indirect contribution to air pollution should not be ignored. This research emphasizes the importance of reducing motor vehicle emissions and increasing forest cover as key strategies to improve air quality in Indonesia.

However, limitations of using secondary data, as mentioned in this study, include potential biases and data gaps that may affect the accuracy and representation of the findings. For example, most of the data used in this study comes from sources such as government annual reports and national surveys. Bias may arise from data collection methods that cannot reach a wider population, such as the technological limitations of each region in Indonesia to measure the percentage of GHG emissions. This may overlook segments of the population that do not have access to technology, potentially providing results that are not fully representative of the entire population.

In addition, secondary data often has limitations in terms of time coverage. In this case, the data used only covers the period of 2023, which may lead to information gaps. Trends or significant changes that occur after the data collection period will not be reflected, impacting the accuracy of the analysis. For example, an increase in the number of motor vehicles or changes in deforestation policies may have affected air quality, but without more recent data, the findings of this study cannot fully represent current conditions.

Therefore, the findings of this study should be interpreted with caution, and it is recommended to combine secondary data from various sources or supplement it with more recent primary data to improve the representation and accuracy of the analysis.

Moreover, when comparing this study to previous research that addressed similar variables, such as the number of motor vehicles and forest cover, it becomes clear that the methods used in this paper provide a more refined analysis. For instance, earlier studies, including the one by Wan Azmi et al. (2024), utilized land use linear regression models to examine the relationship between land use factor and air pollution levels. While these models provided valuable insights into the general trend of air quality deterioration in relation to land use, they failed to capture the nuanced relationship between predictors and different categories of air quality, which is crucial for policy decisions.

In contrast, the use of Ordinal Logistic Regression in this research allows for a more detailed understanding of how factors like motor vehicle density and forest cover influence multiple levels of air quality, from "very good" to "poor." This level of granularity is particularly important in shaping environmental policies because it highlights the specific thresholds at which interventions, such as reducing vehicle emissions or increasing forest cover, may

be most effective. By identifying at what level air quality shifts from "good" to "fair," for instance, policymakers can prioritize targeted measures in regions where air quality is on the verge of significant deterioration.

In conclusion, compared to previous studies that utilized more basic modeling techniques, the Ordinal Logistic Regression employed in this research offers a more robust and policy-relevant framework. It not only identifies significant predictors of air quality but also provides actionable insights on how these predictors affect varying levels of air pollution, making the model more practical for informing SDGs in Indonesia.

To support SDG 11 (Sustainable Cities and Communities) and SDG 13 (Climate Action), the government needs to reduce motor vehicle emissions by promoting environmentally friendly public transportation, electric vehicles, and reducing the use of motor vehicles. In addition, it is necessary to increase forest cover through reforestation, forest protection, and spatial policies that maintain green spaces in cities. These efforts will help maintain air quality and support environmental sustainability in the long run.

5. Conclusion

Based on the results of the discussion and analysis above, it can be concluded that the best ordinal logistic regression model is seen from the small Devian's value and RMSE values so that the ordinal logistic regression model obtained to determine the factors of air quality levels of 34 provinces in Indonesia in 2023 is as follows:

$$Logit[P(Y \leq 1|X)] = -0.746 + 0.344X_1 - 0.469X_3$$

$$Logit[P(Y \leq 2|X)] = 3.765 + 0.344X_1 - 0.469X_3$$

$$Logit[P(Y \leq 3|X)] = 5.350 + 0.344X_1 - 0.469X_3$$

Simultaneous and partial testing also shows that only X_1 and X_3 are significant to the ordinal logistic regression model. So, the factors that affect the level of air quality in 34 provinces in Indonesia in 2023 are the percentage of motor vehicles (X_1) and the percentage of forest cover area (X_3). The positive coefficient on the variable percentage of motor vehicles (X_1) indicates that the higher the number of motor vehicles in a province, then the likelihood of worse air quality is greater. Conversely, the negative coefficient on the variable percentage of forest cover area (X_3) indicates that the larger the forest cover area in a province, then the air quality will be better. From the statistical test results, it was found that these two variables are significant both simultaneously and partially, making them the main factors affecting air quality in Indonesia. This model can be used as a basis in the formulation of environmental policies to control air pollution, with a focus on motor vehicle management and forest preservation.

Thus, based on the ordinal logistic regression model, it is concluded that every 1 unit increase in motor vehicles increases the odds ratio of the AQI from excellent to good, good to fair, or fair to poor by 1.411 times. In other words, provinces with more motorized vehicles have a higher chance of having poor air quality. Conversely, every additional 1 unit of forest cover area decreases the odds ratio of air quality deterioration by 0.626 times, so provinces with more forest cover have a smaller chance of experiencing air quality deterioration.

To advance SDG 11 (Sustainable Cities and Communities) and SDG 13 (Climate Action), the government must mitigate motor vehicle emissions by advocating for sustainable public transportation, electric vehicles, and decreasing reliance on motor vehicles. Moreover, it is essential to enhance forest cover by replanting, forest conservation, and spatial strategies that preserve urban green spaces. These initiatives will preserve air quality and promote long-term environmental sustainability.

Author Contributions

Conceptualization, Restu Arisanti and Yuyun Hidayat; Methodology, Titi Purwandari and Irlandia Ginanjar; software, Aisya Putri Syarnurli and Dianda Destin.; validation, Maharani Rizki Febrianti; formal analysis, Restu Arisanti and Maharani Rizki Febrianti; investigation, Aisya Putri Syarnurli; resources, Yuyun Hidayat and Restu Arisanti; data curation, Dianda Destin and Maharani Rizki Febrianti; writing—original draft preparation, Restu Arisanti, Aisya Putri Syarnurli, Dianda Destin and Maharani Rizki Febrianti; writing—review and editing, Restu Arisanti, Aisya Putri Syarnurli, Dianda Destin and Maharani Rizki Febrianti; visualization, Aisya Putri Syarnurli, Dianda Destin and Maharani Rizki Febrianti; supervision, Restu Arisanti; project administration, Yuyun Hidayat; funding acquisition, Yuyun HIdayat. All authors have read and agreed to the published version of the manuscript.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Acknowledgements

Conflicts of Interest

The authors declare that there are no conflicts of interest related to this research. Although the data taken from BPS and KLHK which are government agencies, does not make the author subjective to the analysis and interpretation of the results. All data sources, literature, and references used have been appropriately acknowledged and no violation of publication ethics has been committed.

References

Anenberg, S. C., Miller, J., Henze, D., & Minjares, R. A. Y. (2019). Pollution-related health impacts of transportation sector emission in 2010 and 2015. *ICCT, Clim. Clean Air Coal.*, 1-55.

Arisanti, R., Purnamawati, S., & Muslim, A. (2024). Determinants of greenhouse gas emissions in the transportation sector in Indonesia: Official statistics and big data approach. *Int. J. Energy Econ. Policy*, *14*(1), 86-97. https://doi.org/10.32479/ijeep.15035.

BPS-Statistics Indonesia. (2022a). *Jumlah Kendaraan Bermotor Menurut Provinsi dan Jenis Kendaraan (unit), 2022*. https://www.bps.go.id/id/statistics-table/3/VjJ3NGRGa3dkRk5MTlU1bVNFOTVVbmQyVURSTVFUMDkjMw==/jumlah-kendaraan-bermotor-menurut-provinsi-dan-jenis-kendaraan--unit---2022.html?year=2022

BPS-Statistics Indonesia. (2022b). *Penduduk, Laju Pertumbuhan Penduduk, Distribusi Persentase Penduduk, Kepadatan Penduduk, Rasio Jenis Kelamin Penduduk Menurut Provinsi, 2022*. https://www.bps.go.id/id/statistics-table/3/V1ZSbFRUY3lTbFpEYTNsVWNGcDZjek53YkhsNFFUUMDkjMw==/penduduk-laju-pertumbuhan-penduduk-distribusi-persentase-penduduk-kepadatan-penduduk-rasio-jenis-kelamin-penduduk-menurut-provinsi.html?year=2022

BPS-Statistics Indonesia. (2022c). *Rekapitulasi Luas Penutupan Lahan Hutan dan Non Hutan Menurut Provinsi Tahun 2014-2022 (Ribu Ha)*. https://www.bps.go.id/id/statistics-table/1/MjExMCMx/rekapitulasi-luas-penutupan-lahan-hutan-dan-non-hutan-menurut-provinsi-tahun-2014-2022--ribu-ha-.html

Chen, Y., Moustaki, I., & Zhang, H. (2020). A note on likelihood ratio tests for models with latent variables. *Psychometrika*, *85*(4), 996-1012. https://doi.org/10.1007/s11336-020-09735-0.

Direktorat Pengendalian Pencemaran Udara. (2022). *Laporan Kinerja.* https://ppkl.menlhk.go.id/website/filebox/1168/230829082359LKJ%20PPU%202022.pdf

Grassi, G., Conchedda, G., Federici, S., Abad Viñas, R., Korosuo, A., Melo, J., Rossi, S., Sandker, M., Somogyi, Z., Vizzarri, M., & Tubiello, F. N. (2022). Carbon fluxes from land 2000-2020: Bringing clarity to countries' reporting. *Earth Syst. Sci. Data*, *14*(10), 4643-4666. https://doi.org/10.5194/essd-14-4643-2022.

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geosci. Model Dev.*, *15*(14), 5481-5487. https://doi.org/10.5194/gmd-15-5481-2022.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons.

IQAir. (2023). *World Air Quality Report 2023*. https://www.iqair.com/world-most-polluted-countries

Jung, Y. & Lee, I. (2021). Optimal design of experiments for optimization-based model calibration using Fisher information matrix. *Reliab. Eng. Syst. Saf.*, *216*, 107968. https://doi.org/10.1016/j.ress.2021.107968.

KLHK. (2019). *Indeks Kualitas Lingkungan Hidup 2019. Kementerian Lingkungan Hidup dan Kehutanan Republik Indonesia*. https://www.menlhk.go.id/cadmin/uploads/1609312579_5f6b7346d1.pdf

Kumar, A., Sinha, C., Singh, K., & Sagdeo, G. D. (2022). COVID-19 era. *Res. Opin. Anesth. Intensive Care*, *9*(4), 375-376. https://doi.org/10.4103/roaic.roaic_58_21.

Kyriazos, T. & Poga, M. (2023). Dealing with multicollinearity in factor analysis: The problem, detections, and solutions. *Open J. Stat.*, *13*(3), 404-424. https://doi.org/10.4236/ojs.2023.133020.

Lee, K. & Greenstone, M. (2021). Indonesia's Air Pollution and Its Impact on Life Expectancy. *Air Quality Life Index*.

McCullagh, P. & dan Nelder, A. (1989). *Generalized Linear Models*. Second Edition. London: Chapman & Hall.

Mehrpooya, M., Golestani, B., & Ali Mousavian, S. M. (2020). Novel cryogenic argon recovery from the air

separation unit integrated with LNG regasification and CO_2 transcritical power cycle. *Sustain. Energy Technol. Assess.*, *40*, 100767. https://doi.org/10.1016/j.seta.2020.100767.

Orr, C. J. & Kish, K. (2022). Sustainability transformations: Emerging pathways toward safe and just futures for people and the planet. *Chall. Sustain.*, *10*(1), 1-2. https://doi.org/10.12924/cis2022.10010001.

Pelaporan Dan Verifikasi. (2022). *Direktorat Inventarisasi Gas Rumah Kaca Dan Monitoring. Sistem Inventarisasi Gas Rumah Kaca Nasional*. https://signsmart.menlhk.go.id/v2.1/app/

Pisică, D., Dammers, R., Boersma, E., & Volovici, V. (2022). Tenets of good practice in regression analysis. A brief tutorial. *World Neurosurg.*, *161*, 230-239. https://doi.org/10.1016/j.wneu.2022.02.112.

Poveda, C. A. (2023). Evaluating the effectiveness of commercially developed appraisal instruments (CDAIs) using composite indices to assess, compare, and rank the liveability, quality of living and sustainability performance of cities and communities. *Chall. Sustain.*, *11*(1), 1-18. http://doi.org/10.12924/cis2023.11010001.

Shidqi, F. S. & Supriyatna, D. (2023). Analysis of mechanical competency and service prices on customer satisfaction at cv. nirmala kn autocare works. *J. Din. Vok. Tek. Mesin.*, *8*(2), 135-143. https://doi.org/10.21831/dinamika.v8i2.68039.

Vamsi, V. (2020). The valuable natural resources of the world, economic growth and the global environment. *Int. J. Eng. Dev. Res.*, *8*(3), 389-421.

Vohra, K., Marais, E. A., Bloss, W. J., Schwartz, J., Mickley, L. J., Van Damme, M., Clarisse, L., & Coheur, P. F. (2022). Rapid rise in premature mortality due to anthropogenic air pollution in fast-growing tropical cities from 2005 to 2018. *Sci. Adv.*, *8*(14). https://doi.org/10.1126/sciadv.abm4435.

Wan Azmi, W. N. F., Pillai, T. R., Latif, M. T., Shaharudin, R., & Koshy, S. (2024). Development of land use regression model to estimate particulate matter (PM10) and nitrogen dioxide (NO_2) concentrations in Peninsular Malaysia. *Atmos. Environ.: X*, *21*, 100244. https://doi.org/10.1016/j.aeaoa.2024.100244.

Wang, B., Wang, B., Lv, B., & Wang, R. (2022). Impact of motor vehicle exhaust on the air quality of an urban city. *Aerosol Air Qual. Res.*, *22*(8), 1-10. https://doi.org/10.4209/aaqr.220213.

Williams, R. A. & Quiroz, C. (2020). *Ordinal Regression Models*. SAGE Publications Limited.

Zamorategui-Molina, A., Gutiérrez-Ortega, N. L., Baltazar-Vera, J. C., Ángel-Soto, J. D., & Tirado-Torres, D. (2021). Carbon monoxide and particulate matter concentrations inside the road tunnels of guanajuato city, mexico. *Aerosol Air Qual. Res.*, *21*(10), 1-17. https://doi.org/10.4209/aaqr.210039.