

Population Prediction Using Multiple Regression and Geometry Models Based on Demographic Data

M. Safii, Rika Setiana

STIKOM Tunas Bangsa, Pematangsiantar, Indonesia

Article Info

Article history:

Received June 05, 2024

Revised June 20, 2024

Accepted July 18, 2024

Keywords:

Demographic Data

Geometry Models

Multiple Linear Regression

Population Prediction

ABSTRACT

Population growth is an important issue because it significantly impacts a country's growth and development. Large population growth can impact potential resources that drive the pace of the economy and national development. On the other hand, it can also be a problem of poverty, hunger, unemployment, education, health, and others. The government needs to control population growth to balance it with good population quality. Data sourced from the Population and Civil Registration Office of Simalungun Regency, Tanah Java sub-district has a high population and continues to increase every year. The impact of the population increase is that it affects the population's welfare, most of whom work as laborers and farmers. To overcome this problem, it is necessary to predict the number of people in the future so that the government can make the right decisions and policies in controlling the population. This study aims to make predictions using two models, namely Multiple Linear Regression, to find linear equations and Geometry Models for population growth projections. This study utilizes multiple regression analysis and geometric models using three independent variables, namely birth rate (X1), migration rate (X2), and death rate (X3), as well as one bound variable, population number (Y). This study's results show that the Tanah Java sub-district population is expected to increase in the next five years (2024-2028). Predictions show that by 2024, the population is expected to reach 61178 people from 59589 in 2023. Based on the results of the study, the conclusion of this study it can be used as a guide for the authorities in planning strategies and resource allocation and making a significant contribution in estimating population development in the Java region so that there will be no population explosion in the future so that it does not have a negative impact.

Copyright ©2024 The Authors.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

M.Safii, 085362234880,

Teknik Informatika,

STIKOM Tunas Bangsa, Pematangsiantar, Indonesia.

Email: m.safii@amiktunasbangsa.ac.id.

How to Cite:

M. Safii and R. Setiana, "Population Prediction Using Multiple Regression and Geometry Models Based on Demographic Data", *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 24, No. 1, pp. 129-140, November, 2024.

This is an open access article under the CC BY-SA license (<https://creativecommons.org/licenses/by-sa/4.0/>)

1. INTRODUCTION

The fourth country with the highest population in the world is Indonesia [1]. The population growth of an area is important for regional development and is a benchmark to assess whether a region can develop well [2]. Population growth has evolved throughout human history, and some of its impacts on the environment and human well-being [3]. Population growth refers to a situation where the number of people decreases and the area increases over time due to births, deaths, migrations, etc [4]. Currently, development is increasing rapidly, and the population is increasing [5]. Population growth affects the economy, society, politics, and defense. Therefore, the government must strive to control population growth so that population growth is balanced and the quality of the population increases [6]. Large population growth can have an impact on potential resources in driving the pace of the economy and national development [7]. On the other hand, it can also be a problem of poverty, hunger, unemployment, education, health, and others. The government needs to control population growth to balance it with good population quality. Data sourced from the Office of Population and Civil Registration of Simalungun Regency, Tanah Java sub-district in 2020 has a population of 54900 people in 2021. It is also seen in the Tanah Java region with a very high population growth of 55495 people and in 2022, as many as 57935, and last year, namely in 2023, as many as 59589; it is concluded that there is an increase every year.

The impact of the population increase is that it affects the population's welfare, most of whom work as laborers and farmers. To overcome this problem, it is necessary to predict the population in the future so that the government can make the right decisions and policies in controlling the population using multiple regression methods and geometric models [8, 9]. By making accurate predictions, it is hoped that the government can take appropriate steps to overcome the negative impact of the increase in population in Simalungun Regency so that there will be no population explosion in the coming year. Population growth is interesting in different fields, such as epidemiology, economics, or biology. Although a wide variety of growth models are available in the scientific literature, their application typically requires advanced knowledge of mathematical programming and statistical inference, especially when modeling growth under dynamic environmental conditions [10]. Accurate forecasting of subnational populations is necessary for sustainable development, including planning for the future, allocating resources, or providing health services [11]. In making population growth predictions, two models are used, namely the Multiple Linear Regression (MLR) method, which will be used to find linear equations, and the Geometry model, which will be used to find population growth projections [12]. In making predictions, several independent variables are needed [13], in this study, 3 variables are used, namely X1, X2, X3, where X1 is the birth number, X2 is the death rate, and X3 is the population movement number, and the bound variable is Y is the number of population.

Some related studies using MLR and geometry include the research by [14]. The regression method is used to analyze the development of the number of individuals, with the results of the analysis estimating that the increase in the number of people in a certain area has reached a very large number. In the next research, study conducted by [15] has different machine learning algorithms used to estimate the population. Regression analysis is one of the most useful tools for academics, although it is a difficult, time-consuming, and expensive effort, especially when it comes to accurately estimating and properly interpreting data [16]. Researchers believe that the findings of MLR produced by Statistical Package for the Social Sciences (SPSS) require a more inclusive and thoughtful interpretation. The results of the analysis using this data mining method will also help the Central Statistics Agency identify factors that affect population growth. Other research conducted by [17] The linear regression method was used to calculate the body mass index, resulting in an excellent height measurement with an accuracy of 98.96% and an average error of 1.04%. **Previous research has not** resolved some gaps, namely the variables of birth, death, and migration rates. This study helps local governments plan appropriate strategies and policies to overcome the negative impacts of population growth. **This** research aims to prevent a population explosion and positively impact the economy and community welfare. **The contribution of this research** is to estimate population development in the Tanah Java sub-district so that there will be no population explosion in the future and that it does not have a negative impact.

2. RESEARCH METHOD

This study uses a quantitative data collection method by applying multiple regression and geometric models [18, 19]. Data was collected from the Population and Civil Registration Office of Simalungun Regency, especially for the Tanah Java region, between 2019 and 2023, with 3 variables, which are part of demographic data: death rates, births, and population displacement rates. Data collection is carried out through surveys and statistical analysis, producing data in the form of numbers. The collected data was then processed using Microsoft Excel and tested with the SPSS application. Predictions were made using the RapidMiner 5.3 application. This approach allows for comprehensive and accurate analysis to predict future population growth. In addition, this study utilizes literature as the main source of information through literature review. The literature review process is carried out by identifying relevant literature, such as academic books and journals, which are used as a theoretical framework to establish research methods and interpret the results obtained.

2.1. Research Framework

This research begins by explaining the first step that needs to be taken in the research. This initial step is problem analysis, where problem identification is carried out to open up a deeper understanding of the background and focus of the investigation. The next stage is a literature review, which provides a strong theoretical foundation. This process involves gathering information from various sources such as scientific journals, books, articles, and other relevant research to understand the progress of science in a particular field. The data collected came from the Population and Civil Registration Office of Simalungun Regency, focusing on research in the Tanah Java region. Data collection involves surveys and statistical analysis, producing data in numbers spanning 5 years, namely in 2019-2023. After the data is collected, determining the right method for making predictions, the data is processed using the Multiple Linear Regression method, which will be used to find MLR equations and the geometry method is used to find test data so that predictions can be made in the next 5 years using the Equation 1 [20, 21]. Y is the dependent variable symbol, and X1 and X2 are independent variable symbols. a, b1, b2 are constants, and a0 is ordinances.

$$y = a_0 + b_1x_1 + b_2x_2 + \dots b_nx_n \quad (1)$$

The determination of A and the performance constants B1 and B2 are sought by applying the Equation 3, 4 and 5. The process determines the training data and searches for the sum values of X_2^2 , X_1^2 , Y^2 , and X_2Y , as well as the search for MLR equations by finding the values a, b1, b2, then determines the test data using geometric formulas with Equation 3.2..

$$a = \frac{(\sum y) - (b_1 * \sum x_1) - (b_2 * \sum x_2)}{n} \quad (2)$$

$$b_1 = \frac{(\sum x_2^2 * \sum x_1y) - (\sum x_2y * \sum x_1x_2)}{(\sum x_1^2 * \sum x_2^2) - (\sum x_1x_2)} \quad (3)$$

$$b_2 = \frac{(\sum x_1^2 * \sum x_2y) - (\sum x_1y * \sum x_1x_2)}{(\sum x_1^2 * \sum x_2^2) - (\sum x_1x_2)} \quad (4)$$

$$\text{Value } r = r = \frac{pt}{po} - 1 \quad (5)$$

The next stage is to look for the difference in the growth of demographic data or r value and then find the pt value or the amount of demographic data in the coming year which will be used as test data with Equation 3.1.. Pt is the Number of births in the t year, and Po is the Number of births in the base year. R is the Growth rate, Birth rate, and T is the Period.

$$PtFormula = Po(1 + r)^t \quad (6)$$

In this study, we will make predictions for the next five years, namely 2024-2028. After performing manual calculations, data testing will be carried out to ensure the quality of the results obtained in the study, using RapidMiner software and SPSS Statistics 27. In Figure 1, a research work design is carried out to predict the population of Tanah Java sub-district.

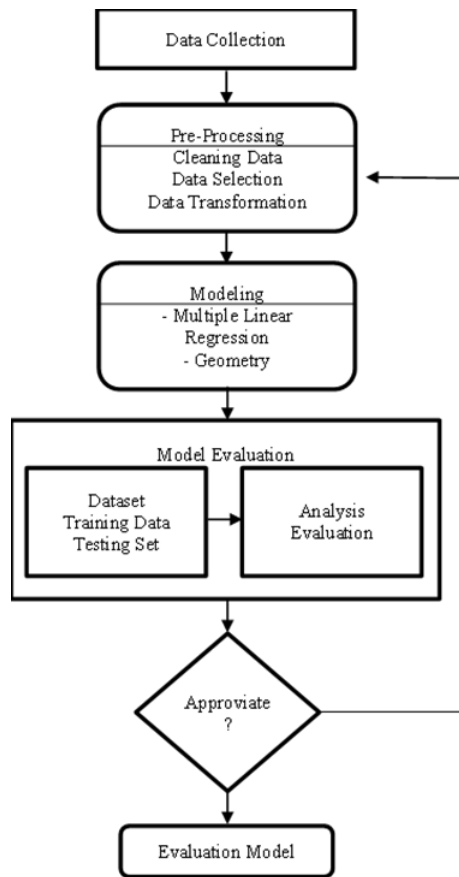


Figure 1. Research Framework

3. RESULT AND ANALYSIS

The findings of this research are predictions made in the next 5 years with a time span of 2024-2028. Making a prediction is carried out using two methods, namely MLR and the Geometry Model. MLR is used to find multiple linear regression equations that will be used to predict population growth, and Geometry Models will be used to search for demographic number testing data. The results of this research are **in line** with or supported by [14] The Regression method is used to analyze the development of the number of individuals, with the results of the analysis estimating that the increase in the number of people in a certain area has reached a very large number.

3.1. Data Processing

In data processing, a manual process of number processing will be carried out using Microsoft Excel. In Table 1, there are 3 independent variables and 1 bound variable, namely X1, X2, and X3 as independent variables and Y as dependent variables. After obtaining the values of the three independent variables, it will be simplified by adding the overall value of each village/sub-district every year; the data will be used as training data (see Table 2) that will be calculated manually.

Table 1. Types of Variables

Variables	Information
X1	Birth Rate
X2	Population Displacement Rate
X3	Mortality
Y	Population

Table 2. Data Training

No	Area	X1					X2					X2				
		2019	2020	2021	2022	2023	2019	2020	2021	2022	2023	2019	2020	2021	2022	2023
1	Bah Jambi II	693	871	937	987	1021	1783	1945	1987	1996	2036	65	89	134	178	231
2	Bah Jambi III	453	631	871	938	1115	1982	2043	2147	2090	2131	78	81	128	152	293
3	Bah Kisat	637	721	789	864	978	2167	2234	2243	2287	2365	64	93	145	161	231
4	Baja Dolok	361	657	661	789	891	1875	1965	2003	2098	2178	56	71	123	147	198
5	Baliju	765	835	921	993	1023	2034	2190	2205	2314	2397	23	54	169	185	241
6	Balimbangan	789	798	815	946	971	1871	1937	2013	1891	2184	80	92	132	164	272
7	Bayu Bagasan	956	1098	1116	1189	1165	2926	2967	2871	2899	2917	53	48	167	181	204
8	Bosar Galugur	458	793	934	978	1021	1559	1675	1730	1793	1832	35	67	145	163	227
9	Maligas Tengah	658	892	915	967	995	1453	1547	1632	1762	1658	97	123	167	191	275
10	Marubun Bayu	782	910	940	1023	1132	1872	1913	1967	2096	2142	89	96	135	172	230
11	Marubun Jaya	967	1073	1126	1176	1195	2317	1934	1989	1987	2092	74	83	187	234	263
12	Mekar Mulia	647	821	898	957	1056	1373	1846	1778	1965	2131	53	78	143	192	291
13	Muara Mulia	889	928	976	1024	1121	2356	1543	1865	1904	2176	47	72	167	228	287
14	Pagar Jambi	451	634	721	871	934	1751	1815	1963	2045	2134	45	69	145	231	264
15	Panambeian Mar-jandi	543	683	598	631	763	1813	1965	1968	2073	1957	39	58	203	372	431
16	Parbalogan	865	931	957	983	1023	1682	1765	1843	1694	2034	48	52	156	165	282
17	Pardamean Asih	859	1034	1088	1192	1441	1692	1860	1770	1971	2019	51	68	161	251	317
18	Pematang Tanah Jawa	673	975	989	1158	1189	2191	2471	2457	2623	2671	76	81	194	339	398
19	Tanjung Pasir	657	853	996	1120	1159	1295	1723	1745	1392	1812	25	56	165	187	335
20	Totap Majawa	898	993	998	1131	1198	1187	1945	1982	2031	2059	78	83	143	165	457
	Σ	14001	17131	18246	19917	21391	37179	39283	40158	40911	42925	1176	1514	3109	4058	5727

The data testing process was carried out using the geometry method based on the training data. In the geometric method for searching test data by determining the r value called the population growth percentage value [22] The following is a manual calculation of the population growth percentage. R is the Birth Rate Growth Presentation, P_t is the Number of births in t year, P_0 is the Number of births in the Primary Year, and T is the Predicted timeframe. After determining the R -value, the value of the birth rate test data is determined by looking for the value of P_t .

$$\text{Value } r = \frac{21391}{14001} \frac{1}{5}^{-1} = 0.088464844$$

$$\text{Year 2024} = 21391 * ((1 + 0.088464844)^1) = 23283$$

$$\text{Year 2025} = 21391 * ((1 + 0.088464844)^2) = 25343$$

$$\text{Year 2026} = 21391 * ((1 + 0.088464844)^3) = 27585$$

$$\text{Year 2027} = 21391 * ((1 + 0.088464844)^4) = 30025$$

$$\text{Year 2028} = 21391 * ((1 + 0.088464844)^5) = 32682$$

The following Mortality Rate Testing Data determination process is carried out in the next stage. After determining the r value, the process of determining the number of deaths in the next 5 years will be carried out. In Table 3, you can see the test data in the next 5 time periods calculated using the geometric method.

$$\text{Value } r = \frac{5727}{1176} \frac{1}{5}^{-1} = 0.372473514$$

$$\text{Year 2024} = 5727 * ((1 + 0.372473514)^1) = 7860$$

$$\text{Year 2025} = 5727 * ((1 + 0.372473514)^2) = 10788$$

$$\text{Year 2026} = 5727 * ((1 + 0.372473514)^3) = 14806$$

$$\text{Year 2027} = 5727 * ((1 + 0.372473514)^4) = 20321$$

$$\text{Year 2028} = 5727 * ((1 + 0.372473514)^5) = 27890$$

Table 3. Data Testing

Years	X1	X2	X3
2024	23283.4	44176.7	7860.16
2025	25343.1	45464.8	10787.9
2026	27585.1	46790.5	14806
2027	30025.4	48154.9	20320.9
2028	32681.6	49559	27889.9

Searching for multiple linear equations is necessary first to carry out the prediction process using the MLR method. In performing the multiple linear search, the following steps must be followed. The first stage will be finding the values of $X1 * Y, X2 * Y, X3 * Y, X1 * X2, and X1 * X3$. In carrying out the calculation process to determine $X1 * Y$, the method that can be used is multiplying the birth rate by the number of people.

$$\begin{aligned}
 \text{Year 2019} &= 14001 * 48004 = 672,104,004 \\
 \text{Year 2020} &= 17131 * 54900 = 940,491,90 \\
 \text{Year 2021} &= 18246 * 55495 = 1,012,561,770 \\
 \text{Year 2022} &= 19917 * 57935 = 1,153,891,395 \\
 \text{Year 2023} &= 21391 * 59589 = 1,274,668,299 \\
 \sum X1Y &= 5,053,717,368
 \end{aligned}$$

In carrying out the calculation process, determine $X2 * Y$ and $X3 * Y$ in the same way as above with the following results. Calculating the value of $X1X2$ can be done by multiplying the value of the birth rate with the value of the population movement number. Likewise, the calculation of the value of $X1 * X3$ with the result of $\sum X1 * X3 = 302,457,767$. In Table 4, the values of $X1 * Y, X2 * Y, X3 * Y, X1 * X2$, and $X1 * X3$ values have been calculated manually. The next stage is to find the values $X2 * X3, X12, X22, X32$. The calculation process can be done in the same way as above. In Table 5, the manual calculation is from the values $X2 * X3, X12, X22, X32$.

$$\begin{aligned}
 \sum X2Y &= 11,097,982,236 \\
 \sum X3Y &= 888,471,692
 \end{aligned}$$

$$\begin{aligned}
 \text{Year 2019} &= 14001 * 37179 = 520,543,179 \\
 \text{Year 2020} &= 17131 * 39283 = 672,957,073 \\
 \text{Year 2021} &= 18246 * 40158 = 732,722,868 \\
 \text{Year 2022} &= 19917 * 40911 = 814,824,387 \\
 \text{Year 2023} &= 21391 * 42925 = 918,208,675 \\
 \sum X1X2 &= 3,659,256,182
 \end{aligned}$$

Table 4. Pembagian data untuk Training dan Testing

X1Y	X2Y	X3Y	X1X2	X1X3
672,104,004	1,784,740,716	56,452,704	520,543,179	16,465,176
940,491,900	2,156,636,700	83,118,600	672,957,073	25,936,334
1,012,561,770	2,228,568,210	172,533,955	732,722,868	56,726,814
1,153,891,395	2,370,178,785	235,100,230	814,824,387	80,823,186
1,274,668,299	2,557,857,825	341,266,203	918,208,675	122,506,257

Table 5. Value $X2X3, X1^2, X2^2, X3^2$

X2X3	X1 ²	X2 ²	X3 ²
43,722,504	196,028,001	1,382,278,041	1,382,976
59,474,462	293,471,161	1,543,154,089	2,292,196
124,851,222	332,916,516	1,612,664,964	9,665,881
166,016,838	396,686,889	1,673,709,921	16,467,364
245,831,475	457,574,881	1,842,555,625	32,798,529

The next step after conducting the process of searching for the values of $X2 * X3$, $X12$, $X22$, and $X32$ will be carried out the process of searching for the values of Matrix A (see Table 6), matrix H, Matrix A1, Matrix A2, Matrix A3, and Matrix A4. A matrix value search can be done using the following formula: after finding the value of matrix A is carried out, the process of finding the value of matrix H will be carried out (see Table 7). The next step after the process of finding the value of the H matrix will be carried out by determining the value of the A1 matrix by replacing Column 1 of the A matrix with the result of the H matrix. can be seen the results from the A1 matrix (see in Table 8).

Table 6. Matrix A Results

n	5	$\sum X1$	90686	$\sum X2$	200456	$\sum X3$	15584
$\sum X1$	90686	$\sum X1^2$	1,677E+09	$\sum X1X2$	3,659E+09	$\sum X1X3$	302457767
$\sum X2$	200456	$\sum X1X2$	3,659E+09	$\sum X2^2$	8,054E+09	$\sum X2X3$	639896501
$\sum X3$	15584	$\sum X1X3$	302457767	$\sum X2X3$	639896501	$\sum X3^2$	62606946

Table 7. Matrix A Results

$\sum Y$	275923
$\sum X1Y$	5,054E+09
$\sum X2Y$	1,11E+10
$\sum X3Y$	888471692

Table 8. Matrix Value A1

275923	90686	200456	15584
5053717368	1676677448	3659256182	302457767
11097982236	3659256182	8054362640	639896501
888471692	302457767	639896501	62606946

In determining the value of the A2 matrix using Column 2, Matrix A is replaced with the result of the H matrix (see Table 9). In the process of determining the value of the A3 matrix by means of Column 3, Matrix A is replaced with the result of matrix H (see Table 10). In the process of determining the value of the A4 matrix by means of Column 4, Matrix A is replaced with the result of matrix H. After the search for the value of the A4 matrix is completed, the search for the determinant values A, A1, A2, A3, and A4 will be carried out (see Table 11). After obtaining the Determinant value of each matrix value, the next step will be the process of determining the value of the MLR, and the results are shown in Table 12.

Table 9. Matrix Value A1

5	275923	200456	15584
90686	5053717368	3659256182	302457767
200456	11097982236	8054362640	639896501
15584	888471692	639896501	62606946

Table 10. Matrix Value A1

5	90686	275923	15584
90686	1676677448	5053717368	302457767
200456	3659256182	11097982236	639896501
15584	302457767	888471692	62606946

Table 11. Matrix Value A4

5	90686	200456	275923
90686	1676677448	3659256182	5053717368
200456	3659256182	8054362640	11097982236
15584	302457767	639896501	888471692

Table 12. Pembagian data untuk Training dan Testing

DET A	87373219029861300000
DET A1	443770637187962000000000
DET A2	180728744323164000000
DET A3	36544761852300400000
DET A4	-117155811157812000000

Determining the B0 value

$$\frac{DETA1}{DETA} = \frac{443770637187962000000000}{87373219029861300000} = 5079.023551$$

Determining the B1 value

$$\frac{DETA2}{DETA} = \frac{180728744323164000000}{87373219029861300000} = 2.06846842$$

Determining the B2 value

$$\frac{DETA3}{DETA} = \frac{36544761852300400000}{87373219029861300000} = 0.418260449$$

Determining the B3 value

$$\frac{DETA4}{DETA} = \frac{-117155811157812000000}{87373219029861300000} = -1.340866371$$

3.2. Results

At this stage, after completing the manual calculation process, a testing process will be carried out using the SPSS application. This testing process aims to see if the manual calculation process follows the software testing process used, and if the results are the same, they are correct. After selecting the analyze menu, a Linear Regression display will appear with the input variable Y in the dependent column and the variables X1, X2, and X3 input in the independent variable column. Because variable Y is a bound variable, variable Y is entered in the dependent column, while because variables X1, X2, and X3 are independent variables, they are entered in the independent column. After the four variables have been entered in their respective columns, then click ok so that the desired result appears. Based on Figure 2, it can be seen that the Coefficient value in the variables X1, X2, and X3 is the same as the result of the value of the MLR equation in manual calculation with the value of B0 (Constant) = 5079.024 and the value of B1 (X1) = 2.068 and the value of B2 (X2) = 0.418 and the value of B3 (X3) = -1.341. After conducting the test process on the value of the MLR equation, the next stage will be a test process on the value of One-Way Variance Analysis (ANOVA), whether the Innova value between the manual calculation and the SPSS test process is the same or not. With this annova value test process, it will be known whether the free variable has a significant effect on the dependent variable (see Figure 3).

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients
		B	Std. Error	Beta
1	(Constant)	5079.024	2736.963	
	X1	2.068	.059	1.317
	X2	.418	.094	.199
	X3	-1.341	.048	-.566

a. Dependent Variable: Y

Figure 2. Linear Regression Test

ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	78699032.27	3	26233010.76	9474.053	.008 ^b
	Residual	2768.932	1	2768.932		
	Total	78701801.20	4			

a. Dependent Variable: Y
 b. Predictors: (Constant), X3, X1, X2

Figure 3. ANNOVA Values

The method of determining the value of Ftable can be done by looking at the table determining the value of df 1 and the value of df 2 in f table 0.05. In Table 13, there is a fixed value of Ftable, which can be used as a tool to determine whether the value of Fhcalculated is greater than that of Ftable. The Ftable can be determined by looking at the values of df 1 and df 2. In this study, the value of df 1 is 3 and the value of df 2 is 1. In the test criterion, F has a test criterion of Fcal < Ftable, then H0 is accepted, and H1 is rejected. If Fcal > Ftable, then H0 is rejected, and H1 is accepted. In the F test has a hypothesis: (a) H0 is There was no significant influence between the independent variable and the bound variable; (b) H1 is There was a significant influence between the independent variable and the bound variable.

Table 13. FTable Determination Value

DF	DF					
	1	2	3	4	5	30
2	161.448	199.5	215.707	224.583	230.162	250.095
3	18.5128	19	19.1643	19.2468	19.2964	19.4624
4	10.128	9.55209	9.27663	9.11718	9.01346	8.61658
5	7.70865	6.94427	6.59138	6.38823	6.25606	5.74588
6	6.60789	5.78614	5.40945	5.19217	5.05033	4.49571
7	5.98738	5.14325	4.75706	4.53368	4.38737	3.80816
8	5.59145	4.73741	4.34683	4.12031	3.97152	3.37581
9	5.31766	4.45897	4.06618	3.83785	3.6875	3.07941
10	5.11736	4.25649	3.86255	3.63309	3.48166	2.86365

In this study, F Table = 215.70735 and F Count = 9474.05, then it is concluded that the value of Fcal is greater than the value of Ftable, then H1 is accepted and H0 is rejected, so it is concluded that the independent variables, namely the birth rate, population displacement rate, and death rate affect the growth of the population of Tanah Java sub-district. After looking for the influence between the free variable and the bound variable, the prediction value will be searched for in the next 5 years, from 2024 to 2028, using the multiple linear regression equation. In the prediction search stage, the multiple linear equations found previously can be used to make predictions using the formula $B_0 + B_1 * X_1 + B_2 * X_2 + B_3 * X_3$. The method can be seen in the following. It can be done using the RapidMiner 5.3 application to prove whether the prediction value is correct. The initial stage will be to determine the training data that will be tested on RapidMiner (see Figure 4). Based on Figure 5, it is obtained that the value of the population in 2024 will increase to a total of 61177 people and in 2025 to a total of 62051,503. In 2026, it will be 61855.55; in 2027, it will be 60079.26; and in 2028, it will be 56011.80, so based on the prediction results that have been obtained, the number of residents of the Tanah Java sub-district will increase (see Figure 5).

$$\begin{aligned} \text{Year 2024} &= 5079.023551 + (2.06846842 * 23283) + (0.418260449 * 44177) + (-1.340866371 * 7860) = 61178 \\ \text{Year 2025} &= 5079.023551 + (2.06846842 * 25343) + (0.418260449 * 45465) + (-1.340866371 * 10788) = 62052 \\ \text{Year 2026} &= 5079.023551 + (2.06846842 * 27585) + (0.418260449 * 46791) + (-1.340866371 * 14806) = 61856 \\ \text{Year 2027} &= 5079.023551 + (2.06846842 * 30025) + (0.418260449 * 48155) + (-1.340866371 * 20321) = 60079 \\ \text{Year 2028} &= 5079.023551 + (2.06846842 * 32682) + (0.418260449 * 49559) + (-1.340866371 * 27890) = 56012 \end{aligned}$$

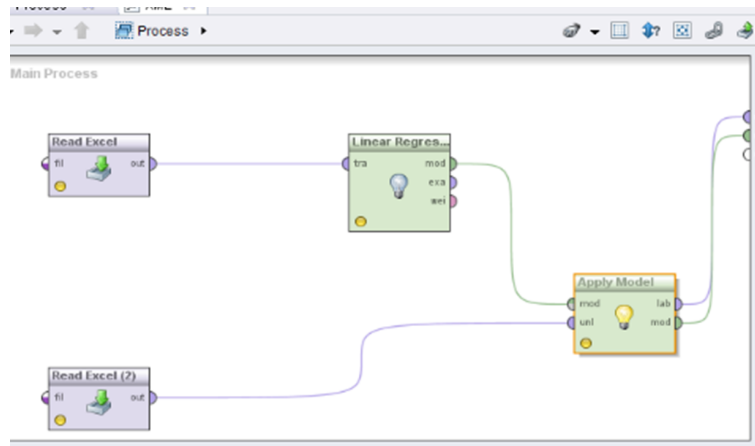


Figure 4. Rapid Miner Process

ExampleSet (5 examples, 3 special attributes, 3 regular attributes)						
Row No.	Tahun	Y	prediction(Y)	X1	X2	X3
1	2024	?	61177.943	23283.400	44176.700	7860.160
2	2025	?	62051.421	25343.100	45464.800	10787.900
3	2026	?	61855.680	27585.100	46790.500	14806
4	2027	?	60079.294	30025.400	48154.900	20320.900
5	2028	?	56011.822	32681.600	49559	27889.900

Figure 5. Prediction Results 2024-2028

4. CONCLUSION

The conclusions obtained after conducting the prediction and testing process using the multiple regression method, commonly known as Multiple Linear Regression and geometric models are data from the Simalungun Regency Civil Registry Office that the population growth of Tanah Java Regency has been increasing every year from 2020 to 2023. Based on the prediction results, the population of the Tanah Java sub-district is expected to increase in the next five years (2024-2028). Predictions show that by 2024, the population is expected to reach 61178 people from 59589 in 2023. The impact of the results of this study is that in the next five years, the number of people significantly with the contribution generated can be objective information for the government to make wise decisions to ensure the population's welfare. In the future, research on the stability model of the population growth rate should be developed to control the population's growth and development in the Tanah Java sub-district, Simalungun Regency.

5. ACKNOWLEDGEMENTS

The Acknowledgments section is optional. Research sources can be included in this section.

6. DECLARATIONS

AUTHOR CONTRIBUTION

Safii Engage in research conceptualization, design methodologies, conduct investigations, and draft manuscripts and contribute to manuscript review, editing, and visualization. Rika Setiana Involved in the conceptualization of research, designing methodologies, collecting data and drafting manuscripts.

FUNDING STATEMENT

This research is carried out independently with personal financing at all stages starting from data collection to the process to publish

the research results.

COMPETING INTEREST

In this study, the authors stated that there were no competing interests.

REFERENCES

- [1] I. H. Kurniasih, M. T. Furqon, and S. Adinugroho, "Prediksi Pertumbuhan Penduduk di Kota Malang menggunakan Metode Extreme Learning Machine (ELM)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 4, no. 2, pp. 509–516, 2020.
- [2] W. A. L. Sinaga, S. Sumarno, and I. P. Sari, "The Application of Multiple Linear Regression Method for Population Estimation Gunung Malela District," *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, vol. 1, no. 1, pp. 55–64, 2022, <https://doi.org/10.55123/jomlai.v1i1.143>.
- [3] V. Skirbekk, "Fertility, Population Growth, and Population Composition BT - Decline and Prosper!: Changing Global Birth Rates and the Advantages of Fewer Children," V. Skirbekk, Ed. Cham: Springer International Publishing, 2022, pp. 329–355, https://doi.org/10.1007/978-3-030-91611-4_18.
- [4] I. Mardiyah, W. Dianita Utami, D. C. Rini Novitasari, M. Hafiyusholeh, and D. Sulistiyawati, "Analisis Prediksi Jumlah Penduduk Di Kota Pasuruan Menggunakan Metode Arima," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 15, no. 3, pp. 525–534, 2021, <https://doi.org/10.30598/barekengvol15iss3pp525-534>.
- [5] F. Fejriani, Hendrawansyah, and L. Muharni, "Forecasting Peningkatan Jumlah Penduduk Berdasarkan Jenis Kelamin Menggunakan Metode Arima," *Kajian, Penelitian dan pengembangan pendidikan*, vol. 8, no. 1, pp. 27–36, 2020.
- [6] D. Anggreini, "Penerapan Model Populasi Kontinu Pada Perhitungan Proyeksi Penduduk Di Indonesia (Studi Kasus: Provinsi Jawa Timur)," *E-Jurnal Matematika*, vol. 9, no. 4, p. 229, 2020, <https://doi.org/10.24843/mtk.2020.v09.i04.p303>.
- [7] M. M. Maja and S. F. Ayano, "The Impact of Population Growth on Natural Resources and Farmers' Capacity to Adapt to Climate Change in Low-Income Countries," *Earth Systems and Environment*, vol. 5, no. 2, pp. 271–283, 2021, <https://doi.org/10.1007/s41748-021-00209-6>.
- [8] S. Li, "Predicting the Total Population Development of China Based on Logistic Blocking Growth Model and Improved Grey GM (1,1) Prediction Model," in *Proceedings of the 6th International Conference on Information Systems Engineering*, ser. ICISE '21. New York: Association for Computing Machinery, 2022, pp. 26–30, <https://doi.org/10.1145/3503928.3503934>.
- [9] A. E. Raftery and H. Ševčíková, "Probabilistic population forecasting: Short to very long-term," *International Journal of Forecasting*, vol. 39, no. 1, pp. 73–97, 2023, <https://doi.org/https://doi.org/10.1016/j.ijforecast.2021.09.001>.
- [10] A. Garre, J. Koomen, H. M. W. den Besten, and M. H. Zwietering, "Modeling Population Growth in R with the biogrowth Package," *Journal of Statistical Software*, vol. 107, no. 1 SE - Articles, pp. 1–51, sep 2023, <https://doi.org/10.18637/jss.v107.i01>.
- [11] N. Alghanmi, R. Alotaibi, S. Alshammari, and A. Mahmood, "Population Fusion Transformer for Subnational Population Forecasting," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, p. 26, 2024, <https://doi.org/10.1007/s44196-024-00413-y>.
- [12] F. Riandari, H. T. Sihotang, and H. Husain, "Forecasting the Number of Students in Multiple Linear Regressions," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 249–256, 2022, <https://doi.org/10.30812/matrik.v21i2.1348>.
- [13] M. Gregorich, S. Strohmaier, D. Dunkler, and G. Heinze, "Regression with Highly Correlated Predictors: Variable Omission Is Not the Solution," *International Journal of Environmental Research and Public Health*, vol. 18, p. 4259, apr 2021, <https://doi.org/10.3390/ijerph18084259>.
- [14] L. Chen, T. Mu, X. Li, and J. Dong, "Population Prediction of Chinese Prefecture-Level Cities Based on Multiple Models," *Sustainability (Switzerland)*, vol. 14, no. 8, pp. 1–23, 2022, <https://doi.org/10.3390/su14084844>.

- [15] F. V. Şahinarslan, A. T. Tekin, and F. Çebi, "Application of machine learning algorithms for population forecasting," *International Journal of Data Science*, vol. 6, no. 4, pp. 257–270, jan 2021, <https://doi.org/10.1504/IJDS.2021.122770>.
- [16] G. Adhikari, "Interpreting the Basic Results of Multiple Linear Regression," *Scholars' Journal*, pp. 22–37, dec 2022, <https://doi.org/10.3126/scholars.v5i1.55775>.
- [17] A. B. Abadi, A. Fadlullullah, S. Sumardi, S. Mahdi, and A. N. Juniar, "Perhitungan Indeks Massa Tubuh Less Contact Berbasis Computer Vision dan Regresi Linear," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 629–638, 2022, <https://doi.org/10.30812/matrik.v21i3.1512>.
- [18] M. Iwasaki, "Multiple Regression Analysis from Data Science Perspective BT - Advanced Studies in Behaviormetrics and Data Science: Essays in Honor of Akinori Okada," T. Imaizumi, A. Nakayama, and S. Yokoyama, Eds. Singapore: Springer Singapore, 2020, pp. 131–140, https://doi.org/10.1007/978-981-15-2700-5_8.
- [19] J. Rana, P. L. Gutierrez, and J. C. Oldroyd, "Quantitative Methods BT - Global Encyclopedia of Public Administration, Public Policy, and Governance," A. Farazmand, Ed. Cham: Springer International Publishing, 2020, pp. 1–6, https://doi.org/10.1007/978-3-319-31816-5_460-1.
- [20] Y. Sun, X. Wang, C. Zhang, and M. Zuo, "Multiple Regression: Methodology and Applications," *Highlights in Science, Engineering and Technology*, vol. 49, pp. 542–548, may 2023, <https://doi.org/10.54097/hset.v49i.8611>.
- [21] H. Nayebi, "Multiple Regression Analysis BT - Advanced Statistics for Testing Assumed Causal Relationships: Multiple Regression Analysis Path Analysis Logistic Regression Analysis," H. Nayebi, Ed. Cham: Springer International Publishing, 2020, pp. 1–46, https://doi.org/10.1007/978-3-030-54754-7_1.
- [22] N. Nagajothi, S. S. Dhenakaran, and S. U. Maheswari, "A Geometric Method for Extracting Images of PDF Files," in *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 2022, pp. 1096–1100, <https://doi.org/10.1109/ICAAIC53929.2022.9792982>.