MISSING DATA IMPUTATION VIA OPTIMIZATION APPROACH: AN APPLICATION TO K-MEANS CLUSTERING OF EXTREME TEMPERATURE

Geovert John D. Labita[1], Bernadette F. Tubo[2]

•

[1]University of Science and Technology of Southern Philippines
[2]Mindanao State University – Iligan Institute of Technology
[1]geovertjohn.labita@g.msuiit.edu.ph

Abstract

This paper introduces an optimization approach to impute missing data within the K-means cluster analysis framework. The proposed method has been applied to Philippine climate data over the previous 18 years (2006-2023) with the goal of classifying the regions according to average annual temperature including the maximum and minimum. This dataset contains missing values which is the result of the weather stations' measurement failure for some time and there is no chance of recovery. As an effect, the regional groupings are greatly affected. This paper adapts a modified method of missing value imputation suitable for climate data clustering, inspired by the work of Bertsimas et al. (2017). The proposed methodology focuses on imputing missing values within observations by finding the value that minimizes the distance between the observation and a cluster centroid in which the Mahalanobis distance is used as the similarity measure. Consequently, the outcomes of clustering obtained through this optimization approach were compared with certain imputation techniques namely Mean Imputation, Expectation-Maximization algorithm, and MICE. The assessment of the derived clusters was conducted using the silhouette coefficient as the performance metric. Results revealed that the proposed imputation gave the highest silhouette scores which means that most of the observations were being clustered appropriately as compared to the results using other imputation algorithms. Moreover, it was found out that most of the areas showing the features of extreme condition are located in the middle part of the country.

Keywords: Optimization, K-Means, Mahalanobis

I. Introduction

The risk of extreme temperature most directly affects health by compromising the body's ability to regulate its internal temperature. Loss of internal temperature control can result in various illnesses including heat cramps, heat exhaustion, heatstroke, and hyperthermia from extreme heat events [7]. Thus, awareness of the climatic differences of a particular region of interest becomes a major concern for the safety of the individual.

In detecting weather phenomena like extreme temperature, it is important to classify or cluster the regions according to their climatic elements. However, the problem of missing climatic data is common in most weather stations which might result from damaged or failure of the weather equipment or instrument. Also, events such as sickness or vacation of the personnel in-charge can create daily missing data values which could affect the climate statistics. If this happens, there will

be no record of measurements for a particular time and could affect the clustering of weather data which is a valuable endeavor in multiple respects. For example, the results can be used in various ways within a larger weather prediction framework or could simply serve as an analytical tool for characterizing climatic differences [4].

From the study of Calvo et al. [6], a new clustering technique was shown aiming to generate a robust regionalization using climate datasets with incomplete information. Their method provided a new approach to cluster time series of different temporal lengths using most of the information contained in heterogeneous sets of climate records. Although they showed that their algorithm is able to generate a climatically consistent regionalization, it must be noted that there is no imputation happened on the missing information. In a sense, the clustering accuracy is somehow questionable.

A common practice for dealing with missing values in the context of clustering is to first impute the missing values, and then apply the clustering algorithm on the completed data [5]. From the study of Bertsimas et al. [3], a flexible framework based on formal optimization to impute missing data was proposed. Specifically, this framework can readily incorporate various predictive models like the k Nearest Neighbors (kNN) for data classification in which the missing data of an observation is imputed by determining the k nearest observations and getting the average of those k observations. However, the imputation for each observation is not based on the possibility that the point belongs to a particular cluster. Thus, the kNN imputation is based purely on the k neighbors without the involvement or intervention of the possible resulting clustering.

Trying to resolve the aforementioned issues or deficiencies, this paper creates an appropriate imputation technique for missing values when dealing with clustering problem. Specifically, this study aims to construct a two-step optimization approach for data imputation in K-means cluster analysis where K is the number of clusters. The first step is to determine the optimal initial cluster centroids which are the K most frequent nearest neighbors from all incomplete observations, that is, the K points with highest densities. The second step is then imputing the missing value of an observation by determining the value that gives the minimum distance from the observation to a cluster centroid. The outcomes of clustering achieved through this optimization approach will be compared with some imputation approaches namely Mean Imputation, Expectation-Maximization algorithm, and Multivariate Imputation by Chained Equations in which the assessment of the derived clusters will be conducted using the silhouette coefficient.

This paper is arranged as follows. Methodology is introduced and discussed in section 2. The model solution is presented and derived in section 3. In section 4, the application of the proposed imputation is illustrated while some concluding remarks are stated in section 5.

II. Methods

This section presents the derivation of the optimization models of the proposed method with imputation algorithm.

Let $X = \{x_i\}_{i=1}^{n}$ be the dataset given with p variables and assume that each data vector x_i contains continuous variables indexed by $q \in \{1, 2, \dots, p\}$. Now, the missing and known values are defined by the following sets:

$$\mathcal{M} = \{(i, q) : x_{iq}\ is\ missing\},$$
$$\mathcal{N} = \{(i, q) : x_{iq}\ is\ known\}.$$

Also, let J be the set of indices of all incomplete observations given by

$$J = \{i :\ x_i\ has\ at\ least\ 1\ missing\ coordinate\}.$$

Let $W \in \mathbb{R}^{n \times p}$ be the matrix of imputed values where w_{jq} is the imputed value for entry x_{jq} for $(j, q) \in \mathcal{M}$. The full imputation for observation x_j is referred to as w_j where $j \in J$. The idea is to consider the missing data problem as an optimization problem in which it optimizes the missing values in all incomplete data points. Thus, the key decision variables are the missing values

$\{w_{jq} : (j,q) \in \mathcal{M}\}$.

As a similarity measure, we can incorporate different types of distance metrics, but we prefer to use the Mahalanobis distance because it takes into account the variances and covariances amongst the variables which is very important in clustering multivariate data. In constructing the Mahalanobis metric, it involves the centroid of the whole dataset which means that the distance actually measures a point from the mean of the distribution. Specifically, according to Ghorbani [8], the Mahalanobis distance measures the number of standard deviations that an observation is from the mean of a distribution.

In using the Mahalanobis distance as a similarity measure, the nearest neighbors of incomplete data are formulated based on the differences of the squared Mahalanobis distances of the two observations. Thus, the nearest neighbor of each w_j, $j \in J$ is the smallest difference $M_j - M_i$ for all $i = 1, 2, \dots, n$, that is, the smallest deviations between w_j and w_i where the squared Mahalanobis distance M_i is given by

$$M_i(w_i, \mu) = [w_{i1} - \mu_1 \quad \dots \quad w_{ip} - \mu_p]\Sigma^{-1}\begin{bmatrix} w_{i1} - \mu_1 \\ \vdots \\ w_{ip} - \mu_p \end{bmatrix}$$

with $\mu = \{\mu_1, \dots, \mu_p\}$ and Σ are the mean and covariance matrix of the whole data respectively which are updated per iteration.

Imputation Model

To obtain the imputed values, the Mahalanobis distance between w_j, $j \in J$ and its appropriate centroid w_{i_l}, $l \in \{1, 2, \dots, K\}$ is minimized. Thus, for each $j \in J$, the goal is to solve the imputation model:

$$\min M_j - M_c \tag{1}$$

subject to

$$w_c \in \{w_{i_l}\} \qquad l = 1, 2, \dots, K \tag{2}$$

$$w_{jq} = x_{jq} \qquad (j,q) \in \mathcal{N} \tag{3}$$

The solution $\{w_{jq}\}$, $(j,q) \in \mathcal{M}$ are regarded as the imputed values for the corresponding $\{x_{jq}\}$. It must be noted that in the objective function (1), we assume that $M_j > M_c$. If $M_c > M_j$, we change the objective to $\max M_j - M_c$ in order to represent the same idea that the value of M_j should be near to M_c. In other words, the objective function ensures that whatever imputed values w_{jq} obtained, the observation w_j is very close to its appropriate cluster centroid w_c which is selected based on constraint (2). These centroids are determined in the assignment model discussed in the next section. The constraint (3) assures that all the observed data are preserved.

Assignment Model

Let K be the number of clusters specified by the analyst. Now, assume that the initial cluster centroids are given by $\{w_{i_l} : l = 1, 2, \dots, K\}$ which are the K most frequent nearest neighbors from all incomplete observations. To obtain the initial centroids, the immediate nearest neighbor for each w_j, $j \in J$ must be determined resulting to the following assignment model:

$$\min \textstyle\sum_{i=1}^{n} z_{ij}(M_j - M_i) \tag{4}$$

subject to

$$\textstyle\sum_{i=1}^{n} z_{ij} = 1 \tag{5}$$

$$z_{jj} = 0 \tag{6}$$

$$z_{ij} \in \{0, 1\}$$

The assignment model assigns each incomplete observation to its immediate nearest neighbor where $z_{ij} = 1$ if w_i is the nearest neighbor of w_j and 0 otherwise. The objective function (4) will

determine which w_i is the nearest neighbor of w_j among all observations. Because of constraint (5), there will only be one immediate nearest neighbor per incomplete observation and an incomplete observation cannot be the nearest neighbor of itself because of constraint (6).

From all of the nearest neighbors, the K most frequent observations can then be formulated as an optimization problem using the binary variables $y_i \in \{0,1\}$ as follows:

$$\max \sum_{i=1}^{n} y_i \sum_{j \in J} z_{ij} \qquad \text{subject to} \qquad \sum_{i=1}^{n} y_i = K \tag{7}$$

The solution $\{y_{i_1}, \dots, y_{i_K}\}$ of model (7) corresponds to the desired initial centroids $\{w_{i_1}, \dots, w_{i_K}\}$. It must be noted that the assignment model will work only on complete data with imputed values. For the first iteration with missing values, the model can be started with mean values as the warm start values for the optimization process. The imputed values from the imputation model are then based on the centroids obtained from the assignment model. In return, the centroids are updated based on the new imputed values making this procedure an iterative process.

Imputation Algorithm

The proposed data imputation algorithm is given in the following steps:

1. **Input:** $X \in \mathbb{R}^{n \times p}$, a data matrix with missing entries $\mathcal{M} = \{(i,q) : x_{iq} \text{ is missing}\}$, warm start $W^0 \in \mathbb{R}^{n \times p}$ and number of clusters K.
2. **Output:** W^*, a full matrix with imputed values, $\mu^* = \{w_{i_1}, \dots, w_{i_K}\}$ initial centroids.
3. **Initialize:** $W^{old} \leftarrow W^0$
4. **repeat**
5. Update mean μ and covariance matrix Σ based on W^{old}.
6. Update the auxiliary variables Z^* using the assignment model.
7. Update the initial centroids μ^* following:
$$\sum_{j \in J} z_{i_l j} > \sum_{j \in J} z_{ij} \quad \forall i \in \{1, 2, \dots, n\}$$
8. Update the imputation W^* using the imputation model.
9. $(Z^{old}, W^{old}, \mu^{old}) \leftarrow (Z^*, W^*, \mu^*)$
10. **until** $\mu^* = \mu^{old}$

III. Results

This section presents the solution of the proposed imputation method using Mahalanobis distance.

Proposition 1. Let $X = \{x_i\}_{i=1}^n$ be a dataset given with p variables where the missing and known values are specified by the sets $\mathcal{M} = \{(i,q) : x_{iq} \text{ is missing}\}$ and $\mathcal{N} = \{(i,q) : x_{iq} \text{ is known}\}$ respectively. If $(j,q) \in \mathcal{M}$, then the solution of the optimization problem (1-3) is given by

$$w_{jq} = \mu_q - \frac{1}{2\sigma_{qq}} \sum_{a:a \neq q}^{p} \sigma_{qa}(w_{ja} - \mu_a)$$

where $\mu_q, \sigma_{qa} \in \mathbb{R}$ and $\sigma_{qq} > 0$.

Proof. Let $(j,q) \in \mathcal{M}$ and consider the optimization problem (1-3). Suppose that $w_c = w_{i_l}$ such that $M_j - M_{i_l} < M_j - M_m$ for all $m \neq l$. Then by considering an unconstrained optimization where we plugin the values of the x_{jq} to the corresponding w_{jq} for all $(j,q) \in \mathcal{N}$ in objective function (1), we can use the concept of relative minimum in calculus to solve for w_{jq} that would minimize $M_j - M_{i_l}$. Since the missing variable w_{jq} is present only in M_j, the problem reduces to differentiating,

$$M_j = \begin{bmatrix} w_{j1}-\mu_1 & \cdots & w_{jq}-\mu_q & \cdots & w_{jp}-\mu_p \end{bmatrix} \Sigma^{-1} \begin{matrix} w_{j1}-\mu_1 \\ \vdots \\ w_{jq}-\mu_q \\ \vdots \\ \lfloor w_{jp}-\mu_p \rfloor \end{matrix}$$

with respect to w_{jq} where $\mu = \{\mu_1, \dots, \mu_p\}$ and Σ are the mean and covariance matrix respectively. Now, suppose that

$$\Sigma^{-1} = \begin{matrix} \sigma_{11} & \cdots & \sigma_{1q} & \dots & \sigma_{1p} \\ \vdots & & \vdots & & \vdots \\ \sigma_{q1} & \cdots & \sigma_{qq} & \cdots & \sigma_{qp} \\ \vdots & & \vdots & & \vdots \\ \lfloor \sigma_{p1} & \cdots & \sigma_{pq} & \cdots & \sigma_{pp} \rfloor \end{matrix},$$

then we have

$$M_j = \sum_{b=1}^{p} \sum_{a=1}^{p} \sigma_{ab}(w_{ia} - \mu_a)(w_{jb} - \mu_b).$$

To differentiate M_j, we have to separate the terms containing w_{jq}, that is,

$$M_j = \sum_{a=1}^{p} \sigma_{qa}(w_{jq} - \mu_q)(w_{ja} - \mu_a) + \sum_{b:b\neq q}^{p} \sum_{a:a\neq q}^{p} \sigma_{ab}(w_{ja} - \mu_a)(w_{jb} - \mu_b)$$

$$D_{w_{jq}}(M_j) = 2\sigma_{qq}(w_{jq} - \mu_q) + \sum_{a:a\neq q}^{p} \sigma_{qa}(w_{ja} - \mu_a).$$

Finally, equating the derivative to zero will solve for the imputed value as follows

$$2\sigma_{qq}(w_{jq} - \mu_q) + \sum_{a:a\neq q}^{p} \sigma_{qa}(w_{ja} - \mu_a) = 0$$

$$2\sigma_{qq} w_{jq} = 2\sigma_{qq}\mu_q - \sum_{a:a\neq q}^{p} \sigma_{qa}(w_{ja} - \mu_a)$$

$$w_{jq} = \mu_q - \frac{1}{2\sigma_{qq}} \sum_{a:a\neq q}^{p} \sigma_{qa}(w_{ja} - \mu_a). \qquad \blacksquare$$

The following theorem will be used to prove the next proposition.

Theorem 1 (Andreasson et al.). Suppose that $f: \mathbb{R}^d \to \mathbb{R}$ is in C^2 on $\mathbb{R}^d$, that is, f is twice differentiable with continuous second partial derivatives. Then $\nabla f(w^*) = 0^{(d)}$ and $\nabla^2 f(w^*)$ is positive definite implies that w^* is a *strict local minimum* of f where $\nabla f(w) = \left(\frac{\partial f(w)}{\partial w_q}\right)_{q=1}^{d}$. For $d = 1$, $f'(w^*) = 0$ and $f''(w^*) > 0$ implies $w^* \in \mathbb{R}$ is a *strict local minimum*.

Proposition 2. The solution w_{jq} given in Proposition 1 is a strict local minimum of the optimization problem (1-3) in an unconstrained setting.

Proof (for the case when $d = 1$). Let $f: \mathbb{R} \to \mathbb{R}$ be defined by the objective function in the optimization problem (1-3) in an unconstrained setting. Following the same argument from the proof of Proposition 1, for any solution w^*, we have

$$f'(w^*) = 2\sigma_{qq}(w^* - \mu_q) + \sum_{a:a\neq q}^{p} \sigma_{qa}(w_{ja} - \mu_a) \quad \Rightarrow \quad f''(w^*) = 2\sigma_{qq}.$$

Since $f'(w^*)$ and $f''(w^*)$ are linear functions, then they are continuous. Also, $f''(w) = 2\sigma_{qq} > 0$ since the diagonal entries of a covariance matrix are positive assuming that the data samples are unique. Now,

$$f'(w_{jq}) = 2\sigma_{qq}\left(\mu_q - \frac{1}{2\sigma_{qq}}\sum_{a:a\neq q}^{p}\sigma_{qa}(w_{ja}-\mu_a) - \mu_q\right) + \sum_{a:a\neq q}^{p}\sigma_{qa}(w_{ja}-\mu_a)$$
$$= 2\sigma_{qq}\left(-\frac{1}{2\sigma_{qq}}\sum_{a:a\neq q}^{p}\sigma_{qa}(w_{ja}-\mu_a)\right) + \sum_{a:a\neq q}^{p}\sigma_{qa}(w_{ja}-\mu_a)$$
$$= -\sum_{a:a\neq q}^{p}\sigma_{qa}(w_{ja}-\mu_a) + \sum_{a:a\neq q}^{p}\sigma_{qa}(w_{ja}-\mu_a)$$
$$= 0.$$

Thus, by Theorem 1, the solution w_{jq} is a strict local minimum. ■

IV. Application

The proposed methodology is applied on the historical Philippine climate data (2006-2023) taken from the 52 weather stations around the country which can be downloaded at https://en.tutiempo.net/climate/philippines.html and shown in Table 2. This dataset of three continuous variables per year (52×54 data matrix) contains actual missing values. This study can be considered as a multivariate time series clustering with the goal of classifying the regions suspected to have extreme temperature conditions.

In doing the experiment, the missing elements among the data are firstly imputed using the different imputation methods, and then the traditional K-means algorithm is applied into the imputed dataset. The experiments with random centroid initialization (mean, MICE, EM) are repeated 100 times with different random seed to reduce the effect of randomness caused by the traditional K-means, and report the best result.

We use the R function "*silhouette()*" from the R package "*cluster*" for obtaining the silhouette scores of the clustering results. Silhouette coefficient or Silhouette score ranging from -1 to +1 is a measure of how similar an object is to its own cluster compared to other clusters. In other words, it is a metric used to calculate the goodness of a clustering [2]. A high value indicates that the object is well matched or having a high relationship to its own cluster. Thus, it acts as the accuracy in the case when the cluster labels are not known.

Table 1 shows the silhouette score results from different number of clusters where the numbers in red are the highest score per case.

Table 1: *Silhouette Scores (%) using different imputation algorithms*

# of Clusters	Proposed Imputation	Mean Imputation	MICE	Expectation-Maximization
K=2	84.78	75.26	61.51	70.98
K=3	72.6	62.7	52.64	58.66
K=4	58.02	36.03	29.39	21.75
K=5	58.02	22	26.38	19.12
K=6	57.99	20.5	33.33	19.04
K=7	41.11	20.09	17.89	18.17
K=8	38.06	17.89	17.66	17.16
K=9	36.56	17.59	17.03	16.65
K=10	36.27	16.88	15.74	17.75

Using the proposed imputation method, we can classify the extreme temperature areas. For example, if we set $K = 10$, results showed that there are two clusters exhibiting extreme temperature having an overall average of at least 28°C. These areas are shown in Figure 1.

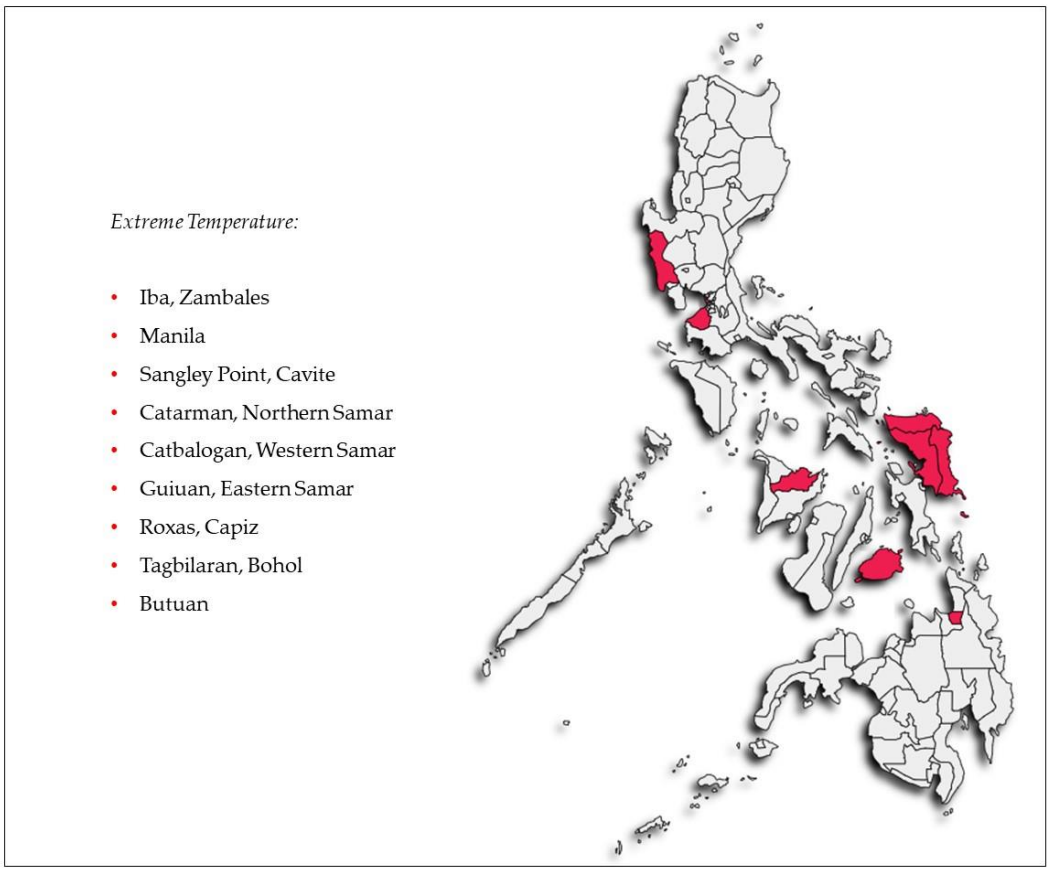

Figure 1: *Philippine map with clustering results from the proposed imputation*

From Figure 1, the areas with red spots are classified with extreme temperature. It can be observed that most of the areas are located in the middle part of the country.

V. Concluding Remarks

This paper presents a missing data imputation algorithm that can handle partitional clustering. It is created out of an optimization approach for imputing missing data and making use of the Mahalanobis distance metric as a similarity measure. Also, it avoids the problem of centroid initialization when performing K-means clustering because the initial cluster centroids are fixed based on the algorithm's generated centroids.

When clustering the Philippine Climate data with 21% actual missing values, we were able to identify 9 places with extreme temperature classification which means that these places must be considered when predicting extreme temperature occurrence. It was found out that the proposed imputation using Mahalanobis distance gave higher clustering performance and is consistent for different number of clusters which means that the proposed optimization approach using Mahalanobis distance is a suitable imputation algorithm in the context of partitional clustering.

Table 2: *Philippine Climate Dataset 2006-2023 (T-average temperature, TM-average maximum temperature, Tm-average minimum temperature)*

Weather Stations	2006			2007			2008			2009			2010			2011			2012			2013			2014		
	T	TM	Tm	T	TM	Tm	T	TM	Tm	T	TM	Tm	T	TM	Tm	T	TM	Tm	T	TM	Tm	T	TM	Tm	T	TM	Tm
Alabat																						27.4	31	24	27.3	30.8	23.8
Ambulong																						27.7	31.3	23.8	27.6	31.5	23.2
Aparri				27.8	32.1	23.8	27.4	31.5	23.4	27.1	30.8	23.5	27.6	31.6	23.7	27	30.6	24.2	27.6	31.1	24.4	27.4	31.1	24.3	26.8	30.7	24
Baguio	19	23.8	15.5	19	23.8	15.5	18.7	23.3	15.3	18.7	23.2	15.4	19.4	24.6	15.3	18.8	23.6	14.5	19.4	23.9	14.8	19.5	24.1	15.3	19	23.3	15
Baler Radar																			26.1	30.2	23.1				26.2	30.6	23
Basco Radar	25.4	27.7	22.1										25.9	28.1	22.2	24.7	26.8	21.1	25.2	27.4	21	25.3	27.3	20.8	25	27.2	22.1
Cabanatuan	27.4	33.2	23.1	27.4	33.3	23.1	27	32.9	23.1	26.9	33.1	23.1	27.2	34	23.4	26.7	32.6	23.4	27.5	32.8	23.6	27.7	33	23.9	27.6	33.1	23.7
Casiguran													26.9	31.7	22	26.1	31	20.3	26.5	31.5	22.4	26.9	30.8	23.3	26.8	31.3	22.8
Clark AB	26.4	31.2	22.3	26.5	31.3	22.2	26.1	30.8	22.3	26.2	30.7	22.2	26.8	32	22.4	26	30.7	22.2	26.5	31.3	22.5	26.6	31.3	22.6	26.7	31.1	23.1
Daet	27.8	31.1	25	27.8	31	24.9	27.4	30.8	24.8	27.4	30.4	24.8	27.7	31.1	24.8	27.3	30.5	24.7	27.6	31	24.4	27	30.9	23.9	27.3	30.7	23.8
Dagupan	27.7	32.2	23.1	27.6	32.1	22.8	27.4	32.1	22.9	27.3	31.9	22.7	28	32.8	23.4	27.3	31.5	23.3	27.8	32.3	23.6	28	32.7	23.8	27.6	32.4	23.5
Iba	27.7	31.5	23.6				27.7	31.5	23.8	27.4	31.3	23.5				27.6	31.3	24.1	28	32.1	24.3	28.3	32.3	24.5	27.8	31.6	24
Infanta	27.3	31.2	24.1				26.9	30.7	23.8	27	31	23.8	27.4	31.6	23.8	26.8	30.7	23.7	27.4	31.6	23.9	27.2	31.2	23.6	27	30.4	23.7
Itbayat										25.4	28.1	21.6				25.1	27.9	21.8	26.1	28.7	22.7	26	28.3	22.5	25.3	28	21.8
Laoag	28.9	32.9	23.1	28.8	32.5	22.8	28.7	32.3	22.9	28.4	32	22.8	28.4	32.7	22.9	28	31.8	23	28	31.6	23	28	31.7	23.1	27.6	31.3	22.5
Legaspi	27.7	30.6	24.9				27.3	30.3	24.6	27.4	30.5	24.8	27.9	31.5	24.7	27.3	30.4	22.7	27.6	30.8	23.3	27.7	31.3	24.2	27.5	30.8	24.5
Manila	28.6	31.7	25.7	28.7	31.8	25.7	28.3	31.4	25.5	28.2	31	25.4	29	32.2	26	28.2	31.2	25.5	28.4	31.6	25.5	28.4	31.8	25.3	28.1	31.8	24.6
Masbate	27.9	32.2	24.9	28.1	32.5	24.9	27.8	31.9	24.8	27.9	31.9	24.8	28.4	32.7	25	27.5	31.7	24.6	27.8	32.1	24.7	28.1	31.8	24.7	27.8	31.6	24.6
NAIA	28	31.6	24.7	28.1	32.1	24.7	27.7	31.5	24.6	27.6	31.3	24.5	28.5	32.7	25	27.7	31.5	24.5	28	31.8	24.7	28.2	32.1	24.8	28.1	32.1	24.7
Puerto Princesa	27.8	32.1	24.5	28	32.3	24.4	27.5	31.5	24.2	27.8	32	24.4	28.1	32.2	24.6	27.6	31.5	24.5	27.9	31.7	24.7	28	31.9	24.8	27.8	31.5	24.5
Romblon	27.7	30.2	25.4	27.6	30.2	25.1	27.4	29.9	25	27.2	29.7	25	28	31	25.4				26.3	28.7	24.4	26.3	28.8	24.3			
Sangley Point	28.8	32	26.1	28.6	32	25.7	28.3	31.5	25.5							28.3	31.6	25.6	28.7	32.1	25.8	28.9	32.3	25.9	28.2	32.1	25.1
Science Garden	27.4	31.8	23.5	27.5	32.1	23.6	27.1	31.9	23.2	27	31.7	23.4	27.7	32.9	23.7	26.8	31.5	23.3	27.2	32	23.3	27.4	32.2	23.4	27.2	32	23.1
Subic Bay	27.8	31.9	24.1	27.9	32.1	23.9	27.5	31.5	23.9	27.6	31.6	23.7	28	33.1	24	27.4	32	23.9	27.7	32.2	24	28.1	33.1	23.8	27.6	32.8	23.1
Tanay				23	26.6	20.1	22.6	25.8	20.2	22.5	25.7	20.1	23.2	27.2	20.4	22.4	25.8	20.2	22.7	26	19.8	22.9	26.3	19.9	22.8	26.4	19.7
Tayabas	26.6	30.4	23.4	26.6	30.7	23	26.2	30.2	22.9	26.2	30.1	22.8	26.7	30.9	23.2	26.2	29.7	22.8	26.5	30.3	23	26.6	30.4	23	26.2	30	22.7
Tuguegarao	27.2	32.5	22.6	27.1	32.6	21.2	26.6	31.8	20.6	26.5	31.4	20.9	27.5	33.3	21.4	26.5	31.6	21.4	26.9	32.2	21.9	26.9	32	22.2	26.8	32.3	22.5
Vigan	27.3	31.1	23.2	27.1	31.1	22.9	27.2	30.9	23	27.2	31	23.2	27.4	31.9	23	27	31.4	22.5	27.3	31.5	23.2	27.3	31.4	23.1	26.8	31.3	22.8
Borongan																									26.9	30.5	23.4
Calapan				27.2	31.2	23.6										26.7	30.5	23.1	27	31.5	23.4						
Catarman	27.6	31.3	23.8	27.6	31.5	23.3	27.2	31.2	23	27.5	31.4	23.3	27.5	31.9	23.2	27.3	30.9	23.4	27.6	31.3	23.7	27.7	31.4	23.7	27.2	30.9	23.5
Catbalogan	28.1	32.3	24.4				27.6	31.9	22.6	28	31.9	24.1	28.1	32.8	24.7	27.6	31.5	24.6	28	32.4	24.7	28.1	32.4	24.6	27.6	32.2	24.3
Dumaguete	27.6	30.6	24.3	27.6	30.8	24.4	27.4	30.7	24.2	27.5	30.9	24.3	27.9	31.3	24.4	27.5	31.1	24.3	27.7	31.4	24.4	28	31.6	24.7	27.6	31.3	24.4
Guiuan	27.6	30	25.4	27.7	30	25.5	27.4	29.5	25.2	27.6	30.1	25.1	27.9	30.7	24.9	27.4	29.9	24.3	27.9	30.5	24.5				27.4	29.8	24
IloIlo	27.9	31.7	23.9	27.9	32	23.4				27.2	30.7	23	28	32.1	23.5												
Maasin				27.3	31.6	24.1	26.9	31.1	23.7	27.1	31.6	23.9	27.1	32.3	23.5	26.5	30.5	21.9	27.2	31.5	21.5	27.3	31.3	23.3	27	30.9	23.7
Mactan	27.9	30.9	24.8	28	31.2	24.5	27.8	31	24.7	27.8	31.4	25	28.4	32.3	24.9	27.4	30.8	24.7	27.8	31.3	24.9	28.1	31.4	25.2	27.8	31.2	24.6
Roxas	27.8	30.8	25.1	28	31	25	27.7	30.6	25	27.8	30.7	25.2	28.2	31.3	25.4	27.6	30.3	25.2	28	30.9	25.5	28	30.9	25.5	27.9	30.8	25.3
San Jose																			28.1	32.3	23.8	28.2	32.4	23.8	27.8	32.4	23.4
Tacloban	27.7	31.3	24.9				27.3	31	24.7	27.6	31.3	24.7	27.6	31	24.9	27.2	30.3	24.8	27.5	31.1	24.8						
Tagbilaran	28.1	32.1	24.3				27.8	32.1	23.9	28.2	32.6	24.2	28.2	33.1	24.1	27.7	32.2	24.1	28.1	32.9	24.3	27.9	31.8	24.3	27.4	31.4	23.9
Virac				27.5	30.8	23.8	27.2	30.6	23.8	27.3	30.7	23.9	27.4	31.1	23.8	27	30.5	23.7	27.4	30.9	24	27.3	30.8	24	27	30.8	23.4
Butuan	27.8	31.9	24.2	27.8	32.2	23.9	27.5	32.1	24	27.8	32.1	24.3	27.9	32.6	24	27.5	31.7	24.1	27.6	32	24	27.5	32.1	23.8	27.6	31.8	23.9
Cotabato																									27.8	32.6	23.3
Davao Airport	28.2	31.8	24.2	28.3	31.7	24.1	28	31.6	23.9	28.5	32	24.2	28.7	32.4	24.3	28.2	31.7	24.1	28.4	32	24.2	28.6	32.3	24.4	28.5	32.4	24.3
Dipolog	27.9	32	24.6	27.8	31.6	24.7	27.3	31.3	24.4	27.5	31.7	24.5	27.7	31.8	24.7	27.4	31.3	24.2	27.5	31.2	24.2	27.6	31.4	24.1	27.6	31.8	23.9
Gen. Santos				27.6	32.7	23.1	27.1	32.1	22.7	27.6	32.8	23.2	27.8	33.1	23.1	27.1	31.9	22.6	27.1	32.1	22.7						
Hinatuan	27.9	31.8	24.2	27.5	31.7	23.9	27.2	31.2	23.7				27.4	31.3	23.8	27.1	31.2	23.6	27.3	31.5	23.9	27.6	31.3	24	27.3	31.3	23.8
Lumbia	26.4	31.6	22.3	26.4	31.6	22.2	26	31.2	21.7	26.4	31.8	22.4	26.6	32.3	22.6	26.1	31.7	22.5	26.4	32.2	22.4						
Malaybalay	24.2	29.2	19.1	24.3	29.6	18.9	24.1	29.5	18.7	24.1	29.3	18.4	24.4	30.8	16.6	23.8	29.8	17.2	23.7	29.7	17.7	23.9	30	18.2	23.9	29.8	18.7
Surigao	28.7	31.4	25.4	28.1	31.6	24.7	27.9	31.3	24.3	27.7	31.3	24.1	27.9	31.7	24	28	31.4	23.5	28.1	31.9	24.2	27.8	31.6	23.6	27.8	31.5	23.5
Zamboanga	28.3	32.4	24	28.3	32.3	24.1	28.2	31.9	23.8	28.5	32.4	23.8	28.7	33.1	21.9	28.4	32.8	23.7	28.7	32.5	24.3	28.5	32.6	24.1	28.8	33.3	23.9

Weather Stations	2015			2016			2017			2018			2019			2020			2021			2022			2023		
	T	TM	Tm	T	TM	Tm	T	TM	Tm	T	TM	Tm	T	TM	Tm	T	TM	Tm	T	TM	Tm	T	TM	Tm	T	TM	Tm
Alabat	27.4	31.1	23.7	27.7	31.4	24	27.4	31.1	23.9	27.5	31.1	24	27.7	31.5	23.7	27.6	31.2	24	27.6	30.8	24.3						
Ambulong	27.8	32	23.2	28.1	32.8	23.7	27.9	32.4	23.8	27.9	32.1	23.8	28	32.4	23.4				27.7	31.8	23.9						
Aparri	27.1	30.4	24.4	27.2	30.2	24.5	26.9	29.4	24.6	27	29.6	24.7	27	29.7	24.8	27.3	30.2	25				27.3	30.4	24.8	27.5	30.6	24.7
Baguio	19	23.3	15.2	19.5	23.7	15.6	19.3	23.7	15.6	19	23.4	15.3	19.1	23.5	15.3	19.5	24.2	15.5				19.4	23.6	16	19.7	23.8	16.4
Baler Radar	26.4	31	23.3	26.7	31	23.6	26.2	30	23.3	26.3	30.5	23.4	26.4	30.8	23.4	26.3	30.4	23.4	26.5	30.6	23.5						
Basco Radar	25.3	27.5	22.4	25.7	27.8	22.9	25.5	27.5	22.5	24.6	27.3	22.3	25.4	27.7	22.9	25.6	27.9	22.7				25.3	27.6	22.7	25.1	27.7	21.9
Cabanatuan	27.5	33.2	23.3	28.1	33.2	24	28	33.1	23.8	27.9	33.2	23.7															
Casiguran	26.5	30.8	22.5	26.8	31.1	23.2	26.4	30.8	23.2	26.8	31.3	23.2	26.9	31.7	23	27	31.4	22.6				26.8	30.9	22.3	27	31.6	22.3
Clark AB	27	31.4	23.5	27.7	32.1	24.2	27.6	32.2	24	27.5	31.9	24	27.4	32.1	23.7	27.9	32.8	23.8	27.7	32.5	23.9	28.1	32.8	24.3	28.1	32.8	24.2
Daet	27.3	30.9	23.5	27.6	31.2	23.5	27.3	30.7	23.9	27.4	30.8	24.5	27.6	31.4	24.2	27.6	31	24				27.4	30	24.3	27.8	31.1	25
Dagupan	27.9	32.5	23.7	28.1	32.6	24.3	28	32.8	24.2	27.9	32.5	23.8	27.8	32	23.8	28.1	32.1	24.4				28.1	32.6	24.5	28.4	32.6	24.8
Iba	27.7	31.7	23.4	27.9	32.1	23.5	28	31.9	24.2	27.8	30.8	24.6	27.7	31.3	23.9	28.1	31.9	24.2	28	32	24.2						
Infanta	27.1	30.7	23.9	27.5	31.1	24	27.1	30.5	22.9	27.2	30.7	22.9	27.5	31.2	23.9	27.4	30.9	23.9									
Itbayat	25.6	28.2	22.2	26	28.5	22.8	25.7	28.3	22.4	25.4	27.9	21.9				26	29.1	22.1									
Laoag	27.6	31.2	22.5	27.9	31.5	23.2	28	31.6	23.2	27.9	31.4	23	27.7	31.1	23	28	31.6	23.3	28	31.7	23.3	28	31.6	23.4	28.2	32	23.7
Legaspi	27.6	30.9	24.7	28.1	31.4	25.3	27.9	31.2	25.1	27.8	31.2	24.9	28	31.6	24.9	28.2	31.8	25.3				27.9	31.2	25.2	27.9	31.3	25
Manila	28.3	32.3	24.8	28.7	32.8	25.4	28.4	32.1	25.2	28.4	31.7	25.4	28.6	32	25.4				28.7	32	25.2						
Masbate	28	32.1	24	28.2	32	24	27.9	31.8	23.6	28.1	31.4	23.6	28.2	31.7	24.8	28.2	31.9	25				28.2	32.2	25.2	28.3	31.8	25.5
NAIA	28.3	32.5	24.7	28.7	32.9	25.2	28.4	32.4	25	28.4	32.3	25	28.6	32.6	25.4	28.8	32.8	25.3	28.7	32.7	25.3	28.5	32.6	25.4	28.7	32.6	25.4
Puerto Princesa	28.1	31.7	24.7	28.9	32.3	25	28.6	31.9	24.9	28.7	31.8	24.9	28.7	32	24.6	28.9	32.3	25.1	28.3	31.7	25	28.5	31.6	25.2	28.8	32.1	25.1
Romblon				26.9	29.5	24.8	26.5	28.9	24.5	26.4	28.8	24.5	26.8	29.6	24.5	26.9	29.4	24.7				26.2	28.5	24.4	26.5	28.9	24.6
Sangley Point	28.9	32.5	25.9	29.1	32.4	26.3	28.9	32	26.3	28.9	32.1	26.2	28.8	32.6	25.9	28.9	32.8	25.9	28.8	32.5	26.1						
Science Garden	27.6	32.4	23.3	28.2	32.5	24.2	27.9	32.1	24	27.8	31.8	23.9	28	32.2	23.9	28.1	32.3	24.1				27.9	31.9	24.4	27.9	31.8	24.4
Subic Bay	27.7	32.8	23.2	28	33.2	23.9	27.7	32.4	24	27.8	32.3	24.1	27.5	31.9	23.7	27.8	32.2	24	27.7	32.2	24	27.7	32	24.1	27.9	32.1	24.2
Tanay	23	26.7	20	23.4	26.4	20.5	23.2	26.4	19.5	23.1	26.1	19.2	23.3	26.3	20.8	23.5	26.7	21	23.1	26.1	20.9						
Tayabas	26.1	30.5	22.7	26.4	30.8	23.3	26.2	30.3	22.7	26.3	30.6	23.3	26.5	31	23.3	26.5	30.8	23.2	26.1	29.9	23						
Tuguegarao	26.8	32.2	22.5	27.4	32.6	23.2	26.6	31.5	23	27	32.5	22.9	26.9	32.3	22.8	27.1	32.4	23.1									
Vigan	27	31.4	23.2	27.1	31.4	23.4	27	31.5	23.3	27	30.7	23.4	26.9	30.6	23.2	27.3	31.2	23.5	27.5	31.3	23.6						
Borongan	27.2	30.6	23.8	27.3	30.8	23.9	27.2	30.8	23.9	27.3	31.4	24.1				27.3	31.4	24.3	27.2	31.2	24.3						
Calapan	27.5	31.7	23.7	27.3	31.4	23.8	27	31.1	23.5	27.1	31.4	23.8	27.7	32.3	23.8	27.8	32.2	24				27.4	31	24	27.7	31.3	24.9
Catarman	27.1	30.9	23.4	27.3	31.3	23.6	27	30.8	23.7	27.3	31.3	24	27.3	31.4	23.6	27.2	31.1	23.9	27.2	30.8	24.1						
Catbalogan	28.2	32.8	24.5	28.5	33	24.9	28.1	32.2	24.8	28.2	32.5	24.8	28.5	33.2	24.6	28.3	33.1	24.8	28.2	32.3	24.8						
Dumaguete	28	31.6	24.6	29	31.8	24.8	28.3	31	24.4	28.9	31.4	24.8	28.9	31.4	24.8	28.9	31.2	25.3	28.6	31.1	25.3	28.8	31.1	25.3	28.8	31.1	25.1
Guiuan	27.6	30.7	24.5	27.7	31.2	24.3	27.3	31.5	24.1	27.3	31.4	24.1	27.6	31.9	24.6	27.7	32.5	24.5	27.4	31.2	24						
IloIlo																											
Maasin	27.2	30.6	24.3	27.5	31.1	24.4	26.9	29.9	24.2	27.1	29.9	24	27.1	30.4	24.1	27.2	30.6	24									
Mactan	28	31.6	24.7	28.3	32.1	25	27.8	31.4	24.8	28.1	31.8	24.9	28.2	32	24.9	28.3	32	25				28	31.2	25.1	28.3	31.4	25.5
Roxas	28.2	31.2	25.4	28.4	31.6	25.6	28.2	31.3	25.5	28.3	31.4	25.7	28.3	31.5	25.6	28.4	31.6	25.8	28.3	31.1	25.8						
San Jose	28.1	32.7	23.6	28.4	32.8	23.5	28.3	32.7	22.9	28.2	32.6	22.2	28.3	33.6	21.6	28.5	33.6	21.3				28.5	33.2	21.8	28.5	33.1	24.7
Tacloban													27.8	31.4	24.7	27.7	31.1	25.2									
Tagbilaran	27.3	31.3	23.8	27.8	31.7	24.3	27.2	31	24	27.1	31.4	23.6	27.1	31.4	23.4	27.6	31.7	23.8	27.6	31.5	23.9						
Virac	27.2	31	23.6	27.7	31	24.2	27.5	30.8	24.1	27.5	31.4	24	27.5	30.9	23.5	27.6	30.9	23.9	27.6	30.7	24.3						
Butuan	28	32.4	23.9	28.2	33.1	24.1	27.7	32.1	24.1	27.9	32.4	23.9	27.8	32.3	23.7	27.8	32.5	23.7	27.7	32.1	23.8						
Cotabato	28.1	32.9	23.3	28.7	33.8	23.5	28.2	33.4	23.1	28.4	34.1	23.1	28.4	34.1	23.2	28.3	34.6	23.5	27.8	34	23.4						
Davao Airport	28.2	32.7	24.2	28.6	33.3	24.5	27.8	32.2	24.2	28.1	32.9	24.3	27.9	32.6	24.2	28.1	32.9	24.4	27.8	32.3	24.3	27.7	32.1	24.3	28	32.9	24.4
Dipolog	28.2	32.3	23.9	28	32.4	23.6	27.6	31.9	23.9	27.9	32.3	24.4	28.1	33.2	24.2	28.5	33.2	24.5				27.9	32.6	24.4	28.1	32.8	24.5
Gen. Santos							28.7	32.6	22.9				28.9	32.7	22.7	28.9	32.9	22.9	27.3	31.5	22.3				28.5	32.3	22.9
Hinatuan	27.3	31.4	23.8	27.4	31.5	23.8	27.3	31.4	24	27.4	31.9	23.8	27.3	31.7	23.7	27.4	31.6	23.8				27.2	31.2	23.8			
Lumbia																											
Malaybalay	24	30.3	18.6	24.6	30.9	18.6	24.2	30.2	19.6	24.5	30.6	19.7	24.6	30.7	19.4	24.5	30.7	18.1	24.2	30.4	18.7						
Surigao	28.1	32.2	23.1	28.2	32.3	24.1	27.7	31.8	24.3	28	32.2	24.4	28.1	32.3	24.5	27.8	31.2	24.3				27.7	31	24.1	28	31.3	24.6
Zamboanga	29.1	33.6	23.6	29	33.3	24.2	28.6	32.7	23.9	28.6	33.2	23.6	28.5	33.4	23.7	28.5	33.7	23.9	28.4	33.7	23.9	28.1	33.3	24	28.6	33.9	24.2

References

[1] Andr´easson, N., Evgrafov, A., & Patriksson, M. (2005). An introduction to optimization: Foundations and fundamental algorithms. *Chalmers University of Technology Press: Gothenburg*, Sweden, 1;1-205.

[2] Bhardwaj, A. (2020). Silhouette coefficient validating clustering techniques. *Towards Data Science*.

[3] Bertsimas, D., Pawlowski, C., & Zhuo, Y.D. (2017). From predictive methods to missing data imputation: an optimization approach. *J. Mach. Learn. Res.*, 18(1);7133-7171.

[4] Beveridge, N.R. (2021). Deep learning for weather clustering and forecasting. *Air Force Institute of Technology*. https://scholar.afit.edu/etd/5082

[5] Boluki, S., Zamani Dadaneh, S., Qian, X., & Dougherty, E.R. (2018). Optimal clustering with missing values. *In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 593-594.

[6] Carro-Calvo, L., Jaume-Santero, F., García-Herrera, R., & Salcedo-Sanz, S. (2021). k-Gaps: a novel technique for clustering incomplete climatological time series. *Theoretical and Applied Climatology*, 143(1-2);447-460.

[7] Ebi, K.L., Capon, A., Berry, P., Broderick, C., de Dear, R., Havenith, G., ... & Jay, O. (2021). Hot weather and heat extremes: health risks. *The Lancet*, 398(10301);698-708.

[8] Ghorbani, H. (2019). Mahalanobis distance and its application for detecting multivariate outliers. *Facta Universitatis, Series: Mathematics and Informatics*, 583-595.