SKEWNESS
Jurnal Statistika, Aktuaria dan Sains Data
Volume 2, No. 1, April 2025

Clustering BMKG Stations in Central Java Based on Meteorological Characteristics Using K-Means Clustering

Tri Ayu Mulyani [1], Agung Prabowo [2*], Niken Larasati [3], Setyo Luthfi Okta Yohandoko [4]

[1,3]*Mathematics Study Program, Mathematics Department, Faculty of Mathematics and Natural Sciences, Universitas Jenderal Soedirman, Purwokerto, Indonesia.*

[2]*Statistics Study Program, Mathematics Department, Faculty of Mathematics and Natural Sciences, Universitas Jenderal Soedirman, Purwokerto, Indonesia.*

[4]*Mathematics Study Program, Mathematics Department, Faculty of Mathematics and Natural Sciences, Universitas Padjdajaran, Sumedang, Indonesia.*

Coresponding author: agung.prabowo@unsoed.ac.id

Abstract. The Meteorology, Climatology, and Geophysics Agency (BMKG) is a trusted provider of weather data. The meteorological characteristics assessed at BMKG stations include elevation, air temperature, humidity, rainfall, and the number of rainy days. This research aims to categorize BMKG stations according to these meteorological characteristics by employing the K-Means Clustering method. An analysis of data from 34 BMKG stations in Central Java, collected between 2019 and 2023, shows an average elevation of 101.6536 meters above sea level, an average air temperature of 22.5965 degrees Celsius, humidity levels at 66.9224 percent, total rainfall measuring 1,848.3924 mm, and an average of 119.5294 rainy days. The clustering process led to the formation of three distinct clusters: Cluster 1, which includes 22 BMKG stations; Cluster 2, containing 6 BMKG stations; and Cluster 3, which also comprises 6 BMKG stations. The assessment of clustering effectiveness using the Davies-Bouldin Index resulted in a DBI value of 1.02575, suggesting that while the clustering outcomes are satisfactory, they are not fully optimized.

Keywords: *meteorological characteristics, bmkg stations, k-means clustering, davies-bouldin index.*

1 Introduction

The Meteorology, Climatology, and Geophysics Agency (BMKG) is pivotal in observing, analyzing, and disseminating information regarding atmospheric, climate, and geophysical conditions in Indonesia. BMKG stations serve as operational units, equipped with instruments to monitor various meteorological factors such as elevation, air temperature, humidity, rainfall, and the number of rainy days. These data are essential for identifying weather patterns, predicting climate changes, and facilitating disaster risk management [1].

Elevation has a significant effect on air temperature, with higher altitudes typically experiencing cooler temperatures. Increased temperatures can lead to higher evaporation rates, while humidity is crucial in influencing the potential for cloud formation and precipitation [2]. Rainfall and the frequency of rainy days are key indicators of precipitation intensity and are vital for understanding water distribution patterns in an area [3]. The interactions among these factors produce unique weather patterns that define specific regions [4].

Meteorological stations located throughout Indonesia, especially in Central Java, are crucial for providing accurate and dependable weather information. The precision of weather data is important because it significantly impacts various daily activities, including public safety, socio-economic operations, agricultural production, plantations, fisheries, aviation, and more in any given region [4]. Central Java features diverse topography, ranging from lowlands to mountains, resulting in a variety of weather patterns [5]. The existence of multiple BMKG stations in this region is vital for effectively monitoring atmospheric dynamics. Nevertheless, each station functions under distinct geographical and meteorological conditions, which can lead to variations in observational results. Thus, clustering BMKG stations based on the meteorological characteristics observed is necessary. This method seeks to develop a more holistic understanding of weather patterns in Central Java and to aid in disaster preparedness and climate change adaptation planning. Data mining is one technique that can be employed for clustering BMKG stations [6].

Data mining can be categorized into six functional groups: description, estimation, prediction, classification, clustering, and association [7]. Clustering entails organizing data into groups with similar characteristics and is essential in various data mining applications, including scientific data exploration, information retrieval and text mining, spatial database applications, and web analysis [7].

There are two main approaches to clustering: hierarchical clustering and non-hierarchical clustering. Hierarchical clustering is used when the number of clusters is unknown, while non-hierarchical clustering is suitable when the number of clusters is predetermined. K-Means Clustering falls within the non-hierarchical clustering category [7].

2 Materials and Methods

2.1 Data and Research Variables

This study utilizes secondary data obtained from the Central Statistics Agency (BPS) of Central Java for the years 2019–2023 [8],[9],[10],[11] and [12]. The data include elevation (measured in meters above sea level), air temperature (in degrees Celsius), humidity (as a percentage), rainfall (in millimeters), and the number of rainy days (in days). The variables involved are as follows:

Table 1. Research variables

Variable	Description
x_{i1}	Data value from i-th BMKG station for elevation variable
x_{i2}	Data value from i-th BMKG station for air temperature variable
x_{i3}	Data value from i-th BMKG station for humidity variable
x_{i4}	Data value from i-th BMKG station for rainfall variable
x_{i5}	Data value from i-th BMKG station for rainy days variable

2.2 Binary Logistic Regression Analysis

This research adopts a literature study approach, collecting information from various sources. The following steps are undertaken to cluster BMKG stations in Central Java using the K-Means Clustering method:

1. Determine the value of (k), representing the number of clusters to be created, using the elbow method.
2. Define initial centroids or cluster centers, selected randomly.
3. Calculate the distance from all data points to each centroid.
4. Assign each data point to the nearest cluster based on calculated distances.
5. Update centroids by averaging the values of each cluster's members.
6. Repeat steps 2–5 until convergence is reached, meaning the members of each cluster no longer change.

The effectiveness of the clustering will be assessed using the Davies-Bouldin Index.

3 Results and Discussion

3.1 Descriptive Statistics and Data Standardization

Descriptive statistics are used to compare the minimum, maximum, average, and standard deviation for each variable. The calculation process for descriptive statistics is as follows:

The average elevation is calculated as follows:

$$\bar{x} = \frac{\sum_{i=1}^{36} x_i}{36} = \frac{7.01 + 77.16 + 55.111 + \cdots + 7.267}{36} = 134.1385$$

The standard deviation for elevation is as follows:

$$s = \frac{\sqrt{\sum_{i=1}^{36}(x_i - \bar{x})^2}}{36 - 1} = \frac{\sqrt{(7.01 - 134.1385)^2 + (77.16 - 134.1385)^2 + \cdots + (7.267 - 134.1385)^2}}{35}$$
$$= 181.9068$$

This calculation method is applied to each variable, and the resulting descriptive statistics can be found in Table 2.

Table 2. Descriptive statistics

Descriptive statistics					
	n	min	Max	average	std. dev
elevation (masl)	36	5.19	794	134.1385	181.90687
air temperature (Degree Celsius)	36	15.5	28.62	22.6994	3.05218
Humidity (%)	36	47.4	83.32	66.3544	7.13705
Rainfall (mm)	36	906.6	3717.5	1879.7428	711.72314
Rainy Days (day)	36	60	182	120.2222	29.16663
Valid N (*listwise*)	36				

The data standardization process relies on the information presented in Table 2, including mean values and standard deviations. The z-score values help identify outliers within the dataset. The z-score calculation yields z_x_{it}, representing the z-score for the (i)-th BMKG station concerning the (t)-th variable, where (i = 1, 2, 3, …, 36) and (t = 1, 2, 3, …, 5). The z-score calculation for each variable is as follows:

The z-score for the Cilacap Meteorological Station is calculated as:

$$z_{x_{11}} = \frac{7.01 - 134.1385}{181.9068} = -0.69887;$$

$$z_{x_{12}} = \frac{7.01 - 22.6994}{3.05218} = 1.44178;$$

$$z_{x_{13}} = \frac{7.01 - 66.3544}{7.13705} = 2.30425;$$

$$z_{x_{14}} = \frac{7.01 - 1{,}879.7428}{711.72314} = 1.897;$$

$$z_{x_{15}} = \frac{7.01 - 120.2222}{29.16663} = 2.1181.$$

Using the descriptive statistics and the data standardization process, two outliers are identified: the elevation data for SMPK Wadaslintang in Wonosobo, which has a z-score of 3.62749, as well as the Rowoseneng BMKG station with a z-score of 3.20662.

These outlier data points are removed, resulting in a total of 34 stations available for analysis.

3.2 Data Suitability Testing

1. Multicollinearity Testing

To detect multicollinearity, the Tolerance values for each variable are assessed. The multicollinearity test for the research data was performed using SPSS 25 on each variable. For example, the results for the elevation variable are displayed in Table 3.

Table 3 Nilai uji multicollinearity test value for elevation

Coefficients[a]							
Model	*Unstandardized Coefficients*		*Standardized Coefficients*	T	Sig.	*Collinearity Statistics*	
	B	*Std. Error*	Beta			Tol	VIF
(Constant)	-1.06×10^{-15}	0.168		0.000	1.000		
z_*score*:x_{i2}	0.14	0.201	0.14	0.695	0.493	0.725	1.38
z_*score*:x_{i3}	-0.324	0.238	-0.324	-1.36	0.184	0.515	1.94
z_*score*:x_{i4}	-0.01	0.25	-0.01	-0.04	0.968	0.467	2.142
z_*score*:x_{i5}	0.483	0.277	0.483	1.742	0.092	0.38	2.634

a. *Dependent Variable*: z_*score*: x_{i1}

According to Table 3, the Tolerance values exceed 0.1, indicating that multicollinearity is not present in the dataset.

2. KMO and Bartlett's Test

This test evaluates the adequacy of the sample for overall analysis while ensuring that the variables included in the analysis are significantly related. The outcomes of the KMO and Bartlett's Test for the research data are summarized in Table 4.

Table 4 Results of KMO and Bartlett's Test

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.565
Bartlett's Test of Sphericity	Approx. Chi-Square	48.986
	df	10
	Sig.	0.000

As shown in Table 4, the KMO test value is greater than 0.5, and the significance value from Bartlett's Test is 0.000. This suggests a significant relationship among the variables, supporting the adequacy of the sample for the analysis conducted.

3.3 Determining the Optimal Number of Clusters

One effective approach for establishing the optimal number of clusters in cluster analysis is the elbow method. This method identifies the number of clusters ((k)) that achieves optimal data separation by observing the decline in Sum of Squared Errors (SSE) as the number of clusters increases. The SSE is computed by summing the squared distances between each data point and its corresponding cluster centroid. The analysis will consider values of (k) ranging from 2 to 10.

Table 5. SSE value for each cluster

Number of Cluster	Number of Iteration	SSE	SSE Value Decrease
2	7	115.11722	-
3	**3**	**79.95795**	**35.15927**
4	3	74.4159	5.54203
5	3	55.60554	18.81036
6	4	38.58384	17.02171
7	2	43.22187	-4.63804
8	3	29.43734	13.78453
9	3	24.96053	4.47682
10	3	24.12132	0.8392

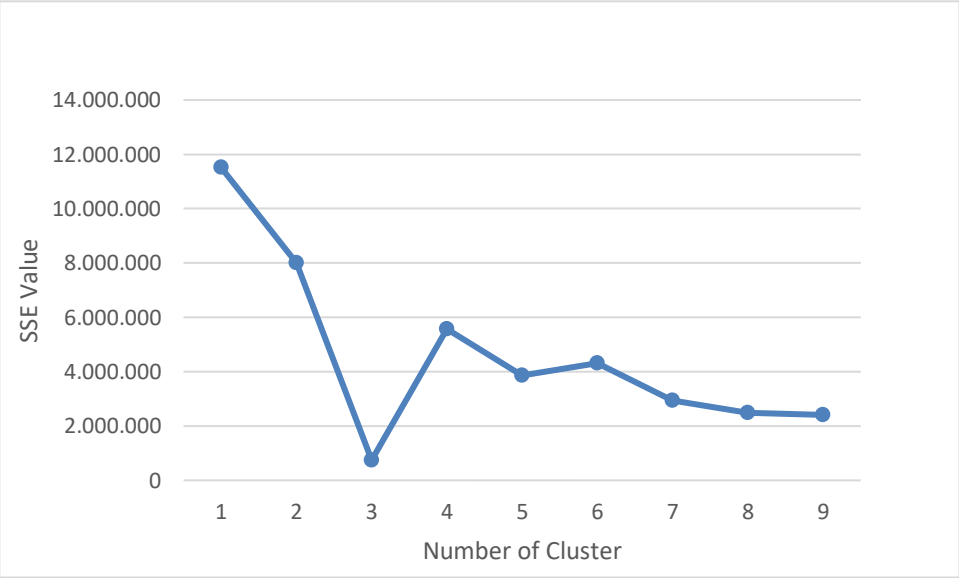

Figure 1. SSE Value Decrease Graph

Table 5 shows that at (k = 3), there is a significant reduction in the Sum of Squared Errors, as depicted in Figure 1. This indicates that (k = 3) (three clusters) is the optimal number of clusters, as evidenced by the clear slope of the lines.

3.4 Clustering Process for the Optimal Number of Clusters

According to the elbow method, the optimal cluster formation consists of three clusters. The clustering process begins by randomly selecting initial centroids from the 34 data points being analyzed. The initial centroids, determined using IBM SPSS 25, are shown in Table 6.

Tabel 6 Initial Centroid Values

Initial Cluster Centers			
	Cluster		
	1	2	3
z_*score*: Elevation	-0.76776	2.69737	1.56875
z_*score*: Air Temperature	-0.29171	1.79841	0.45381
z_*score*: Humidity	-0.35498	-0.50783	2.50641
z_*score*: Rainfall	-1.24916	0.24719	2.29270
z_*score*: Rainy Days	-1.92713	-0.18523	2.02565

3.4.1 Iteration Process

The iteration process starts after the random selection of initial centroids. The subsequent step involves calculating the Euclidean distance of each data point from each centroid. Using the initial centroid values in Table 6, the calculated Euclidean distances during the first iteration are as follows:

The Euclidean distance between the Cilacap Meteorological Station and centroid 1 is calculated as:

$$(z_{X_{1t}}, c_{1t}) = \sqrt{\begin{pmatrix} (-0.77736-(-0.76776))^2+(1.49887-(-0.29171))^2+ \\ (2.42693-(-0.35498))^2+(1.95117-(-1.24916))^2+ \\ (2.09265-(-1.92713))^2 \end{pmatrix}}$$

$$= 6.0285$$

The Euclidean distance between the Cilacap Meteorological Station and centroid 2 is calculated as:

$$(z_{X_{1t}}, c_{1t}) = \sqrt{\begin{pmatrix} (-0.77736-2.69737)^2+(1.49887-1.79841)^2+ \\ (2.42693-(-0.50783))^2+(1.95117-0.24719)^2+ \\ (2.09265-(-0.18523))^2 \end{pmatrix}}$$

$$= 5.2532$$

The Euclidean distance between the Cilacap Meteorological Station and centroid 3 is calculated as:

$$d(z_{X_{1t}}, c_{1t}) = \sqrt{\begin{pmatrix}(-0.77736 - 1.56875)^2 + (1.49887 - 0.45381)^2 + \\ (2.42693 - 2.50641)^2 + (1.95117 - 2.2927)^2 + \\ (2.09265 - 2.02565)^2\end{pmatrix}}$$

$$= 2.5148$$

A second iteration is necessary to explore the potential movement of BMKG stations into different clusters. In this second iteration, new centroids are established by averaging the z-score values of the data within the same cluster. The calculation of the new centroid values for the elevation variable is as follows:

Iteration 2

$$c_{11} = \frac{-0.20118 + (-0.2588) + (-0.0194) + \cdots + (-0.7753)}{22} = -0.47009$$

$$c_{21} = \frac{1.2071 + 2.6974 + 0.5170 + 1.7859 + 1.7859 + 2.0740}{6} = 1.6779$$

$$c_{31} = \frac{-0.7774 + (-0.3823) + 1.5688 + (-0.6082) + 1.2071 + (-0.7356)}{6}$$

$$= 0.0452$$

Following this, Euclidean distances are recalculated, with the minimum distance selected to assign data to specific clusters. The iteration process continues until the third iteration is complete, resulting in the final centroid values presented in Table 7.

Table 7. Final Centroid

Final Cluster Centers			
	Cluster		
	1	2	3
z_*score*: Elevation	-0.43736	1.79294	-0.18930
z_*score*: Air Temperature	-0.22363	-0.22625	1.04623
z_*score*: Humidity	-0.42167	-0.33969	1.88583
z_*score*: Rainfall	-0.40930	0.33005	1.17073
z_*score*: Rainy Days	-0.46844	0.54615	1.17145

3.4.2 Characteristics of Clustering Results

In the K-Means clustering analysis, each formed cluster demonstrates unique characteristics. Based on the results from the three iterations, three clusters have been identified, with the membership count for each cluster displayed in Table 8.

Table 8. Number of Members in each cluster

Number of Cases in each Cluster		
Cluster	1	22.000
	2	6.000
	3	6.000
Valid		34.000
Missing		0.000

The characteristics of the clustering results can be analyzed by comparing the centroid values to the average values of each variable. A significant deviation of a cluster's centroid from the average indicates that the cluster possesses distinct characteristics compared to the overall dataset. The average values for each variable are as follows: elevation at 115.2852 meters above sea level, air temperature at 22.7434 degrees Celsius, humidity at 66.896 percent, rainfall at 1849.044 mm, and the number of rainy days at 119.5294 days. Based on the final centroid values in Table 8, the characteristics of each cluster can be summarized as follows:

1. Cluster 1 comprises BMKG stations characterized by meteorological features that are below average for elevation, air temperature, humidity, rainfall, and the number of rainy days.
2. Cluster 2 consists of BMKG stations with significantly above-average elevation, while air temperature and humidity are below average; and both rainfall and the number of rainy days are above average.
3. Cluster 3 contains BMKG stations that, while having below-average elevation, exhibit above-average values for air temperature, humidity, rainfall, and the number of rainy days.

3.4.3 Evaluation of Clustering Results

1. Sum of Square Within Cluster (SSW)

The calculation of SSW requires determining the nearest distances from each data point to the centroids obtained in the final iteration.

$$SSW_1 = \frac{0.6827 + 2.0891 + 0.815 + \cdots + 0.7211}{22} = 1.26429$$

$$SSW_2 = \frac{2.3821 + 1.491 + 2.3425 + 0.8062 + 0.4343 + 1.0535}{6} = 1.41826$$

$$SSW_3 = \frac{1.517 + 2.0268 + 1.3888 + 2.4115 + 2.8887 + 1.3258}{6} = 1.92641$$

2. Sum of Square Between Cluster (SSB)

The SSB calculation requires the centroids identified during the final iteration. The methodology for calculating the SSB is as follows:

$$SSB_{1,2} = \sqrt{\begin{pmatrix}(-0.43736 - 1.79294)^2 + \left(-0.22363 - (-0.22625)\right)^2 + \\ \left(-0.42167 - (-0.33969)\right)^2 + (-0.40930 - 0.33005)^2 + \\ (-0.46844 - 0.54615)^2\end{pmatrix}}$$

$$= 2.56066$$

$$SSB_{1,3} = \sqrt{\begin{pmatrix}(-0.43736 - (-0.18930))^2 + (-0.22363 - 1.04623)^2 + \\ (-0.42167 - 1.88583)^2 + (-0.40930 - 1.17073)^2 + \\ (-0.46844 - 1.17145)^2\end{pmatrix}}$$

$$= 3.49061$$

$$SSB_{2,3} = \sqrt{\begin{pmatrix}(1.79294 - (-0.18930))^2 + (-0.22625 - 1.04623)^2 + \\ (-0.33963 - 1.88583)^2 + (0.33005 - 1.17073)^2 + \\ (0.54615 - 1.17145)^2\end{pmatrix}}$$

$$= 3.40575$$

3. Cluster Ratio

After determining the values of SSW (Sum of Squares Within) and SSB (Sum of Squares Between), the next step is to calculate the ratio between the clusters. The results of the inter-cluster ratios are presented in Table 9.

$$R_{1,2} = \frac{1.26429 + 1.41826}{2.56066} = 1.047599$$

$$R_{1,3} = \frac{1.26429 + 1.92641}{3.49061} = 0.914083$$

$$R_{2,3} = \frac{1.41826 + 1.92641}{3.40575} = 0.982065$$

Table 9 R-Max Matrix Results

Ratio between x-th *cluster*	1	2	3	R-*Max*
1	0	1.047599	0.914083	1.047599
2	1.047599	0	0.982065	1.047599
3	0.914083	0.982065	0	0.982065

Based on the calculations, the maximum ratios are as follows: $R_{1,2}$=1.047599; $R_{1,2}$=1.047599; $R_{2,1}$=1.047599; $R_{2,1}$=1.047599; and $R_{2,3}$=0.982065.

4. Davies-Bouldin Index

Once the inter-cluster ratio values have been established and the maximum ratio selected, the next step is to calculate the Davies-Bouldin Index (DBI).

$$DBI = \frac{1}{3}(1.047599 + 1.047599 + 0.982065) = 1.02575$$

The calculated Davies-Bouldin Index for the three formed clusters yields a value of 1.02575. This indicates that the accuracy level of clustering using the K-Means method into three clusters is considered quite optimal.

4 Conclusion

Based on the research findings and discussions regarding the clustering of 36 BMKG stations in Central Java using the K-Means Clustering method based on meteorological characteristics, the following conclusions can be drawn:

1. There are two outlier data points with meteorological characteristics that significantly exceed the average elevation, specifically BMKG Wadasalintang Wonosobo and BMKG Rowoseneng Temanggung stations.
2. The average values for each meteorological characteristic are as follows: elevation 115.2852 meters above sea level, air temperature 22.7434 degrees Celsius, humidity 66.896 percent, rainfall 1849.044 mm, and rainy days 119.7143 days.

3. The clustering of BMKG stations resulted in three clusters as follows:
    a. Cluster 1: Consists of 22 BMKG stations with meteorological characteristics below average for elevation, air temperature, humidity, rainfall, and rainy days.
    b. Cluster 2: Comprises 6 BMKG stations with an elevation significantly above average, while air temperature and humidity are below average, and rainfall and rainy days are above average.
    c. Cluster 3: Contains 6 BMKG stations with elevation below average, while air temperature, humidity, rainfall, and rainy days are above average.

The following recommendations can be made for future research:

1. Apply the K-Means Clustering method while eliminating outliers by modifying extreme values to maximum or minimum limits.
2. Implement the K-Medoids method, which can effectively handle outliers by using median values rather than means as cluster centers.
3. Incorporate methods for evaluating the internal validity of the clustering results to compare and enhance the accuracy of the derived clusters.

References

[1] BMKG. (2024). Stasiun dan UPT BMKG. Diakses Oktober 2024, dari https://iklim.bmkg.go.id

[2] Rohmana, S. F., Rusgiyono, A., & Sugito. (2019). Penentuan faktor-faktor yang mempengaruhi intensitas curah hujan dengan analisis diskriminan ganda dan regresi logistik multinominal. Jurnal Gaussian, 8(3), 398–406.

[3] Nugroho, S., Febriamansyah, R., Ekasaputra, E. G., & Gunawan, D. (2019). Analisis iklim ekstrem untuk deteksi perubahan iklim di Sumatera Barat. Jurnal Ilmu Lingkungan, 17(1), 7–14.

[4] Puspitasari, N., & Haviludin. (2016). Penerapan metode K-Means dalam pengelompokan curah hujan di Kalimantan Timur. Seminar Nasional Riset Ilmu Komputer (SNIRK), 1(1), 20–25.

[5] BMKG Provinsi Jawa Tengah. (2024). Iklim dan topografi Jawa Tengah. Diakses Oktober 2024, dari https://iklim.bmkg.go.id

[6] Chandra, A. (2016). Pembangkitan itemset untuk aturan asosiasi dengan algoritma Apriori data mining. *Teknologi Data*, 2(1).

[7] Han, J., Pei, J., & Tong, H. (2023). *Data mining: Concepts and techniques* (4th ed.). Cambridge: Morgan Kaufmann Publishers.

[8] Badan Pusat Statistik. (2020). *Provinsi Jawa Tengah dalam angka 2020*. Jawa Tengah: Badan Pusat Statistik Provinsi Jawa Tengah.

[9] Badan Pusat Statistik. (2021). *Provinsi Jawa Tengah dalam Angka 2021*. Jawa Tengah: Badan Pusat Statistik Provinsi Jawa Tengah.

[10] Badan Pusat Statistik. (2022). *Provinsi Jawa Tengah dalam Angka 2022*. Jawa Tengah: Badan Pusat Statistik Provinsi Jawa Tengah.

[11] Badan Pusat Statistik. (2023). *Provinsi Jawa Tengah dalam Angka 2023*. Jawa Tengah: Badan Pusat Statistik Provinsi Jawa Tengah.

[12] Badan Pusat Statistik. (2024). *Provinsi Jawa Tengah dalam Angka 2024*. Jawa Tengah: Badan Pusat Statistik Provinsi Jawa Tengah.