A NOVEL APPROACH FOR CLIMATE CLASSIFICATION USING AGGLOMERATIVE HIERARCHICAL CLUSTERING

SRI SANKETH UPPALAPATI

201416127

Email: sanketh.uppalapati@research.iiit.ac.in

Thesis submitted in partial fulfillment

of the requirements for the degree of

Master of Science

in

Building Science and Engineering by Research

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY, HYDERABAD

JUNE 2024

International Institute of Information Technology

Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled “A Novel Approach for Climate Classification using Agglomerative Hierarchical Clustering” by Sri Sanketh Uppalapati (201416127), has been carried out under my supervision and is not submitted elsewhere for a degree.

Date: Adviser: Dr. Vishal Garg

ACKNOWLEDGEMENTS

I am immensely grateful to Prof. Vishal Garg whose guidance and mentorship have been paramount in shaping this thesis. Prof. Jyotirmay Mathur, Prof Vikram Pudi, Aviruch Bhatia, and Raj Gupta deserve my sincere thanks for their valuable assistance, which significantly enriched the content and depth of my research. Special appreciation goes to Pavan Ramapragada for his unwavering support throughout this journey. To my friends and family, your constant encouragement and understanding have been my pillars of strength. This thesis is a culmination of the collective efforts of these remarkable individuals, and I extend my heartfelt thanks for their indispensable contributions.

ABSTRACT

Climate classification plays a significant role in the development of building codes and standards. It guides the design of building envelope and systems by considering their location's climate conditions. ASHRAE standard 169, as well as established climate classification systems such as Köppen and Trewartha, employ meteorological parameters like temperature, humidity, solar radiation, and precipitation for such classifications. However, the above approaches often fall short in acknowledging the relation between climatic conditions and the energy consumption of buildings, a critical consideration for comprehensive energy efficiency assessments.

This research employs clustering techniques to group cities into different climate zones based on the number of similar days. Two days are considered similar when the absolute difference between their meteorological parameters falls within a specified threshold range. A similarity matrix for the given set of cities is created using three key meteorological parameters: mean daily temperature, mean daily relative humidity, and mean daily solar radiation. This matrix, indicating the number of similar days between each pair of cities based on these meteorological characteristics, is used to cluster cities into distinct climate zones.

To assess the quality of clustering, the simulated annual energy consumption of a standard building model for each city is used. The analysis focuses on three annual energy parameters: sensible cooling, latent cooling, and heating. The quality of the clustering is evaluated using the silhouette score method, which uses annual energy consumption data. The silhouette score method considers both inter- and intra-cluster distances, with the best value being 1, the worst -1, and values near 0 indicating overlapping clusters.

The application of the proposed methodology to 786 U.S. cities' meteorological datasets has shown that the clustering, evaluated using silhouette score, computed across a set of threshold values (7 °C for daily mean temperature, 45 % for daily mean relative humidity, and 35 Wh/m² for daily mean solar radiation), has given a better clustering over the prevalent ASHRAE Standard 169 classification. The clustering achieves a better score (0.113) than ASHRAE Standard 169 (0.054).

TABLE OF CONTENTS

Acknowledgements ..... **iv**

List of Figures ..... **vii**

LIST OF TABLES ..... **viii**

LIST OF ABBREVIATIONS ..... **ix**

1 Introduction **10**

1.1 Background *10*

1.2 Motivation *12*

1.3 Aim and research question *12*

1.4 Approach for clustering *13*

1.5 Organization of thesis *14*

2 Literature Review **15**

3 Methodology **17**

3.1 Overview *17*

3.2 Climate classification parameters *18*

3.3 Method for clustering *21*

3.4 Validation method *27*

4 Analysis of Climate Classification of USA **29**

4.1 Preparation of Data *30*

4.2 Clustering of cities of USA *31*

4.3 Calculation of score *32*

5 Results and Discussion **34**

5.1 Results *34*

5.2 Limitations *36*

6 Conclusion **37**

References **38**

Appendix A: Extraction of weather data files **41**

Appendix B: Converting data to required format **42**

Appendix C: Clustering Algorithm **43**

Appendix D: Silhouette Score **44**

List of Publications **45**

LIST OF FIGURES

Figure No. Title	Page No.
Figure 1-1 ASHRAE Standard 169 climate zone definitions	12
Figure 3-1 Methodology for climate classification	19
Figure 3-2 EnergyPlus simulation engine	23
Figure 3-3 Different linkage strategies	24
Figure 3-4: AHC algorithm-based clustering process	25
Figure 3-5: Hierarchical clustering dendrogram	25
Figure 3-6: Flowchart depicting the process of generating a matrix indicating the count of similar days between cities	27
Figure 3-7: Bipartite matching	28
Figure 4-1 United States climate zone map based on ASHRAE-169	30
Figure 4-2 Number of cities in ASHRAE Standard 169 classification zones for USA	31
Figure 4-3 EnergyPlus model	32
Figure 4-4: Weather data clusters with proposed method	33
Figure 5-1 ASHRAE climate zones and spread for sensible cooling	35
Figure 5-2 ASHRAE climate zones and spread for total cooling	36
Figure 5-3 Spread with proposed clustering – sensible cooling and total cooling	36
Figure 5-4 Spread with proposed clustering – heating	37

LIST OF TABLES

Table No. **Title** **Page No.**

Table 3-1 Sample values from weather file 27
Table 3-2 Matrix of number of similar days between sample cities 28
Table 4-1 Building model parameters 32
Table 4-2 Threshold combinations and scores 34

LIST OF ABBREVIATIONS

ASHRAE	American Standard for Heating Refrigerating and Air-conditioning Engineers
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
CDD	Cooling Degree Days
CZ	Climate Zone
DBSCAN	Density-based spatial clustering of applications with noise
DOE	Department of Energy
EPW	EnergyPlus Weather
GHI	Global Horizontal Irradiance
HDD	Heating Degree Days
HVAC	Heating Ventilation and Air Conditioning
IECC	International Energy Conservation Code
NSRDB	National Solar Radiation Database
ORNL	Oak Ridge National Laboratory
PNNL	Pacific Northwest National Laboratory
TMY	Typical Meteorological Year

1 INTRODUCTION

1.1 Background

Climate classification categorizes regions based on their long-term weather patterns and meteorological conditions, providing insights into typical temperature, humidity, precipitation, and seasonal variations. It is essential for optimizing urban planning, agriculture, energy efficiency, disaster preparedness, and environmental conservation by tailoring strategies to specific climatic conditions. In building design, it is useful because similar climates often require similar strategies to achieve energy efficiency and thermal comfort.

Climatic classification introduced by Wladimir Köppen in 1990s was based on vegetation studies and in later improvements included precipitation and air temperature. Köppen initially categorized global climates into five vegetation zones. Subsequently, Rudolf Geiger undertook revisions to Köppen's world map leading to Köppen-Geiger classification world map [1].

Another important climate classification method is presented in ASHRAE Standard 169 [2]. Conceived in 2004 through surface observations, the collaboration between ASHRAE and the Department of Energy in the United States resulted in the development of climate zone maps at the Pacific Northwest National Laboratory (PNNL), integral to the formulation of building codes. It was included in the 2004 Supplement to the International Energy Conservation Code (IECC) and later incorporated into ASHRAE Standard 90.1.

The IECC climate zone map serves to categorize the United States into eight distinct temperature-oriented climate zones. Further granularity is achieved by subdividing these zones into three moisture regimes, designated as A, B, and C. Consequently, the IECC climate zone map facilitates a classification system, offering up to 24 potential climate designations as shown in Figure 1-1. These climate zone definitions, aligned with IECC and ASHRAE standards, promote consistent regional building codes for energy-efficient and sustainable construction [2].

International Climate Zone Definitions

Zone Number	Zone Name	Thermal Criteria (I-P Units)	Thermal Criteria (SI Units)
1A and 1B	Very Hot –Humid (1A) Dry (1B)	9000 < CDD50°F	5000 < CDD10°C
2A and 2B	Hot-Humid (2A) Dry (2B)	6300 < CDD50°F ≤ 9000	3500 < CDD10°C ≤ 5000
3A and 3B	Warm – Humid (3A) Dry (3B)	4500 < CDD50°F ≤ 6300	2500 < CDD10°C < 3500
3C	Warm – Marine (3C)	CDD50°F ≤ 4500 AND HDD65°F ≤ 3600	CDD10°C ≤ 2500 AND HDD18°C ≤ 2000
4A and 4B	Mixed-Humid (4A) Dry (4B)	CDD50°F ≤ 4500 AND 3600 < HDD65°F ≤ 5400	CDD10°C ≤ 2500 AND HDD18°C ≤ 3000
4C	Mixed – Marine (4C)	3600 < HDD65°F ≤ 5400	2000 < HDD18°C ≤ 3000
5A, 5B, and 5C	Cool-Humid (5A) Dry (5B) Marine (5C)	5400 < HDD65°F ≤ 7200	3000 < HDD18°C ≤ 4000
6A and 6B	Cold – Humid (6A) Dry (6B)	7200 < HDD65°F ≤ 9000	4000 < HDD18°C ≤ 5000
7	Very Cold	9000 < HDD65°F ≤ 12600	5000 < HDD18°C ≤ 7000
8	Subarctic	12600 < HDD65°F	7000 < HDD18°C

Marine (C) definition – Locations meeting all four of the following criteria:

1. Mean temperature of coldest month between 27°F (-3°C) and 65°F (18°C)
2. Warmest month mean < 72°F (22°C)
3. At least four months with mean temperatures over 50°F (10°C)
4. Dry season in summer. The month with the heaviest precipitation in the cold season has at least three times as much precipitation as the month with the least precipitation in the rest of the year. The cold season is October through March in the Northern Hemisphere and April through September in the Southern Hemisphere.

Dry (B) definition – Locations meeting the following criteria:

Not marine and

$P < 0.44 \times (T - 19.5)$ [I-P units]

$P < 2.0 \times (T + 7)$ [SI units]

Where:

P = annual precipitation in inches (cm) and

T = annual mean temperature in °F (°C).

Moist (A) definition – Locations that are not marine and not dry.

Figure 1-1: ASHRAE Standard 169 climate zone definitions [2]

In the scope of building energy efficiency, climate zoning finds application in three distinct categories: performance-based requirements, prescriptive-based requirements, and passive design guidelines. Prescriptive methods involve specific thermal attributes related to building envelope components, offering simplicity in execution and potential energy savings. Performance-based techniques consider overall building metrics, fostering versatility and improvement in energy-saving initiatives. Passive design guidelines leverage natural resources

to reduce energy consumption, focusing on building geometry, orientation, envelope features, and passive heating/cooling approaches [3].

1.2 Motivation

Initially most of the climate classifications were not developed on building energy efficiency standpoint. Most well-known climate classifications like Köppen were related to vegetation cover and later improved including patterns of seasonal precipitation and temperature. Subsequently, many classifications were developed for agriculture and other specific domains [3]. Currently the most popular criteria used for climate classification for building energy efficiency is based on ASHRAE Standard 169 which considers Heating Degree Days (HDD) and Cooling Degree Days (CDD) along with annual precipitation and annual mean temperature. Current climate classifications do not consider relative humidity and solar radiation which have a significant impact on building energy consumption. Relative Humidity and Solar Radiation play an important role in thermal comfort, thereby effecting the building energy consumption patterns and is important to consider them for classifying the climates. Relative Humidity is considered as a main factor that affects latent cooling energy consumption of a building. Solar radiation is another important meteorological parameter influencing climate classification for building energy efficiency. The amount of solar radiation that reaches a location depends on factors such as latitude and local weather conditions. Variation in solar radiation can increase or decrease the energy consumption in buildings.

Despite the widespread usage of climate classifications, there isn't a consensus regarding the optimal method for implementing climate zones within the context of building energy efficiency programs. Climate classification for building energy efficiency poses challenges due to a mix of numerous independent factors [4].

1.3 Aim and research question

The aim of this research is to use clustering technique to cluster a set of cities based on similar days to explore under-addressed link between climate classification and building energy consumption.

The proposed methodology is applied to weather dataset of U.S. cities. A range of threshold values were employed for critical meteorological parameters (daily mean temperature, daily

mean relative humidity, and daily mean solar radiation) to gauge the method's robustness across varying climatic conditions. The comparative analysis with the established ASHRAE Standard 169 classification serves as a benchmark, revealing performance of the clustering methodology.

1.4 Approach for clustering

The method developed in this research is clustering using similar days. Unlike traditional climate classification approaches, which primarily rely on meteorological parameters such as temperature, humidity, and solar radiation, the proposed method centres on the identification of similar days between cities. The key parameters considered for this analysis include mean daily temperature, mean daily relative humidity, and mean daily solar radiation.

The work in the thesis involves the following steps:

Data Preparation: Meteorological data for each city, encompassing mean daily temperature, mean daily relative humidity, and mean daily solar radiation, is collected.

Similarity Calculation: A matrix is constructed to quantify the frequency of similar meteorological conditions between pairs of cities. Similarity is based on the identified parameters.

Agglomerative Hierarchical Clustering: A matrix is used as input for agglomerative hierarchical clustering, a technique that systematically groups cities based on their similarities. This results in the formation of distinct climatic zones.

Scoring Mechanism: A scoring mechanism is developed to evaluate the quality of the clustering. This mechanism assesses the uniformity of energy consumption distribution within each identified climatic zone.

Application to U.S cities: The developed methodology is applied to diverse U.S. cities' meteorological datasets, and scores are computed for various threshold values of critical meteorological parameters.

Comparison with ASHRAE 169: The performance of the proposed method is benchmarked against the widely adopted ASHRAE 169 classification.

1.5 Organization of thesis

The thesis is organized into six chapters.

Chapter 1 gives an overview of the research. Issues and research questions relevant to the research are identified.

Chapter 2 provides a literature survey to identify current methods and the gaps that have been identified.

Chapter 3 proposes a methodology for climate classification

Chapter 4 is about employing the proposed methodology for US cities

Chapter 5 is regarding Results and Discussion.

Chapter 6 consists of Conclusion.

2 Literature Review

The construction sector plays a vital role in meeting climate objectives as it ranks as the second-largest consumer of electricity. To enhance energy efficiency in buildings, many countries have enforced mandatory regulations. These building energy regulations hinge on climate zones, and accurately classifying a city into the appropriate climate zone is crucial. In their work, Franciso Jose et al. [5] established a connection between climate zoning and its application to the energy performance of Spanish buildings. Over the past few years, several methods have been proposed for climate classification that can be applied in programs aimed at enhancing building energy efficiency. The Köppen-Geiger climate classification system [1] is the most widely adopted and referenced model globally for delineating climate zones. The Köppen climate classification categorizes climates into five primary groups, each further divided based on seasonal precipitation and temperature patterns. However, it's worth noting that this classification does not consider certain weather elements such as winds, precipitation intensity, amount of cloudiness, and daily temperature extremes.

Briggs R. S. et al. [6, 7] used Heating Degree Days (HDD) and Cooling Degree Days (CDD) based approach for classification. ASHRAE [2] has come up with a climate classification system which classifies localities into climate zones based on temperature and precipitation basis. The thermal climate zone of a locality can range from 0-8. The moisture climate zone can be Marine, Dry or Humid. Monjur Mourshed [8] has shown the importance of degree days that is used in ASHRAE classification.

There are many other classification approaches proposed such as Bansal and Minke [9] who have developed climate classification for India using mean monthly temperature and humidity values. Mayank B. et al. [10] have shown the classification of Indian cities using ASHRAE Standard 169 and compared with Bansal and Minke's classification. ORNL researchers [11] used similar methods for reimaging climate zones of the US. Cao Jingfu et al. [12] have considered cooling energy consumption to provide an efficient climate index for China. Zscheischler et al. [13] showed the value of unsupervised clustering for climate classifications. Hudson et al. [14] have provided climate classification for Columbia k-means clustering with multivariate climate data. Sathiaraj et al. [15] used k-means, DBSCAN, and BIRCH techniques for climate classification.

It was reported that DBSCAN shows less accuracy and effectiveness when applied for climate classification purposes.

Xiong Jie et al. [16] used hierarchical climate zoning for China. Shin M et al. [17] have suggested using enthalpy based CDD instead of conventional CDD value that is based on outdoor dry-bulb temperatures that neglect the influence of latent heat on the total energy consumption. Giovanni Pernigatto [18] provided a classification of European cities using cluster analysis. Walsh A et al. [19] reported that most of the current classifications are oversimplified and not fit for building energy efficiency programs. One out of six areas analyzed was mis-classified while using ASHRAE classification criteria [20].

It is noticed among all the different methodologies, there isn't any significance of building energy consumption taken into consideration when classifying, apart from the climate-derived parameters like dry-bulb temperature, humidity, precipitation data, HDD, CDD etc, when one of the objectives is proposing building energy regulations. A building situated in a location within a climate zone should have a similar thermal energy consumption in comparison with another building in another location within the same climate zone. The hypothesis idea for the proposed methodology, where the spread in thermal energy consumption of buildings within a climate zone should be minimized and differ from other zones.

This study aims to develop a methodology to classify climate using hierarchical clustering method based on the number of similar days between cities. The upcoming sections will outline the methodology, followed by the data analysis for cities in United States, the presentation of results and ensuing discussion, finally the conclusion.

3 Methodology

3.1 Overview

The idea of this study is mainly divided into three main stages: preparation of data, clustering and scoring.

In the first step, weather files (EPW format) are taken for the cities of the selected region. EPW files are historical weather files used in climate analysis of cities and for use in simulation purposes. From all the weather files in the selection region. mean daily temperature, mean daily relative humidity, and mean daily solar radiation data is calculated which will be used in step two. A building model was built to perform energy simulations for all the weather files available.

In step two, the daily mean data of all the cities are used to identify the number of similar days between each city based on the maximum bipartite matching considering different threshold values.

The similar days method considers daily data, allowing for a fine-grained analysis of climate patterns compared to traditional classification methods that often rely on monthly or annual averages.

Agglomerative hierarchical clustering [21] is used for climate clustering based on the number of similar days present between the cities for all the threshold values as shown in Figure 3-1.

In step three, the clustering is evaluated using Silhouette score method where the score ranges from -1 to +1. The building thermal load and the clustering labels for all the cities are taken as input for this method. Based on the score achieved by each threshold value, the clusters with the highest score were selected for further analysis. Also, the scores of clusters that were selected are compared with the scores achieved by the ASHRAE Standard 169 method. Details of the applied methods are presented in the following sections.

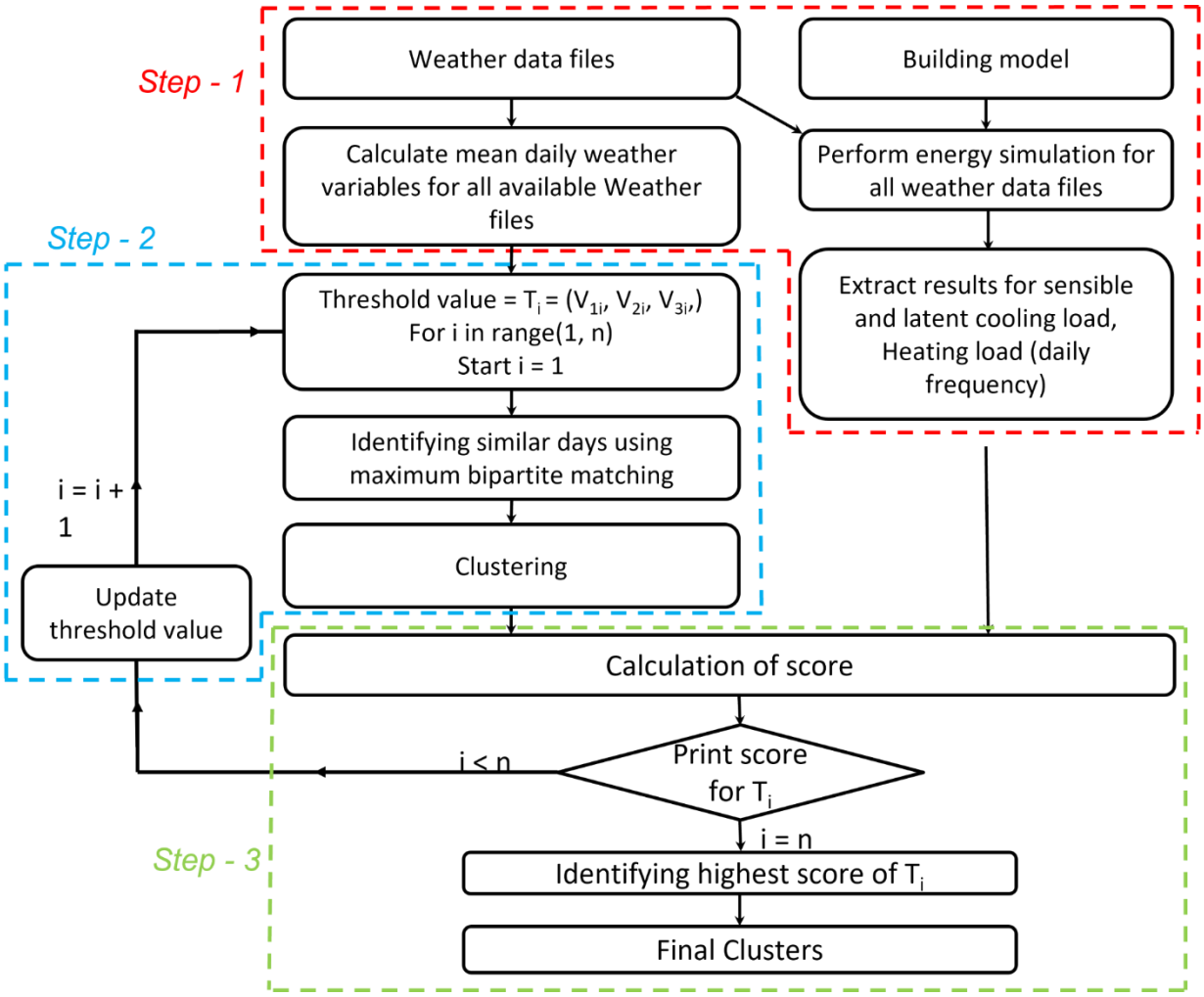

Figure 3-1: Methodology for Climate Classification

3.2 Climate classification parameters

3.2.1 Weather files

This methodology necessitates the utilization of Typical Meteorological Year (TMY) [22] data sourced from diverse geographical locations to conduct a rigorous analysis. The dataset encompasses essential meteorological parameters, including but not limited to dry bulb temperature, dew point temperature, solar radiation, relative humidity and wind speed. After data acquisition, the focus sharpens on the extraction of daily mean dry-bulb temperature, daily mean relative humidity and daily mean solar radiation from the TMY files. These refined metrics serve

as the basis for computing ‘similar’ days between two cities. The quantification of similar days establishes a metric defining the ‘closeness’ between any pair of cities, a pivotal parameter employed in the subsequent clustering approach.

3.2.1.1 Typical meteorological year (TMY)

The inception of Typical Meteorological Year (TMY) data files traces back to a period when computing resources were constrained, characterized by slower processing speeds and limited memory capacities compared to contemporary standards. These files originated from the long-term data within the National Solar Radiation Database (NSRDB) and were specifically designed to facilitate the analysis of building performance. During this era, users sought a condensed 1-year dataset that could effectively replicate the outcomes derived from utilizing the extensive 30-year data available in the NSRDB.

Crucially, the decision to focus on a 1-year dataset arose from the recognition that certain meteorological parameters wielded a more significant influence on building performance than incident solar radiation alone. Thus, the TMY datasets were crafted to encapsulate the “typical” meteorological data present in the NSRDB. Beyond just addressing computational constraints, these datasets emerged as valuable representations of the meteorological conditions crucial for understanding and simulating diverse building performance scenarios.

Each TMY data file represents a comprehensive year of data, constructed from 12 months selectively chosen as the most typical from the years constituting the NSRDB. The original files, developed by Sandia National Laboratory, employed a method where a typical month was chosen based on nine daily indices. Theses indices encompass the maximum, minimum and mean dry bulb and dew point temperatures, as well as the maximum and mean wind velocity and the total Global Horizontal Irradiance (GHI). This meticulous selection process ensured that the resulting TMY datasets not only addressed computational limitations but also remained faithful to the essential meteorological characteristics influencing building performance. Through the strategic use of daily indices, these datasets offer a condensed, yet representative snapshot of the meteorological nuances embedded in the broader NSRDB proving invaluable for a range of applications in building design, energy management and climate impact assessment.

3.2.1.2 Typical meteorological year 2 (TMY 2)

The TMY2s are datasets of hourly values of solar radiation and meteorological elements for a 1-year period. TMY2 files were created from the 1961-1990 NSRDB, in which 93% of the values were modeled. For TMY2 data files, the DNI was added to the weighting indices. This improved the comparison of annual average DNI in TMY file to long-term average DNI in the NSRDB files by an approximate factor of 2. The weighting for wind speed was reduced and the criteria for persistence were altered slightly in TMY2 and later TMY3 data files. For TMY2 files, the months from May 1982 through December 1984 were excluded from the analysis because the aerosols from the eruption of El Chichon in Mexico differed significantly from typical values [22].

3.2.1.3 Typical meteorological year 3 (TMY 3)

The TMY3s are datasets of hourly values of solar radiation and meteorological elements for a 1-year period. TMY3 data files were created from 1991-2005 NSRDB data and 1961-1990 NSRDB data if they existed for that location. For TMY3 files, the months from June 1991 to December 1994 were excluded because the aerosols from the eruption of Mount Pinatubo in the Philippines were atypical. As a result of the exclusion, 83% of TMY3 files were derived using 11.5 years of data.

Note that for the TMY2 and TMY3 datasets, half of the weight was placed on solar irradiance values and the other half on meteorological parameters.

Their intended use is for computer simulations of solar energy conversion systems and building systems to facilitate performance comparisons of different system types, configurations and locations because they typically represent rather typical than extreme conditions. They are not suited for designing systems to meet the worst-case conditions occurring at a location [22].

3.2.2 Building model

To calculate the thermal load characteristics of office buildings across the study area, a building model was first developed based on the relative building envelope present across the study area. Thermal load simulation for the model can be performed by any simulation software or tool such as EnergyPlus [23], IES-VE [24] and eQUEST [25] or any tool that provides hourly outputs based on weather files. Sensible cooling, latent cooling, and heating energy are calculated using

simulation and used for the calculation of score, considering similar internal load, and load generated using similar occupancy.

3.2.2.1 EnergyPlus

EnergyPlus software is used for the building modeling and simulation. EnergyPlus™ is a whole building energy simulation program developed US Department of Energy to model energy consumption-for heating, cooling, ventilation, lighting and plug, and process loads. EnergyPlus is a console-based program that reads input and writes output to text files. EnergyPlus uses the conduction transfer function module for the calculation of conduction through walls. Figure 3-2 shows the modules and managers used in EnergyPlus.

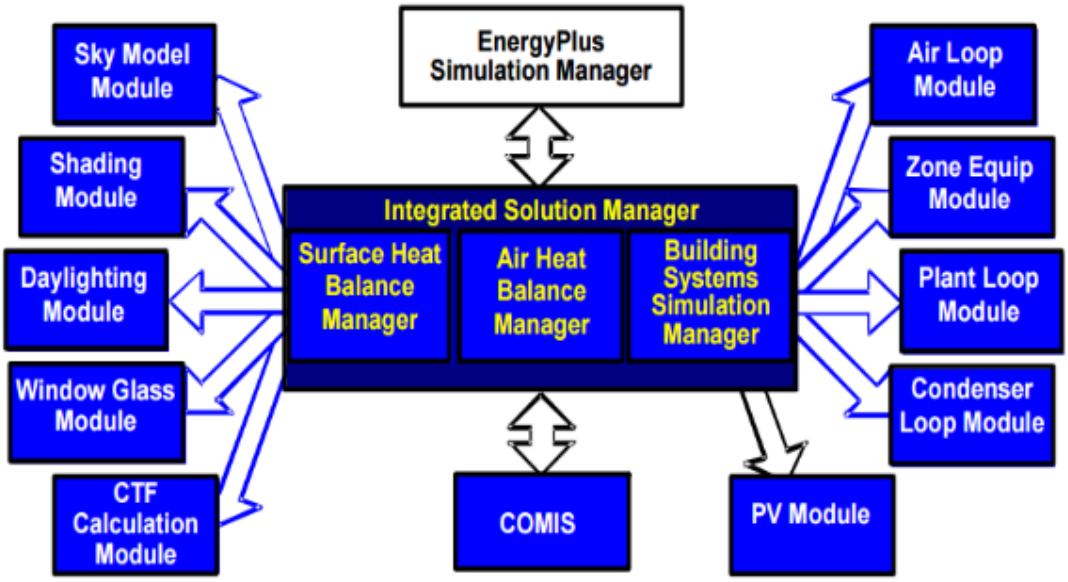

Figure 3-2: EnergyPlus simulation engine [23]

3.3 Method for clustering

As the classification labels were not known prior, the unsupervised machine learning technique is used for the classification of climates. Unsupervised machine learning algorithms discover hidden patterns or data groupings from the given dataset. In this context, the agglomerative hierarchical clustering (AHC) algorithm was utilized. AHC is a connectivity-based algorithm that groups data points together based on their proximity or closeness to one another.

3.3.1 Agglomerative Hierarchical Clustering

The algorithmic process commences by treating each location as an autonomous cluster, establishing an initial state of individualized clusters. Subsequent iterations involve the combining of clusters that exhibit marginal distinctions in climate conditions, thereby forming distinct clusters. The merging criterion in this proposed method predicates that the cities with the highest count of similar days are conjoined to create a cohesive cluster. This process iterates until specific termination criteria are met, signaling the conclusion of cluster merging. The linkage criteria determinates the metric used for the merge strategy as shown in Figure 3-3:

Ward linkage minimizes the sum of squared differences within all clusters. It is a variance minimizing approach

Maximum or complete linkage minimizes the maximum distance between observations of pair of clusters.

Average linkage minimizes the average of the distance between all the observations of pairs of clusters.

Single linkage minimizes the distance between the closest observations of pairs of clusters.

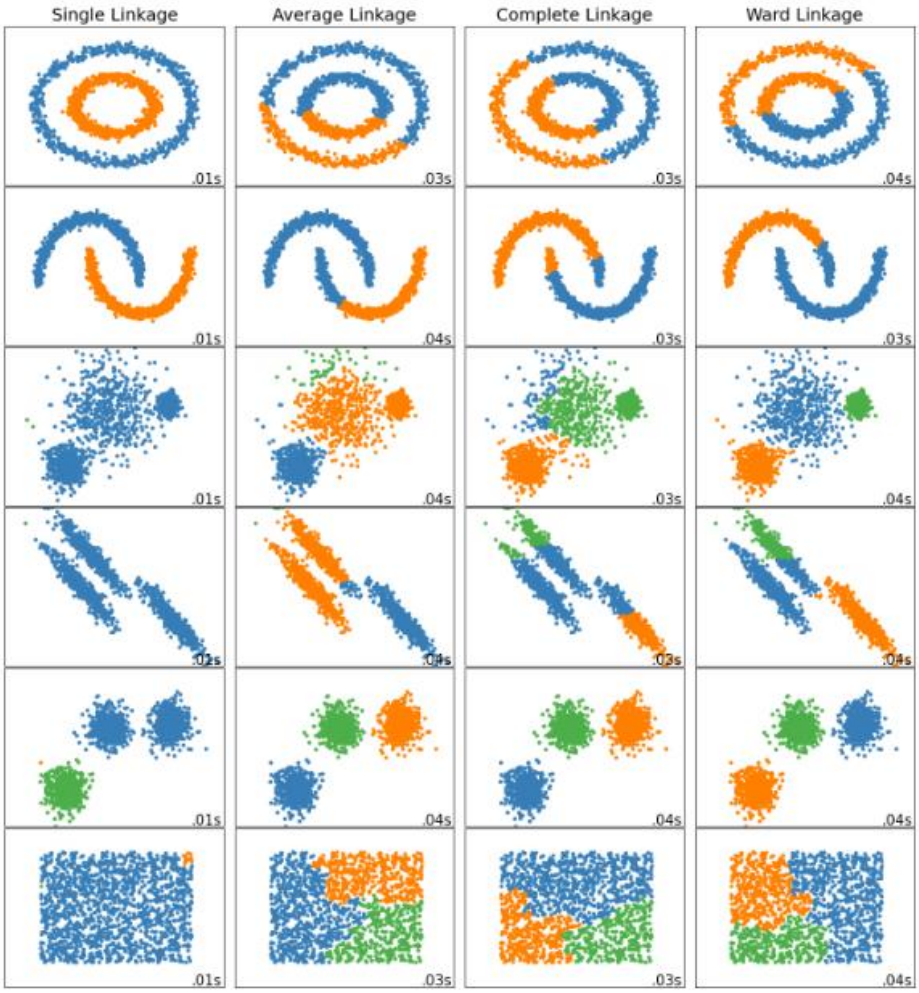

Figure 3-3: Different linkage strategies [30]

The procedural workflow of this Agglomerative Hierarchical Clustering (AHC) algorithm is visually represented in Figure 3-4. Hierarchical cluster analysis, as elucidated in reference [21], generates a distinctive array of nested clusters through sequential pairing based on pre-defined criteria. Illustrated in the form of a dendrogram as shown in Figure 3-5. Since the idea is to reduce the variance between cities in a cluster when classifying into climate zones, 'Ward' linkage criteria is used.

Given the necessity in the proposed method to ascertain similar days between two cities with a focus on unique matching and maximizing such matches, the methodological approach incorporates the utilization of maximum bipartite matching. This technique ensures a meticulous and efficient matching process, aligning with the overarching objective of discerning and optimizing the similarity between the climatic profiles of distinct cities.

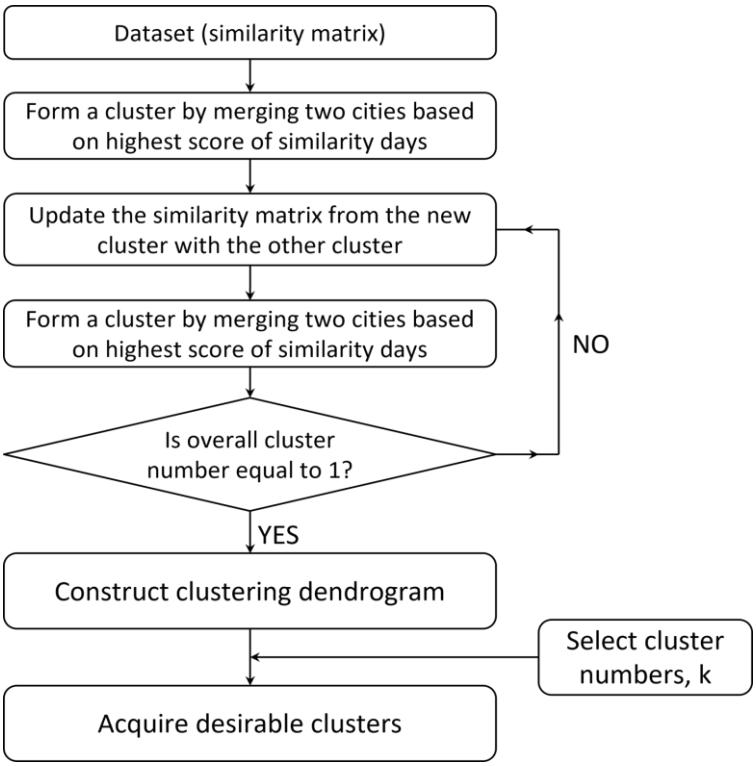

Figure 3-4: AHC algorithm-based clustering process

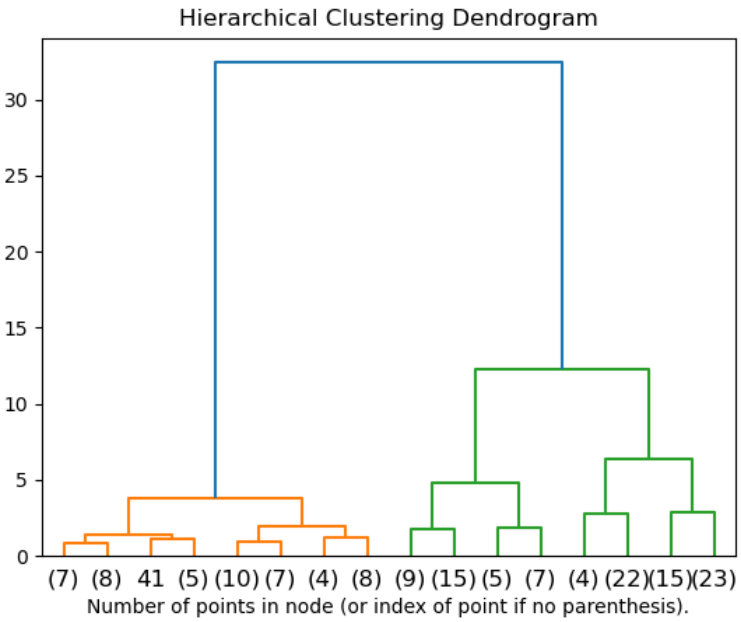

Figure 3-5: Hierarchical clustering dendrogram [21]

3.3.2 Maximum bipartite matching

Bipartite matching, as delineated in reference [26], entails the selection of a set of edges in a graph in such a way that no two edges within that set share an endpoint. The quest for maximum matching involves determining the highest count of such edges. In the context of the proposed method, a threshold value is assigned to each variable, a pivotal step in identifying the number of similar days. The intricacies of this process are encapsulated in the visual presentation provided in Figure 3-6, outlining the systematic steps involved in calculating similarity between cities utilizing maximum bipartite matching.

To elucidate, consider a concise example featuring five weather files denoted from “a” to “e”. Each file encompasses data for 5 days pertaining to two weather variables, designated as “V1” and “V2”, as expounded in Table 3-1. The determination of the maximum count of similar days between all cities involves setting specific thresholds for “V1” and “V2”, established at 5 and 50, respectively. The ensuring outcomes are organized in a square matrix with dimensions n x n where, in this instance, n = 5.

Upon executing the maximum bipartite matching, the resulting matrix assumes a 5x5 format exemplified in Table 3-2. When comparing a weather file (WF) to itself, all days inherently match. For instance, a comparison of “a” with “b” reveals that only 2 days meet the prescribed threshold for similarity. Similarity, the comparison of “d” with “e” yields 4 days that satisfy the threshold for similarity, as elucidated in Figure 3-7.

This resultant matrix serves as input for subsequent Agglomerative Hierarchical Clustering (AHC). In this illustrative scenario, the number of clusters is predefined as three (k = 3). Executing the AHC process yields generates labels for the five cities: [2 1 0 0 0]. This labelling scheme signifies that weather files “c”, “d” and “e” form a cluster, which “a” and “b” belong to distinct clusters. The systematic application of these methodologies provides a robust framework for discerning and categorizing climatic similarities among cities.

Table 3-1: Sample values from Weather file

	B1	B2	B3
A1	0.1	0.2	0.3
A2	…	..	.
A3	…	..	.

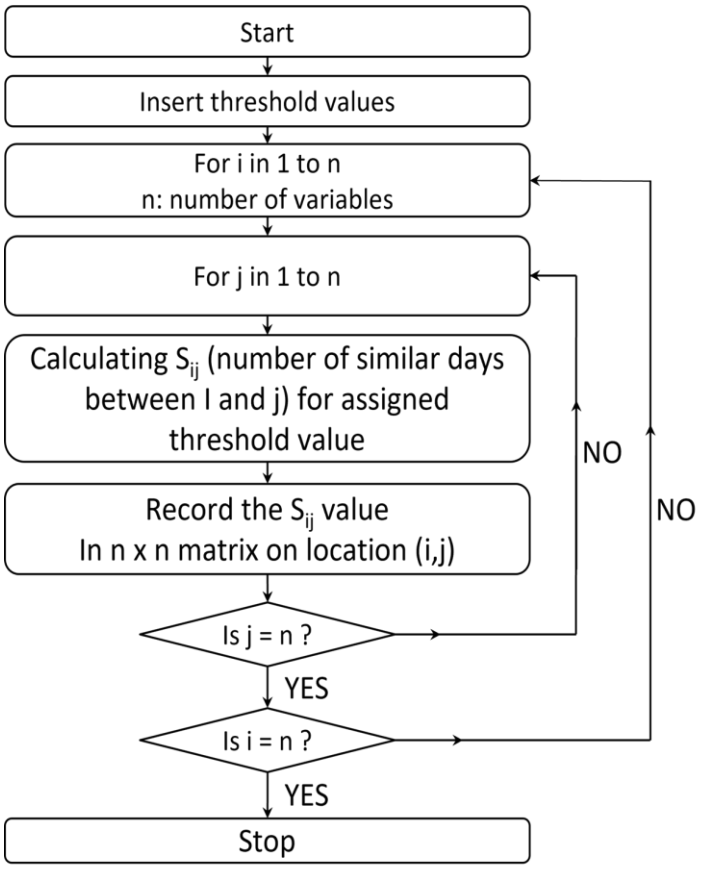

Figure 3-6: Flowchart depicting the process of generating a matrix indicating the count of similar days between cities

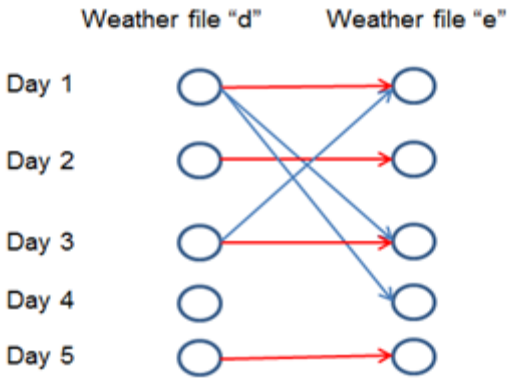

Figure 3-7: Bipartite matching

Table 3-2: Matrix of number of similar days between sample cities

	WF "a"	WF "b"	WF "c"	WF "d"	WF "e"
WF "a"	5	2	0	0	0
WF "b"	2	5	0	0	0
WF "c"	0	0	5	4	4
WF "d"	0	0	4	5	4
WF "e"	0	0	4	4	5

3.4 **Validation method**

Given the inherent variability in the number of cities or sites within each cluster, a silhouette score method is employed to evaluate the clustering.

3.4.1 Silhouette score

Silhouette score [30] is a metric used to evaluate how good clustering results are in data clustering. This score is calculated by measuring each data point's similarity to the cluster it belongs to and how different it is from other clusters.

$$s(i) = b(i) – a(i)/\max\{a(i), b(i)\}$$

where,

i = No of Iteration

a () – Mean intra-cluster distance

b () – Mean nearest-cluster distance

s () – Silhouette score

A higher Silhouette score indicates more consistent and better clustering results, while a low score may indicate that data points are assigned to incorrect clusters or that the clustering algorithm is not suitable for the data.

4 Analysis of Climate Classification of USA

The methodology described in Chapter 3 has been implemented on 786 USA cities. The ASHRAE Standard 169 is referred for the climate classification of the USA as shown in Figure 4-1. The EPW files for the USA were extracted from EnergyPlus weather data source. These weather files are in typical meteorological year format and arranged by the World Meteorological Organization. For the analysis, the daily mean dry-bulb temperature, daily mean relative humidity, and daily mean solar radiation were extracted from the weather files for each city. These variables serve as inputs for the analysis. Figure 4-2 displays the geographic distribution of cities across the USA, representing their respective climate zones, which have been considered for the analysis.

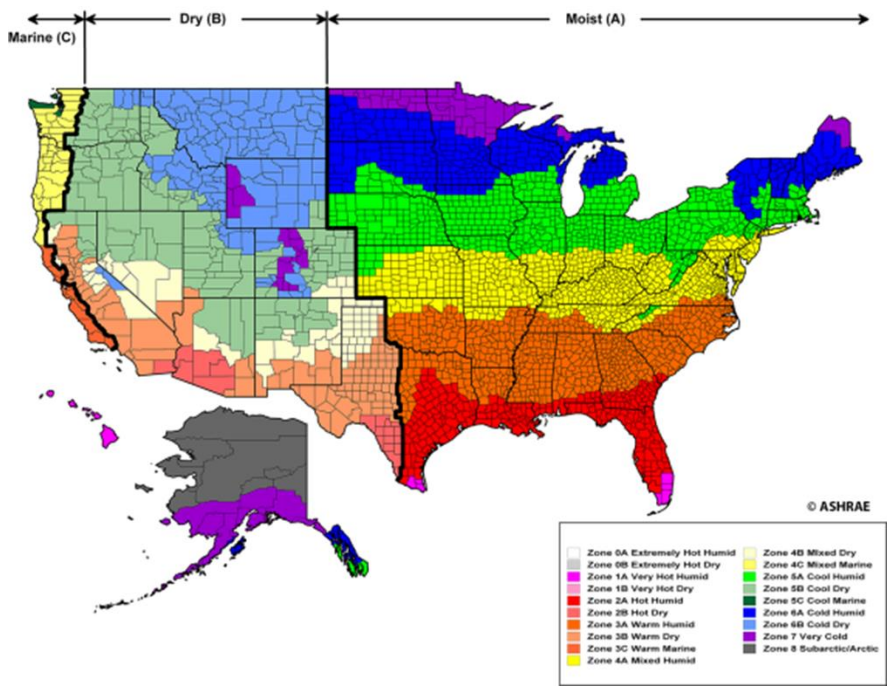

Figure 4-1: United States climate zone map based on ASHRAE-169 [2]

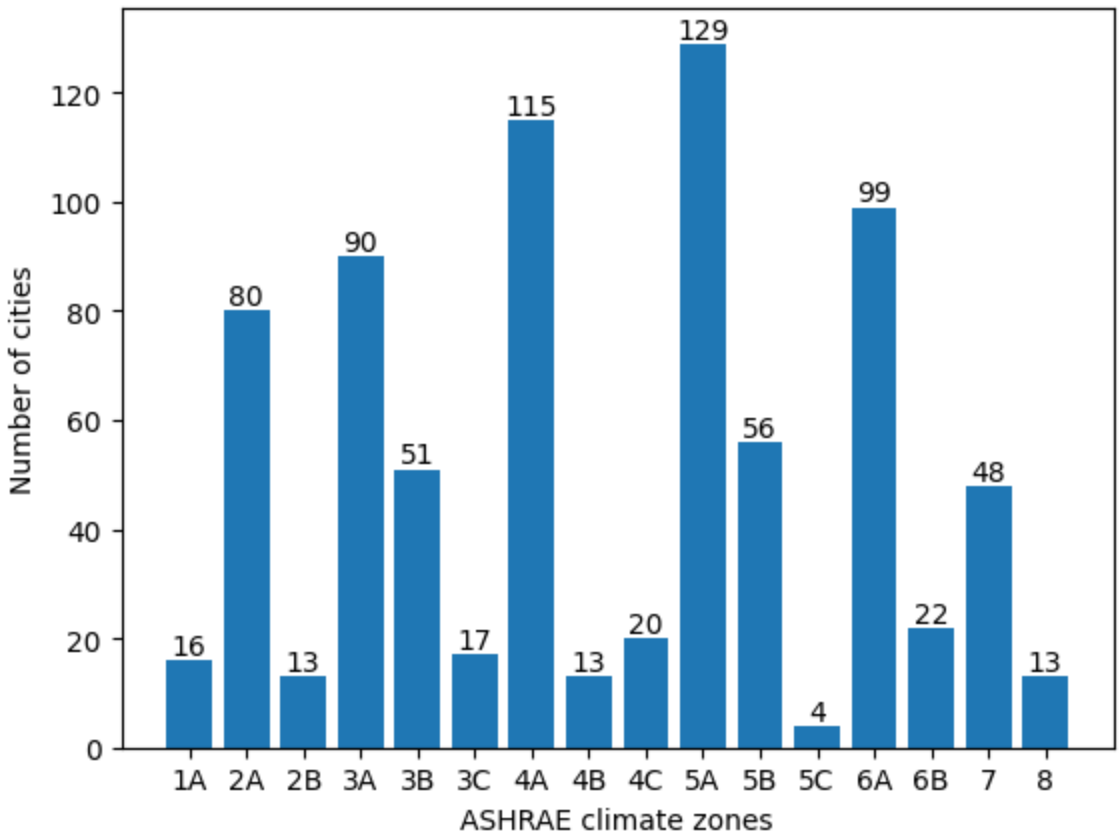

Figure 4-2: Number of cities in ASHRAE Standard 169 classification zones for USA

4.1 Preparation of Data

4.1.1 Weather data files

4.1.1.1 Extraction of weather data files

The required weather data EPW files were extracted from EnergyPlus website using a Python script [27]. The EPW files were in TMY, TMY2 and TMY3 formats. The relevant code is provided in Appendix A.

4.1.1.2 Converting data to required format

After extracting the EPW files, using a Python script [27], the files were read individually and the required daily mean dry-bulb temperature, daily mean relative humidity and daily mean global horizontal radiation were extracted in a .CSV format file. The relevant code is provided in Appendix B.

4.1.2 Building model

The EnergyPlus software was utilized to conduct building energy simulations. A typical core-perimeter (4 perimeter zones and a core zone) zone office building with 400 m^2 floor area was prepared and used for the simulations. The developed model of the building is shown in Figure 4-3. The simulations were performed for all 786 weather files in the USA. From the simulation data, the energy consumption details are extracted i.e., daily latent cooling energy, daily sensible cooling energy, and daily heating energy into a .CSV format file as shown in Appendix B.

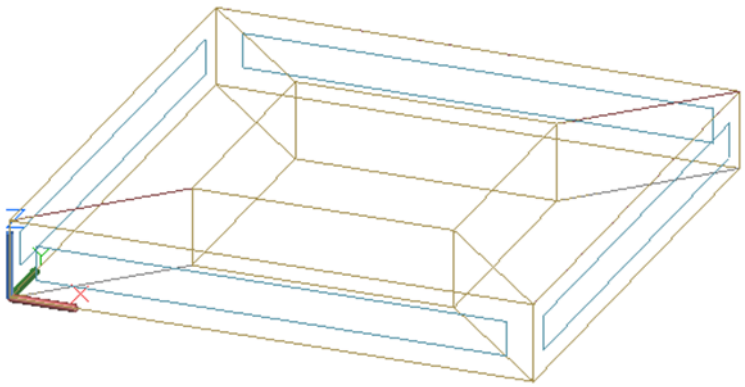

Figure 4-3: EnergyPlus model [31]

Table 4-1: Building model parameters

S No.	Parameter	Unit	Value
1	Building Type		Office
2	Schedule		9 am to 6 pm
3	Occupancy	m^2/person	10
4	Window-to-Wall Ratio		40%
5	Number of Floors		1
6	Floor Area	m^2	400

4.2 Clustering of cities of USA

Daily values of the three parameters are utilized to calculate the 'similar' days between the cities, based on the maximum bipartite matching as per the flow diagram shown in Figure 3. Scipy module [28] in Python was used for maximum bipartite matching. For the considered cities in the USA, the output matrix for similar days is of dimension 786 x 786. Various sets of threshold

values were considered for calculating the number of similar days and some of the threshold values are listed in Table 4-2. Pandas module in Python was used for data analysis [29]. Then the clustering analysis was performed for all the cities and divided the cities into 16 zones (Same as total zones of USA by ASHRAE Standard 169) and the cities were labeled with numbers ranging from 0 to 15 as seen in Appendix C. Figure 4-4 shows the weather data clusters with the proposed method. Figure 8 shows the frequency of cities in ASHRAE Standard 169 classification zones for the USA. Sk-learn module in Python was used for agglomerative clustering [30].

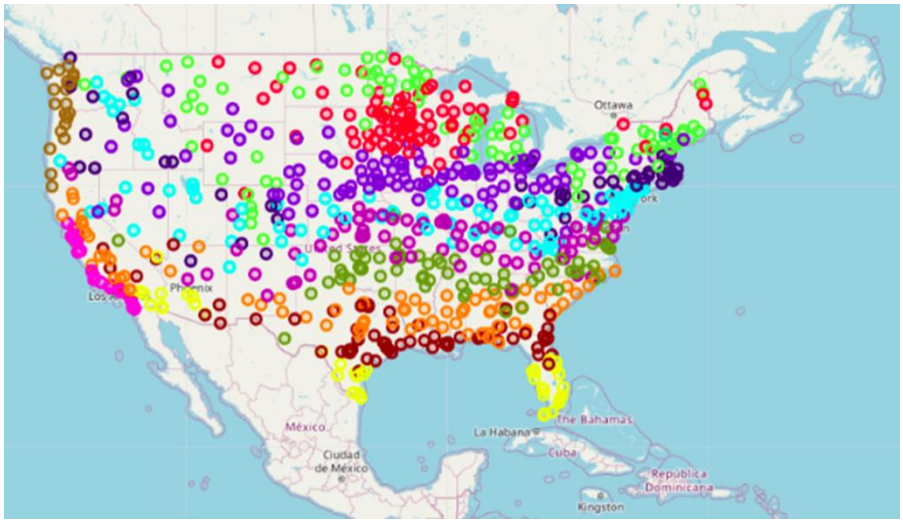

Figure 4-4: Weather data clusters with proposed method

4.3 Calculation of score

The score was calculated for both ASHRAE zones and zones generated with the proposed approach based on number of similar days for different threshold values. The silhouette score method is used to evaluate the clustering upon all the threshold combinations as shown in Table 4-2. The calculation of silhouette score, Sk-learn module [30] in Python is used as shown in Appendix D. Table 4-2 has been arranged in descending order based on the scores, as a higher score indicates a better clustering.

Table 4-2: Threshold combinations and scores

Threshold combinations [DBT (ºC), RH (%), GHR (Wh/sqm)]	Scores
7, 45, 35	0.113
6, 45, 35	0.103
5, 45, 35	0.091
8, 45, 35	0.066
ASHRAE	0.054
4, 40, 30	0.047
9, 40, 35	0.035

5 Results and Discussion

5.1 Results

The findings demonstrate that the proposed method, utilizing the number of similar days and scoring techniques, achieves higher score compared to ASHRAE Standard 169. Upon examining Table 4-2, it is evident that the best clustering is achieved when employing threshold values of 7 °C for the daily mean dry-bulb temperature, 45% for the daily mean relative humidity, and 35 Wh/m^2 for the daily mean global horizontal radiation. Figure 5-1 and 5-2 shows the spread of sensible cooling and total cooling respectively for ASHRAE Standard 169 classification. Figure 5-3 and 5-4 shows the spread of sensible cooling, total cooling and heating for the proposed classification.

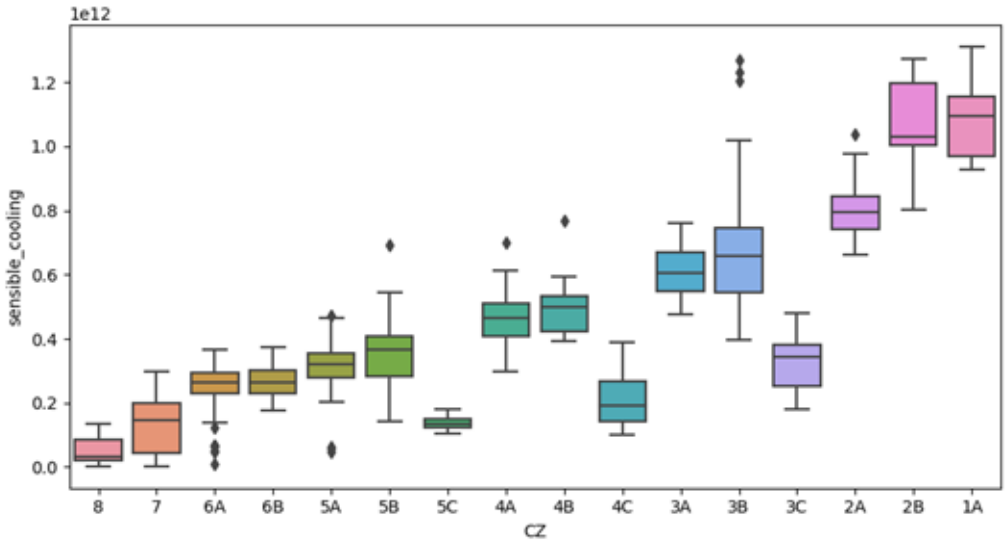

Figure 5-1: ASHRAE climate zones and spread for sensible cooling

It can be observed from Figures 5-1 and 5-2 that although Climate zone (CZ) 1 to 8 represents extreme hot to arctic, the mean cooling energy values are not in decreasing order. The climate zones 3C, 4C, and 5C of ASHRAE Standard 169 represent warm marine, mixed marine and cool marine respectively are not in order.

The zones identified using unsupervised clustering are shown in Figure 5-3 and 5-4. Zone 13 is having highest cooling energy consumption can be referred to as an extreme hot climate and

Zone 0 can be an arctic zone. Upon looking at the distribution of cities in ASHRAE zones and the selected cluster classification based on the proposed method, more than 50% overlap was observed between the two. But a significant number of cities are distributed uniquely in the new classified zones.

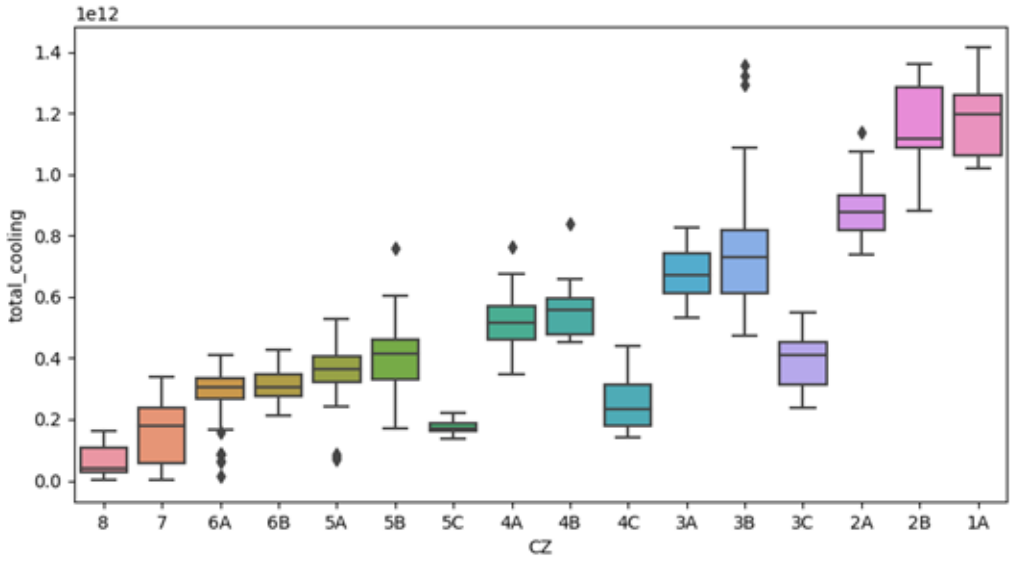

Figure 5-2: ASHRAE climate zones and spread for total cooling

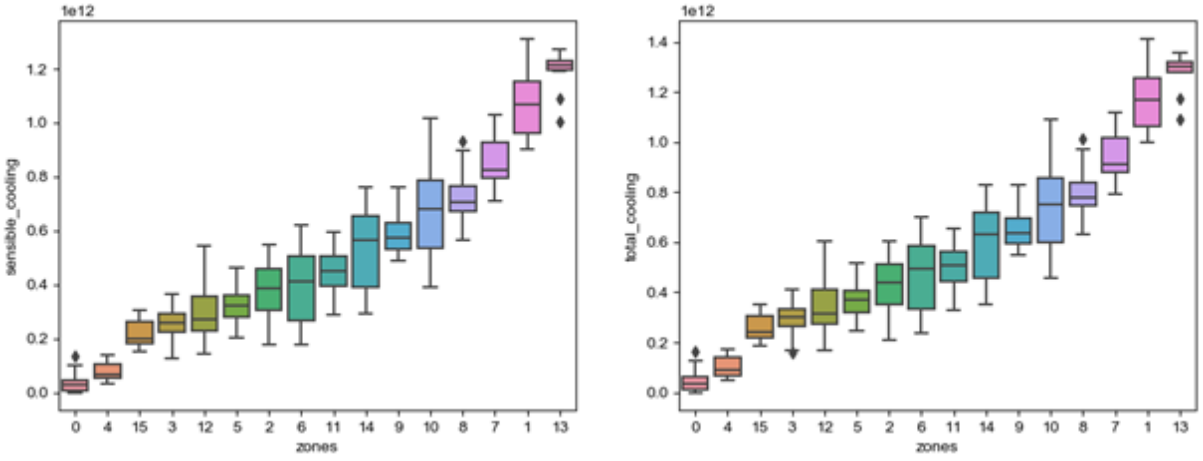

Figure 5-3: Spread with proposed clustering– sensible cooling and total cooling

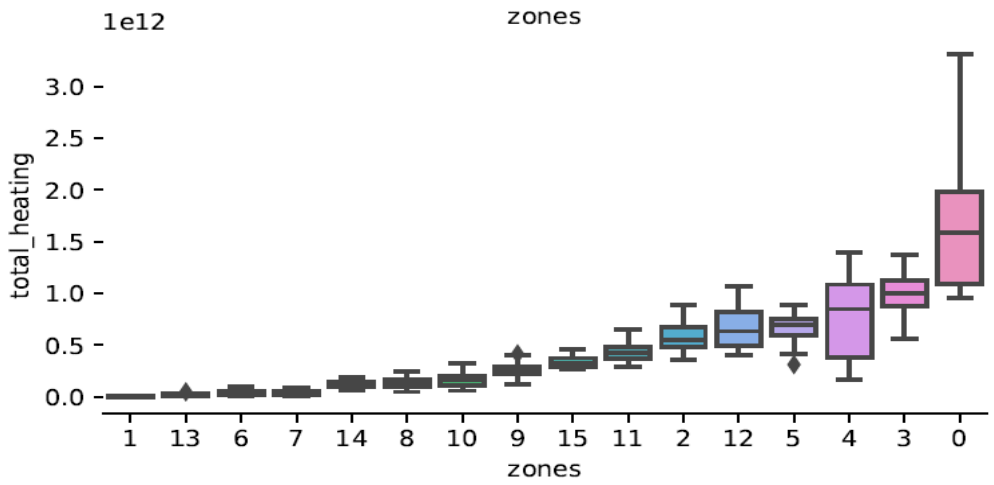

Figure 5-4: Spread with proposed clustering– heating

5.2 **Limitations**

The study relied on the availability of weather data for the selected U.S. cities. Limitations in data coverage and quality may influence the results. The choice of thresholds was based on careful consideration, but they may not be universally applicable to all regions or building types. The study did not explicitly account for potential climate change effects, which may alter long-term climate patterns and impact building design considerations.

6 CONCLUSION

This study introduces a new approach for climate classification of diverse cities by utilizing the number of similar days and evaluating quality of clustering based on building energy consumption. The method identifies the climate zoning among various combinations using different threshold values. Unsupervised learning was employed, utilizing mean daily weather data, to discover similarities between cities. The clustering was then scored based on building energy consumption, calculated through simulation tools. Silhouette scores was utilized to check the quality of clustering. A higher score indicates a better clustering in this context.

To test the proposed method, available weather files from the USA were employed. The climate zones of the USA were divided into 16 clusters using the developed method outlined in this research. Silhouette scores were computed for different combinations of threshold values to obtain improved zoning. The classification method proposed in this study exhibits a better score of 0.113, as compared to the ASHRAE Standard 169 classification score of 0.054. Furthermore, the method developed in this thesis has the potential to be applied to other countries as it operates by identifying similarities among weather data.

REFERENCES

[1] Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F., "World Map of the Köppen-Geiger climate classification updated, Meteorol. Zeitschrift," vol. 15, no. 3, pp. 259–263, Jul. 2006.

[2] ANSI/ASHRAE Standard 169 Climate data for Building Design Standards.

[3] A. Walsh, D. Costola, L.C. Labaki, "Review of methods for climatic zoning for building energy efficiency programs," Build. Environ. 112 (2017) 337–350.

[4] R Gupta, J Mathur, V Garg - Energy and Buildings, 2023, Assessment of climate classification methodologies used in building energy efficiency sector, https://doi.org/10.1016/j.enbuild.2023.113549

[5] Flor, F. J. S., Dom´ınguez, S. A., F´elix, J. L. M., Falc´on, R. G., "Climatic zoning and its application to Spanish building energy performance regulations," Energy and Buildings, vol. 40, no. 10, pp. 1984–1990, 2008.

[6] Briggs, R. S., Lucas, R. G., Taylor, Z. T., "Climate classification for building energy codes and standards: Part 1-development process," Trans. Soc. Heat. Refrig. Air Cond. Eng., vol. 109, no. 1, pp. 109–121, 2003.

[7] Briggs, R. S., Lucas, R. G., Taylor, Z. T., "Climate classification for building energy codes and standards: Part 2 - Zone definitions, maps, and comparisons," ASHRAE Trans., vol. 109 PART 1, pp. 122–130, 2003.

[8] Mourshed, M., "Relationship between annual mean temperature and degree-days," Energy and Buildings, vol. 54, pp. 418–425, 2012.

[9] Bansal, N. K., Minke, G., "Climate Zones and Rural Housing in India, General Indian Corporation, 1988.

[10] Bhatnagar, M., Jyotirmay, M., Garg, V., "Reclassification of climate zones for Indian Cities," ISHRAE, 2016.

[11] Kumar, J., Hoffman, F. M., New, R. J., Sanyal, J., "Reimagining Climate Zones for Energy Efficient Building Codes." [Online]. Available: https://web.eecs.utk.edu/~jnew1/presentations/2014_ASHRAE_Weather.pdf /

[12] Cao, J., Li, M., Zhang, R., Wang, M., "An efficient climate index for reflecting cooling energy consumption: Cooling degree days based on wet bulb temperature," Meteorol. Appl., vol. 28, no. 3, pp. 1–10, 2021.

[13] Zscheischler, J., Mahecha, M. D., Harmeling, S., "Climate classifications: The value of unsupervised clustering," Procedia Comput. Sci., vol. 9, pp. 897–906, 2012.

[14] Hudson, R., Velasco, R., "Thermal Comfort Clustering; Climate Classification in Colombia," pp. 590–595, 2018.

[15] Sathiaraj, D., Huang, X., Chen, J., "Predicting climate types for the Continental United States using unsupervised clustering techniques," Environmetrics, vol. 30, no. 4, pp. 1–12, 2019.

[16] Xiong, J., Yao, R., Grimmond, S., Zhang, Q., Li, B., "A hierarchical climatic zoning method for energy efficient building design applied in the region with diverse climate characteristics," Energy and Buildings, vol. 186, pp. 355–367, Mar. 2019.

[17] Shin, M., Do, S. L., "Prediction of cooling energy use in buildings using an enthalpy-based cooling degree days method in a hot and humid climate," Energy and Buildings, vol. 110, pp. 57–70, 2016.

[18] Pernigotto, Giovanni & Gasparella, Andrea & Hensen,. "Assessment of a weather-based climate classification with building energy simulation," Jan 2021, 10.26868/25222708.2021.30765.

[19] Walsh, A., C´ostola, D., Labaki, L. C., "Comparison of three climatic zoning methodologies for building energy efficiency applications," Energy and Buildings, vol. 146, pp. 111–121, 2017.

[20] Walsh, A., C´ostola, D., Labaki, L. C., "Validation of the Degree-Days Method for Climatic Zoning– Initial Results Based On the Mean Percentage of Misplaced Areas," in uSIM, 2018.

[21] Bridges, C. C., "Hierarchical Cluster Analysis," Psychol. Rep., vol. 18, no. 3, pp. 851–854, Jun. 1966.

[22] Frank E., et al, "Solar Energy Forecasting and Resource Assessment" 2013 pp. 97-131.

[23] *EnergyPlus* [Online]. Available: https://energyplus.net/.

[24] *IES-VE*, [Online]. Available: https://www.iesve.com.

[25] James J. Hirsch & Associates, "The Quick Energy Simulation Tool (eQUEST)," 2016.

[Online]. Available: http://www.doe2.com/equest/.

[26] Hopcroft, J. E., Karp, R. M., "An $n^{5/2}$ Algorithm for Maximum Matchings in Bipartite Graphs," SIAM J. Computer., vol. 2, no. 4, pp. 225–231, Dec. 1973.

[27] Python Software Foundation [Online]. Available: https://www.python.org/ .

[28] Scipy [Online], Available: https://scipy.org/.

[29] *Pandas – Python Data Analysis Library* [Online]. Available: https://pandas.pydata.org.

[30] Sk-learn [Online]. Available: https://scikit-learn.org.

APPENDIX A: EXTRACTION OF WEATHER DATA FILES

```
from bs4 import BeautifulSoup as bs
import requests
import os

def extract_epw_files(home_dir):

    response = requests.get(home_dir)

    epws_dir = os.path.join(os.getcwd(), home_dir.split('/')[-1])
    if(os.path.isdir(epws_dir) == False):
        os.mkdir(epws_dir)
    os.chdir(epws_dir)

    soup = bs(response.content, 'html.parser')

    states = soup.find_all('a')[15:-1]

    for state in states:
        state_dir = 'https://energyplus.net' + state['href']
        response = requests.get(state_dir)
        soup = bs(response.content, 'html.parser')

        cities = soup.find_all('a')
        #city = cities[16]
        for city in cities:
            if not city.contents == []:
                if 'TMY3' in city.contents[0]:
                    city_dir = 'https://energyplus.net' + city['href']
                    response = requests.get(city_dir)
                    soup = bs(response.content, 'html.parser')
                    files = soup.find_all('a')
                    epw_link = 'https://energyplus.net' + files[-4]['href']
                    response = requests.get(epw_link)

                    city_name = '_'.join(city.contents[0].split(' ')[:-1])

                    epw_file = open(city_name + '.epw', 'w')
                    epw_file.write(response.content)
                    epw_file.close()
                    print(city_name)


if __name__ == "__main__":

    home_dir = 'https://energyplus.net/weather-region/north_and_central_america_wmo_region_4/USA'
    extract_epw_files(home_dir)
```

APPENDIX B: CONVERTING DATA TO REQUIRED FORMAT

```
import pandas as pd
import glob

files = sorted(glob.glob('weather_files/*'))

total_heating_energy_vals = []
sensible_cooling_energy_vals = []
latent_cooling_energy_vals = []
total_cooling_energy_vals = []
ids = []

for file in files:

    id = file.split('_')[-1]
    ids.append(id)
    data_file = file + "/ModelIDF.csv"
    data = pd.read_csv(data_file)


    total_heating_energy = 0
    sensible_cooling_energy = 0
    latent_cooling_energy = 0
    total_cooling_energy = 0
    for column in data.columns:
        if 'Total Heating Energy' in column:
            total_heating_energy = total_heating_energy + data[column].sum()
        elif 'Sensible Cooling Energy' in column:
            sensible_cooling_energy = sensible_cooling_energy + data[column].sum()
        elif 'Latent Cooling Energy' in column:
            latent_cooling_energy = latent_cooling_energy + data[column].sum()
        elif 'Total Cooling Energy' in column:
            total_cooling_energy = total_cooling_energy + data[column].sum()

    total_heating_energy_vals.append(total_heating_energy)
    sensible_cooling_energy_vals.append(sensible_cooling_energy)
    latent_cooling_energy_vals.append(latent_cooling_energy)
    total_cooling_energy_vals.append(total_cooling_energy)

final_data = pd.DataFrame()
final_data['filename'] = files
final_data['ID'] = ids
final_data['total_heating'] = total_heating_energy_vals
final_data['sensible_cooling'] = sensible_cooling_energy_vals
final_data['latent_cooling'] = latent_cooling_energy_vals
final_data['total_cooling'] = total_cooling_energy_vals

final_data.to_csv('final_energy_vals.csv', index = False)
```

APPENDIX C: CLUSTERING ALGORITHM

```
import pandas as pd
import numpy as np
from scipy.sparse import csr_matrix
from scipy.sparse.csgraph import maximum_bipartite_matching
from sklearn.cluster import AgglomerativeClustering


class ClusteringAlgo:

    def __init__(self, n_clusters = 2, threshold = {}):
        ''' Initialises required parameters '''
        self.__threshold = threshold
        self.__n_clusters = n_clusters
        self.__files = None
        self.labels = None
        self.__no_of_cities = 0
        self.__city1 = None
        self.__city2 = None
        self.__parameters = None
        self.__matrix = None

    def compare_cities(self, city1, city2):
        ''' Returns number of matched days between two cities '''
        self.__parameters = ['mean_dbt', 'mean_rh', 'mean_ghr']
        self.__city1 = city1[self.__parameters]
        self.__city2 = city2[self.__parameters]

        threshold = pd.Series(self.__threshold)[self.__parameters].to_numpy()
        distance_matrix = abs(self.__city1.to_numpy()[:, None] - self.__city2.to_numpy()[None, :])/threshold
        distance_matrix = (distance_matrix >1).__eq__(False).all(axis = 2).astype(int)
        graph = csr_matrix(distance_matrix)


        matched_days = np.less(0,1 + maximum_bipartite_matching(graph, perm_type='column')).sum()

        return matched_days

    def fit_predict(self, data):
        ''' Clusters data and returns respective labels for data '''
        return self.fit(data).labels

    def __iterate(self, combination):
        ''' Clusters two clusters '''
        i = combination[0]
        j = combination[1]
        if(i>j):
          self.__matrix[i][j] = self.__matrix[j][i]
        elif(i==j):
          self.__matrix[i][j] = 365
        else:
            city_i_data = pd.read_csv(self.__files[i].split('/')[-1])
            city_j_data = pd.read_csv(self.__files[j].split('/')[-1])
            matched_days = self.compare_cities(city_i_data, city_j_data)
            self.__matrix[i][j] = matched_days
        print(i, j)


    def fit(self, data):
        ''' Clusters data and returns self '''
        self.__files = data
        self.__no_of_cities = len(data)
        self.labels = [None]*self.__no_of_cities
        self.__matched_days = [0]*self.__no_of_cities

        i_range = range(self.__no_of_cities)
        j_range = range(self.__no_of_cities)

        self.__matrix = np.empty([self.__no_of_cities, self.__no_of_cities], dtype = int)

        combination_array = pd.DataFrame(np.array(np.meshgrid(i_range, j_range)).T.reshape(-1, 2))
        combination_array.apply(self.__iterate, axis = 1)

        clustering = AgglomerativeClustering(n_clusters = self.__n_clusters)
        self.labels = clustering.fit_predict(self.__matrix)

        return self
```

APPENDIX D: SILHOUETTE SCORE

```python
from sklearn.metrics import silhouette_score


def calculate_silhouette_score(energy_data, labels):

    # `energy_data` is a Pandas Dataframe with columns latent_cooling, sensible_cooling and totat_heating values for all the cities
    # labels is a list of cluster labels for classification

   score = silhouette_score(energy_data, labels)
```

List of Publications

Sri Sanketh Uppalapati et al. “A Novel Approach for Climate Classification using Agglomerative Hierarchical Clustering” E.I.A 2023: Energy Informatics pp 152-167