NAZARBAYEV UNIVERSITY

ENHANCING CLIMATE MAPPING METHODOLOGIES: A NOVEL PERFORMANCE-BASED FRAMEWORK FOR KAZAKHSTAN'S BUILDING CLIMATE ZONING

by

Alexey Remizov

Submitted in partial fulfillment
of the requirements for the degree of Doctor
of Philosophy in Civil Engineering

April, 2024

DEVELOPMENT OF BUILDING PERFORMANCE-BASED CLIMATE MAPS OF KAZAKHSTAN

by

Alexey Remizov

Submitted in partial fulfillment of the requirement for the degree of
Doctor of Philosophy in Civil Engineering

School of Engineering and Digital Sciences
Civil and Environmental Engineering Department
Nazarbayev University

April, 2024

Supervised by
Shazim Ali Memon

Declaration

I declare that the research contained in this thesis, unless otherwise formally indicated within the text, is the author's original work. The thesis has not been previously submitted to this or any other university for a degree and does not incorporate any material already submitted for a degree.

Signature: 

Date: 10.04.2024

Abstract

Accurate climate zoning is crucial in the construction sector for both building thermal performance and energy efficiency, playing a vital role in achieving energy reduction targets, and facilitating early-stage design decisions to minimize energy consumption while maintaining occupant comfort. Traditional climate classification methods, which mainly rely on climatic data, often fail to consider the energy consumption of buildings, leading to a disconnect between climate classification and building energy performance. In Kazakhstan, this mismatch is evident as existing standards and climate maps do not adequately inform about potential building energy consumption in different climate zones, which leads to a high energy usage rate, emphasizing the urgent need for effective energy efficiency measures as the building stock expands. Moreover, despite currently low energy prices and sufficient energy resources in Kazakhstan, primarily non-renewable sources like coal and gas, there's a looming concern over future energy sustainability and affordability. The uncertainty arises from the significant increase in energy prices that are already being noticed. Considering the low average household income, proactive energy-efficient practices are imperative to mitigate future energy challenges.

The study's innovative methodology merges clustering and spatial analysis, utilizing simulated building energy consumption data to develop a more accurate energy performance-based climate classification system. The core objective of this study is the creation of a new buildings climate zoning map in Kazakhstan, which aims to provide compelling arguments on the intricate relationship between climate, buildings, and their energy consumption in the specific geographical region. A novel aspect of this research is the introduction of a new climate classification quality metric - a novel climate zoning misclassification index (CZMI), designed to assess the effectiveness of climate classification. CZMI operates on the idea that each climate zone should have a distinctive climate, leading to specific energy needs for a particular building type with minimal overlaps between climate zones. It assesses the level of overlap between climatic zones by comparing the overlap of buildings' energy performance distributions using kernel density estimation curves, highlighting the distinctiveness of each climate zone. CZMI also provides a framework for evaluating the accuracy and reliability of both the proposed and existing official climate maps of Kazakhstan. By juxtaposing the proposed building performance-based climate map with the official map, the study highlights key discrepancies and misclassifications, advocating for a more holistic and accurate climate zoning approach. Additionally, by combining CZMI with spatial analysis principles the study examines how spatial constraints affect building` climate zoning accuracy.

Through the comprehensive analysis of 47 climate factors influencing building energy consumption in Kazakhstan, implementing correlation analysis and machine learning algorithms, latitude, cooling degree days, heating degree days, and global horizontal irradiance were identified as the most influential variables. Next, multivariate cluster analysis was employed, resulting in a set of building climate zoning maps (12 climate-based and 4 performance-based), having from 3 to 8 climate zones as the optimal numbers. For each map, the level of misclassification was determined using CZMI. Performance-based classifications, outperformed other methods with the lowest CZMI percentages, indicating higher accuracy and lower misclassification. However, specific climate-based approaches can show similarly high effectiveness matching the accuracy of performance-based methods. By carefully examining the impact of spatial constraints (latitude) on the classification results, which had not been explored before, it was observed that there is no significant advantage when including spatial constraints in minimizing the CZMI. Ultimately, precise values were established to measure the misclassification of the official climatic zoning map, revealing its significant lack of practical use.

The potential impact of this research is substantial. It aims to create the link between climate, buildings, and their energy performance thus propelling Kazakhstan towards modern energy standards. The study's implications extend beyond technological advancements to encompass economic and social benefits, including possible reduced building operational costs and enhanced environmental sustainability. By informing evidence-based energy policies, this research marks a pivotal step towards revolutionizing climate zoning in Kazakhstan, optimizing energy performance, and transitioning towards advanced energy efficiency standards, with wide-reaching implications for the construction sector, policy development, and environmental sustainability.

Acknowledgments

To begin with, I just want to say how much I am grateful to my life for providing me with all the opportunities, strength, and determination to reach this point and complete my Ph.D. program.

I would like to express my profound appreciation to Professor Shazim Ali Memon, my distinguished advisor, whose help, constant encouragement, and kindness have played a crucial role in these remarkable 4 years. There aren't many people in my life who have had such an enormous impact on who I am; Professor Shazim Ali Memon is certainly one of those people, who now occupies a special place in my life.

My deepest gratitude goes out to Professor Jong Kim, who not only helped me finish my thesis on time but also provided me with invaluable resources. Professor Hongzhi CUI, who served as my external supervisor, also has my deepest gratitude for all of the valuable feedback and helpful suggestions he provided me during my research.

The people who have been there for me every step of the way—my mother, my wife, and her mother—deserve the utmost praise and appreciation. My lovely spouse, to whom I am eternally grateful, was my solid foundation during the four years that I was busy too much with my PhD. No matter what, her unwavering dedication, love, and care have always remained truly unbelievable.

In addition, I'd like to thank Assemgul, Abrar, and Saleh, my hardworking lab mates, for being such great friends and supportive allies along this challenging journey.

Contents

List of Tables

List of Illustrations

List of Symbols

Random forest regression

MDI	Mean Decrease Impurity
$p_{n,t}$	Fraction of observations falling in node *t*
$\{\mathcal{T}_\ell\}_{1\leq\ell\leq M}$	Collection of trees in the forest
$\left(j_{n,t}^*, z_{n,t}^*\right)$	Split that maximizes the empirical criterion in node *t*

Gradient Boosting

$\Psi(y,f)_{L_2}$	L_2 loss function
y	Observed value
f	Predicted value
$h(x, \theta)$	Base-learner model
$(x,y)_{i=1}^N$	Input data
M	Number of iterations
ρ_t	Gradient descent step size
$\widehat{f_0}$	Initial guess.

Extreme Gradient Boosting

$\mathcal{L}^{<t>}$	Objective function at the t-th iteration
i	The $i-th$ sample to be predicted
N	Total number of samples
t	$t-th$ iteration
$\Psi(y_i, \widehat{f_i})$	Loss function between the true label y_i and the predicted label $\widehat{f_i}$;
$h(X_i)$	Base-learner model
Ωf_t	Regularization term to avoid over-fitting

The optimal number of climate zones

$WCSS$	Within Cluster Sum of Squares
C	Cluster centroids
d	Data point in each cluster

K-means clustering

$J(V)$	K-means squared error function
c_i	The number of data points in an i_{th} cluster

c	The number of cluster centers

Hierarchical clustering

$D(Z,Y)$	Hierarchical clustering complete linkage function

Uniqueness of clusters

Unq	The uniqueness indicator

Dispersion of clusters

Dsp	The dispersion of clusters
MAE	The mean absolute error
k	The value of the building performance indicator within a specific cluster
μ	The mean value of the building performance indicator within the cluster

The Silhouette Score

SS_i	The Silhouette score for each sample i
$a(i)$	The average distance from the i^{th} sample to the other points in the same cluster
$b(i)$	The lowest average distance from the i^{th} sample to points in a different cluster, minimized over all clusters

Climate zoning validation with building performance

$\hat{z}_h(x)$	Estimated probability density function at point x
n	Number of data points
h	Bandwidth, a smoothing parameter that controls the width of the kernel
$K(\cdot)$	Kernel function, in this study Gaussian (normal distribution)
K_t	Total KDE (integral of the KDE function over its entire range)
$\bar{I}$	The overlap of two clusters
$\hat{z}_h^i(x)$	The probability density function of the KDE curve of $i-th$ cluster
$\hat{z}_h^j(x)$	The probability density function of the KDE curve of $j-th$ cluster
O	The overlap percentage
$CZMI$	Climate zoning misclassification index

$CZMI_{corr}$	Intra-cluster distances corrected climate zoning misclassification index

Adjusted Rand Index

RI	Rand Index
ARI	Adjusted Rand Index
$E[RI]$	Expected Rand Index

List of Acronyms and Definitions

A

ALT	Altitude
ARI	Adjusted Rand Index
ASHRAE	The American Society Of Heating Refrigerating And Air-Conditioning Engineers

B

BES	Building Energy Simulation
BCM	Bioclimatic Charts Method

C

CA	Cluster Analysis
CDD	Cooling Degree-Day
CSIM	Climate Severity Index Method
CZ	Climate Zoning
CZB	Climate Zoning for Buildings
CZMI	Climate Zoning Misclassification Index

D

GB	Gradient Boosting
DBT	Dry Bulb Temperature
DD	Degree-Day
DDM	Degree-Days Methods
DPT	Dew Point Temperature

G

GIS	Geographic Information System
GHI	Global Horizontal Irradiance
GHIllim	Global Horizontal Illuminance

H

HC	Hierarchical clustering
HDD	Heating Degree-Day
HIRIS	Horizontal Infrared Radiation Intensity from Sky

HVAC	Heating Ventilation And Air Conditioning Systems

I

IJM	Interval Judgment Method

K

KC	K-means clustering
KDE	Kernel Density Estimation
KGM	Köppen-Geiger Method

L

LAT	Latitude

M

MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MFB	Multi-family building
MPMA	Mean Percentage of Misclassified Areas
MSE	Mean Squared Error

N

NA	Northern archetype

O

ONC	The optimal number of clusters
ONCZ	The optimal number of climate zones

P

PCA	Principal Component Analysis

R

RFR	Random Forest Regression
RH	Relative Humidity
RI	Rand Index
RLF	Residential Load Factor

RMSE	Root Mean Square Error

S

SA	Southern archetype
SCHC	Spatially Constrained Hierarchical Clustering
SCKC	Spatially Constrained K-means Clustering
SFB	Single-family building
SR	Solar Radiation

T

TMY	Typical meteorological year

X

XGBoost	Extreme Gradient Boosting

W

WCSS	Within Cluster Sum of Squares
WS	Wind Speed

Preface

Mr. Alexey Remizov received his Bachelor's degree in civil engineering from the Innovative University of Eurasia, Pavlodar, Kazakhstan, in 2012. He earned his Master's degree in geodesy and remote sensing from the Siberian State University of Geosystems and Technologies, Novosibirsk, Russia in 2020. With more than 10 years of professional experience, he has a solid background in the industry. He worked in geomatics and related fields at "National Company KazMunayGas" JSC before beginning his scientific career. There, he proved his expertise in data collecting, analysis, and mapping. Later, Alexey worked with leading civil engineering companies in Astana, where he introduced advanced surveying tools, such as remote sensing and 3D photogrammetry systems. In addition, he has an expertise in GIS and spatial analysis. Currently, he is pursuing a Ph.D. dealing with building climate zoning at Nazarbayev University.

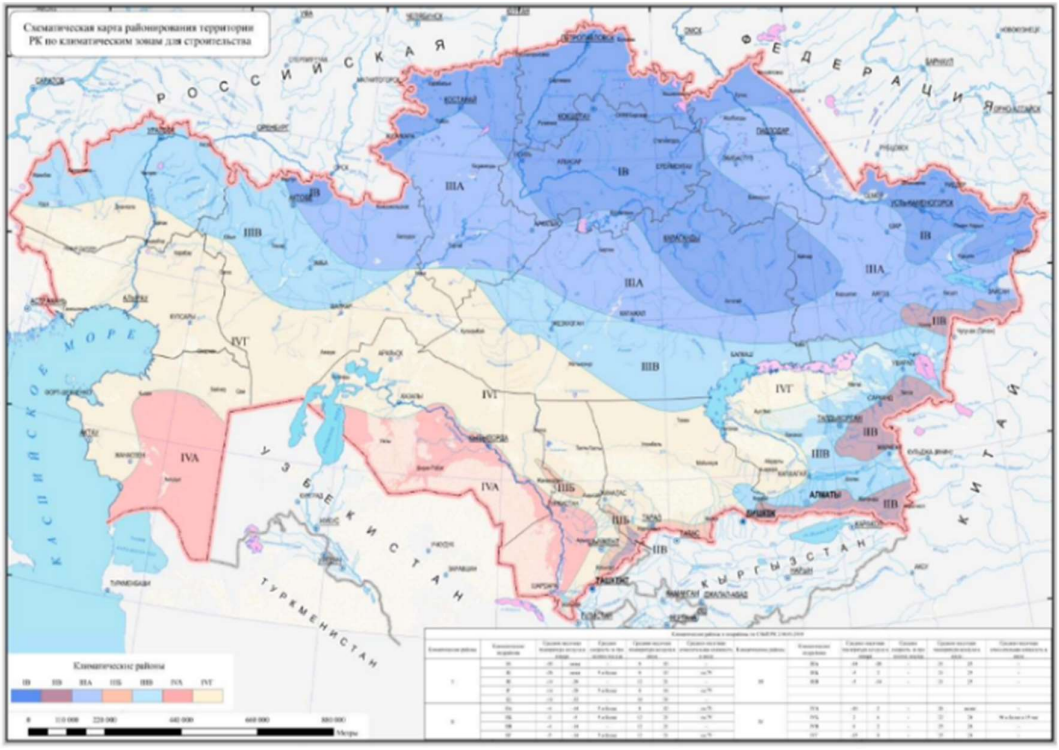
Chapter 1: Introduction

1.1. Background

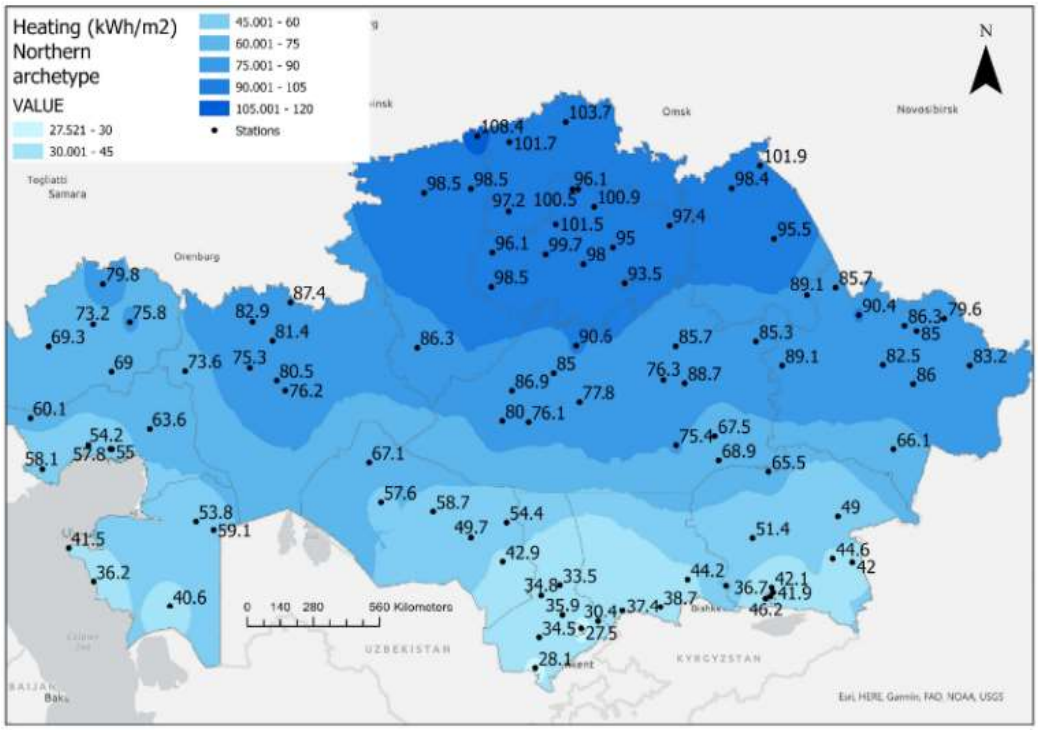
It is known that the thermal performance of buildings is influenced by the climate and the accurate climate zoning definition is essential for the construction sector. Proper climate classification is a key component of most building energy efficiency initiatives [1] emphasizing the importance of climate zoning in meeting growth goals for energy security and reducing greenhouse gas emissions. Providing a required level of occupants' thermal comfort, precise climate categorization may significantly reduce heating and cooling energy consumption [2]. Building specialists utilize climate classifications to precisely determine the thermal performance criteria for their design, taking into account the necessary heating or cooling demands. It helps the production of buildings that are specifically optimized to efficiently handle the fluctuating climatic conditions nevertheless achieving the energy efficiency requirements of the building regulations. However, traditional climate classification techniques (based on climatic data) can hardly be accurately used for the needs of sustainable building design and construction since they do not take the energy consumption of buildings into account [3-6]. In other words, there is no consistent link between climate classification and the thermal performance of buildings - a fundamental aspect of building energy efficiency.

A significant number of articles were published on that problem [3, 7-14]. In the previous 20 years, a considerable number of countries modified or implemented climate zoning of their territories [1]. Kazakhstan has officially approved climatic maps for construction in 2017 [15]. However, the method used to classify climate is identical to those used in the climatology codes of the USSR in the 1980s [16], which is based on average air temperature (AT), wind speed (WS), and relative humidity (RH). Kazakhstan's climate zoning for buildings (CZB) utilizes the interval judgment method (IJM), which has a primary focus on enhancing the thermal insulation of buildings rather than optimizing their energy needs. It only uses a variety of climate variables (including AT and RH, among others) along with predefined threshold values to determine a region's climate zoning (CZ) [17] with no direct connection to the energy needs of buildings in these regions. However, the utilization of IJM, which purely relies on climatic data presents challenges in accurately addressing the climate classification to design energy-efficient buildings. It constantly faces criticism due to its inadequacy in meeting modern energy efficiency standards for buildings [3, 18]. This is primarily due to the absence of a consistent connection between climate classification and the energy performance of buildings, which is a crucial determinant of building energy efficiency.

The situation in Kazakhstan highlights the limitations of focusing solely on environmental factors in the context of CZB. Reliable CZB guarantees that buildings with comparable energy consumption are grouped together in the same CZ [4]. However, the observed disparity between Kazakhstan's official CZB [15] and the energy consumption pattern of buildings is evident in Figure 1.1, where the official map (a) is shown alongside a heating energy performance map of a single-family building (SFB) (b). The discrepancy in official CZB and building energy performance patterns is easily apparent. Preliminary results also show a very high overlap of energy consumption among climate zones (CZs) on the official map (Figure 1.2).



(a)



(b)

Figure 1.1: The discrepancy in the official climate zoning map of Kazakhstan [15] (a) and the building heating energy performance map of SFB (b).

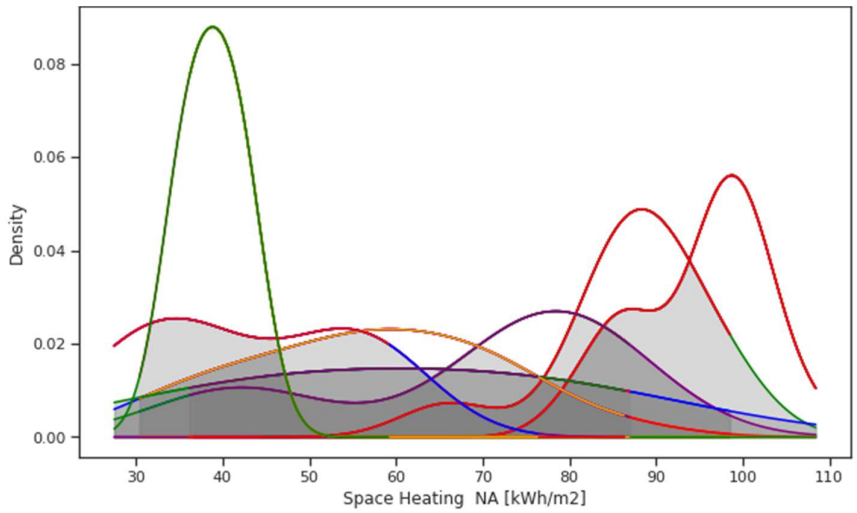


Figure 1.2: SFB's space heating energy needs overlap based on Kernel Density Estimation (KDE) among CZs in the official climate map for the construction of Kazakhstan.

The existing standards in Kazakhstan, and in particular the climate maps in its current conditions are not able to provide the necessary, accurate information for engineers, researchers, economists, policymakers, and other involved parties in terms of possible building energy consumption in each particular climate zone. Considering the increased energy efficiency requirements of buildings in recent years [19], and taking into account the

measures taken by the government and the president of the country [20, 21], the current climate mapping approach applied in the republic should be substantially improved. However, there is a lack of substantial research on the intersection of climate, energy, and buildings in Kazakhstan. Individual initiatives focused on energy-efficient buildings [22-24] lack the necessary level of quality and depth to be seriously considered or practically implemented. Other research efforts explore specific building parameters and their effects on energy consumption [25], lacking a comprehensive and holistic approach to the topic. This scarcity of nationwide, comprehensive studies highlights a significant problem that must be addressed to develop accurate and effective climate zoning and energy efficiency strategies for buildings across Kazakhstan. The pressing need for this is underscored by the remarkably high average energy usage of residential buildings in Kazakhstan (around 270 kWh/m²), surpassing that of the European Union (120 kWh/m²) and Russia (210 kWh/m²) [26]. As the building stock expands, effective solutions for energy efficiency become crucial.

In response this research introduces an innovative methodology, combining multivariate clustering with spatial analysis to classify Kazakhstan's territory into CZB using building energy consumption data based on the idea that CZB should follow the corresponding range of building energy performance. The originality of this research involves spatial analysis to examine the impact of spatial phenomena on the outcomes and quality of CZB, marking the first application of spatial analysis within the CZB domain. Enhanced with the introduction of a novel climate zoning misclassification index (CZMI) this study establishes a critical connection between climate characteristics, geography, and building energy consumption in Kazakhstan, offering a holistic understanding of this complex relationship.

This research will methodically examine a set of crucial research requests, each intended to offer significant arguments on the identified topic:

- What are the most common building archetypes in Kazakhstan, and what are their basic energy consumption levels?
- What are the spatial patterns of energy consumption of the dominant building types in Kazakhstan?
- What are the primary climate variables that exert the most significant influence on building energy consumption in Kazakhstan?
- What is the optimal number of CZs for Kazakhstan?
- How to estimate the accuracy and reliability of CZB?

- How do conventional climate-based and contemporary performance-based techniques for CZB compare in terms of outcomes, and what index can be established to determine the most efficient methodology for CZB development?
- Is proposing a CZB directly using building energy consumption data, excluding climate variables, more efficient? If so, which classification technique yields the best results?
- How do spatial constraints affect the outcomes of classification techniques in CZB, and to what extent do they improve the accuracy and consistency of climate zoning?
- What are the discrepancies and misclassifications between the proposed building performance-based climate map and the official building climate map of Kazakhstan?

The research questions are methodically designed to address the identified gaps and their interconnected aspects. For example:

1. Identifying the most common building archetypes and their energy consumption levels provides a foundational understanding of the energy demands within different building types across Kazakhstan, which is crucial for developing accurate performance-based CZB.
2. Analyzing the spatial patterns of energy consumption offers insights into regional variations, informing the classification process and ensuring that zoning captures these variations.
3. Determining the primary climate variables that affect building energy consumption refines the climate zoning methodology, ensuring it includes the most critical factors for accurate zoning.
4. Establishing the optimal number of climate zones is essential for creating a balanced and effective zoning framework that aligns with observed energy consumption patterns.
5. Estimating the accuracy and reliability of climate zoning methodologies ensures that the proposed system is robust and applicable in real-world scenarios.
6. Comparing conventional climate-based and contemporary performance-based techniques for climate zoning highlights the advantages and limitations of each, guiding the selection of the most effective method.
7. Investigating the efficiency of a CZB system based solely on building energy consumption data versus including climate variables helps determine the most practical and accurate approach.
8. Assessing the impact of spatial constraints on classification outcomes enhances the precision and consistency of the climate zoning framework.

9. Identifying discrepancies and misclassifications between the proposed building performance-based climate map and the official map of Kazakhstan validates the new methodology and underscores its improvements.

1.2. Research objectives

This study introduces an innovative approach that classifies Kazakhstani territory into CZs based on both metrological factors and spatial constraints while ensuring its quality by validating it with the building energy performance. It aims to create a direct link between CZB and building energy consumption in Kazakhstan.

In more detail, the objectives of the research are as follows:

- Propose a novel building energy performance-based CZB map for Kazakhstan, including the identification of key climate variables influencing building energy consumption and the application of multivariate cluster analysis to define CZs for local building archetypes.
- Verify the proposed maps using a novel performance-based CZB misclassification index, and compare them with the official CZB map of Kazakhstan, evaluating the accuracy and reliability of the official map in predicting building energy performance patterns.

1.3. Significance of the study

Considering the official predictions and the anticipated rapid expansion of house development, the prioritization of energy efficiency emerges as a crucial national strategy for Kazakhstan [20, 21, 27]. This research holds substantial significance as it aims to revolutionize CZB in Kazakhstan. Employing innovative spatially constrained multivariate cluster analysis, the study will generate advanced climate maps, which have a great potential to serve as a foundational resource for driving advancements in buildings, construction, energy, and other related fields within the country. This research promises to catalyze advancements in science and technology by offering a precise and comprehensive understanding of the connectivity between climate and building performance. In practice, the results of the study can be used for the following technological purposes:

- To establish uniform guidelines for envelope characteristics (U-values, windows-to-wall ratio, solar heat gain coefficient, etc.) for different climate regions of Kazakhstan;
- To assist policymakers in developing energy policy based on scientific data-driven evidence and set energy-saving targets for Kazakhstan;

- To assess the energy performance of existing buildings for maintenance and renovation purposes.

Beyond technological advancements, the economic and social implications in Kazakhstan are noteworthy. The introduction of improved climate zoning in Kazakhstan is expected to have far-reaching economic and social impacts. It will foster the development of high-standard building designs, particularly in energy efficiency. This is especially relevant given that the potential for energy savings in buildings in Kazakhstan can reach up to 50% [27]. One significant factor that could contribute to this is the implementation of new CZB standards, which would replace the outdated Soviet building codes that lack energy-saving requirements. The proposed framework can play a pivotal role in this transition. Understanding the optimal CZB enables the establishment of critical early-stage design parameters, which facilitates reduced energy consumption, lower operating costs, and enhanced environmental sustainability. Ultimately, this research has the potential to shape the future of building practices, modernize energy policies, and positively influence the well-being of Kazakhstan's society and economy.

1.4. Outline

Chapter 1 consists of the background of the research, objectives, and significance of the research. The rest of the thesis is organized as follows:

Chapter 2 provides a detailed literature review of various CZB methods, including traditional approaches and modern techniques. The differentiation between climate-based and performance-based approaches for CZB is established. The chapter also explores the interplay between CZ and building energy performance.

Chapter 3 outlines the methodology employed, exploring a research framework, the dual-phase structure of the research, the study area description, and the source of weather data. It further clarifies the building archetypes selection process, performance simulation, and the identification of the most important climate variables. The chapter culminates in the description of classification methodology (multivariate clustering) and classification results validation metrics.

Chapter 4 delves into the double-passed development of CZB for Kazakhstan, showcasing results from building performance simulations and spatial patterns analysis, highlighting key climate classification variables based on correlation analysis and

machine learning algorithms. Phase 1, exploring the climate-based classification approach, determines the optimal number of CZs, presents clustering results, assesses clustering quality, provides validation through building performance-based misclassification analysis, and offers an overview of the results. Phase 2, investigating the performance-based climate classification approach, identifies its specific optimal number of CZs, presents clustering results, evaluates clustering quality, establishes the reference levels of misclassification, and summarizes findings. In the end, the chapter contains a comparative analysis of the results obtained from both phases. It evaluates discrepancies and misclassifications between the two approaches and compares the findings with traditional climate maps.

Chapter 5 presents conclusions drawn from the research, acknowledges its limitations, and presents recommendations for future investigations in the realm of CZB.

Chapter 2: Literature Review

The objective of this chapter is to critically examine the existing literature to comprehend the extent and depth of current knowledge regarding CZB. This exploration is pivotal for understanding the multifaceted approaches employed in climate zoning and their connection with building energy performance, which ultimately guides the focus of this research. Section 2.1 focuses on an overview of existing CZB methods. It delves into the details of specific CZB techniques, such as the degree-days (Section 2.1.1), machine learning (Section 2.1.2), building energy simulation (Section 2.1.3), the interval judgment method (Section 2.1.4), bioclimatic charts (Section 2.1.5), the Köppen-Geiger climate classification (Section 2.1.6), and others. Moreover, section 2.1.8 analyzes the combination of these techniques to emphasize the complexity and versatility built into CZB practices. Section 2.2 transitions into a critique of traditional approaches, challenging their limitations and setting the stage for innovative solutions that address identified gaps. Section 2.3 focuses on building energy performance in CZB, emphasizing the critical role of building energy simulation in climate zoning optimization. Section 2.4 investigates the energy-saving potential of proper CZB, underscoring the significance of tailored climate zoning in reducing energy consumption and enhancing sustainability. Finally, Section 2.5 concludes the chapter with a summary of the findings, synthesizing the insights gained from the literature review and highlighting the contributions this research makes towards advancing the field of CZB.

2.1. Overview of existing Buildings climate zoning methods

Since ancient times, attempts have been made to categorize territories based on their climate. There are several recorded attempts by Greek philosophers (Pythagoras, Aristotle, Plutarch, and Ptolemy) to map and categorize the climate [28]. The Torrid Zone, two temperate zones, and two cold zones were all suggested by ancient Greeks for a spherical globe. Aristotle later identified the Tropics using astronomy and geography. Based on the duration of the longest day, Ptolemy (90-168 A.D.) established a more detailed classification with seven CZs [29]. The earliest reported maps that used AT and, later, precipitation data mark the beginning of the contemporary period of climate classification, which dates back to the 19th century. Von Humboldt [30] created the first isothermal map in 1817. Supan [29] created the first climatic map based on mean annual and warmest month temperatures in 1879. Later Köppen attempted to develop methodologies that would determine CZs linked to vegetation characteristics and published his initial scheme in 1900.

The Köppen map, which was presented in its latest version by Geiger in 1961 is continually being updated and improved upon today, it is still the most often used map for

classifying climates [31-35]. Nonetheless, there has been substantial criticism of the Köppen-Geiger categorization when implemented in CZB [36]. Vegetation-based climate classifications may be useful for predicting agricultural potential, but they can hardly be applied to other purposes like assessing the effectiveness of energy efficiency measures in buildings or providing insight into people's comfort in varying climatic conditions. The first climate classifications dedicated to buildings appeared in building codes and regulations in the early half of the twentieth century [37-39]. The criteria of construction standards and codes progressively increased over time. Energy efficiency requirements were increasingly being included in addition to weather protection and interior comfort. The early 1970s oil supply crisis accelerated energy efficiency regulations for buildings in several countries. During the 1980s and 1990s, most OECD (Organisation for Economic Co-operation and Development) countries introduced or expanded energy efficiency rules. For example, the United States has had building energy efficiency requirements since 1975, with ASHRAE Standard 90-75 [40], Germany since 1977. This new regulation was developed in response to the Kyoto Protocol and other CO2 emission reduction or stabilization targets [41].

Ideally, a building should create comfortable interior conditions while consuming as little energy and resources as possible. Many factors affect how much energy a building uses, but one of the most significant is the environment or climate [13, 42]. Changes in climatic conditions have an impact on building energy usage when all other elements (such as socioeconomic circumstances and building attributes) remain constant. Given that it accounts for 17% of worldwide CO2 emissions and 27% of global energy consumption, the residential sector ought to be a major player in the effort against climate change. As a result, the energy use of buildings and the climate are intimately related. Reliable climate zoning is extremely important to most building energy-saving initiatives, highlighting its importance in meeting the desired goal for energy security and decarbonization [43, 44]. Correct climate classification saves energy while maintaining occupants' thermal comfort within acceptable limits [2]. With this in mind, the purpose and benefits of CZB become clear. The significant number of articles published on that problem in the previous 15 years demonstrates the topic's relevance to the public, government, and scientists [1, 3-5, 8-11, 14, 17, 18, 45-56].

Two review articles on CZB [1, 13] were discovered in the reviewed literature. In 2017 Walsh et al. [1] conducted a comprehensive assessment of the climatic classification of buildings and energy-saving measures taken by 54 countries. This research encompassed local and international standards, regulations, scientific publications, and other relevant materials. However, the study predominantly adopted data from national and international building rules, with 90% of the cases derived from normative sources. The authors' principal

conclusions, among others, are: the development of CZB serves multiple purposes, with the majority aimed at facilitating performance-based and/or prescriptive-based building regulations; over the years, there has been a rise in the utilization of building performance simulation to aid in the establishment of CZB, and this shift has occurred from a climate-based methodology to a performance-based methodology; only a combination of methods (building performance simulation and cluster analysis) can deliver tools for tracking connections between the climate and buildings. In 2021 Verichev et al. [13] examined the climate-related research in buildings between 1979 and 2019. The review addressed a broad spectrum of subjects, both directly and indirectly related to climate, CZs, and buildings. Enhanced with bibliographic analysis, this work is significant in providing a methodology for categorizing buildings-related climate-oriented studies. The researchers conclude that 88% of all climate-related studies fall under the overall theme of energy conservation, and realizing how energy conservation is related to climate and buildings can help build climate-appropriate housing around the world.

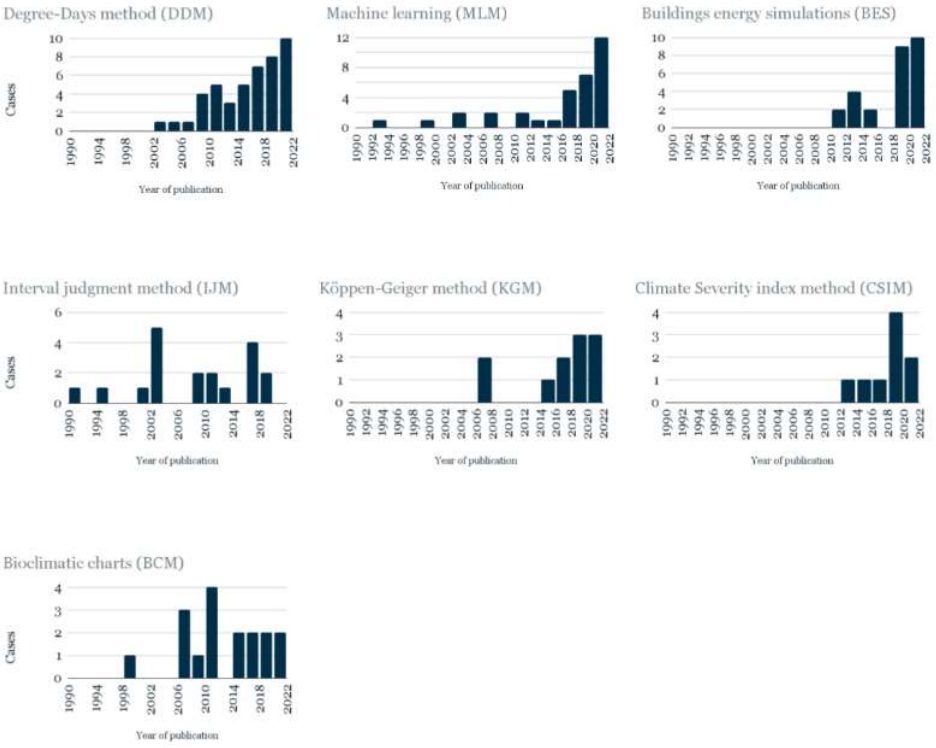
In this review, 156 scientific sources and national codes spanning from 1990 to 2022 and covering 66 countries were analyzed. CZB techniques were categorized based on earlier research (Table 2.1). Here is a detailed analysis of the most often used methods, the number of methods utilized simultaneously, and which methods are used more often in national codes and academic publications will be given. Along with "method", we will employ "approach," "technique," "strategy," and "mechanism" terms to prevent lexical repetition.

The variety of global CZB techniques is seen in Figure 2.1. Building energy simulation (BES), machine learning method (MLM), and degree days method (DDM) are three of the most often used strategies. Furthermore, Figure 2.1 (a) provides time-series histograms of the publishing years according to the approach together with data on the average and median dates of release of papers devoted to various methods. Figure 2.1 (b) may serve as proof of which techniques are more current or applicable today and which were more common in the past. Peak dates for IJM and BCM were 2009 and 2012, respectively. Two of the most modern approaches are BES and MLM.

Table 2.1: CZB methods definitions.

#	Name of a Method	Abbreviations	Criteria
1	Machine learning methods	MLM	Classification is based on clustering techniques, neural networks, support vector machines, sensitivity analysis, or principal component analysis
2	Degree-days/-hours methods	DDM	Classification is based on DD values only. OR If several variables are used in the classification, then DDs should be the primary variable
3	Building energy simulation	BES	Classification is based on BES results
4	Bioclimatic charts method	BCM	Classification is based on Givoni, Lamberts, Milne, and Olgay charts with the combination of temperature and humidity as the main variables
5	Köppen–Geiger method	KGM	Classification is followed by the Köppen–Geiger system and is based on seasonal precipitation and temperature pattern
6	Climate severity index method	CSIM	Classification is based on the climate severity index (a site-specific value that defines the severity index of a specific climate) according to Formulas 4 and 5
7	Interval judgment method (the complex combination of climate variables based on the repeatability of their elements)	IJM	The classification is based on a combination of different variables with established limits (threshold) of variables for each zone
8	Frequency distribution of climate variable	FDV	Classification is based on the different types of probability distributions of variable(s)
9	Mahoney method	MM	Classification is based on Mahoney tables
10	Thornthwaite climate classification method	TCCM	Classification is based on Thornthwaite climate classification
11	Existing building stock performance method	EBSM	Classification is based on actual data of building stock performance
12	Heating or cooling index	HCI	Classification is based on heating and cooling indexes. This index is commonly used to determine how ambient temperature, relative humidity, and radiation affect human comfort
13	Other		Quitt's climate classification Roriz method The World Health Organization (WHO) classification method Administrative division Approximation and interpolation method (AIM) Camargo climatic classification

(a)



(b)

Figure 2.1: Different methods in climatic zoning with mean and median years of publication (a), and publication years distribution by the method (b).

Often only one method is used for climate zoning. Of 138 documents, 102 sources (74%) utilized one method, 28 (20%) used two, and 8 (6%), three or more (Figure 2.2). Subsection 2.2.8 discusses combining approaches in more detail. Figure 2.3 shows how

frequently methods are employed alone or in combination, as well as their distribution between scientific sources and national standards. In academic research, preference is given to MLM and BES; DDM and IJM are much more widely adopted in national codes and standards (Figure 2.3).

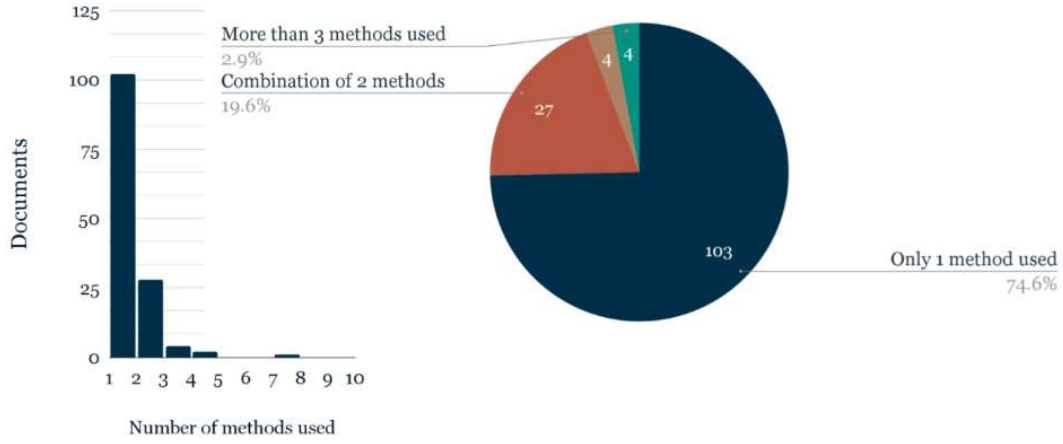
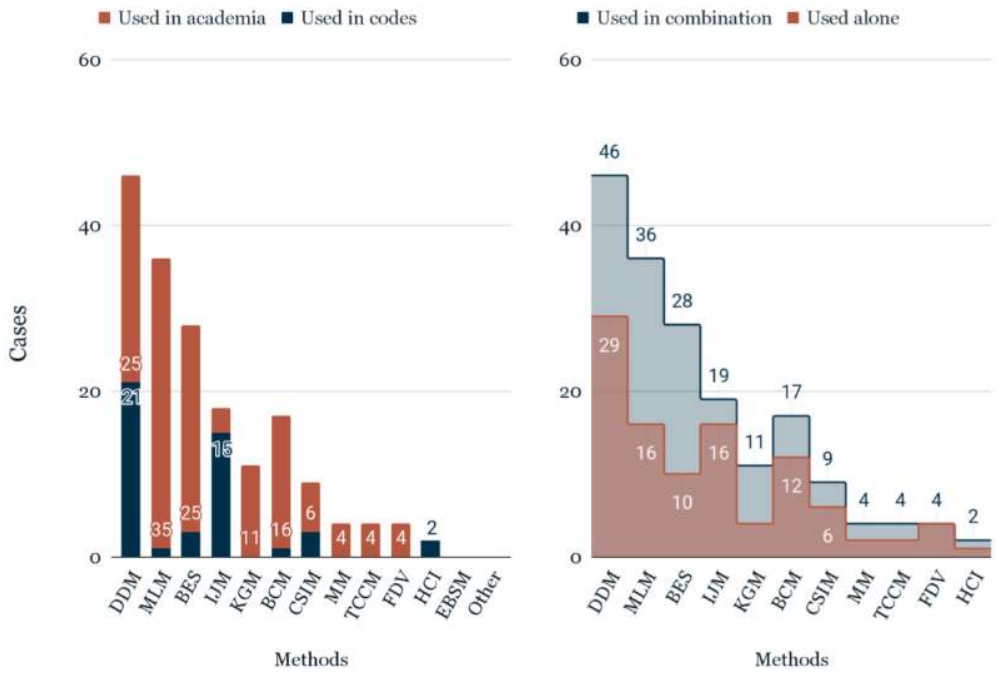


Figure 2.2: The number of CZB methods used and their frequency.



(a) (b)

Figure 2.3: Usage of various CZB methods: broken down by national codes and academic publications categories (a); used alone or combined (b).

Next, the characteristics and essence of each particular method will be given in order from most to least common.

2.1.1. **Degree-days method (DDM)**

For decades, climate zoning has been accomplished with the use of a well-known DDM [57]. Degree days (DDs) determine the climatic conditions by capturing events when the outdoor AT drops below or goes above a specific threshold in a particular year, requiring the use of heating or cooling systems. DDs are commonly described as the total temperature variations between the average outside AT during 24 hours and a base temperature each day. The base temperature refers to the ambient temperature at which HVAC systems do not have to operate to keep the building's internal climate comfortable. The DDM is widely used for CZB, in part because it is straightforward to grasp, requires little computation, and closely links energy usage, particularly in cold climates [57]. There is also a strong correlation between DD data and the use of natural gas, electricity, and heating [58, 59]. Additionally, HDD is a trustworthy measure of household energy use [60, 61].

Pusat & Ekmekci [62] applied DDM for the climatic zoning of Turkey. In contrast to Turkey's official CZB, which solely considers heating degree days (HDD), the authors used a combination of heating and cooling degree days (CDD). A new technique developed from typical meteorological year (TMY) files was compared to current Turkish regulations. Six primary climatic zones were identified using the proposed technique instead of the official code's four. The authors propose reclassifying the country's climate in terms of both heating and cooling requirements. The findings emphasize the need for cooling load consideration in degree-day climate zone classification. Noh et al. [63] used DD data from 255 South Korean cities to propose a new climate classification. DDs were estimated using data from the Korean Meteorological Administration for outdoor dry-bulb temperatures (DBT) from 1981 to 2010, and CZs were visualized using ArcGIS. The authors used the HDD calculation to classify the country into three, four, and five CZs, according to ASHRAE guidelines. As a result of the distribution and significant variations in DDs, four CZs were identified as the best solution. Abebe & Assefa [45] seek to regionalize Ethiopia's climatic zones and estimate the energy consumption in different zones using DDM. HDD and CDD values were computed using AT data from 952 National Center for Environmental Prediction stations for 34 years (1979–2013). Ethiopia was regionalized into five CZs using ArcGIS interpolation and intersect tools based on ASHRAE guidelines. In addition, the total energy required for indoor comfort was determined for the whole country, considering the growing population and living standards. Based on the average yearly cost of energy use, Ghedamsi et al. [64] divided the Algerian region into seven CZs. Annual heating and cooling needs of buildings in 48 different locations of Algeria were evaluated using the DDM. The base temperature for HDD was chosen at 18°C and that for CDD at 26°C. Using GIS, a CZB

map was created. Each area's annual energy use for heating, cooling, and home appliances was determined. Consequently, the final climatic map of Algeria displayed the variety of costs associated with thermal energy demands for cooling and heating in each zone.

With a demonstrated relationship to building energy, DDM can provide quality climatic zoning for a variety of purposes. However, DDM essentially operates with outside AT only, leaving out other significant climatic elements that impact a building's energy use. Also, the selection of the proper base temperature is critical. According to this review, the most widely utilized base temperatures for HDD are 18°C and 10°C for CDD.

2.1.2. **Machine Learning Methods (MLM)**

MLM stands for several multivariate data segmentation and classification methods that are effective when used for CZB. It can incorporate a variety of geographic and climatic factors at the same time, or it can be coupled with certain construction attributes, preventing oversimplification and yielding more insightful outcomes [65]. Aside from cluster analysis (CA), additional ML techniques for CZB (neural networks, principal component analysis, sensitivity analysis, etc.) were disclosed in the literature study. As a result, CA was combined with other ML algorithms [66-69] in this review. Nevertheless, CA remains the most often used approach for classifying climates among all conceivable MLM (Figure 2.4).

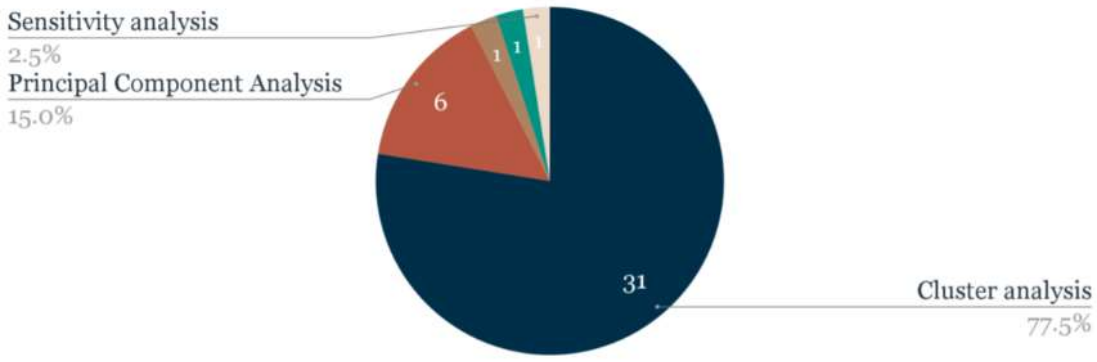
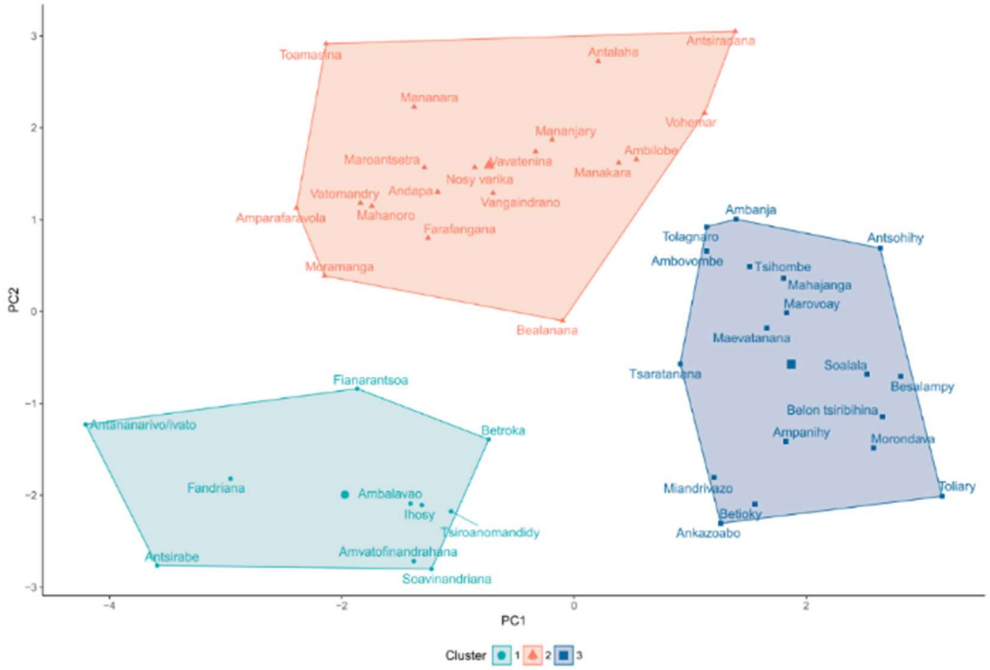


Figure 2.4: ML algorithms for CZB.

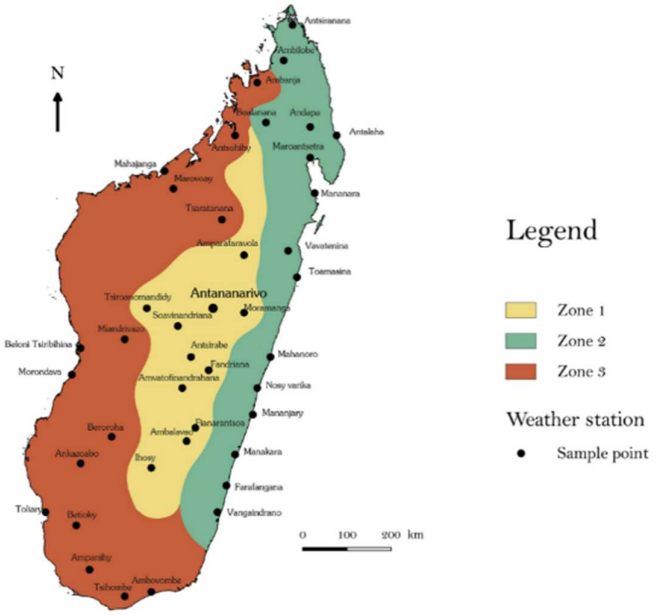
Of 138 documents with known methodology, 36 (26.1%) cases of applying ML techniques were found. MLM is widespread and actively studied in the scientific environment. All 35 cases on MLM for CZB are scientific publications.

One of the earliest attempts to implement MLM for climate zoning is the work from 1993 by Fovell et al. [70], where hierarchical clustering (HC) with the Euclidean and Mahalanobis metrics on AT and precipitation data was used to classify the conterminous United States. The country was classified into 14 CZs by using a combination of principal

component analysis (PCA) and HC. The authors noticed cluster overlap and buffer zone inconsistencies, as main negatives. For a more accurate classification near the study area's borders, it is necessary to have data points outside the boundaries of the area itself, which were not included in the analysis. Praene et al. [53] relied on hierarchical clustering on principal components (HCPC) with spatial interpolation to classify the climate of Madagascar. 47 weather stations and 9 climatic variables were used in the PCA matrix. Figure 2.5 (a) shows the PCA findings for the weather stations with three distinct clusters (CZs). GIS was used for mapping a CZ pattern for Madagascar's region (Figure 2.5 (b)). The relationship between defined climatic zones and the thermal comfort of traditional building archetypes was examined. The findings show that some archetypes provide a guarantee of a higher yearly comfort level in particular climatic zones.



(a)



(b)

Figure 2.5: HC results for the weather stations (b), and final CZ results (c).

Lau et al. [71] presented a study using data from 123 measuring stations in China to make a climate map based on solar radiation (SR). The monthly average daily clearness index was used as the main climate variable for CA. Ward's technique helped create a hierarchical tree to find distinct clusters. Five major solar CZs have been recognized. Official climate classification and the dominant topography were compared to the proposed solar climates. The authors claim that the produced map can be used by architects and designers in the preliminary stages of a building project to establish passive methods. Understanding the potential and profitability of solar power conversion technologies is important for energy policymakers. To divide the 661 weather stations in China into distinct CZs, Shi & Yang [72] proposed a novel spectral clustering-based technique. In the first step, the connections between five weather parameters (average daily AT, average RH, sunshine hours, diurnal temperature range, and atmospheric pressure (AP)) were established. The advantages of k-nearest-neighbour and sparse subspace representation were then used to produce a similarity matrix. The classification's logic was supported by the similarity matrix blocking effect. Researchers examine climatic classification using a single variable and multiple variables for a different number of clusters. The proposed method turned out to be consistent with single or multiple-view classification. The authors suggested a method for calculating the number of clusters and conducted a sensitivity analysis on various parameters. Yang et al. [55] built a supervised classification approach using data collected from 701 national weather stations from 1984 to 2013 to better categorize China's climate. Twenty-two

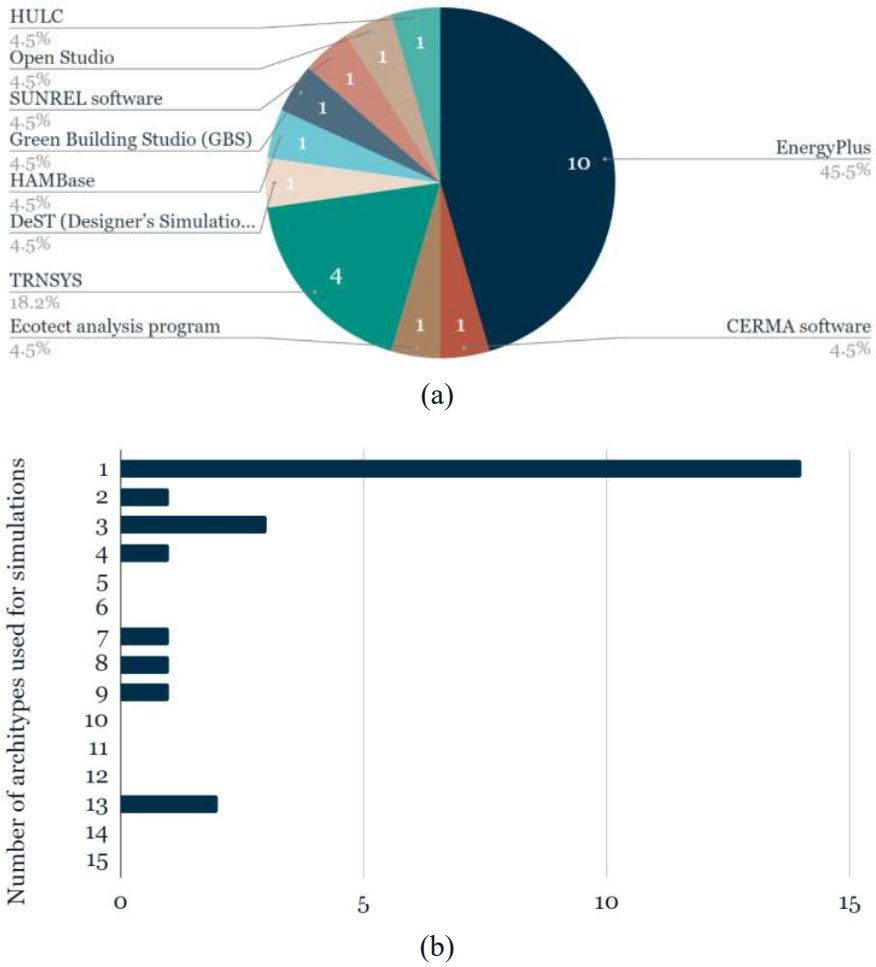
classification algorithms: decision trees, discriminant analysis, support vector machines, nearest neighbour classifiers, and ensemble classifiers were used. All of the available classification algorithms were compared. "Bagged Trees" and "Subspace Discriminant" had the highest prediction accuracy. The confusion matrix of different variables was composed. The Mahalanobis distance was used to compare clusters' climates, and a seven-zone final climate map was created. The authors proved that the supervised classification algorithm is better than China's existing CZB system, which was formed in 1993 and based on IJM. Energy-efficient building design solutions in Israel were proposed by Erell et al. [73] using a climate map derived from a CA. The HC method was used to pick the first set of climate variables. Four "core" variables were selected for the winter and summer seasons. The authors suggest that AT data alone may not be sufficient to reach the full potential of energy-efficient design. CZB needs to take into account other climate factors like RH and SR. The authors provided maps depicting 7 winter and 4 summer CZs, as well as the mixed classification. The proposed approach may help to define spatial boundaries for certain performance criteria or climate-conscious design techniques. The authors also concluded that CZB should differentiate both winter and summer periods.

Computer power and software have made complex data processes and calculations faster and more efficient. This makes sophisticated methods popular and easy to utilize. ML allows scientists to investigate problems more deeply and find more accurate solutions. MLM appears in many climatic classifications research studies, which is not surprising given its primary advantages: analyzing large amounts of data, ability to use both meteorological and building features, etc. MLM is often (61% of cases) used with BES to acquire more reliable results or to validate them.

2.1.3. **Building energy simulation (BES)**

The use of BES in climate zoning has demonstrated significant promise [1, 4]. The BES approach predicts how a building and its parts will use energy in real-world situations. To accomplish this, a digital model is employed to simulate events within a virtual environment. The BES climate categorization approach makes performance maps that illustrate how certain metrics, such as energy consumption or thermal comfort, vary throughout a region (country or area) for the designated buildings using meteorological data from a typical year file [4, 56, 74]. Next, this performance data is classified into CZs. The performance of a building model within a single climate zone is assumed to be almost the same in this technique.

Of 138 documents, 29 cases of applying BES techniques from 19 countries were revealed (21.0%). Two countries (Australia [75] and Morocco [76]) have an existing policy that officially utilizes the BES method for supporting its CZB. EnergyPlus software is the most commonly used for BES (Figure 2.6 (a)), and most often during simulations, only one building archetype is used (Figure 2.6 (b)). Among all publications studied in the review, BES was used as the basic method in 6% of cases.



(a)

(b)

Figure 2.6: The most commonly used software for BES (a), and the number of building archetypes used during BES (b).

Semahi et al. [77] provided updated climate zoning maps for Algeria based on the thermal energy consumption and indoor-discomfort hours of the residential building archetype, using a climate dataset of 74 meteorological stations and a calibrated residential building model. Figure 2.7 shows the study's conceptual framework, which includes climate data collection and model generation, building energy performance simulations, and GIS-based climate mapping. The findings offer a climatic categorization of Algeria's 9 CZs. An

energy need for heating and cooling as well as hours of discomfort have been calculated for each climatic zone.

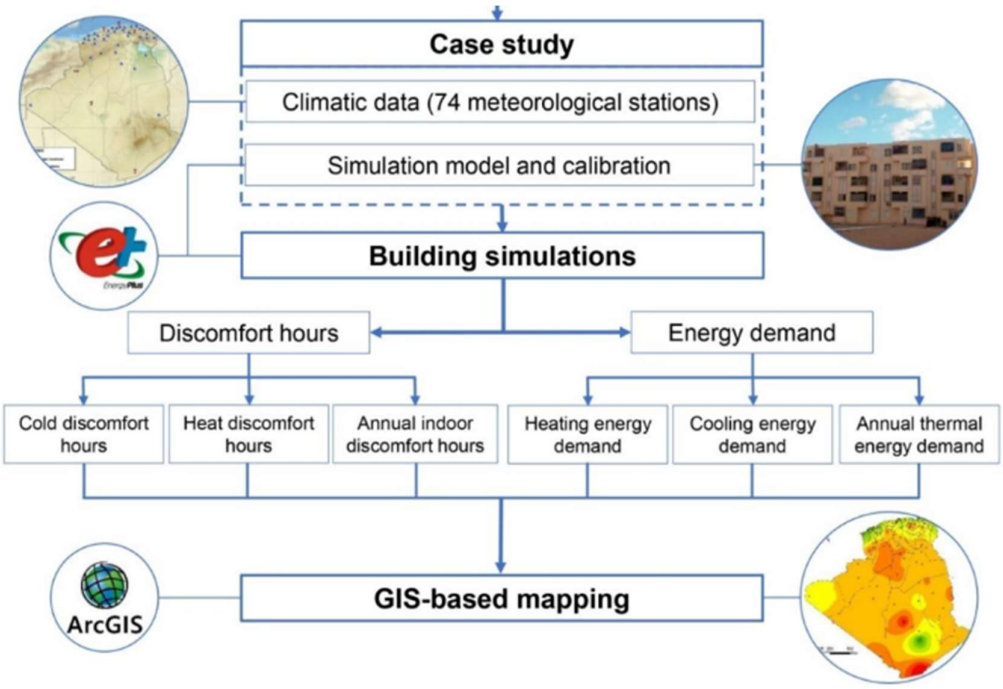


Figure 2.7: Conceptual framework of climate mapping based on BES.

Using meteorological data from 1997 to 2013, Bai & Wang [47] proposed to create updated CZs for China's using the BES of a 24-story office skyscraper to account for climatic change over the last several decades. According to the authors, revising China's thermal climatic zones based on post-1997 climate data is far more realistic. The official thermal climate zoning of China [78] is based on 1951-1985 meteorological data. Next, the "climate jump" phenomenon was explored using a moving t-test. After the inadequacies of the current climatic zones were identified, 41 cities were moved from colder to warmer regions. According to the study, if the analyzed building were constructed in the designated climatic zone, has the potential to conserve a greater amount of energy. Schijndeln & Schellen [74] produced building performance maps of similar buildings that are virtually spread over the whole of Europe. The authors mainly focused on interior climate performance factors, noting that BES maps can help study regional climate influence on building performance indicators including energy use and indoor climate. Meteonorm 2011 was used to create approximately 130 weather files for different cities. The Bestest was implemented as building simulation tools. For BES, the authors utilized a 48 m^2 single-story building model with a 2.7 m floor-to-floor height. The mean annual heating and cooling power, peak heating and cooling, the mean indoor AT, and RH maps were produced based on simulation results. Consequently,

the authors showed potential for applying such mapping tools to visualize potential building measures over Europe.

When it comes to national standards, in addition to Australia's National Construction Code (NCC) with eight CZs, the country's territory is divided into 69 subzones by the Nationwide House Energy Rating Scheme (NatHERS) [75], which divides the country's territory into 69 subzones based on BES and local geographic features like wind patterns and height above the sea. The CZs are delineated by postcodes. NatHERS software simulates Australian homes' thermal comfort and rates them 0 to 10 stars. The more stars, the less energy is needed to keep inhabitants comfortable. The software generates computer simulations determining how well a building fits into a particular climate zone helping to optimize dwellings' thermal performance during the design stage. Another country, that utilizes the BES for CZB is Morocco. However, in Morocco, the official regulations [96] use a more complex approach with the combination of DDM and BES.

BES is also a reliable data source for validating traditional climate classification. It ensures that the main classification is correct and reduces misclassified areas at climatic zone borders, making it more robust and accurate. However, BES usage in climatic zonings is limited by the need to pre-define a design hypothesis based on building type, occupational patterns, and HVAC systems. Meteorological data is also needed, which isn't always available [49]. According to numerous studies, detailed climatic data and building modeling could benefit climatic categorization [3]. Despite having substantial advantages, the approach's adoption in official standards throughout the world is still limited, and BES is considerably more popular in the scientific context.

2.1.4. **The interval judgment method (IJM)**

As was already mentioned in the introduction, to identify a region's climatic zone, IJM uses a range of zoning factors and threshold values, such as the AT and RH levels. There can be modifications in variables and thresholds employed between different countries and target sets [55]. Numerous post-Soviet nations, including Armenia, Belarus, Georgia, Kazakhstan, Kyrgyzstan, Mongolia, Russia, Ukraine, and Uzbekistan, have been found to employ this technique. The 1983-approved SNiP 2.01.01-82 "Building climatology and geophysics" [16] provided the framework for contemporary post-Soviet countries' climatic zoning regulations, which can be easily traced.

In this literature review, 19 cases using IJM were discovered out of 138 documents. It was discovered in 19 countries' documents, with 18 having an existing policy that uses the IJM to establish their CZB. IJM was utilized as the primary method in 9% of the publications

examined in the review. Since this method is mostly utilized (95% of cases) in national standards, in this section we will focus only on these cases.

In Kazakhstan, SP RK 2.04-01-2017 "Building climatology" [15] is a collection of climatic data for large cities of the country, which is derived using information gathered from the National Hydro-Meteorological Service of Kazakhstan (KAZHYDROMET) network of weather stations over extended periods of time. For some climate variables, the calculation's observation periods are 1971–2010 and 1981–2010, respectively. Among other things, a differentiation table for climate zoning (Table 2.2) is displayed accompanying the republic's official climate-zoning map (Figure 2.8). The mean monthly RH level in July, the mean monthly AT of the air in January and July, and the mean WS for the three winter months are used to determine the type of climate. CZB are divided into four primary zones and seven subzones.

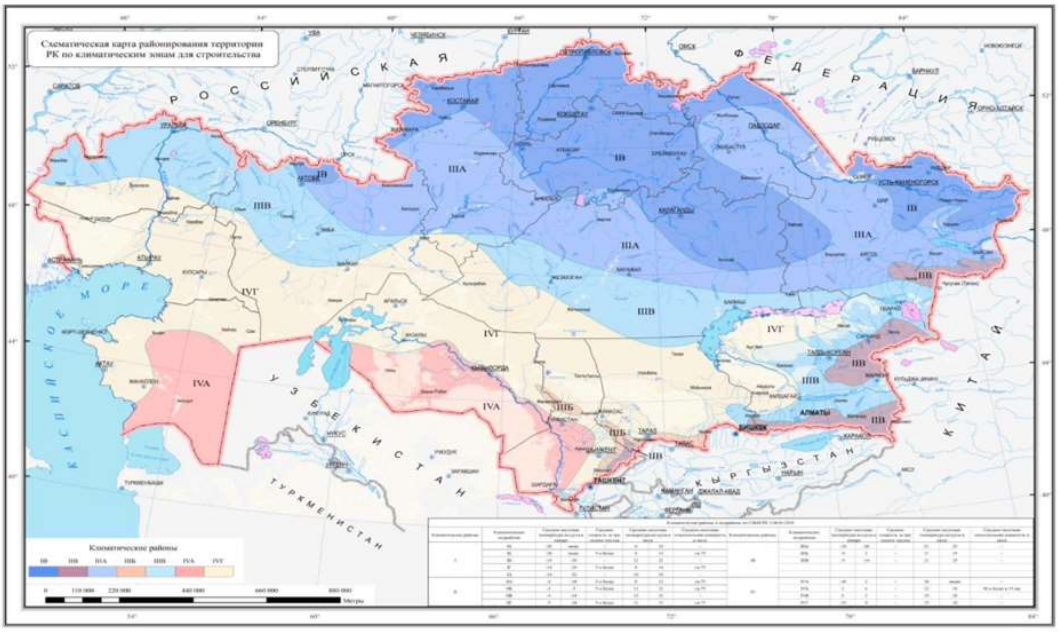


Figure 2.8: Official CZB map of Kazakhstan.

Table 2.2: The Republic of Kazakhstan's official climatic zoning differentiation norms for buildings.

Climate zones	Climate subzones	Average monthly air temperature in January, °C	Average wind speed for three winter months, m/s	Average monthly air temperature in July, °C	Average monthly relative humidity in July, %
I	IA	From -32 and below	-	From 0 to 19	-
	IБ	From -28 and below	5 and more	From 0 to 13	More than 75
	IB	From -14 to -28	-	From 12 to 21	-
	IГ	From -14 to -28	5 and more	From 0 to 14	More than 75
	IД	From -14 to -32	-	From 10 to 20	-
II	IIA	From -4 to -14	5 and more	From 8 to 12	More than 75
	IIБ	From -3 to -5	5 and more	From 12 to 21	More than 75
	IIB	From -4 to -14	-	From 12 to 21	-
	IIГ	From -5 to -14	5 and more	From 12 to 21	More than 75
III	IIIA	From -14 to -20	-	From 21 to 25	-
	IIIБ	From -5 to 2	-	From 21 to 25	-
	IIIB	From -5 to -14	-	From 21 to 25	-
IV	IVA	From -10 to -2	-	From 28 and above	-
	IVБ	From 2 to 6	-	From 22 to 28	50 and more at 3 pm.
	IVB	From 0 to 2	-	From 25 to 28	
	IVГ	From -15 to 0	-	From 25 to 28	

China [78] uses a similar approach, while somewhat different, to determine its CZB. The zoning requirements (environmental variables and intervals) are shown in Table 2.3. There are five primary climate regions (SCZ, CZ, HSCWZ, MZ, and HSWWZ) and their eleven subzones. The average AT in the hottest month (Thot), the mean AT in the month with the lowest temperatures (Tcold), the number of days with an average temperature below 5°C (D5), and the number of days with an average temperature over 25°C (D25) are the primary variables in Chinese IJM. Moreover, DDs at predetermined intervals are used to divide subzones. China, in contrast to Kazakhstan and Russia, employs a hybrid approach for CZB, enhancing its IJM with DDM.

Table 2.3: China's official climatic zoning differentiation standards for buildings.

Zone name		Zoning criteria		
		Main zone		Subzone
		Main criteria	Complementary criteria	
Severe cold	A	Tcold ≤-10°C	D5≥145 days	6000≤HDD18°C
	B			5000≤HDD18°C<6000
	C			3800≤HDD18°C<5000
Cold	A	Tcold =0-(-10)°C	D5=145-90 days	2000≤HDD18°C<3800
				CDD26°C≤90
	B			2000≤HDD18°C<3800
				CDD26°C>90
Hot summer and cold winter	A	Tcold =0-10°C	D5=90-0 days	1200≤HDD18°C<2000
	B	Thot =25-30°C	D25=40-110 days	700≤HDD18°C<1200
Hot summer and warm winter	A	Tcold >10°C	D25=100-200 days	500≤HDD18°C<700
	B	Thot =25-29°C		HDD18°C<500
Mild	A	Tcold =0-13°C	D5=90-0 days	CDD26°C<10
		Thot =18-25°C		700≤HDD18°C<2000
	B			CDD26°C<10
				HDD18°C<700

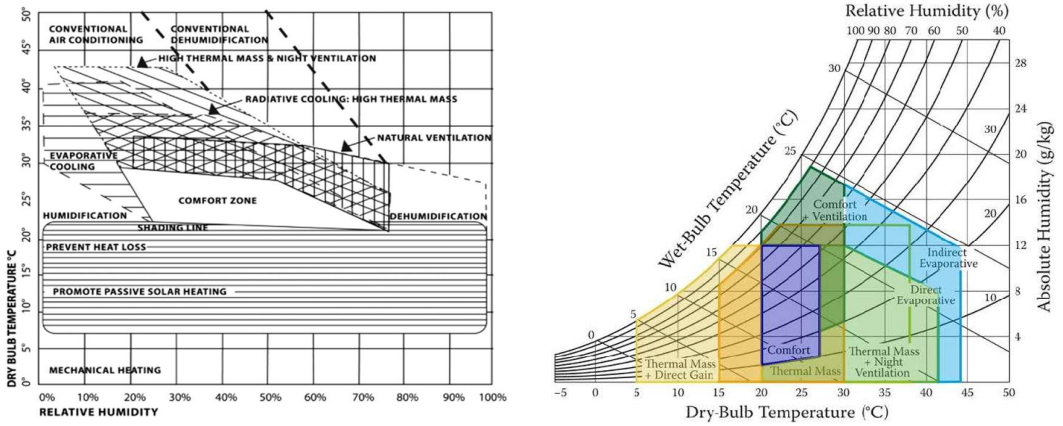
In Austria [79], climate classification also relies on IJM and is based on the monthly mean values of the outside AT. The mean monthly AT in the country is mainly determined by the sea level. The federal territory is divided into seven different regions with a corresponding mean vertical temperature gradient. Among European countries, the IJM is also used by Lithuania [80], Poland [81], and Romania [82]. In Southern America, Colombia uses IJM for CZB [83] and is divided into four climatic zones (warm-humid, hot-dry, temperate, and cold) according to altitude, air temperature, and relative humidity intervals.

With very few modifications, IJM has been in use in multiple countries from the 1980s of the 20th century [15, 78, 84]. IJM is frequently criticized for being out of step with the much stricter current standards for building energy efficiency, particularly when compared to BES or MLM [17, 47, 55]. There's no consistent link between IJM climate categorization and building energy needs in its current forms. Still, it can be applied for CZB, with a narrow range of variables that correlate strongly with the overall energy consumption of the buildings, like AT, SR, and altitude [85-88]. There are now generally accepted, considerably more reliable procedures that are founded to be more precise and transparent classification strategies and yield more robust outcomes.

2.1.5. **Bioclimatic charts method (BCM)**

In recent decades, there have been numerous attempts to develop bioclimatic charts (BC) that would be customized to human needs and building design. It is proved that bioclimatic zone analysis can help modify existing or create new climate zoning [18, 89-91]. The BC analysis often results in the identification of passive design techniques that can maintain interior thermal comfort while also making the built environment more energy-efficient. It is typically simpler to assess the climatic elements of a certain region on the chart because BC shows the combination of both humidity and AT at each given moment. This review includes Givoni, Lamberts, Milne, and Olgay and other psychometric chart methods in this category. We found 17 cases of use of this method among 138 (12%) in this review. As a stand-alone method, BCM was used in 9% of cases. Only one country (Argentina [92]) officially utilizes BCM for CZB.

Olgyay [93] was the first to suggest a systematic bioclimatic building design strategy. Cool, temperate, hot and arid, and hot and humid were the four key climate groups he discovered in his research on the impact of climate on building design concepts around the world. Human tolerance ranges as a combination of DBT and RH were employed, which he derived from a bioclimatic chart (Figure 2.9 (a)). Not only was the average radiant temperature taken into account, but also the WS and SR. Later, Milne and Givoni [94] created BC based on standard psychrometric charts commonly used to analyze moist air properties (Figure 2.9 (b)). A zone in the middle of psychrometric charts establishes the range of conditions people find comfortable in different situations (such as summer and winter). Hot, warm-temperate, cool-temperate, and cold are the four basic climate types defined by Givoni [95], with eleven additional sub-climatic types. His research had promising implications for the development of HVAC system designs and the improvement of building energy efficiency by examining the impact of climate on occupant comfort and thermal adaptiveness.



(a) (b)

Figure 2.9: Bioclimatic charts: Victor Olgyay chart (a), and Givoni chart (b).

Lam et al. [89] used BC to explore passive design strategies in China. 18 cities representing five CZs were examined. Solar heating, natural ventilation, thermal mass, night-time ventilation, and evaporative cooling were all considered passive design solutions. The summer and winter passive design solutions were separated into nine distinct CZs using a bioclimatic approach. The study's findings can inform architects' and designers' decisions about passive design strategies, particularly in the preliminary stages of project development. Nevertheless, the study's limitations include a small sample size of only 18 cities, potentially overlooking cities with diverse microclimates not covered by the nine passive design zones. Rakoto-Joseph et al. [96] describe the climate categorization of Madagascar island using the method established by Lam et al. [89]. This classification was based on meteorological data collected in several cities throughout the country over 29 years. Using BCM, three primary climatic zones were found. Finally, specific design strategies for each climate zone were proposed, such as solar heating, natural ventilation, and thermal mass. In the same way, Singh et al. [90] reclassified the climate of the northeast region of India into three distinct climatic zones using BCM. This classification is based on 30 years of weather data. Psychometric charts were established for each bioclimatic zone to examine solar passive design for dwellings. According to the authors, psychometric charts can quickly depict a region's climate, and new climate classification helps identify passive design aspects for traditional buildings.

Although achieving comfortable interior conditions under thermal comfort criteria is the main objective of BCs, its implementation also results in a reduction in energy consumption, which raises the building's energy efficiency. A direct link between bioclimatic potential and annual building energy usage is also established [18, 52]. Nonetheless, several resources recommend utilizing BC in conjunction with BES [6, 18, 52, 91] to produce secure results in terms of CZB.

2.1.6. **Köppen-Geiger climate classification method (KGM)**

KGM was stated at the beginning of this review as being among the oldest and most well-established. The KGM only needs a monthly record of average AT and precipitation data. Due to its enormous popularity and availability, KGM is often used for CZB without being directly linked with buildings' energy performance [97-99]. We found 11 cases of use of this method among 138 documents (8%), separately from other KGM used in 4 cases (3%). In 7 cases, the KGM was used in combination with other methods. Besides, not a single country adopted this method for CZB officially.

Acknowledging its frequent usage by academics in a variety of fields as a foundation for climatic regionalization and global climate models, as well as its extensive use in teaching high school and university courses on climate, Peel et al. [100] developed the updated KG map based on a worldwide long-term monthly AT and precipitation time series data. The revised KG world climate map uses climate variables data from 1951 to 2000, interpolated individually. Each station's climate variables were interpolated using a two-dimensional (latitude and longitude) thin-plate spline on a 0.1x0.1 degree grid. The authors evaluated the results continent by continent, investigating the challenges of dealing with areas not uniquely classified by the KG system. Another high-resolution KG classification (1 km resolution) was published by Beck et al. [32] for current (1980–2016) and predicted future scenarios (2071–2100). The map was created using four high-resolution, geographically-corrected climate maps. The future map was created by superimposing 32 climate model forecasts (scenario RCP8.5). The authors estimated classification confidence for both periods, providing valuable indications of the classifications' trustworthiness. Achieved maps are more detailed and accurate than earlier maps, especially in areas with steep spatial or elevation gradients. Sarricolea et al. [99] proposed the up-to-date (1950–2000) KG climate map for continental Chile as a 1x1 km map. This investigation includes 200 weather stations and climate surfaces. Raster images of monthly precipitation and temperature data, as well as bioclimatic variables, were used to describe climatic surfaces. The landforms were categorized using the KG system. The classification was validated by comparing the KG surface classification to FAO Clim 2.0 station classification. 80% of the FAO Clim 2.0 map matches KG station categorizations. According to the results, the majority of Chile's continental regions have arid, temperate, and polar climates. In a publication by Ali & Szalay [97] devoted to the overheating problem and the thermal comfort of buildings in Sudan, the classification of the territory of the country is indicated according to the KGM classification and divided into three climatic zones (Zone I: warm desert climate, Zone II: warm semi-arid climate and Zone III: tropical savanna climate).

The Köppen map is the most extensively utilized climatic classification map, and it still undergoes continuous updates and improvements [4]. However, while vegetation-based climate classifications can indicate agricultural potential, they can't characterize building energy behaviour [101]. MLM and BES outperform KG classification, according to multiple sources [36, 53, 56, 102]. In recent years, other comparisons have shown that KG did not allow the accumulation of precise information required to address the issue of building design and thermal comfort [103, 104].

2.1.7. **Other methods**

The climate-classifying techniques identified throughout the literature analysis extend beyond the seven approaches outlined in the preceding sections. Here, the less popular techniques will be presented, among which, are the climate severity index method (CSIM) [105-108], Mahoney method [109], Thornthwaite climate classification [110, 111], heating or cooling index [112], Quitt's climate classification [113], existing building stock performance [114, 115], the frequency distribution of AT and RH [116, 117].

There exist several indices that can describe the severity of climates, and as a result, there are various methods for calculating these indices [106, 107, 118-120]. A given climate's severity index may be determined by combining a set of climate variables into a single site-specific value. The data may be analyzed in the same manner as other meteorological parameters to detect patterns, create applications unique to certain regions, and evaluate annual or seasonal climatic challenges to better understand their severity. Verichev & Carpio [121] examined three specific areas in the southern part of Chile and revised the average values of HDD on an annual basis by utilizing data from meteorological observations over the last ten years (2011-2015), and revised CZs using the CSIM. Three climatic zones were discovered. Finally, implementing CSIM the relative heating and cooling energy consumption of buildings were examined. The investigation showed the discrepancy between the new and official CZB, where the cost of energy for cooling purposes during the summer might vary by 50% within the boundaries of a single climate zone. Díaz-Lopez et al. [118] used CSIM, to update the CZs of peninsular Spain with modern climate data (2015-2018), and then adapted these zones to the RCP4.5 and RCP8.5 future scenarios to create future climate maps. It demonstrates that the current and future climatic conditions do not match the official Spanish climatic zones for building. Two-thirds of Spanish cities design buildings using outdated climate data that doesn't account for current or future climate realities.

Ogunsote & Prucnal-Ogunsote [122] analyzed current definitions of climatic design zones in Nigeria and used revised Mahoney tables to propose a new one. Local architects frequently refer to Mahoney tables to aid in designing buildings tailored to their specific climates. There are six tables: two for reading off relevant design criteria and four for entering climate data and comparing it to requirements for thermal comfort. The authors noted that the country's official southern and northern classifications are oversimplified and lack a scientific basis. Six design (climate) zones were proposed as a result of the work.

Using the Thornthwaite categorization system [123], Izzo et al. [111] divided the Dominican Republic into five distinct CZs based on thirty-year averages of precipitation and

AT collected from 115 sites. Thornthwaite's climate classification method uses the atmospheric humidity index, the aridity index, and the moisture (humidity) index. The authors selected the humidity for the analyses. Based on this, nine main climatic types (from humid to arid) were identified for the Dominican territory.

Gangolells et al. [114] compiled existing building energy performance using energy performance certificate data for existing buildings in Spain. An analysis of 129,635 energy performance certificates estimated baseline energy usage for existing Spanish residential sector buildings and its correlation with the country's official climate classification. The findings provide new information on the energy efficiency of current Spanish dwellings. The results could be utilized to select energy conservation measures depending on building types, construction periods, and climatic zones. Hjortling et al. [115] analyze energy performance certificates issued for commercial buildings by the Swedish National Board of Housing, Building, and Planning. The authors correlate building energy consumption data with official design standards and CZB. It was discovered that climate zone has less impact on energy use than building types. Also, the measured energy consumption of modern buildings has decreased compared to earlier buildings due to new building regulations and tighter energy performance requirements. Despite this, the energy consumption of new structures is frequently higher than what is required by local building codes.

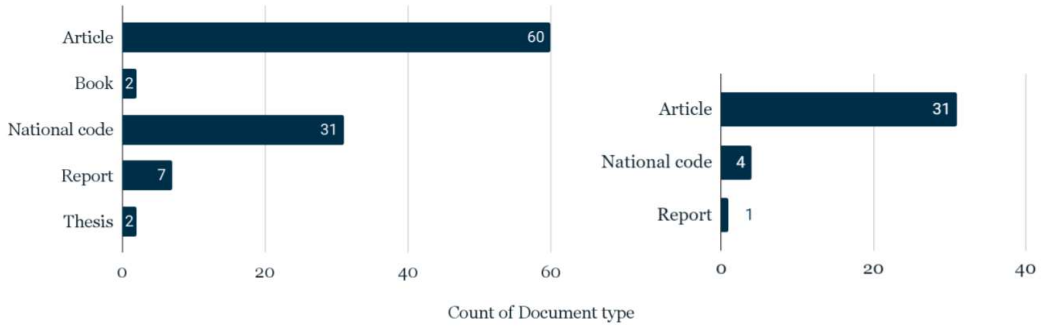
Khedari et al. [117] presented new climate maps, which classify Thailand's locations into zones of relatively similar AT and RH based on a simple statistical analysis (frequency distribution, regular frequency distribution, and relative frequency distribution). The data from 18 years of ambient AT and RH from the 73 observation stations were used. Each observed value, according to the authors, should be placed in one of the classes, which should not overlap, because the frequency distribution is a basic function for estimating the probability of occurrence. The series should be run with the same class interval as much as possible. The highest and lowest readings of variables were used to calculate the expected intervals. The completed map features three distinct temperature zones and four distinct humidity levels. Following the same methodology, Joan Felix et al. [116] composed climate maps of the Dominican Republic for building performance analysis and decision-making regarding energy efficiency.

The methods presented in this section are the least popular among the studied publications, but this does not indicate their poor quality or lack of potential. Based on building energy consumption translated into the conditional index, and incorporating DDs, and SR, CSIM seems an effective and successful solution for CZB. However, the calculation of the regression coefficients is not entirely clear, which in turn makes it difficult to apply

the method in other regions outside Spain. The actual energy consumption of buildings is valuable information that can show how certain parameters (including climate) affect a building's final energy consumption. Techniques such as existing building stock performance [114, 115] or the frequency distribution of variables [116, 117] can be successfully applied to create new climate zoning, to revise or enhance the existing one. However, to date, there are no works where these methods would be compared with others to assess their quality. Additionally, problems arise with data availability, analysis, normalization, etc., when archetypes, envelope parameters, building orientation, and people's behaviour, etc. are considered.

2.1.8. **Combinations of methods**

The combinations of various methods are often used for CZB purposes (Figure 2.10), especially in research. National standards are usually much simpler and are limited for the most part to one method, very rarely resorting to using two at the same time. More than 26% of sources use multiple methods (Figure 2.11). The most popular combination of the two methods is BES/MLM (11 cases). The second and third most popular combinations are DDM/BES and DDM/MLM, with 9 and 8 cases, respectively (Figure 2.12). KG/MLM is the fourth most popular combination with 6 cases. Other combinations are represented by the following connections: DDM/IJM, BES/BCM, BES/CSIM, and DDM/BCM. The most popular combination of the three methods is DDM/BES/MLM.



(a) (b)

Figure 2.10: Histograms of analyzed documents with only one method used (a), and with two or more methods used (b).

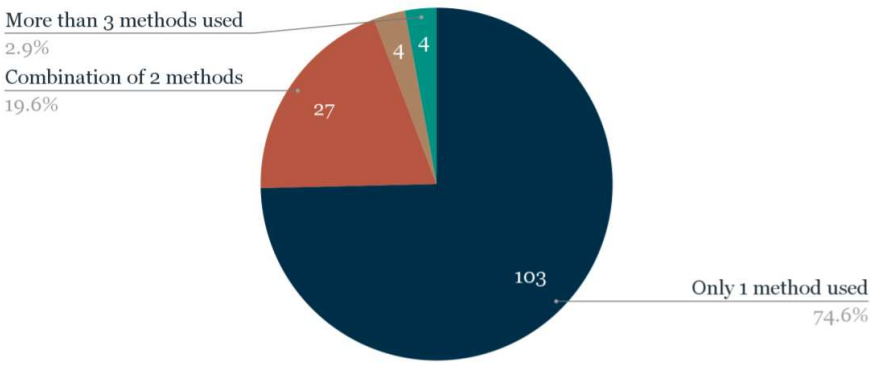


Figure 2.11: The number of methods used for climate zoning.

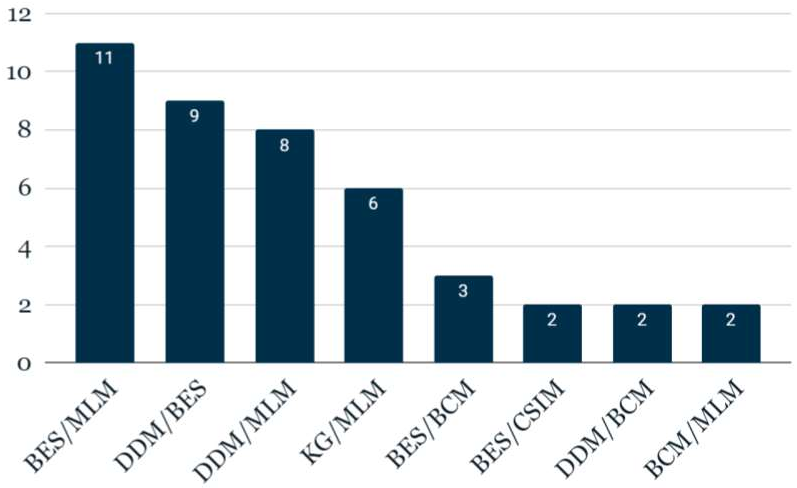


Figure 2.12: The most popular combinations of two methods.

A study by Mazzaferro et al. [51] is an example of the most popular combination of 2 methods (BES and MLM) for CZB. The authors developed a data-driven climatic zoning methodology to increase the robustness of climatic zoning principles. EnergyPlus simulations were performed with 3 archetypes in 411 Brazilian locations. Building performance was analyzed and clustered. A set of climate variables (DBT, RH, and GHI) was subjected to correlation with sensible and latent cooling loads. The optimal number of clusters, based on the Hubert index, was defined as five, and the k-means algorithm was used with selected climatic variables for classification. Building thermal load values within each climatic zone, obtained by clustering climatic variables, were then compared to validate the climatic zones. According to the authors, the combination of BES, and clustering methodologies has benefited the creation of climatic zoning methodology considerably. Tsikaloudaki et al. [124] presented an approach for defining climatic zones in Europe based on the number of DDM and BES results. The actual energy needs for heating and cooling a reference building unit in typical cities within the defined climatic zones facing the four

main directions were compared with the DD classification. It was demonstrated that the energy loads for heating and cooling in both simulations matched the distribution of degree days for each. The article concludes that using HDD and CDD together can produce the most realistic classification. Walsh et al. [54] attempted to demonstrate that there is no scientific evidence connecting building energy performance to ASHRAE CZs. The main objective was to evaluate the energy performance gap between buildings and the expected energy performance in the climate zone they located. The study relies on BES and GIS to conduct a performance-based assessment. Using the Energy Plus software, climatic zoning performance indicators were gathered based on the energy usage of 13 typical U.S. buildings and meteorological data. Thirteen building models used during BES in this study are the most significant number of archetypes we encountered during this review. Additionally, the proposed Mean Percentage of Misclassified Areas (MPMA) index introduces a new concept, which asserts that every climatic zone should have its unique climate conditions, resulting in a specific type of building having a unique energy performance. This way, the performance of similar buildings situated in different zones shouldn't overlap. The difficulties in CZB to accommodate different building stypes are also demonstrated by the results, as each typology has a unique sensitivity to climate.

2.2. Critique of traditional approaches

The critique of traditional approaches in CZB, such as DDM, KGM, IJM, etc, focuses on their limitations in accurately reflecting the complex interactions between climatic conditions and building energy needs [1, 4, 5].

The DDM approach exclusively uses outside AT, eliminating other environmental variables that affect a building's energy consumption. The base temperature, which signifies the threshold at which HVAC systems stop functioning to maintain comfort within a building, holds critical importance. The incorrect base temperature will lead to inaccurate DDs [125]. The KGM remains the most widely used climate classification map, which is still constantly updated and refined [31]. However, while vegetation-based climate classifications can indicate agricultural potential, they can't tell how humans would feel in different climates or characterize building energy efficiency [104]. In recent years, other comparisons have shown that KG did not allow the accumulation of precise information required to address the issue of building design and thermal comfort [103]. Frequently, particularly when compared to BES or MLM, IJM encounters criticism for its inconsistency with the much improved contemporary standards for the energy efficiency of buildings [17, 47]. There's no consistent link between IJM climate categorization and building energy in

its current forms. But still, it can be applied for CZB, with a narrow range of variables that correlate strongly with the overall energy consumption of the buildings, like AT and SR [85, 126].

There are now generally accepted, considerably more reliable procedures like MLM and BES that are founded to be more precise and transparent classification strategies and yield more robust outcomes. These advanced methods are appreciated for their ease of use and reliability, marking a shift towards performance-based climate zoning approaches that better align with the dynamic nature of climate impact on buildings.

2.3. Building energy performance in CZB

BES stands out as the most accurate method for predicting the thermal performance of buildings, offering insights into buildings' energy behaviour. Originating in the 1960s, BES has evolved to accurately assess a wide range of solutions, enabling detailed, predictive, and performance-based criteria for energy efficiency programs. In addition, the significance of using BES to validate CZB was proved by several publications [3, 4, 54]. However, limitations include the need for specific design hypotheses and detailed climate data, which may not always be available, making climate zoning via simulation challenging and error-prone. In recent years, there has been a noticeable transition from a climate-based methodology to a performance-based approach in utilizing building performance simulation for CZB establishment, indicating a growing trend despite the absence of a standardized framework for simulation application.

2.4. Energy-saving potential of proper CZB

According to several studies, accurate CZB is a pivotal factor in achieving substantial energy savings, as it deeply influences the tailored strategies for climate-specific architectural design, material utilization, and the deployment of HVAC systems suitable for distinct regional weather patterns. The amount of energy conserved can vary significantly, potentially reaching savings of over 50% in optimized conditions [127].

According to Chen et al. [128] adjusting climate zoning in China can lead to a 6.4% reduction in the annual cumulative building energy load, without the need for increased insulation costs. A study by Thornton et al. [127] demonstrated that incorporating advanced energy efficiency measures (climate-specific design) leads to a minimum of 50% reduction in energy consumption, with an average reduction of 56.6% compared to the ANSI/ASHRAE/IESNA Standard 90.1-2004 in different CZ across the United States. By implementing climate-specific solutions such as envelope retrofitting and energy system

replacements, it is possible to achieve primary energy savings of 42-44 kWh/m²a in certain climates for existing buildings [129]. This method also leads to substantial decreases in polluting emissions (11.5 ÷ 12.0 kgCO2-eq/m^2a) and cost savings (5 ÷ 78 €/m^2). Gillingham et al. [130] indicate that proper CZB can also lead to reductions in carbon dioxide and particulate matter emissions, improving outdoor air quality and human health.

However, incorrect CZB can have adverse effects, leading to suboptimal implementation of energy-saving design principles. This can reduce the potential for energy savings and negatively impact thermal comfort in buildings, underscoring the importance of accurate climate classification [17, 55].

2.5. Chapter Summary

CZB has emerged as a critical area of study within the realm of buildings, underlining the necessity for precise climate classification to optimize their design and operation. Twelve widely used methods were identified in CZB, which includes a wide variety of applicable methodologies. MLM, DDM, and BES stand out as the most common approaches. For CZB, a single technique is used in most situations (65%), two methods combined are used in 28% of cases, and at least three approaches are integrated at the same time in 7% of occurrences. Interestingly, MLM, DDM, BES, and BCM are the most often selected techniques when implementing a single technique. MLM offers distinct advantages for CZB, as it generates more reliable and previously unattainable outcomes. Its ability to analyze vast datasets and identify complex patterns allows for more accurate classification. Nonetheless, it is typical for MLM to be used in conjunction with other techniques, such as BES, to improve the accuracy of CZB outcomes. DDM, which is well-known for its established relationship with building energy, offers superior CZ for a range of uses. However, DDM is mostly dependent on external AT and may miss other important meteorological factors affecting a building's energy use. It should be noted that for more precise DDs calculation, the mean daily degree hours approach [131] is preferred when complete meteorological data, especially hourly ambient AT, is available [132].

The discussion on global CZB is framed around the use of two predominant climate zoning systems: the KG map and the ASHRAE Standard's degree day-based map. While the KG map is widely acknowledged for its comprehensive climate classification, it falls short in providing the ability to evaluate the energy performance of buildings. Conversely, the ASHRAE Standard, although being the sole global solution offering climate zone data pertinent to buildings, is critiqued for its reliance solely on DDs. ASHRAE Standard

overlooks other environmental variables that significantly impact a building's energy usage, resulting in an oversimplified zoning approach that may have negative consequences.

Emerging methodologies, particularly BES and MLM, have demonstrated considerable promise in refining the process of CZ. These advanced approaches advocate for a transition from a climate-centric to a performance-based zoning paradigm, highlighted by their simplicity and reliability in delineating CZ. Despite their potential, the application of BES and MLM remains constrained to specific geographic areas, with their integration into global standards being a focal point of academic discourse.

Climate zoning methods can be employed either individually or in combination. Combining approaches is more widespread in scientific publications than in regulations. The most popular combinations of the two methods are BES/MLM, DDM/BES, and DDM/MLM. Multiple methods are most utilized together to improve climate zoning or validate its results.

Moreover, the documented potential for significant energy savings through accurate CZB underscores the importance of adopting climate-specific strategies in architectural design and HVAC system implementation. Studies have illustrated that adjusting climate zoning could result in substantial reductions in energy consumption and emissions, thereby contributing to environmental sustainability and public health.

In conclusion, the academic narrative surrounding CZB is evolving, with a marked increase in research over the past decade signaling a growing commitment to advancing the field. The prospective integration of BES and MLM into official standards across various countries holds the promise of improving building energy efficiency on a global scale, aligning with broader efforts to address climate change. The findings of this chapter are summarized below:

- DDM focuses primarily on outside AT for climatic zoning, missing other key climatic factors affecting building energy use, with 18°C and 10°C as the most common base temperatures for HDD and CDD respectively.
- MLMs are notable in climatic classification for their capacity to process large data volumes and incorporate building characteristics, enhancing CZB accuracy. However, proper cluster number determination is crucial, often achieved through the Elbow method or Hubert index.
- BES is valued for its accuracy in thermal performance prediction and its role in climatic classification, despite its limitations such as the necessity for specific design hypotheses and detailed meteorological data.

- IJM has been a longstanding method since the mid-20th century, beneficial under low energy efficiency standards for protecting buildings from climatic effects. However, its relevance to building energy performance is questioned, with a lack of clarity in selecting climate variables and thresholds.
- KGM is deemed inadequate for assessing building energy performance due to its inability to gather precise data for CZB, with methods like unsupervised clustering (k-means) and BES providing superior classification accuracy.
- CSIM relies on building energy usage for CZB, offering a straightforward calculation of climatic severity indexes and climate zone classification using degree days and sun-hours data. Nonetheless, the method's regression coefficients calculation remains ambiguous, limiting its applicability beyond its initial region, Spain.
- Emerging techniques like EBS and FDV show promise for developing or refining CZs, leveraging real building energy consumption data to elucidate the impact of various factors on energy usage. Challenges include data availability, processing, and the consideration of building diversity in design and use.

Based on the literature review, the following explores potential answers to the identified research gaps in CZB for Kazakhstan:

- Despite a lack of direct scientific sources on the most common building archetypes and their energy consumption in Kazakhstan, this information can be obtained through national statistics [133] and by conducting energy simulations of representative building types.
- Determining the optimal number of climate zones (ONCZ) can be achieved using established methods like the Silhouette and Elbow techniques [134, 135], which offer visual guidance on cluster quality based on data distribution.
- Identifying the primary climate variables influencing energy consumption can be addressed through correlation analysis and PCA [8, 53]. Additionally, exploring the potential of ML [36, 136-138] to discover complex relationships within data can be valuable. It's worth noting that while AT is often the most commonly studied variable, and the combination of AT and RH is frequently examined, other factors can also play a significant role. Therefore, a comprehensive approach that considers various climate variables and explores their potential interactions through ML holds promise for a more nuanced understanding of energy consumption patterns.
- While evidence suggests that contemporary performance-based CZB methods outperform conventional climate-based classifications [1, 5, 17, 48], further research

is necessary, due to the limited number of works where these methods were compared relative to each other [3, 5].

- While clustering quality metrics like Silhouette Score (SS) are used for evaluation, a performance-based validation index specifically designed to connect CZB to actual building energy performance is needed. Ideally, this index should ensure that zones within the CZB correspond to similar building energy performance ranges, and address the limitations of the existing MPMA index [4].
- Notably, existing literature on CZB lacks investigation into the impact of spatial constraints [73, 139, 140]. This gap presents a valuable opportunity to incorporate spatial analysis techniques alongside traditional clustering methods, potentially leading to more accurate and consistent climate zoning in Kazakhstan.

Chapter 3: Methodology

This chapter outlines the approach used to accomplish the thesis objective, which is organized within the established structure. The path starts with an introductory part (3.1), offering a thorough step-by-step workflow, exploring two main phases of the research, followed by an examination of the study area in section 3.2. Section 3.3 provides a detailed weather data description. Section 3.4 covers the choice of building archetypes, which sets the stage for further discussion on building performance simulations (3.5) and its verification (3.6). Section 3.7 identifies key building performance indicators. Section 3.8 focuses on the careful and thorough identification of the most important climate variables with comprehensive correlation analysis (3.8.1), and ML methods (3.8.2), (3.8.3), (3.8.4). The focal point of Section 3.9 is the identification of the optimal number of CZs, which subsequently leads to the utilization of a multivariate clustering approach in Section 3.10. Section 3.11 covers assessment metrics and validation methodologies, specifically focusing on clustering quality analysis (3.11.1), climatic zoning validation with building performance (3.11.2), and validation using the Adjusted Rand Index for comparative analysis with official local CZB and ASHRAE map.

3.1. Framework

The methodology of the study is partitioned into coherent stages, each serving as a foundational component of the research narrative.

Weather data collection from 94 meteorological stations across Kazakhstan and identification of the most typical building archetypes based on the analysis of state statistics data have formed the data collection stage at the beginning of the research. In the simulation stage, the energy simulations are methodically conducted using digital models of predetermined building archetypes across various regions. The data collected from these simulations provide key performance indicators to link climate variables with building performance. Statistical and ML approaches are then used to identify the most significant climate variable in the following stage, setting up the dual-phased mapping process.

The decision to use a dual-phase approach arises from the need to thoroughly evaluate and compare conventional (climate-based) and contemporary (performance-based) methods for CZB. A split is motivated by the necessity to determine if these different techniques provide comparable outcomes with CZB, and consequently, to determine the most efficient way for CZB development. Before the start of clustering runs, the optimal number of climatic zones was determined for each phase and each dataset.

In phase 1, emphasis is placed on the traditional climate-based approach, wherein CZB relies purely on the key climate variables influencing building energy needs and using them for clustering. This conventional method has long been a cornerstone in CZB development.

Phase 2, in turn, introduces a paradigm shift towards a performance-based CZB approach, leveraging BES and building performance indicators. This modern methodology is motivated by the recognition that BES offers a nuanced understanding of building thermal performance, thereby enhancing the robustness of CZB. The climate classification in Phase 2 is purely based on building energy performance indicators. In both phases, the classification process is executed utilizing the same multivariate clustering methodology with and without spatial constraints.

Next, by using a novel CZMI and examining the overlap of CZs determined by KDE, the research aims to verify and discern the efficacy of both phases. The degree of KDE overlaps becomes a quantitative measure of the effectiveness of each methodology. This analysis also ultimately addresses the pivotal question: Is it more efficient to propose a CZB directly using building energy consumption data, eliminating reliance on climate variables, or do traditional methods still offer appropriate results? By delving into this comparative evaluation, the research aims to contribute insights that can inform future practices in CZB development. The culmination of these two phases is a comparative analysis and synthesis.

Based on the analysis and verification, a final climate zoning map is proposed based on the results of the best CZB method. The official building climate map of Kazakhstan and the ASHRAE map of the region will also be compared with the resulting new map, and possible misclassifications will be analyzed. The comparative framework is predicated upon both the CZMI and the Adjusted Rand Index (ARI), facilitating a nuanced definition of misclassification. The culmination of the research is illustrated by the generation of a final CZB map, which represents a synthesized and more accurate representation of CZs.

The step-by-step workflow is visually depicted in Figure 3.1. Subsequent sections will provide an in-depth exploration of each stage in the research, offering a comprehensive understanding of the methodology applied in this study.

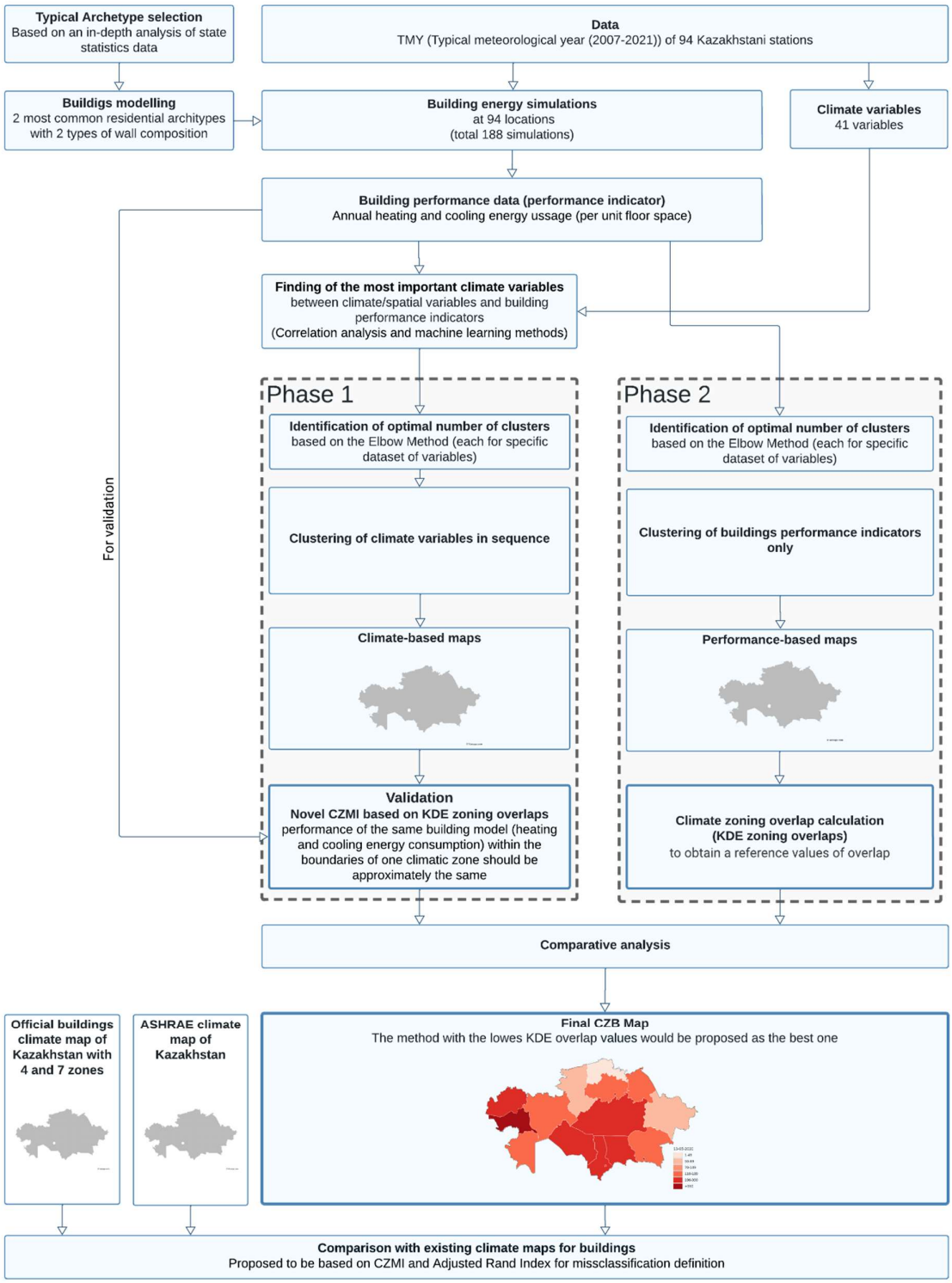


Figure 3.1: A novel performance-based framework for Kazakhstan's CZB.

3.2. Study area

Situated in Central Asia, Kazakhstan stands as the ninth-largest country globally in terms of land area, showcasing a diverse topography that encompasses expansive steppes, arid deserts, and rugged mountains. This varied geography contributes to the distinctive climate

experienced in the region. Kazakhstan witnesses considerable temperature fluctuations throughout the year, characterized by chilly winters in the northern regions and predominantly warm summers in the southern areas [141].

The annual average extraterrestrial radiation - a measure of solar energy received, varies inversely with latitude, underscoring the fundamental role of SR in shaping the climatic patterns of Kazakhstan. Since the country is landlocked in the center of the Eurasian continent there is no other geographical aspect influencing the climate. With a mean value of approximately 294 Wh/m², SR exhibits a strong correlation with both AT and DDs. This SR not only governs the ambient temperatures but also significantly influences other climatic factors such as RH and WS.

In the winter months of January and February, average AT in the northern parts can plummet to as low as -16°C, while in the southern regions, average AT generally stays above -7°C. Contrarily, the average temperature in July exhibits a range from 20°C in the northern and north-eastern regions to 29°C in the southern areas near the Uzbekistan border. The annual mean temperature in Kazakhstan is observed to be around 7.43°C, with the minimum and maximum average temperatures recorded at 2.26°C and 15.27°C, respectively. Figure 3.2 illustrates the map of the average annual DBT. The climatic nuances of Kazakhstan present notable challenges in the design and operation of buildings. To ensure consistent occupant comfort throughout the year, buildings must incorporate efficient heating and cooling systems and be adequately insulated, given the substantial temperature variations and wind velocities experienced in the region.

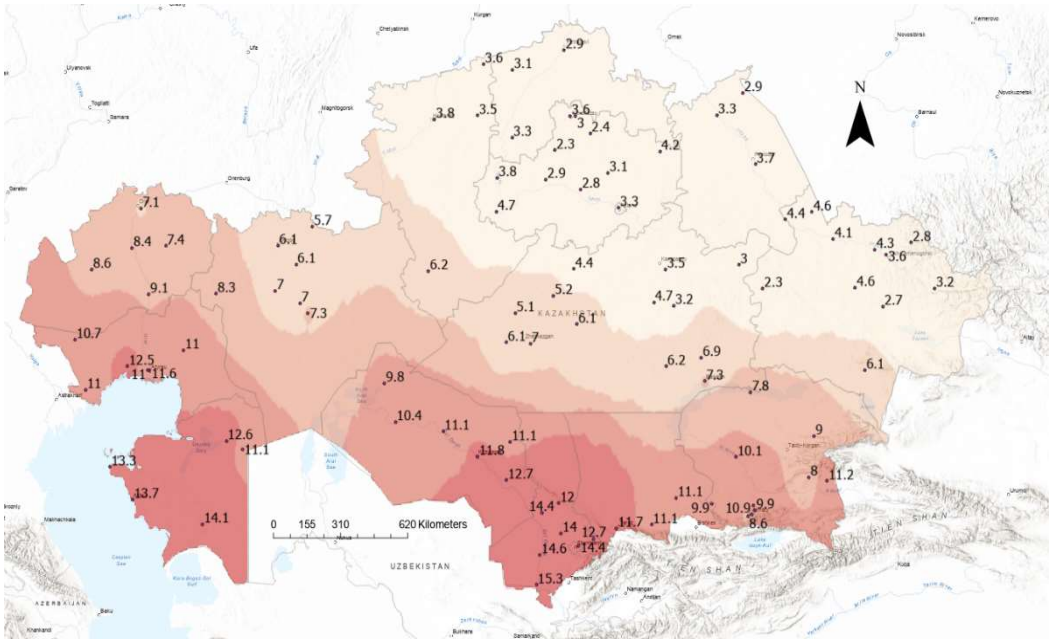


Figure 3.2: Kazakhstan's average annual dry bulb temperature.

Central to understanding Kazakhstan's climate is the concept of DDs – both HDD and CDD, which are pivotal in quantifying energy requirements for heating and cooling. The mean HDD stands at approximately 4546, indicative of the substantial heating requirement during colder periods, particularly in northern regions. Conversely, the CDD, representing the need for cooling, averages around 687, with a noteworthy peak at 1424, highlighting the warmer conditions experienced primarily in southern Kazakhstan. The analysis reveals a strong interplay between these metrics and the latitudinal positioning of various locations within Kazakhstan. The WS and RH across various regions exhibit distinct ranges, reflecting the diverse climatic conditions of the country. For WS, the data indicates a range from a minimum of around 1.16 m/s to a maximum of approximately 5.51 m/s. The observed range of RH in the dataset spans from a minimum of about 44% to a maximum of 71%. Figure 3.3 shows the distributions of climate variables.

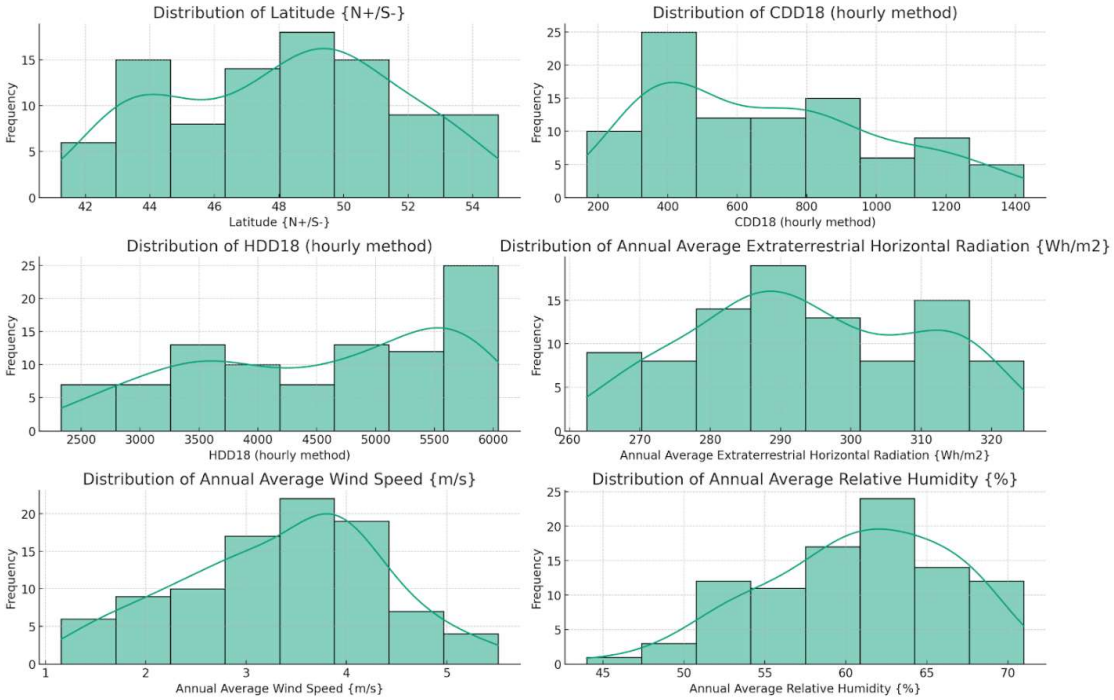


Figure 3.3: Kazakhstan's climatic variable distributions.

Kazakhstan's climate is predominantly heating-dominated. The HDD values are generally higher across the country, indicating a greater need for heating, especially in the northern regions. This is further supported by the average temperature range and the latitude of Kazakhstan, which align with a climate that experiences colder conditions for a significant part of the year. To ensure consistent occupant comfort throughout the year, buildings in Kazakhstan must incorporate efficient heating and cooling systems and be adequately

insulated, given the substantial temperature variations and wind velocities experienced in the region.

3.3. Weather data

For any climate-related research, weather files for the selected location are required. This research utilized weather data obtained from the typical meteorological year (TMYx) dataset sourced from https://climate.onebuilding.org/ [142]. The TMYx dataset was employed to supply necessary meteorological data for conducting building energy simulations using EnergyPlus and DesignBuilder. This allowed for the assessment of energy consumption patterns, analysis of correlations between climatic and spatial variables, and subsequent application of clustering techniques to develop a CZB. A dataset consisting of 94 meteorological stations with observations spanning from 2007 to 2021 was utilized.

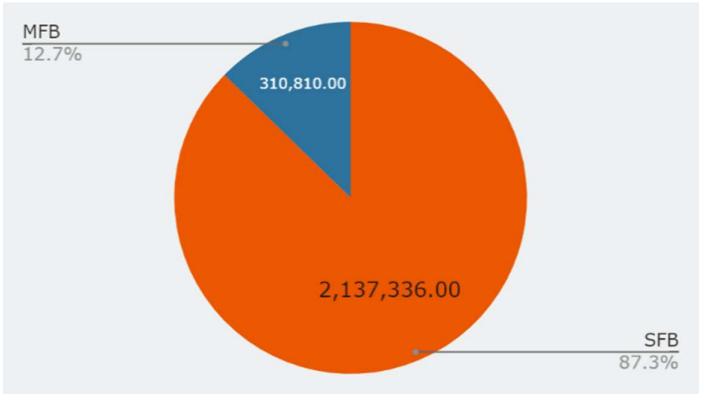
TMYx represents a recent addition to the TMY series of datasets (TMY, TMY2, and TMY3), which comprises standard meteorological information obtained from hourly weather data collected between 2007 and 2021, sourced from the Integrated Surface Database of the United States National Oceanic and Atmospheric Administration. The dataset was derived using the TMY/ISO 15927-4:2005 methodologies [143]. A significant enhancement involves the utilization of the ERA5 satellite-derived dataset for SR [142]. Historically, the utilization of the National Solar Radiation Database (NSRDB) necessitated the derivation of SR data through the reliance on other climatic parameters [144]. With ERA5, all TMYx files contain data that is satellite-derived, which is preferred over calculations that make assumptions based on climate variables [142, 145].

The TMY dataset is generated by aggregating and analyzing meteorological data from many years to produce a typical year that represents the prevailing climatic conditions of the area [146-148]. The data consists of climatic variables such as AT, RH, SR, WS, and PR, recorded on an hourly or sub-hourly basis [144]. It is calculated by averaging weather data from a specific time frame in the past. The most typical months are chosen through statistical analysis [147, 149, 150]. It is vital to recognize that TMYx datasets do not include extreme weather events or uncommon climatic phenomena [146].

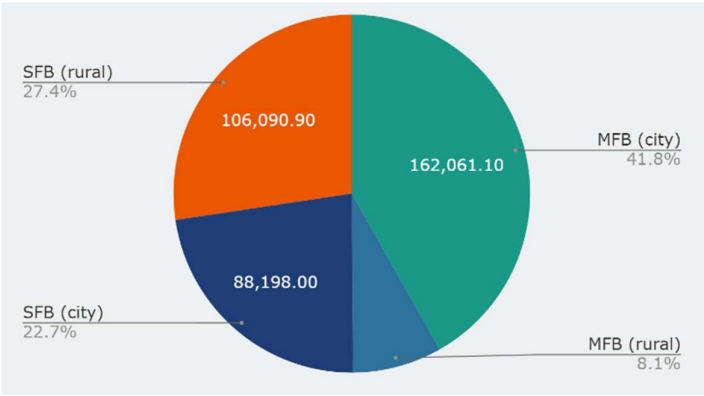
3.4. Building archetype selection

The process of selecting archetypes for simulations was based on the analysis and interpretation of statistical data on residential buildings obtained from the Bureau of National Statistics of Kazakhstan [133]. Governmental data categorize all residential buildings in the country into two primary groups: Single-family buildings (SFB) and Multi-

family buildings (MFB). Examination of the data revealed that the predominant or typical residential archetype in Kazakhstan is the SFB, as illustrated in Figure 3.4. SFBs in Kazakhstan exhibit varying sizes, ranging from 48 to 205 m^2, with a noteworthy majority (71.8%) falling within the range of two to four rooms and an area of 52 to 95 m^2, as depicted in Figure 3.5.



(a)



(b)

Figure 3.4: Kazakhstan's total number of single-family and multi-family buildings (SFB and MFB, respectively) (a). The percentage (in thousand m^3) of the entire MFB and SFB living area (b).

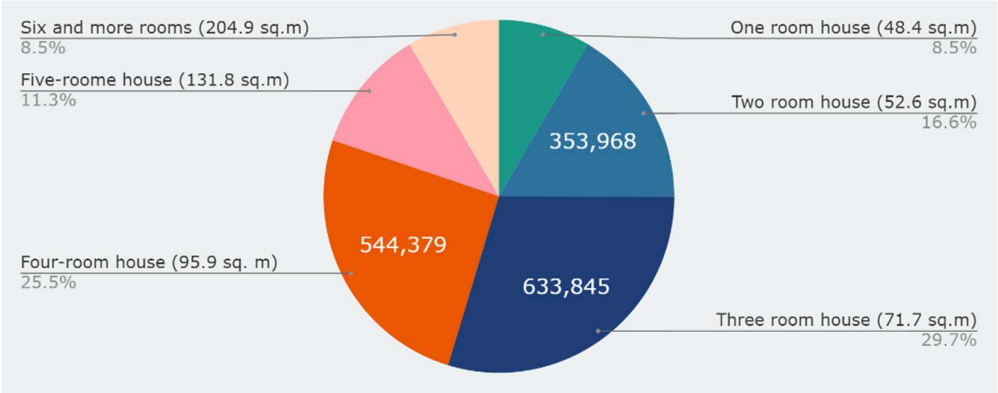
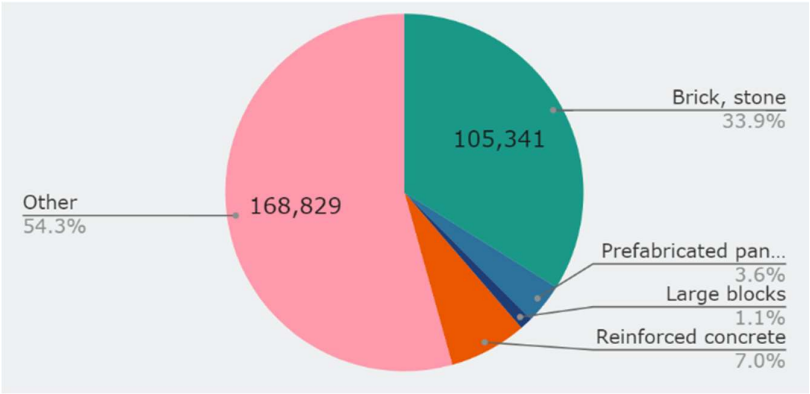
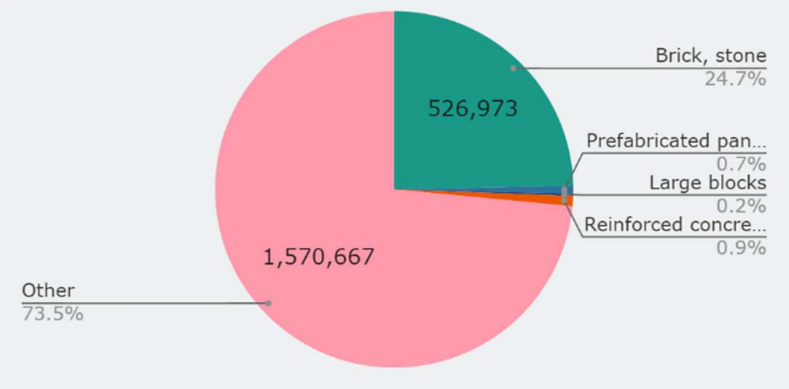


Figure 3.5: SFB distribution based on area and number of rooms (units).

Various construction materials, including brick (stone), concrete, and alternative options, are employed in the construction of walls. Specifically, 24.7% of SFBs feature exterior walls made of brick or stone, while this percentage is slightly higher at 33.5% for MFBs. Notably, the utilization of prefabricated panels, large-block, and reinforced concrete in residential buildings is relatively low, constituting 1.8% of the total number of SFBs and 11.7% of the total number of MFBs. Among MFBs, those with walls made of reinforced concrete comprise 7.0% of the total. Statistics indicate that a significant proportion, 73.5% of SFBs and 53.4% of MFBs, have external walls constructed with materials classified as "other materials." It is plausible that wood and adobe/mudbrick fall within the classification of "other materials," as depicted in Figure 3.6.



(a)



(b)

Figure 3.6. The quantity of MFB based on the materials used for the outer walls (units) (a). The quantity of SFB based on the materials used for the outer walls (units) (b).

In addressing the composition of walls and roofs, the standards set by the republic predominantly prioritize resistance to heat conduction, quantified by the R-value (m^2K/W). These standards refrain from specifying particular materials but mandate that the R-value must meet or surpass predefined thresholds. Given the extensive climatic diversity across the country, R-values exhibit considerable variation from the southern to northern regions.

Specifically, R-values for external walls range from 3.20 to 2.40 m²K/W, while for roofs, they vary from 4.00 to 2.40 m²K/W [19].

SFBs make up more than 71.8% of residential dwellings in Kazakhstan, with most of them having a living space between 52 to 95 m^2. This information was used as the basis for creating the base case digital model. The model consists of two zones: the major living space, covering 88 m^2, and the main entrance block (hallway), which occupies 6 m^2. The main living area is protected by an unoccupied pitched roof, as seen in Figure 3.7. Two archetypes, the northern archetype (NA) and the southern archetype (SA), were created to improve accuracy and address significant climate fluctuations in the research area. Every archetype has the same size and shape but includes unique insulating properties. The wall and roof compositions were calculated following local rules and regulations regarding the thermal resistance of residential buildings [19], as specified in Table 3.1.

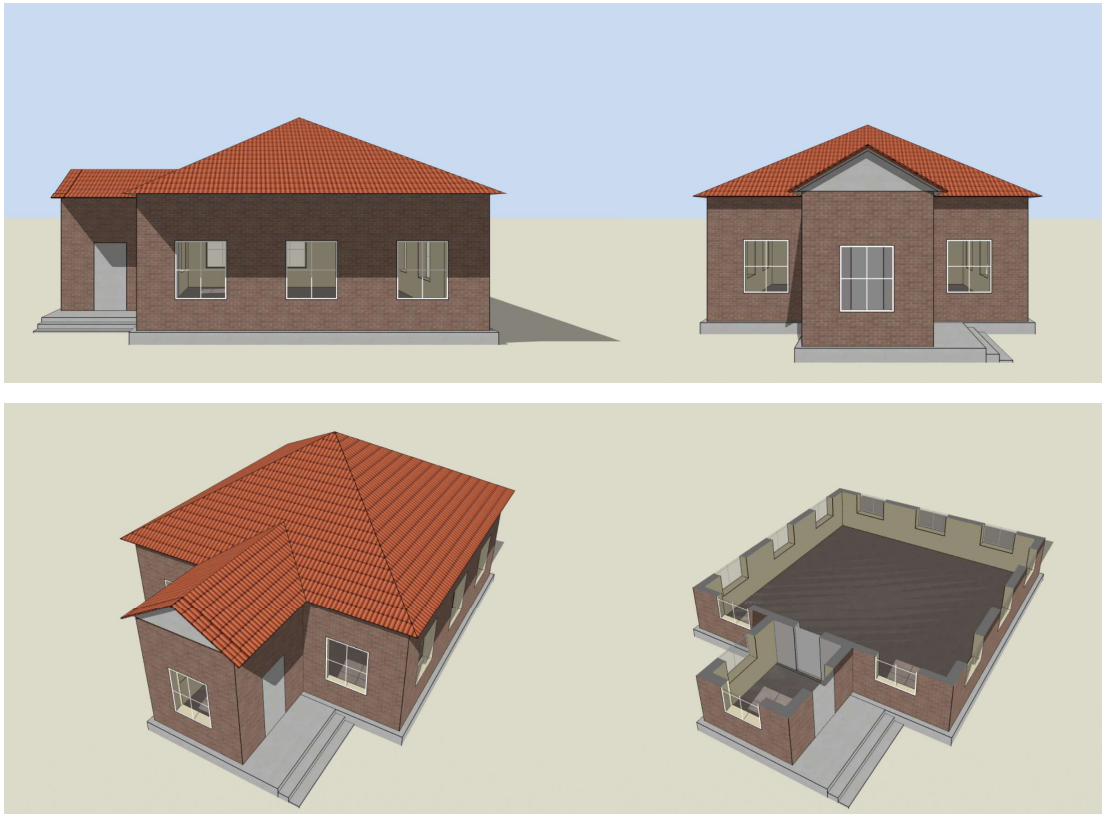


Figure 3.7: Base case building model.

Table 3.1: General information of the base case building model.

Aspects	Description	
Orientation	The building's entrance is oriented to the south	
Building` floors	1	
Heights	3 m/storey	
Dimensions	10 x 10 m	
Aspect ratio	1.00	
Floor areas	Total floor area 92.30 m^2	
Window-to-wall ratio	20 %	
External walls R-value	Northern archetype 3.20 (m^2K/W)	Southern archetype 2.40 (m^2K/W)
Semi-exposed ceilings R-value	Northern archetype 4.00 (m^2K/W)	Southern archetype 2.40 (m^2K/W)

By incorporating two archetypes, the study thoroughly examines diverse climate variations and demonstrates how the proposed CZB technique can engage several building types at the same time. Figure 3.8 delineates the insulating characteristics employed in the base case building model. The HVAC system of the base case building model utilizes a district heating system, while an electric-powered split air-conditioning system is used for cooling. Further information regarding HVAC can be found in Tables 3.2 and 3.3.

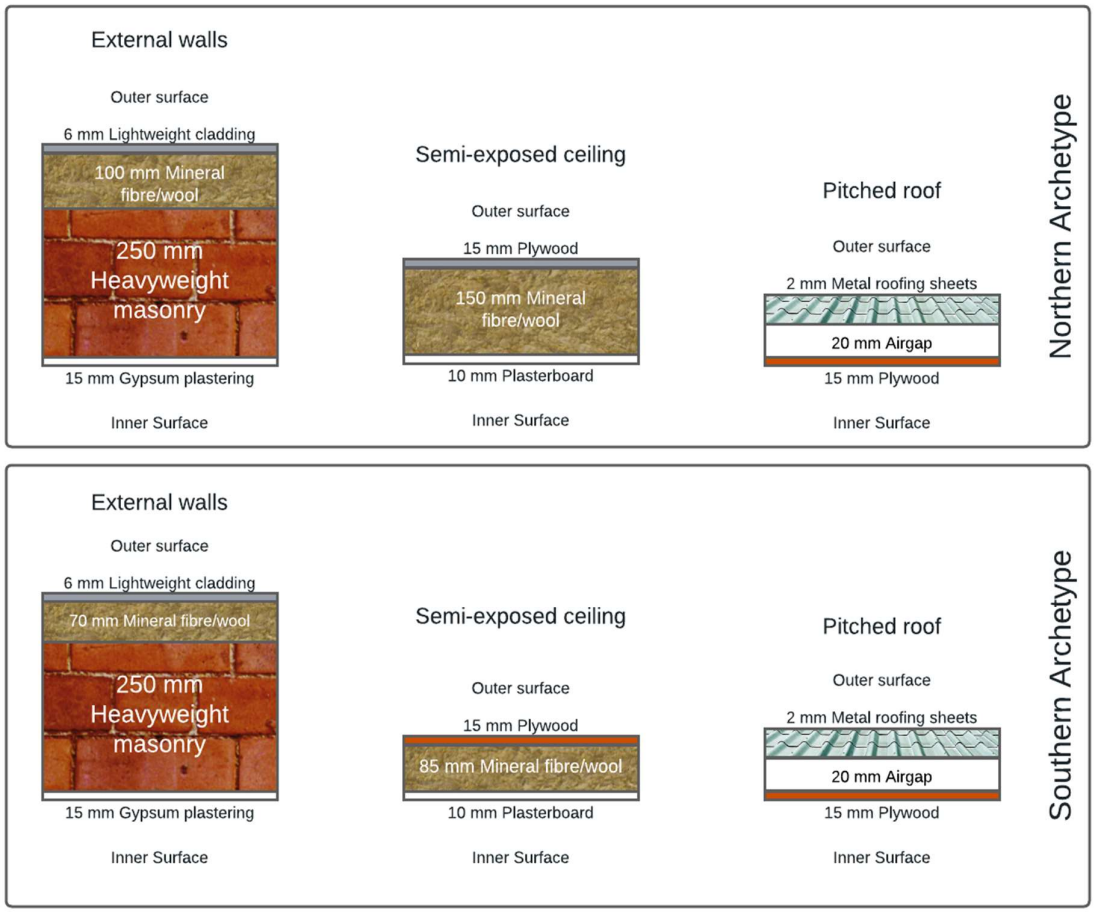


Figure 3.8: The insulation parameters used in the base case building model.

Table 3.2: Key details about the HVAC system of the base case building model.

	Heating	Cooling
System approach	Radiator heating, Boiler HW (district heating)	Mixed mode Nat Vent, Local comfort cooling (split air-conditioning system (electricity))
Setpoint temperatures	20 °C	26 °C
Setback temperatures	16 °C	32 °C

Table 3.3: The HVAC schedule for the base case building model.

Time of Day	Weekdays Summer Design Day	Winter Design Day	Weekends	All Other Days
00:00 - 06:00	0.50	1.00	1.00	0.70
06:00 - 18:00	1.00	1.00	1.00	0.70
18:00 - 24:00	0.70	1.00	1.00	0.70

The building's living area is planned to house up to four people, with each person allocated around 23 m^2 of living space, based on the national average [133]. Table 3.4 displays the occupancy schedule.

Table 3.4: Occupational schedule of the base case building model.

Time of Day	Weekdays	Saturday	Sunday/Holidays
00:00 - 05:00	1.00	1.00	1.00
05:00 - 06:00	0.80	0.80	0.80
06:00 - 08:00	0.75	0.75	0.75
08:00 - 09:00	0.50	0.50	0.50
09:00 - 15:00	0.43	0.43	0.43
15:00 - 18:00	0.50	0.50	0.50
18:00 - 21:00	0.75	0.75	0.75
21:00 - 24:00	0.80	0.80	0.80

The archetypes chosen for this study were selected based on their representativeness of typical residential building configurations, thus ensuring that the conclusions drawn are relevant to the most common building practices. However, the findings derived from the selected building archetypes provide a robust foundation for extrapolating climate zoning insights to a broader range of building types. Generalizing these findings to other building types involves considering the fundamental principles of thermal dynamics and energy consumption that are common across various buildings [151]. For instance, the impact of climate variables like temperature, humidity, and solar radiation on energy needs tends to follow similar patterns across different building types, even though the magnitude of impact

might vary [86, 152, 153]. This consistency allows for the application of learned insights to optimize CZB in a wider array of buildings, not strictly limited to the studied archetypes.

Additionally, the employed methodology offers a flexible framework that can be adapted to other buildings. By adjusting the model parameters to reflect the unique characteristics of different building types, such as varying occupancy patterns, construction materials, and architectural designs, the proposed CZB's applicability extends beyond the initial archetypes.

3.5. Building performance simulations

A total of ninety-four locations in Kazakhstan underwent energy simulations using two archetypes in the EnergyPlus 9.2 program. EnergyPlus is a widely used and recognized program among building energy researchers. It is a highly capable software tool that enables the modeling of various energy flows, including cooling, heating, ventilation, and lighting [154]. The building thermal zone calculation method employed by EnergyPlus utilizes a heat balance model that operates under certain assumptions. These assumptions include the one-dimensional nature of heat conduction across surfaces, the diffuse nature of surface irradiation, the constancy of temperature for each surface, and the uniform heating of air within the thermal zone [155]. EnergyPlus has been widely employed in multiple studies to evaluate the energy consumption of buildings and to create CZB [3, 4, 14, 17, 77, 124]. Sections 4.1 and 4.2 will provide further information on energy simulation findings, including precise energy consumption numbers and ranges.

3.6. Verification of EnergyPlus simulations

Verifying EnergyPlus results is crucial to ensure the accuracy of the complex simulations, which integrate various features such as building design, HVAC systems features, occupant behaviour, etc. This verification provides confidence that the simulation accurately reflects the interactions between these components, contributing to the reliability of extracted building performance data and further climate zoning classification.

For the verification of EnergyPlus simulation results, an approach utilizing the ASHRAE residential cooling and heating load calculations technique with the Residential Load Factor (RLF) method [156] was implemented, albeit with certain simplifications. Overall the RLF method offers an effective and quick approach to load calculations. This method, which can be easily implemented, serves two primary applications: education and training, where its transparency and simplicity make it ideal for approximate load estimates. In this study, the RFL calculations involved a comprehensive breakdown of the Heating

Load and Cooling Load into distinct components, including Envelope, Infiltration/Ventilation, Internal Gain, and Distribution Loss.

Following the calculation process, a rigorous comparison was conducted between the simulated results generated by EnergyPlus and those obtained through the ASHRAE technique. To quantify the accuracy of the simulations, a suite of evaluation metrics was employed, including Root Mean Square Error (RMSE) [157, 158], Mean Absolute Error (MAE) [157], Mean Squared Error (MSE) [159], and Mean Absolute Percentage Error (MAPE) [157]. Table 3.5 provides metrics used for comparison between the EnergyPlus and ASHRAE energy needs results with their detailed description. These metrics provided a nuanced understanding of the disparities between the simulated and reference data, allowing for thorough assessment and validation of the EnergyPlus simulations.

Table 3.5: Verification metrics

Metric	Formula	Acceptable range	Description
Coefficient of Variation of the Root Mean Square Error	$CV-RMSE = 100\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{n}(xi-yi)^2}}{x^-}$	±15%	The coefficient of Variation Root Mean Squared Error (CV RMSE) is a standardized measure of prediction error expressed as a percentage of the mean of the main dataset values, providing a comparative assessment of the predictive performance relative to the scale of the data
Mean Absolute Error	$MAE = \frac{1}{N}\sum_{i=1}^{n}\lvert xi-yi\rvert$	$(0,+\infty)$	Mean Absolute Error (MAE) measures the average absolute differences between the validation values and the main dataset values. It is calculated by taking the mean of the absolute differences between each pair of corresponding values
Mean Squared Error	$MSE = \frac{1}{N}\sum_{i=1}^{n}(xi-yi)^2$	$(0,+\infty)$	Mean Squared Error (MSE) measures the average of the squared differences between the validation values and the main dataset values. It is similar to RMSE but does not take the square root of the mean
Mean Absolute Percentage Error	$MAPE = 100\frac{1}{N}\sum_{i=1}^{n}\left\lvert\frac{xi-yi}{xi}\right\rvert$	$(0\%,+\infty)$	Mean Absolute Percentage Error (MAPE) calculates the average absolute percentage difference. It is calculated by taking the mean of the absolute percentage differences between each pair of corresponding values

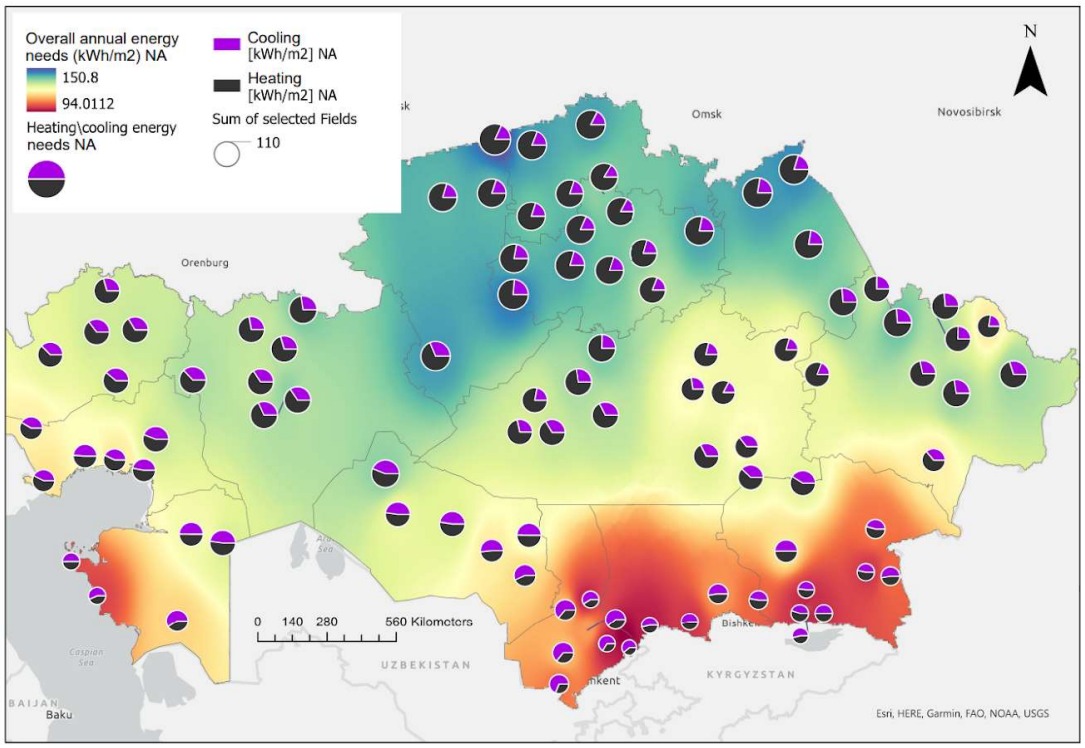
3.7. Building performance indicators

A diverse array of building performance indicators can be applied in the context of CZB, with the selection often contingent on specific CZB objectives, though overall annual energy needs per unit area are commonly utilized [1, 3, 4, 56]. It is crucial to underscore the lack of consistency in indicator selection, with climate and CZB goals exerting a significant influence on their relevance. The factors shaping the choice of performance indicators in this study include the distinct cold and warm seasons experienced in Kazakhstan throughout the year, necessitating substantial heating and cooling requirements that impact the overall annual energy consumption of buildings.

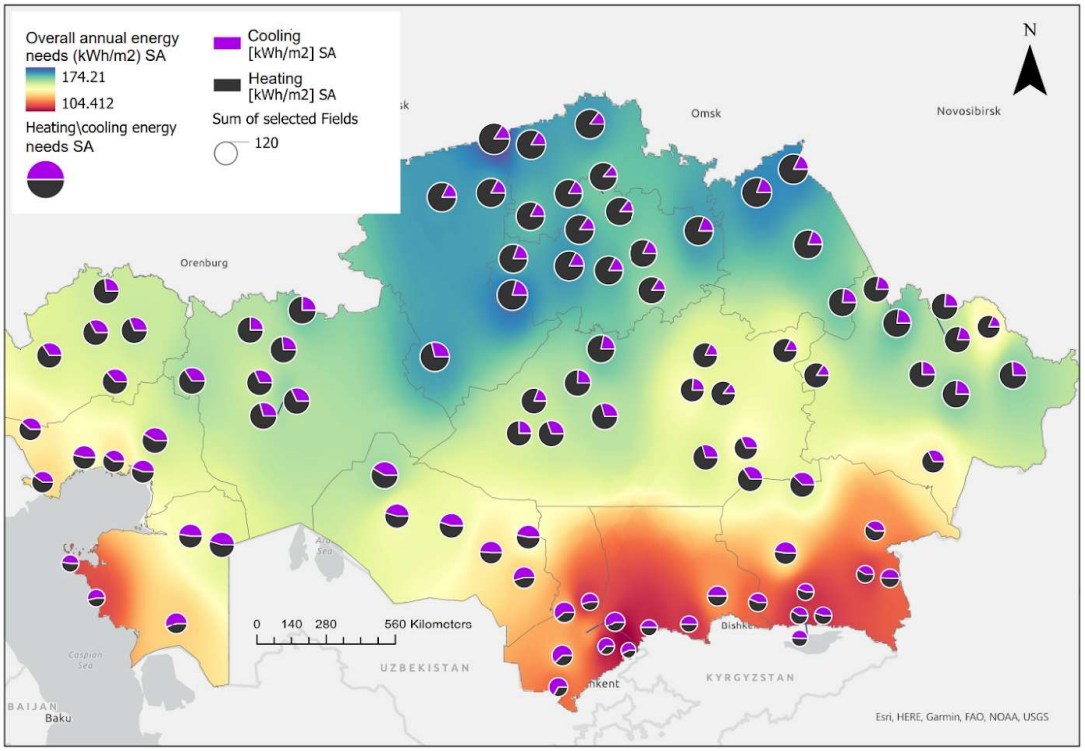
In this study, the term "energy consumption" or "energy needs" refers solely to "delivered energy." Due to temperature differences, SFB's yearly energy consumption in the south (where air conditioning is the primary energy source) and in the west and central parts of Kazakhstan (where heating is the primary energy need) may be comparable. The performance map of Kazakhstan's buildings, illustrated in Figure 3.9, shows the color-coded gradient pattern of energy needs. Whereas red areas, which represent cooling-dominated regions, indicate lower energy demands, blue areas, representing heating-dominated regions, indicate greater overall yearly energy needs. This study focuses on analyzing energy usage by dividing it into two main components. On the maps in Figure 3.9, the purple sections of the pie charts indicate the amount of yearly cooling energy needs, while the black sections represent the annual heating needs for NA (a), and SA (b). The main building performance indicators used for analysis in this research are yearly space heating (kWh/m^2) and annual space cooling (kWh/m^2) needs. Nevertheless, terminology like heating and cooling will also be used throughout the text for ease. Furthermore, total yearly energy requirements (kWh/m^2) data were gathered.

By specifically focusing on the heating and cooling combination, the clustering method becomes more aligned with the particular energy consumption of each climate zone. The study demonstrates also that the suggested method can simultaneously incorporate multiple variables for climate classification.

The simulation conditions imply that district heating provides energy for heating, while electricity from the grid is used for cooling purposes.



(a)



(b)

Figure 3.9: Heating and cooling energy needs of the SFB. NA (a). SA (b).

3.8. Selection of the most important variables for classification

Climate variables are crucial factors in the formation of CZB since they impact the thermal performance of buildings and regional energy consumption patterns. Determining the most relevant climate-related variables is a crucial stage in developing a CZB that is both reliable and precise [1, 5, 53, 72]. This research aims to gain a thorough understanding of how climate affects building energy consumption in specific Kazakhstani conditions. To achieve this, the study uses an enhanced approach that includes traditional correlation analysis supported by advanced ML techniques such as Random Forest Regression (RFR), Gradient Boosting (GB), and Extreme Gradient Boosting (XGBoost) to capture both linear and non-linear dependencies within the dataset.

The integration of advanced ML techniques with classical correlation analysis for variable selection offers a sophisticated and thorough approach. By combining the traditional statistical understanding of correlation with the nuanced predictive capabilities of ML methods, this approach opens a more efficient identification of the key variables that influence the phenomena under study. The goal is to establish a strong basis for the latter stages of CZB development and improvement. The main advantages of using ML techniques in combination with correlation analysis are:

- ML algorithms naturally capture the relationships between variables. They can recognize scenarios in which the collective impact of multiple variables surpasses their individual impacts. Correlation analysis may fail to consider such interactions;
- ML approaches are frequently more resilient to outliers in comparison to correlation analysis. Outliers have a substantial influence on correlation coefficients, which might result in misleading outcomes. ML, especially tree-based models, exhibit reduced sensitivity towards individual outliers;
- ML is typically more resilient to multicollinearity, which refers to a strong correlation between independent variables, as compared to traditional regression models. Even in cases when variables are linked, they are nevertheless capable of producing precise outcomes, hence preventing problems such as excessive standard errors.

The choice of this set of ensemble tree-based methods (RFR, GB, XGBoost) was guided by several key considerations aligned with the project's objectives:

1. Tree-based methods inherently provide clear metrics on feature importance, which are derived from the structure of the trees themselves. Variables that more frequently split nodes at higher levels of the tree tend to be more significant in predicting the target variable, making these methods highly effective for feature selection. This

interpretability is crucial for identifying key climate variables that impact building energy consumption;

2. Tree-based methods are powerful for modeling complex and nonlinear relationships between variables. These techniques incrementally build on weak predictive models to create a final, highly accurate prediction model. Their ability to learn from previous errors makes them particularly effective in refining predictions across iterative training rounds;
3. Selected methods, by combining multiple trees, reduce the variance and avoid overfitting, which is a common challenge with single decision trees. These ensemble approaches provide more stable and reliable results, essential for making consistent feature selection across various datasets;
4. While techniques such as neural networks or support vector machines also offer powerful predictive capabilities, they do not inherently provide straightforward mechanisms for ranking feature importance without additional processing. Furthermore, they require extensive data preprocessing and parameter tuning, which can be resource-intensive and less transparent in terms of model decisions.

The next section examines each approach utilized for the identification of the most significant variables.

3.8.1. **Correlation Analysis**

In the initial phase, correlation analysis was employed as a valuable statistical tool for examining the interrelationships between variables to ascertain the impact of meteorological and geographic variables on building energy usage. This classical statistical approach provides insights into the strength and direction of linear associations, shedding light on potential dependencies that may guide CZB development. While climate factors and their effects on building performance are one of the main parts of this research, investigating their relationship was crucial for a comprehensive understanding within the specific context of Kazakhstan.

Given the consensus about the impact of climatic factors, specifically AT, on building energy use, it was necessary to investigate the variations of this impact across various building archetypes. Besides climate variables, the dataset was augmented with DDs as they are important and often used parameters in the development of climate classifications. The DD values were computed using the hourly and ASHRAE techniques. The analysis also facilitated an approximate examination of how the official and ASHRAE CZ of Kazakhstan correlate with the building's energy use, incorporating information on these CZs into the

dataset. This approach aimed to provide a more comprehensive understanding of the intricate relationship between climatic and spatial factors and their impact on the energy performance of the archetype.

3.8.2. **Random forest regression (RFR)**

RFR is an ensemble learning approach that utilizes the combined predictions of numerous decision trees to perform regression problems [160, 161]. This process involves training each decision tree on separate bootstrapped samples of the dataset, which introduces diversity. The technique increases diversity by including random subsets of features at each tree split. Random forests employ the Mean Decrease Impurity (MDI) method [160] to assess the significance of variables in regression or classification tasks.

Set $\mathbf{X} = (X^{(1)}, \ldots, X^{(p)})$. The MDI of the variable $X^{(j)}$ in a forest formed by the combination of M trees is determined by:

$$\widehat{MDI}\left(X^{(j)}\right) = \frac{1}{M}\sum_{\ell=1}^{M}\sum_{\substack{t \in \mathcal{T}_\ell \\ j_{n,t}^{*}=j}} p_{n,t}\, L_{reg,n}\left(j_{n,t}^{*}, z_{n,t}^{*}\right) \tag{3.1}$$

where

$p_{n,t}$ is the fraction of observations falling in node *t*,

$\{\mathcal{T}_\ell\}_{1\le\ell\le M}$ is the collection of trees in the forest,

$\left(j_{n,t}^{*}, z_{n,t}^{*}\right)$ is the split that maximizes the empirical criterion in node *t*.

Therefore, the MDI calculates the weighted reduction in impurity that results from dividing the data based on the variable and then takes the average of this value over all trees [162, 163].

RFR exhibits robustness against overfitting, owing to its ensemble nature and the incorporation of random feature selection [136, 164]. Its ability to effectively handle non-linear relationships in data and scalability to large datasets with high-dimensional features make it suitable for various applications. Additionally, the model provides an inherent measure of feature importance, aiding in the identification of key variables [164]. However, while random forests are extensively utilized in many applications, the influence of tree depth on the statistical effectiveness of the method remains uncertain [160]. Also, the computational complexity can increase with a higher number of trees, and there is a risk of assigning importance to noisy or irrelevant features in the dataset.

3.8.3. **Gradient Boosting (GB)**

GB is an ensemble technique that builds a sequence of decision trees, with each tree aiming to correct the errors of its predecessor. The main concept underlying this technique is to create new base learners that are highly linked with the negative gradient of the loss function, which is connected with the entire ensemble [165, 166].

Loss functions in GB can be categorized based on the nature of the response variable y. Boosting algorithms have been developed for several types of response tasks, including regression, classification, and time-to-event analysis [165]. The squared-error L_2 loss is a widely used loss function in GB practices. For the L_2 loss function, the derivative is the difference between the observed and predicted values as shown in Equation (2). This means that the GB method effectively does residual refitting. The rationale behind this loss function is to impose a penalty on significant deviations from the desired outputs while disregarding minor discrepancies [166].

$$\Psi(y, f)_{L_2} = \frac{1}{2}(y - f)^2 \tag{3.2}$$

where

y is the observed value;

f is the predicted value.

A wide range of base-learner algorithms have been introduced in the existing literature. The often employed base-learner models may be categorized into three distinct groups: linear models, smooth models, and decision trees. The specific structure of the particular GB method, along with its associated equations, is greatly influenced by the design decisions made for loss function $\Psi(y, f)$ and base-learner model $h(x, \theta)$. In summary, the whole structure of the GB method, as initially suggested by Friedman [167] can be expressed as follows:

Inputs:

- input data $(x, y)_{i=1}^{N}$
- number of iterations M
- choice of the loss function $\Psi(y, f)$
- choice of the base-learner model $h(x, \theta)$

Algorithm:

1: initialize $\widehat{f_0}$ with a constant
2: **for** $t = 1$ to M **do**
3: compute the negative gradient $g_t(x)$
4: fit a new base-learner function $h(x, \theta_t)$
5: find the best gradient descent step size ρ_t:

$$\rho_t = \arg min_p \sum_{i=1}^{N} \Psi\left[y_i, \hat{f}_{t-1}(x_i) + ph(x_i, \theta_t)\right]$$

6: update the function estimate:

$$\widehat{f_t} \leftarrow \hat{f}_{t-1} + p_t h(x, \theta_t)$$

7: **end for**

where
$\widehat{f_0}$ is the initial guess.

GB excels in predictive accuracy and is adept at capturing complex relationships and interactions within the data. The model's ability to provide a quantitative measure of feature importance enhances interpretability. Furthermore, GB exhibits robustness to outliers in the dataset. Conversely, GB is prone to overfitting, especially without meticulous hyperparameter tuning. The computational complexity of the algorithm may pose challenges for large datasets, and successful implementation often requires careful tuning of hyperparameters.

3.8.4. **Extreme Gradient Boosting (XGBoost)**

XGBoost is an optimized and scalable version of GB. Developed by Chen et. al. [168] this method, is currently employed as the definitive categorization and learning system. Similar to GB, XGBoost calculates the contribution of each variable to the reduction in the chosen loss function. However, XGBoost differs from random forest in its approach to adding predictors. It utilizes a parallel tree learning technique, enabling it to execute in parallel and achieve optimal performance [168, 169]. While random forest adds several predictors simultaneously, XGBoost adds models sequentially. Each new model ($\mathcal{L}^{<t>}$ at iteration t) is developed by specifically targeting the mistakes made by the preceding predictors [170], to reduce the loss function as given in Equation (3.3).

$$\mathcal{L}^{<t>} = \sum_{i=1}^{N} \Psi\left(y_i, \left(\widehat{f_i^{t-1}} + h_t(X_i)\right)\right) + \Omega f_t \quad (3.3)$$

where

i denotes the $i-th$ sample to be predicted, N is the total number of samples;

t denotes the $t-th$ iteration; $\Psi(y_i, \widehat{f_i})$ is the loss function between the true label y_i and the predicted label $\widehat{f_i}$;

$h(X_i)$ is the base learner added at the $t-th$ iteration, X_i denotes the features for the $i-th$ sample;

Ωf_t is the regularization term to avoid over-fitting;

$\mathcal{L}^{<t>}$ denotes the objective function at the t-th iteration.

The selection of base learners is another crucial aspect of the ensemble learning process. XGBoost utilizes the second-order Taylor Expansion to estimate the value of loss functions. It leverages both the first-order derivative and the second-order derivative to aid in the selection of the base learner [168].

XGBoost stands out for its computational efficiency, being optimized for performance and speed. The incorporation of regularisation terms aids in preventing overfitting, and the model can handle missing data seamlessly [137]. Similar to RF, XGBoost provides a clear measure of feature importance. Nonetheless, XGBoost is sensitive to hyperparameter settings, and its performance is contingent on appropriate tuning [170].

For each target variable, feature importance was extracted from the trained models, and the top 10 influential features were identified. The importance of each feature was quantified by the algorithms and subsequently visualized through horizontal bar charts. Additionally, cumulative bar charts were computed based on a summation of feature importance metrics to assess the collective impact of features across all target variables. These visualizations not only provided a nuanced understanding of the significance of individual features but also highlighted commonalities and distinctions in feature importance across the ensemble methods.

3.9. An optimal number of climate zones

Finding the ONCZ is essential to the clustering process since it has a significant impact on how the data should be interpreted [171, 172]. It might be challenging to extract useful insights from the data if the incorrect ONCZ is used since this can produce excessively

simple or complicated clustering findings [1]. The literature has provided several methods for assessing the ONCZ [134, 135]. The Elbow approach was selected to determine the ONCZ in this study as the most popular and reliable option [8, 10, 51, 135].

The process involves assessing the clustering performance at each stage by continuing to repeatedly increase the number of clusters from a starting value of K=2 to a maximum value that is defined [135]. For every K, the Within Cluster Sum of Squares (WCSS) is computed as the distances of each data point in all clusters to their respective centroids. Next, the ONCZ is obtained using an Elbow graph in which the WCSS is shown for each K value. The WCSS rapidly drops to the called WCSS peak value before achieving ONCZ, and after surpassing ONCZ it continues to climb with the called WCSS peak value almost unaltered, which sets the ideal K apart.

$$WCSS = \sum_{C_k}^{C_n} (\sum_{d_i in C_i}^{d_m} distance\ (d_i, C_k)^2) \tag{3.4}$$

where

C is the cluster centroids and *d* is the data point in each cluster.

3.10. Multivariate Clustering Methodology

Complex data computations and operations are now faster and more efficient because of improvements in computer power and software capabilities. As a result, sophisticated clustering techniques have grown in popularity and accessibility for a variety of applications, such as CZB. CA provides a wide range of algorithms that effectively segment and classify multivariate data [102, 173]. K-means clustering (KC) and Hierarchical clustering (HC) are the two algorithms that were utilized in this research.

3.10.1. **K-means clustering**

Based on the similarity of the observations, KC classifies the data into a predefined number of clusters or, in our case, CZs. Data points are allocated to clusters using an iterative procedure, and cluster centroids are updated until convergence is reached [174, 175]. When using KC, the squared Euclidean distance between the points of data and cluster centroids is commonly used as the similarity metric. This distance metric is used to allocate the data point to the cluster with the closest centroid. The squared error function, or sum of squared distances between each data point and its allocated cluster centroid, is what the k-mean method aims to reduce. It is defined as follows:

$$J(V) = \sum_{i=1}^{C} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2 \tag{3.5}$$

where

$\|x_i - v_j\|$ is the Euclidean distance between x_i and v_j;

c_i is the number of data points in an i_{th} cluster;

c is the number of cluster centers.

3.10.2. **Hierarchical clustering**

HC establishes a hierarchical arrangement of clusters by continuously combining or dividing them according to their similarity or dissimilarity [17, 176, 177]. During each iteration of agglomerative HC, the two groups with the smallest distance between them are merged. The concept of "shortest distance" distinguishes between several agglomerative clustering approaches such as single, complete, average, weighted, centroid, median, and ward. This study examined seven hierarchical clustering linkage approaches using the cophenetic correlation coefficient [178]. The cophenetic correlation of a cluster dendrogram is the linear correlation coefficient between the cophenetic distances from the dendrogram and the original distances used to create the dendrogram [179]. Complete-linkage clustering for example involves linking two clusters based on the maximum distance between any pair of components, one from each cluster. The shortest link remaining at each stage leads to the fusing of the two clusters with the components involved. The complete linkage function may be mathematically defined by the following expression:

$$D(Z, Y) = d(z, y) \tag{3.6}$$

where

d(z,y) is the distance between elements $z \in Z$, $y \in Y$;

Z and Y are two sets of elements (clusters).

The complete linkage distance update formula for $d(Z \cup Y, V)$ can be represented as:

$$max(d(Z,V), d(Y,V)) \quad (3.7)$$

where

d(Z, V) is the distance between clusters;

Z, Y, and V are clusters.

To correctly capture the complex linkages and interactions that drive CZB, a multidimensional approach is necessary. Selected clustering algorithms can handle multidimensional datasets with multiple variables, enabling in-depth analysis of climatic and building-related aspects [65]. However, every clustering method has its limitations.

Depending on the initial settings, KC delivers different results depending on how sensitive it is to the initial values of the cluster centers. Furthermore, the spherical or circular structure of clusters implied by KC may limit its application to datasets with more complex cluster topologies [174]. KC clustering is not ideal for dealing with noisy or outlier data points, as they can greatly influence the clustering results [175]. Nevertheless, HC also possesses various drawbacks. Because the method has to compute the distances between every pair of data points, it can be computationally costly for large datasets [176]. HC is prone to noise and outliers, potentially resulting in erroneous grouping outcomes. Determining the optimal number of clusters might be challenging due to the dendrogram's inability to divide groupings at a certain level. Although they have limitations, KC and HC approaches are well recognized and often utilized in CZB studies [7, 10, 51, 53, 180].

3.10.3. **Spatial constraint in cluster analysis**

Integrating geographical elements into a dataset aims to analyze how spatial phenomena affect the results and quality of CZB. While the investigation of climate classification is location-specific [140, 181], none of the CZB publications explore or utilize the concepts of spatial analysis and its influence on the results. Spatial analysis is now commonly employed in many fields of study, including sociology, ecology, tourism, and epidemiology [182-184]. The First Law of Geography, which posits that "everything is related to everything else, but closer things are more related than distant things", serves as the foundation for the essential ideas of spatial autocorrelation and geographical dependency [185]. Spatial objects and phenomena should be examined largely based on their specific locations and connections. This remark emphasizes the importance of space in CZB, enhancing the understanding of spatial concepts and patterns. Recognizing the spatial factor in CZB is expected to generate unique and more reliable results, therefore preventing incorrect assumptions [139].

3.11. Evaluation Metrics and Validation

This section of the thesis is dedicated to outlining the methodologies employed for evaluating and validating the CZB developed. Section 3.11.1 introduces specific clustering quality metrics (uniqueness, compactness, Silhouette Score) which are essential for assessing the effectiveness of the climate zoning clustering. Section 3.11.2 shifts the focus to CZB validation with building performance, presenting methodologies for correlating the defined CZs with actual building energy consumption and performance metrics using CZMI. Section 3.11.3 discusses the validation using the ARI, a statistical measure used to evaluate the similarity between two clustering or the agreement between the proposed CZs and official classifications.

3.11.1. **Clustering quality metrics**

Clustering quality assessment plays a crucial role in evaluating the effectiveness of clustering algorithms and determining the meaningfulness of their outputs [186]. Given the absence of clear standards in clustering, objective quality assessments become essential for validating the results. These assessments provide insights into the reliability and appropriateness of the clusters formed by the algorithm. By employing various cluster validity metrics or internal and external indices, practitioners can quantitatively evaluate the cohesion and separation of clusters, aiding in the selection of optimal clustering solutions [187]. This process ensures that the clustering output aligns with the underlying structure of the data, contributing to more informed decision-making in various fields, including data analysis, pattern recognition, and machine learning.

To assess the efficacy and performance of the investigated clustering outcomes, dispersion, uniqueness, and the Silhouette Score (SS) indicators were computed [3, 171, 188, 189]. The same three indicators were also applied to methods of the local official climate map, and ASHRAE climate map to assess the quality of their climate classification.

At first, based on a few literature sources [3, 51] the selection of the most effective clustering method was guided by two primary metrics: the uniqueness of clusters and the dispersion of clusters. These metrics were chosen to ensure that the clusters formed were both distinct from each other and internally cohesive. However, as the research progressed, it became evident that the SS could effectively encapsulate the qualities assessed by both uniqueness and dispersion metrics. More on that in subsection 3.11.1.3. “The Silhouette Score”.

3.11.1.1. **Uniqueness of clusters**

Uniqueness or separation is a crucial factor in evaluating the quality of clustering [177, 188]. The metric calculates the proportion of data points from other clusters that exhibit building performance indicators beyond the range found in the cluster being studied. Higher values are preferable for achieving uniqueness. The clustering method has effectively identified separate groups with low overlap, leading to a more significant and dependable clustering result [3]. The uniqueness indicator is presented as follows:

$$Unq = \left(\frac{N}{T}\right) * 100 \tag{3.8}$$

where

N is the quantity of data points in different clusters, except the one now under evaluation;

T is the total number of data points in the dataset.

3.11.1.2. **Dispersion of clusters**

The dispersion or compactness is determined by calculating the MAE for each building performance indicator in every cluster [3, 188, 189]. With this quality metric, it can be determined if a certain cluster showed considerable dispersion, suggesting the necessity for more subdivisions. The formula for determining the dispersion of a building performance indicator within a cluster may be expressed as follows:

$$Dsp = MAE = \frac{\sum(|k - \mu|)}{n} \tag{3.9}$$

where:

k is the building performance indicator's value inside a certain cluster;

μ symbolizes the building performance indicator's mean value within that cluster;

n indicates how many observations or data points are in the cluster.

3.11.1.3. **The Silhouette Score**

SS is a well-regarded metric for evaluating the quality of clusters in a dataset [134]. It measures the degree of confidence in the clustering assignments by assessing how similar an object is to its own cluster compared to other clusters, thus encapsulating both cohesion within clusters and separation between them. The SS for each sample is calculated as:

$$SS_i = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3.10}$$

where for each sample i: $a(i)$ is the average distance from the i^{th} sample to the other points in the same cluster; and $b(i)$ is the lowest average distance from the i^{th} sample to points in a different cluster, minimized over all clusters.

In the SS calculation, the distances referred to are typically Euclidean distances, although other distance metrics can be used depending on the nature of the data and the specific requirements of the analysis. SS for a set of samples is the mean of SS for each sample. The SS goes from -1 to +1, with a high number indicating good alignment with its own cluster and poor alignment with nearby clusters. When the majority of items possess a significant value, the clustering setup is suitable. When several points possess a low or negative value, the clustering arrangement could result in an excessive or insufficient number of clusters.

The SS, which measures both the cohesion within clusters and the separation between them, emerged as a particularly comprehensive metric for evaluating the performance of clustering methods in the context of CZB. By adopting SS as the primary metric, the methodology not only simplified the evaluation process but also ensured that the chosen clustering method met the critical requirements of distinctiveness and internal consistency within clusters.

3.11.2. **Climate zoning validation with building performance**

The successful implementation of CZB occurs when there is a noticeable variation in a set of performance indicators for identical buildings in different CZs, but minimal variation between identical buildings within the same climate zone [4]. Based on that idea, buildings that are identical in various CZ will exhibit highly discernible and unique performance qualities, whereas buildings that are identical within the same zone will demonstrate comparable performance. With that, the accuracy of the resultant CZB can be increased significantly if there is performance data accessible for all pertinent locations inside each CZ, taking into account a collection of pertinent buildings.

As far as the authors are aware, there is just one metric available in the literature to evaluate the quality of a specific CZB. The Mean Percentage of Misclassified Area (MPMA), as proposed by Walsh et. al. [4, 54], is a quality metric used to validate CZB outcomes. MPMA may aggregate the percentages of misclassified regions for all archetypes,

yielding a singular figure that represents the overall concordance between CZ and performance for certain building stock. Nevertheless, the fundamental mathematical nature of this index remains rather uncertain.

This section introduces a novel CZMI that is based on the assumption that in an ideal situation, each CZ should have a distinct climate, resulting in distinctive energy needs for a certain archetype with the smallest overlaps across CZ. The CZMI measures the degree of overlap across CZ based on the intersection of KDE curves [190], thus indicating the uniqueness of the climate in each zone.

The KDE is a non-parametric way to estimate the probability density function of a random variable [191]. Given a set of data points $x_1, x_2, \ldots, x_n$, the KDE is typically estimated as:

$$\hat{z}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{3.11}$$

where:

$\hat{z}_h(x)$ is the estimated probability density function at point x;

n is the number of data points;

h is the bandwidth, a smoothing parameter that controls the width of the kernel;

$K(\cdot)$ is the kernel function, in this study Gaussian (normal distribution).

CZMI involves the calculation of the total KDE function over the entire range of each performance indicator and the total overlapping area of clusters for each building performance indicator within each CZB result. An example of overlapping KDE of 3 clusters is presented in Figure 3.10.

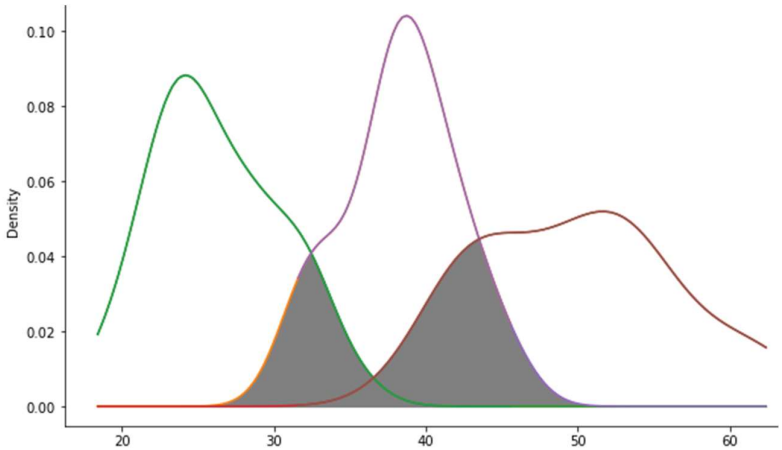


Figure 3.10: KDE overlap of three clusters.

Total KDE represents the integral of the KDE function over its entire range and can be calculated as:

$$K_t = \int_{x\,min}^{x\,max} \hat{z}_h(x)\,dx \tag{3.12}$$

The overlap ($\bar{I}$) is then computed for each CZB method and can be expressed as:

$$\bar{I} = \int_{x\,min}^{x\,max} min(\hat{z}_h^i(x), \hat{z}_h^j(x))dx \tag{3.13}$$

where

$\hat{z}_h^i(x), \hat{z}_h^j(x)$ are the probability density functions of the $i-th$ and $j-th$ KDE curves of two clusters, respectively.

For each pair of clusters within each clustering result and each performance indicator, the overlap percentage (O) is calculated as:

$$O = \left(\frac{\int_{x\,min}^{x\,max} min(\hat{z}_h^i(x), \hat{z}_h^j(x))dx}{\int_{x\,min}^{x\,max} \hat{z}_h(x)\,dx}\right) \tag{3.14}$$

Having multiple archetypes and multiple performance indicators, the base CZMI calculation can be formed to consider all combinations of archetypes and indicators as follows:

$$CZMI_{base} = \frac{\sum_{i=1}^{A}\sum_{j=1}^{P} O_{archetype_{i,\ perf.indicator_j}}}{A * P} \tag{3.15}$$

where:

A is the number of archetypes;

P is the number of performance indicators for each archetype;

i iterates over the archetypes;

j iterates over the other performance indicators.

The base CZMI is then adjusted using the normalized intra-cluster distances (D_n) to take into account the separation between clusters. Larger distances (better separation) lead to a greater reduction in the corrected CZMI. The Euclidean distance (D_E) between the centroids of two clusters A and B is calculated using the formula:

$$D_E(A,B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \tag{3.16}$$

where
A_i and B_i are the coordinates of the centroids in the multidimensional space, and
n is the number of dimensions.

After calculating the D_E for all pairs of clusters for each clustering method, the maximum distance (D_{max}) within that method is determined:

$$D_{max} = \max(D_E(A,B) for\ all\ unique\ pairs\ A,B) \tag{3.17}$$

Each D_E is then normalized by dividing by the maximum distance in that cluster column as:

$$D_{norm}(A,B) = \frac{D_E(A,B)}{D_{max}} \tag{3.18}$$

This normalized distance (D_{norm}) is a value between 0 and 1, where 0 means no distance (completely overlapping clusters), and 1 represents the maximum observed distance in that cluster column. If normalization is done, the distances are scaled relative to the largest distance in the set, ensuring comparability. The final corrected CZMI is calculated using the formula:

$$CZMI = CZMI_{base} * (1 - D_{norm}(A,B)) \tag{3.19}$$

Using performance-based validation, this study not only assesses whether the level of variances within and across CZ is deemed acceptable or not, but it also presents a measure

to streamline the process of decision-making. Among all explored CZB methods, the method with the smallest CZMI will be considered the best (proposed) CZB.

Additionally, the proposed CZB will be comprehensively compared with the official CZB map of Kazakhstan to identify any disparities in the current map in use. The same KDE-based index will be implemented in the official CZB map. The possible high overlap values of the official CZB map will show the level of inaccuracy of the official CZB map concerning the proposed method.

3.11.3. **Validation using the Adjusted Rand Index**

The most effective performance-driven method identified in this study will be meticulously compared to all the other results, as well as to the official CZB map and ASHRAE climate map of Kazakhstan, to identify disparities. The primary objective of this comparison is to assess the precision and reliability of the current official maps in accurately depicting building energy performance patterns. ARI will be utilized to measure the disparities and inconsistencies in the classification results. ARI is a commonly used statistical measure for assessing the similarity of two datasets, providing a robust evaluation of the level of agreement or disagreement [192-194]. ARI is a corrected-for-chance modification of the Rand Index (RI) that considers the anticipated agreement between the two classifications. [193, 195].

To determine the RI, it is essential to first define the divisions. Thus, a partition of N objects with R subsets is defined as $\mathcal{R} = \{R_1, \dots, R_R\}$, *where* $R_1 \cup R_2 \cup \cdots R_R = \mathcal{R}$ and $R_i \cap R_j = \emptyset, \forall i \neq j$ (indicating that all subsets [e.g., clusters] are mutually exclusive). Given two partitions, $\mathcal{R}$ and $\mathbb{C}$, with R and C subsets respectively, RI can be calculated using four different basic types of pairs, from the $\binom{N}{2}$ possible object pairs according to the following formula [192]:

$$RI = Rand\ Index = \frac{(a+b)}{(a+b+c+d)} \tag{3.20}$$

where:

a are objects in the pair placed in the same cluster in $\mathcal{R}$ and the same cluster in $\mathbb{C}$;
b are objects in the pair placed in the same cluster in $\mathcal{R}$ and different clusters in $\mathbb{C}$;
c are objects in the pair placed in different clusters in $\mathcal{R}$ and the same clusters in $\mathbb{C}$;
d are objects in the pair placed in different clusters in $\mathcal{R}$ and in different clusters in $\mathbb{C}$.

The values of a, b, c, and d may be calculated using the following equations [196]:

$$a = \frac{\sum_{r=1}^{R}\sum_{c=1}^{C} t_{rc}^2 - N}{2} \tag{3.21}$$

$$b = \frac{\sum_{r=1}^{R} t_{r+}^2 - \sum_{r=1}^{R}\sum_{c=1}^{C} t_{rc}^2}{2} \tag{3.22}$$

$$b = \frac{\sum_{c=1}^{C} t_{+c}^2 - \sum_{r=1}^{R}\sum_{c=1}^{C} t_{rc}^2}{2} \tag{3.23}$$

$$d = \frac{\sum_{r=1}^{R}\sum_{c=1}^{C} t_{rc}^2 + N^2 - \sum_{r=1}^{R} t_{r+}^2 - \sum_{c=1}^{C} t_{+c}^2}{2} \tag{3.24}$$

where:

t_{rc} representing the number of objects that were classified in the *rth* subset of partition $\mathcal{R}$ and the *cth* subset of partition $\mathbb{C}$.

To correct the RI for chance, ARI was proposed with the general equation of [193]:

$$ARI = \frac{\big(RI - E[RI]\big)}{\big(\max(RI) - E[RI]\big)} \tag{3.25}$$

where:

$$E[RI] = \frac{(\sum_{r=1}^{R} t_{r+}^2 - N)(\sum_{c=1}^{C} t_{+c}^2 - N) + (N^2 - \sum_{r=1}^{R} t_{r+}^2)(N^2 - \sum_{c=1}^{C} t_{+c}^2}{2N(N-1)} \tag{3.26}$$

ARI takes into account both the true positive and true negative categories, which makes it very suitable for comparing the proposed approach with the ASHRAE and official CZB maps. ARI streamlines comparisons by providing a single numerical value that condenses the overall alignment between the proposed methods and the benchmark maps. The range of ARI is from -1 to 1. Negative values represent a lack of agreement, whilst positive numbers suggest a level of agreement that surpasses chance [193]. This simplifies the interpretation, supporting the measurement of discrepancies and misclassifications, and allowing for an adequate assessment of the efficacy and superiority of the proposed technique compared to current alternatives. It allows for an accurate and unbiased evaluation, minimizing the chance of subjective interpretations or biases.

3.12. Chapter Summary

The methodology proposed in this study is outlined below, organized into two principal phases aimed at fulfilling the research objectives. The initial step involves outlining the study area to provide the geographical context of the Kazakhstan region. It is planned to meticulously gather and incorporate weather data in the form of TMYx files for 94 locations, which will serve as a fundamental basis for simulations and climate analysis. The selection of building archetypes is offered through a systematic analysis of the national building stock statistics of Kazakhstan to ensure the samples for simulations are representative. Building energy performance simulations are planned, with two archetypes and comprehensive verification processes to ensure the results are accurate and reliable. Key performance indicators are proposed for identifying energy needs and further climate zoning classification.

A strategy is offered to employ correlation analysis and a set of ML techniques to uncover the most significant climate variables connected with building energy needs, and the Elbow method to suggest the optimal number of CZs before classification. Multivariate clustering methodologies, such as KC and HC, are proposed to distinguish CZs with similar characteristics. Evaluation metrics like uniqueness, compactness, and SS are planned to be utilized for assessing the quality of the clustering results. Furthermore, the validation of climate zoning through building performance is proposed, utilizing a specially developed CZMI, to validate the effectiveness of all involved approaches. Finally, the consistency and accuracy of the official building climate maps are planned to be validated using the Adjusted Rand Index against the best-proposed method, to check the reliability of the official CZB.

Chapter 4: Results and Discussion

This chapter delves into the findings and analysis of the proposed performance-driven CZB for Kazakhstan, where in section 4.1 the results of the base case building energy performance will be presented. Section 4.2 verifies the simulation results obtained from EnergyPlus, ensuring their reliability and accuracy for further analysis. In section 4.3, the spatial patterns of building energy performance across Kazakhstan are examined.

A detailed investigation into the variables crucial for climate classification is undertaken in section 4.4, employing a series of analytical techniques. This section starts with correlation analysis results (4.4.1), moving through RFR (4.4.2), GB (4.4.3), and XGBoost (4.4.4), before summarizing the findings in section 4.4.5 to identify the most significant predictors of climate impact on building energy use.

The narrative then shifts to the introduction of Phase 1 (climate-based CZB) in section 4.5, which outlines the optimal number of CZs (4.5.1) and discusses the clustering results (4.5.2). The quality of the proposed CZB is thoroughly assessed (4.5.3) through metrics such as uniqueness, compactness, and the SS. The validation results of climate-based classification through a novel CZMI are introduced (4.5.4), culminating in a summary of Phase 1 findings (4.5.5). Phase 2 (performance-based CZB), described in section 4.6, builds upon the development of CZB based only on building performance. Similar to Phase 1, it includes determining the optimal number of CZs (4.6.1), presenting clustering results (4.6.2), and assessing clustering quality (4.6.3). This phase introduces climate zoning overlap (4.6.4) as an additional evaluation metric, leading to a comprehensive summary of Phase 2 findings in section 4.6.5.

A comparative analysis and synthesis in section 4.7 integrate insights from both phases. This includes an analysis of mean overlap percentages and the CZMI (4.7.1.1), as well as the ARI (4.7.1.2), providing a holistic understanding of the methodologies' effectiveness in comparison with the official CZB map of Kazakhstan. The chapter concludes with a summary (4.8), synthesizing the key insights and contributions of the research to the field of building energy performance and CZB in Kazakhstan.

4.1. Building Performance Simulations

This section delves into the results of building energy performance simulation results. Examining overall annual energy needs, the values for NA range from 131.3 to 150.8 kWh/m^2, while the SA demonstrates a broader range of 140.5 to 174.2 kWh/m^2. Exploring the specific components of energy consumption, the annual space cooling and heating values for the NA remain within the range of 18.4 to 62.4 kWh/m^2 and 27.5 to 108.4 kWh/m^2, respectively. The SA, on the other hand, demonstrates higher energy consumption, ranging from 18.5 to 69.4 kWh/m^2 for cooling and 34.4 to 132.7 kWh/m^2 for heating. Figure 4.1 shows the simulation results.

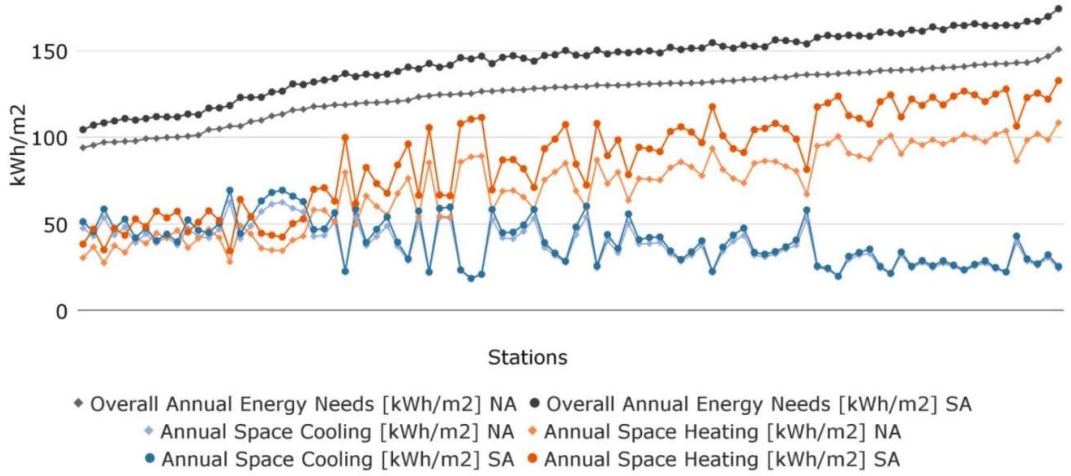


Figure 4.1: Used archetypes energy consumption levels.

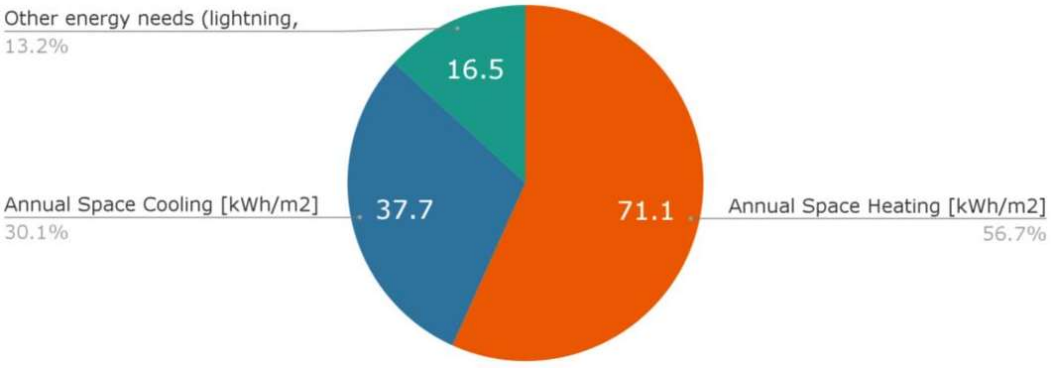
There is a noticeable difference in energy usage between the two archetypes being used. The insulation in NA is more substantial (higher R-value) since the wall and roof composition is specifically tailored for colder areas of the country. On the other hand, SA has a less insulated envelope that is built for the warmer climates of southern regions. Although the difference in cooling variance may be small, it is clear that the lack of sufficient insulation in SA results in much higher energy consumption for heating compared to NA, especially in northern regions. Increased insulation in NA results in an average decrease in energy usage of around 17 kWh/m^2 for heating and 2.8 kWh/m^2 for cooling.

Averaging the values it is seen that the NA has a lower mean space cooling needs (37.7 kWh/m^2) compared to the SA (40.5 kWh/m^2). Similarly, the mean space heating for the NA (71.1 kWh/m^2) is lower than that of the SA (88.1 kWh/m^2). The mean overall annual energy needs for the NA is 125.3 kWh/m^2, while the SA results in a slightly higher mean of 143.4 kWh/m^2 (Table 4.1). The examination of mean energy needs confirms that the local climate

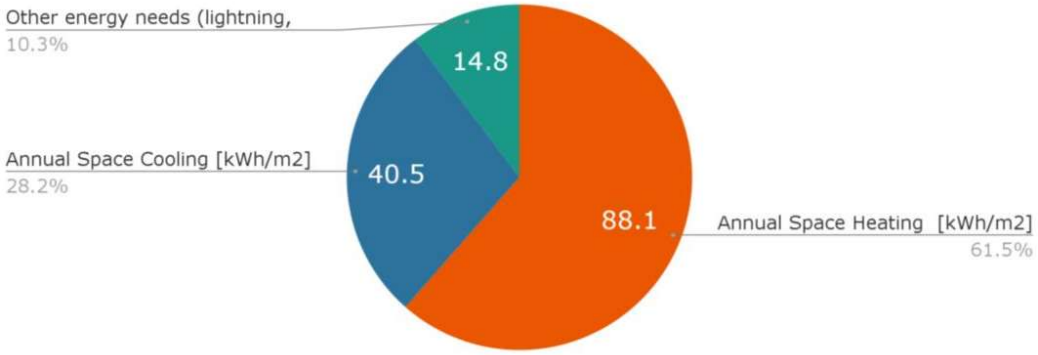
is predominantly heating-dominated, with mean heating needs accounting for around 60% of the total energy consumption (Figure 4.2).

Table 4.1: The mean values of buildings simulation results

Metric	NA [kWh/m^2]	SA [kWh/m^2]
Mean Overall Annual Energy Needs	125.3	143.4
Annual Space Cooling Mean	37.7	40.5
Annual Space Heating Mean	71.1	88.1



(a)



(b)

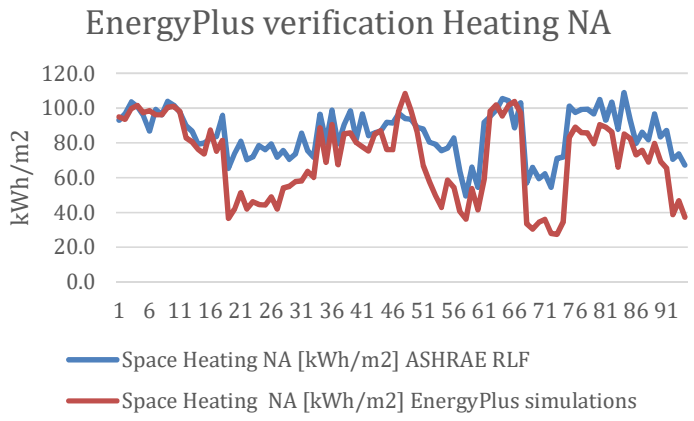
Figure 4.2: Mean energy needs of NA (a), and SA (b).

4.2. Verification of EnergyPlus simulation results

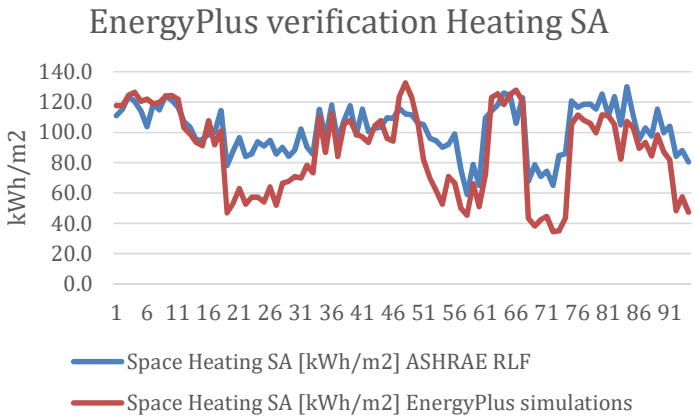
The verification of EnergyPlus simulation results was conducted using Python (Appendix A). First energy load calculations for heating were performed utilizing the ASHRAE residential cooling and heating load calculations technique with the RLF method. Utilizing thermal resistance values for walls and ceilings, along with temperature differentials and other building parameters (size, window-to-wall ratio, etc.), the code computed heating and cooling loads per square meter for each city and archetype. Next, the comparison was made

through the utilization of a suite of evaluation metrics, including RMSE, MAE, MSE, and MAPE, and simulation accuracy information was achieved, thereby affirming the validity of the EnergyPlus results within the context of the study.

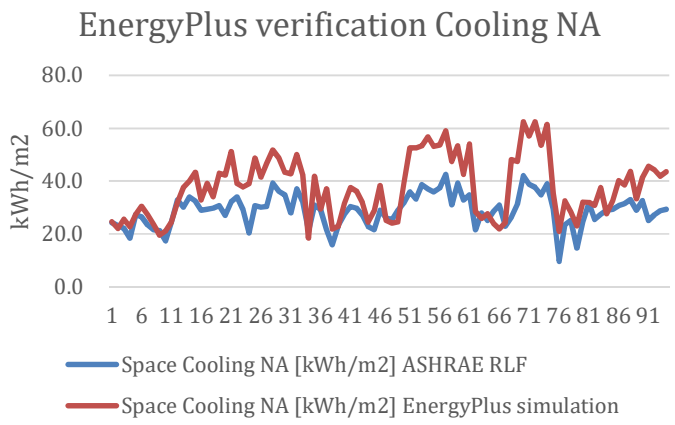
As can be seen in Figure 4.3 (a, b, c, d), where the blue line represents the ASHRAE RLF values, and the red line depicts the EnergyPlus simulations' results for NA and SA, both approaches exhibit a similar pattern over time, with some fluctuations. However, the EnergyPlus simulations generally show lower energy needs compared to the ASHRAE RLF values for cooling and higher values for heating.



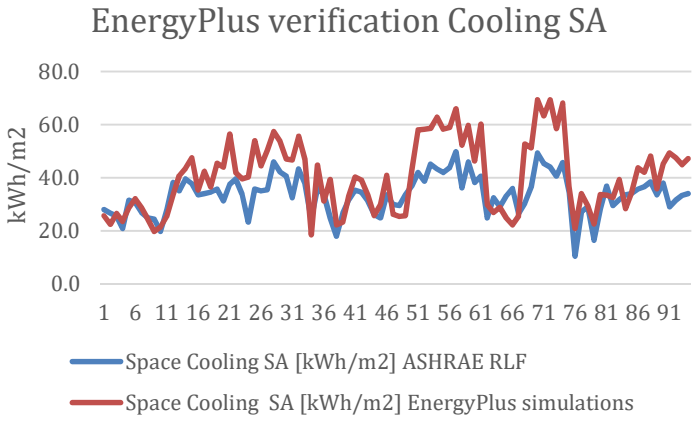
(a)



(b)



(c)



(d)

Figure 4.3: Comparison of EnergyPlus simulations results with ASHRAE RFL calculations. Heating NA (a), heating SA (b), cooling NA (c), and cooling SA (d).

For space cooling, the verification demonstrates relatively close results, with RMSE values of 11.31 and 10.77% for NA and SA, respectively, falling within an acceptable range (Table 3.5). Conversely, for space heating, the ASHRAE RFL exhibits higher RMSE values, indicating less accurate predictions, with 18.67 and 20.53% for NA and SA, respectively. Despite these variations, the average RMSE of 15.32% suggests an overall acceptable level of accuracy, albeit with a slightly higher discrepancy in heating predictions. MAE values range from 8.68 to 16.49, suggesting a relatively small average difference between approaches. MSE values range from 116.05 to 421.42 while MAPE ranges from 19.91% to 30.04%. As a percentage of actual values, MAPE errors might seem high. However, the combined analysis of all metrics suggests the ASHRAE RFL exhibits reasonable alignment in predicting both space cooling and heating needs, considering all its limitations. All evaluation metrics can be seen in Table 4.2.

Table 4.2: EnergyPlus verification evaluation results.

Performance indicators	CV RMSE	MAE	MSE	MAPE
Space Cooling NA	11.31	9.29	128.02	22.43
Space Heating NA	18.67	15.26	348.43	30.04
Space Cooling SA	10.77	8.68	116.05	19.91
Space Heating SA	20.53	16.49	421.42	26.31
Average	15.32	12.43	253.48	24.67

While there's always room for improvement, the ASHRAE RFL values appear to be reasonably close to the EnergyPlus values, with average errors being on the upper boundary of the acceptable range. However, it's important to acknowledge that the RLF method's simplifications in representing the full spectrum of building energy components might contribute to these results.

4.3. Buildings Energy Performance Patterns

An examination of the energy performance patterns of buildings offers useful insights into how energy demands are distributed over the area of the Republic of Kazakhstan. A detailed understanding of the variations in energy consumption patterns across the country was achieved by mapping building energy performance indicators, focusing on heating and cooling energy needs.

Upon analyzing the heating energy needs, it was noted that areas situated in the northern portion of Kazakhstan, which experience a prolonged and intense winters, demonstrated higher energy usage for heating (Figure 4.4). The regions of Akmola, Kostanay, North Kazakhstan, and Pavlodar have the greatest needs for heating. The outcome is consistent with predictions based on the higher heating requirements in colder areas. Nevertheless, the research also emphasized specific differences within these locations, indicating the necessity for more comprehensive CZ. In contrast, the analysis of cooling energy needs revealed that the southern areas of Kazakhstan, including Turkistan and Mangystau, had the highest cooling energy consumption. Once again, it was seen that there were differences in energy consumption patterns in the southern areas, highlighting the need to accurately represent the local climatic details. Furthermore, the areas with the greatest demand for heating may not always align with the areas that utilize the biggest amount of energy for cooling, and vice versa. This finding highlights the need to separately evaluate heating and cooling when examining patterns of building energy performance. Moreover, the examination of building energy performance patterns uncovers spatial disparities in energy usage that may be ascribed to specific local geographical conditions. The Caspian

sea to the west, the Tien-Shan mountains to the south, and the Altay mountains to the east play a role in creating diverse climatic zones inside the country. Overall, the energy consumption pattern has a strong correlation with variations in latitude.

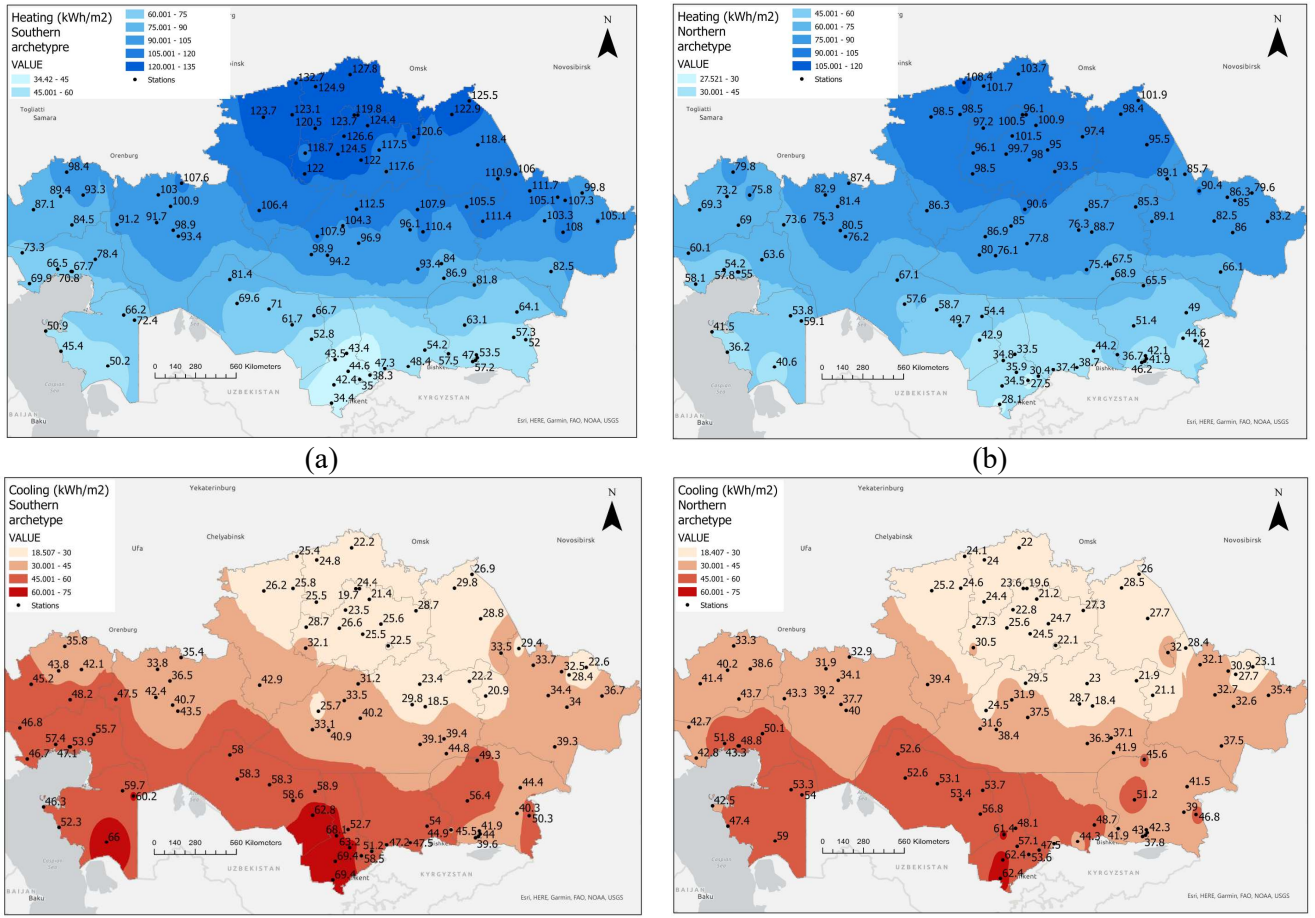


Figure 4.4: Patterns of building energy performance. The heating energy consumption of SA (a), the heating energy consumption of NA (b), the cooling energy consumption of SA (c), and the cooling energy consumption of NA (d). All with a 15 kWh/m2 interval.

4.4. The most important climate variables

This section will explore the identification of the most influential climate variables through a comprehensive analysis, including correlation analysis, RFR, GB, and XGBoost techniques. The results will provide insights into the key climatic factors that significantly impact the energy dynamics of the studied locations. 47 variables were analyzed, with 36 extracted from the TMYx file (including latitude, longitude, elevation, annual average dry bulb temperature, annual average dew point temperature, etc.) and 5 derived and added to the dataset later (CDD10 (hourly method), CDD18 (hourly method), HDD18 (hourly method), CDD10 (ASHRAE), HDD18 (ASHRAE)). Six other variables were included in the dataset following the completion of building energy simulations: total energy needs, space cooling, and space heating for both archetypes.

In the context of correlation analysis, the Python pandas library [197] was employed to calculate the Pearson correlation coefficient [198] for all variable pairs. The calculation of feature importance for all three MLs was also conducted in Python using standard ensemble learning techniques provided by the scikit-learn library [199] and XGBoost package [168]. Each ML algorithm was configured with 100 estimators to ensure robust and stable results.

4.4.1. **Correlation Analysis Results**

The correlation analysis findings reveal substantial statistical correlations between climatic, geographical, and performance features, offering insights into the linked factors shaping building energy consumption patterns. Figure 4.5 displays the results of the correlation study, highlighting the top 20 variables with the highest Pearson's correlation coefficients.

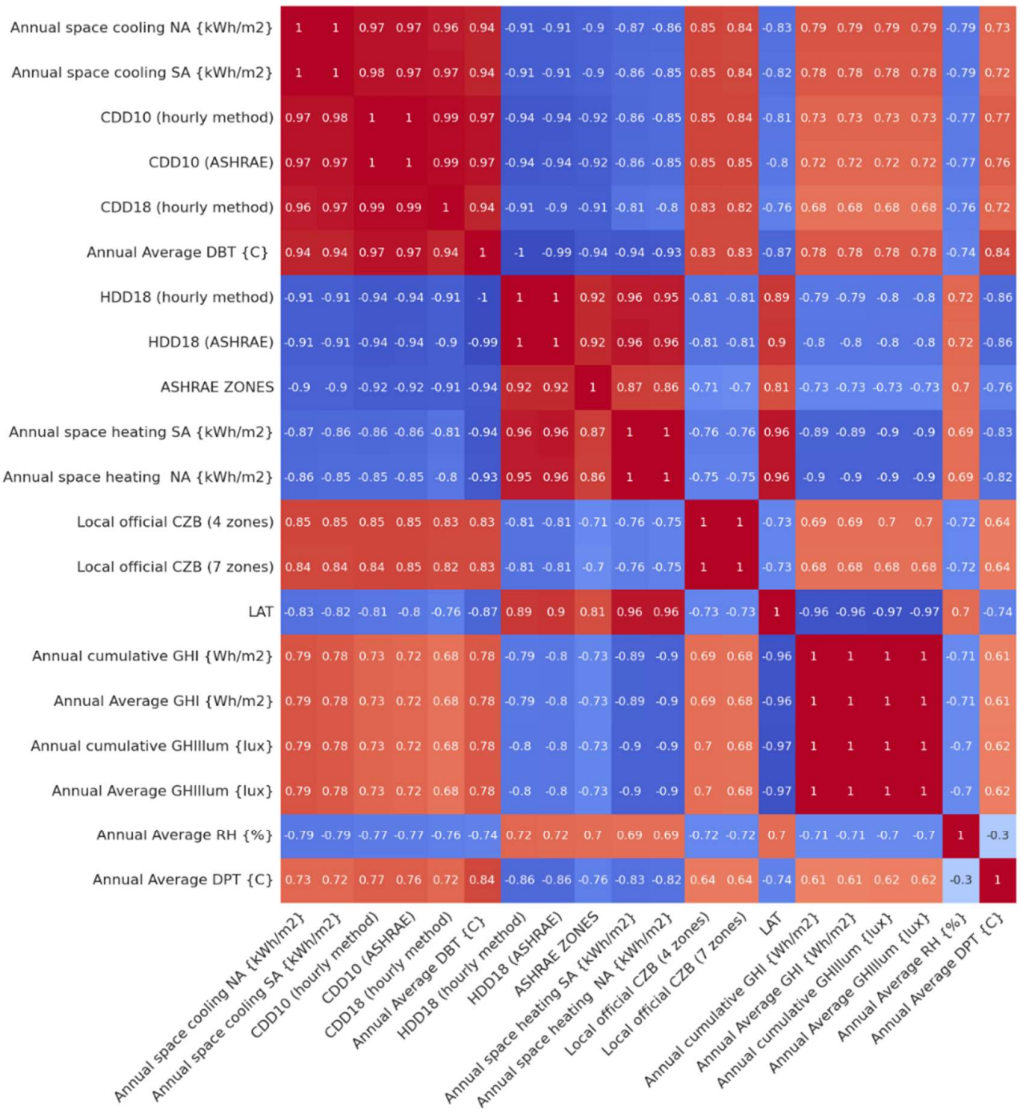


Figure 4.5: The correlation matrix comprises the top 20 variables, which encompass performance indicators, official local and ASHRAE CZB data, as well as meteorological and geographic data.

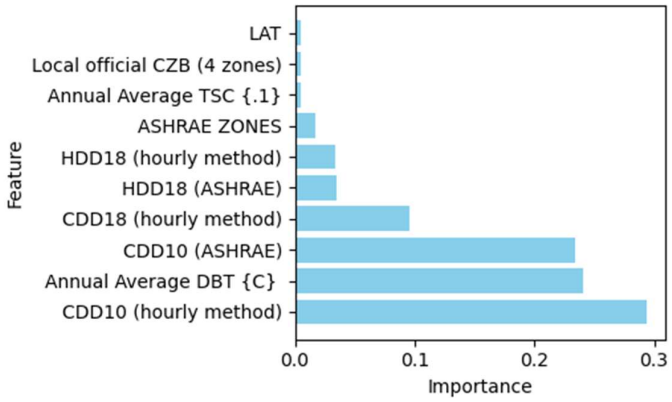
The energy consumption of the archetypes strongly correlates with the yearly average DBT, as anticipated. Similarly, all categories of DDs have significant correlations with heating and cooling energy consumption. The correlation coefficient for total energy demands is greater for HDD (0.79-0.83) compared to CDD (0.57-0.62). Furthermore, SR components like GHI and global horizontal illuminance (GHIllum) have a significant association with heating loads, with a correlation value of around -0.9. The correlation coefficient for RH ranges from 0.69 to 0.7. The correlation between WS, AP, and building energy use in Kazakhstan is rather moderate, with coefficients ranging from 0.35 to 0.47 for WS and 0.20 for AP.

In addition to climatic variables, the ASHRAE CZ information and the official CZB map of Kazakhstan are of significant interest. Correlation analysis was also used to evaluate the degree to which they match the buildings' energy patterns of the country. The ASHRAE classification shows a high correlation of 0.86-0.87 for heating, 0.90-0.92 for cooling, and 0.66-0.69 for overall energy consumption, demonstrating a substantial agreement between this climatic classification and buildings' energy. The official CZB map of Kazakhstan shows a moderate correlation, with coefficients of 0.75 for heating, 0.85 for cooling, and 0.53-0.57 for overall energy consumption, suggesting a limited alignment with building energy demands.

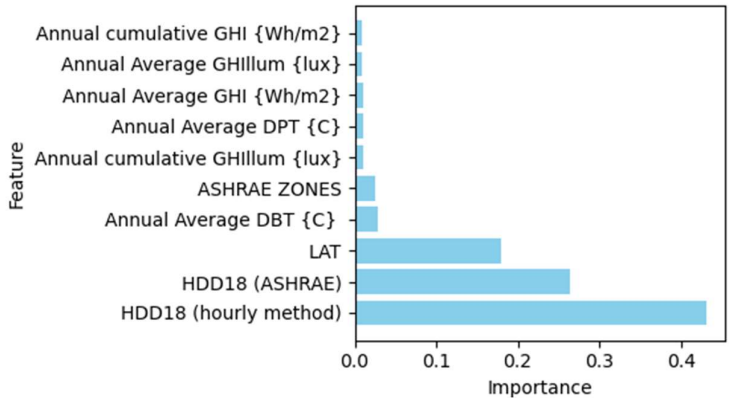
4.4.2. **Random Forest Regression Results**

An evaluation was conducted to determine the significance of climatic variables for each individual indicator and archetype. The variations in archetypes had minimal impact on the significance of the variables and were essentially identical across both archetypes. Considering this, further, in this section all figures and graphs will be provided just for the NA.

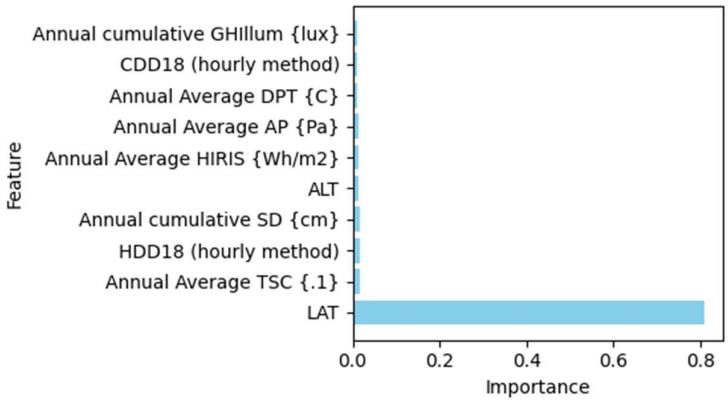
For cooling, CDD10 (hourly method), annual average DTB, CDD10 (ASHRAE), followed by CDD18 (hourly method) have emerged as the most important variable, with an importance index of up to 30% (Figure 4.6 (a)). It also reveals that the key contributors to heating energy needs are the HDD18 (hourly method), HDD18 (ASHRAE), and LAT, with their respective importance scores of 43%, 26%, and 18% (Figure 4.6 (b)). The examination of the overall annual energy needs reveals that latitude emerges as the sole and dominant variable, with a coefficient above 80% (Figure 4.6 (c)). To improve interpretation, all coefficients were consolidated by summing the significance of climatic variables for each indicator and archetype and presented in Figure 4.7 in the form of a bar chart.



(a)



(b)



(c)

Figure 4.6: Bar charts for the top 10 important variables for cooling NA (a), heating NA (b), and overall energy needs NA (c) based on the RFR method.

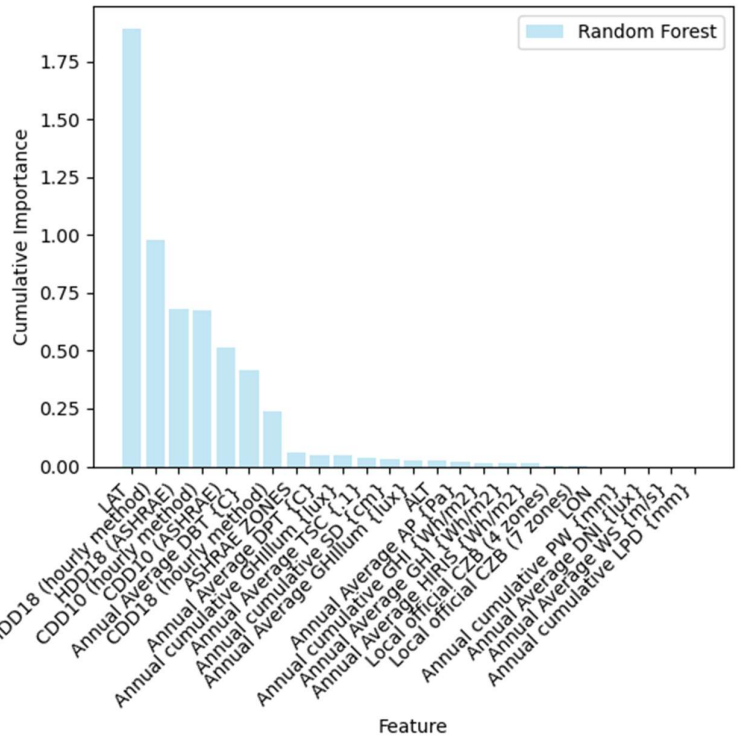
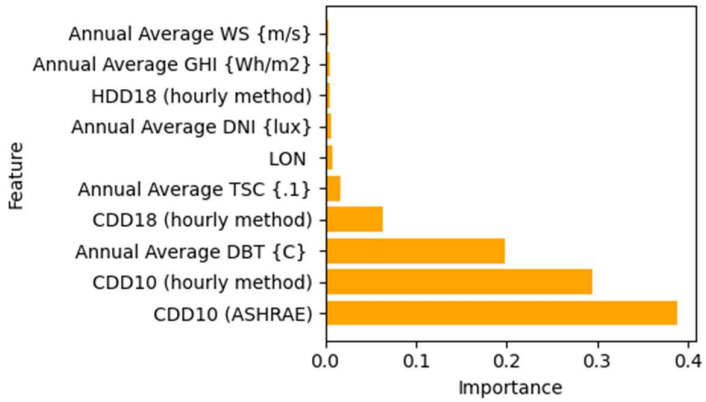


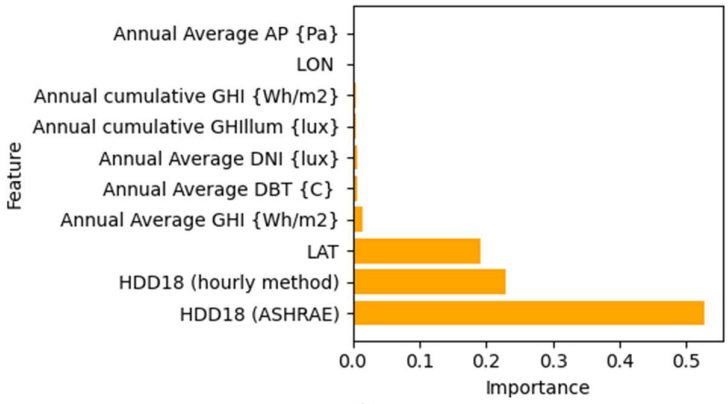
Figure 4.7: Cumulative bar chart of the most important variables for buildings' energy needs based on RFR.

4.4.3. **Gradient Boosting results**

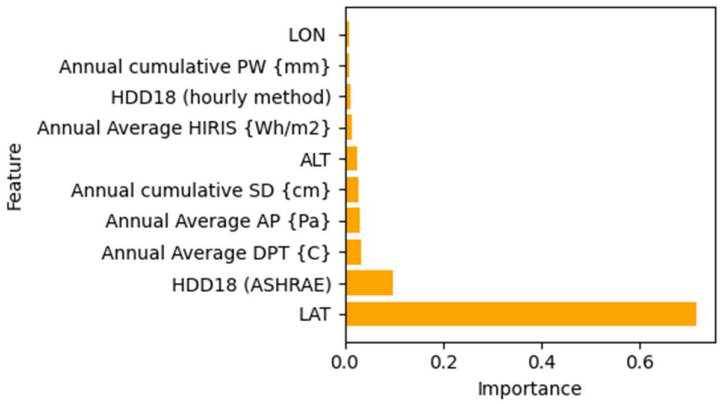
The gradient boosting approach has identified several variables that are directly linked to cooling, including CDD10 (hourly method) and CDD10 (ASHRAE) with an index of up to 39%, as well as the yearly average DTB with an index of 20% (Figure 4.8 (a)). Conversely, for heating energy needs, the predominant influencers include HDD18 (hourly method), HDD18 (ASHRAE) with up to 52% of relevance, and LAT (around 20%) (Figure 4.8 (b)). In the same way as in RFR results, in GB overall annual energy needs are strongly connected with LAT, which stands out as the exclusive and dominant variable, exhibiting a coefficient exceeding 60% (Figure 4.8 (c)). The importance coefficients revealed by the GB method were combined and shown in the form of a cumulative graph, as shown in Figure 4.9.



(a)



(b)



(c)

Figure 4.8: Bar charts for the top 10 important variables for cooling NA (a), heating NA (b), and overall energy needs NA (c) based on the GB method.

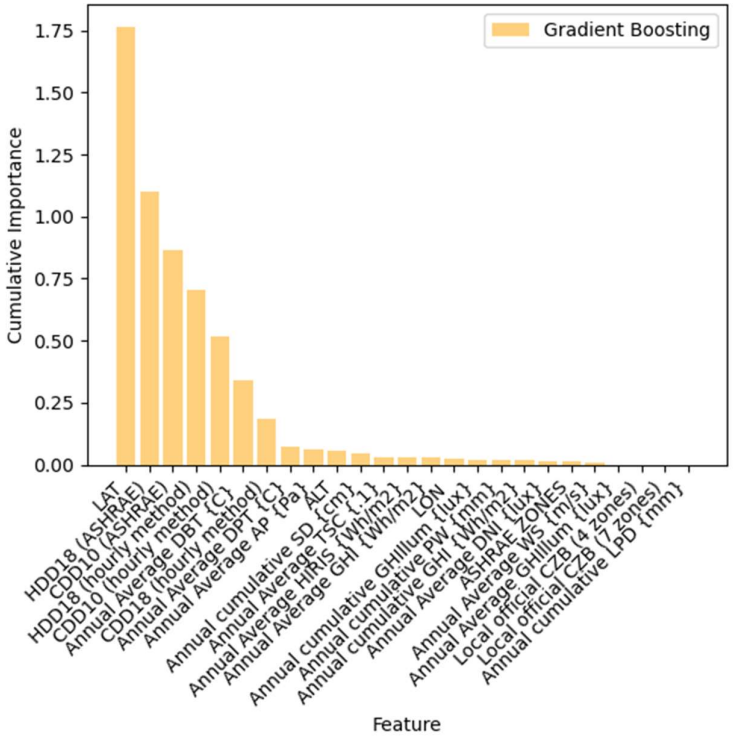
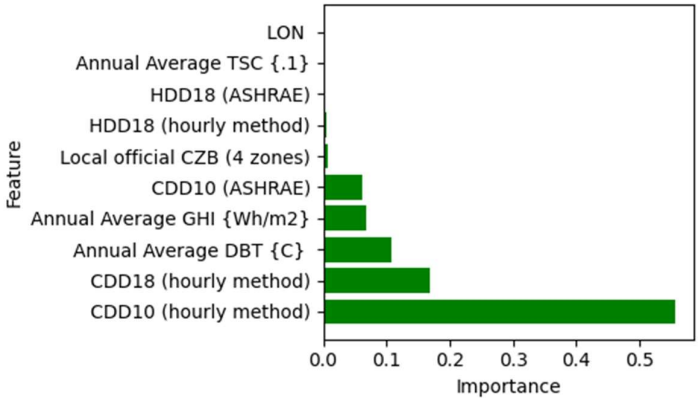


Figure 4.9: Cumulative bar chart of the most important variables for buildings' energy needs based on GB.

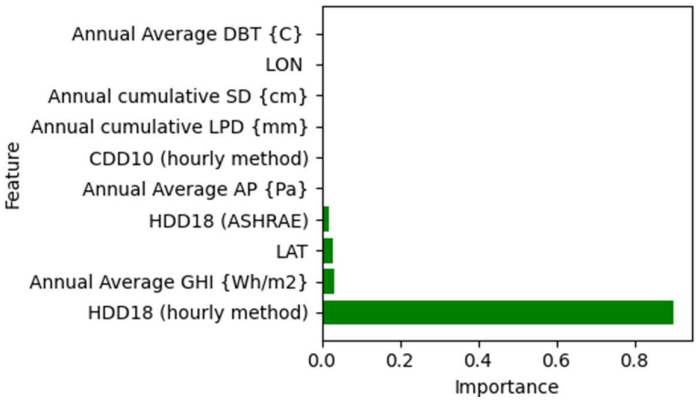
4.4.4. **Extreme Gradient Boosting results**

In the analysis for the space cooling target variable, the XGBoost identified key contributors such as CDDs, annual average DBT, and GHI, with CDD10 (hourly method) and CDD18 (hourly method) having the highest importance at 56% and 16% respectively (Figure 4.10 (a)). For the annual space heating needs, the most influential variable was HDD18 (hourly method) with an importance score of 90% (Figure 4.10 (b)). For the overall energy needs,

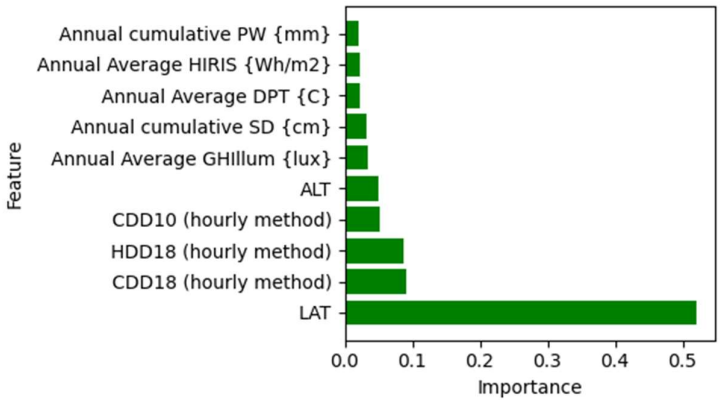
the analysis also highlighted LAT as the most influential factor, accounting for over 52% of the model's predictions (Figure 4.10 (c)). The importance coefficients revealed by the XGBoost method were combined and shown in the form of a cumulative graph in Figure 4.11.



(a)



(b)



(c)

Figure 4.10: Bar charts for the top 10 important variables for cooling NA (a), heating NA (b), and overall energy needs NA (c) based on the XGBoost method.

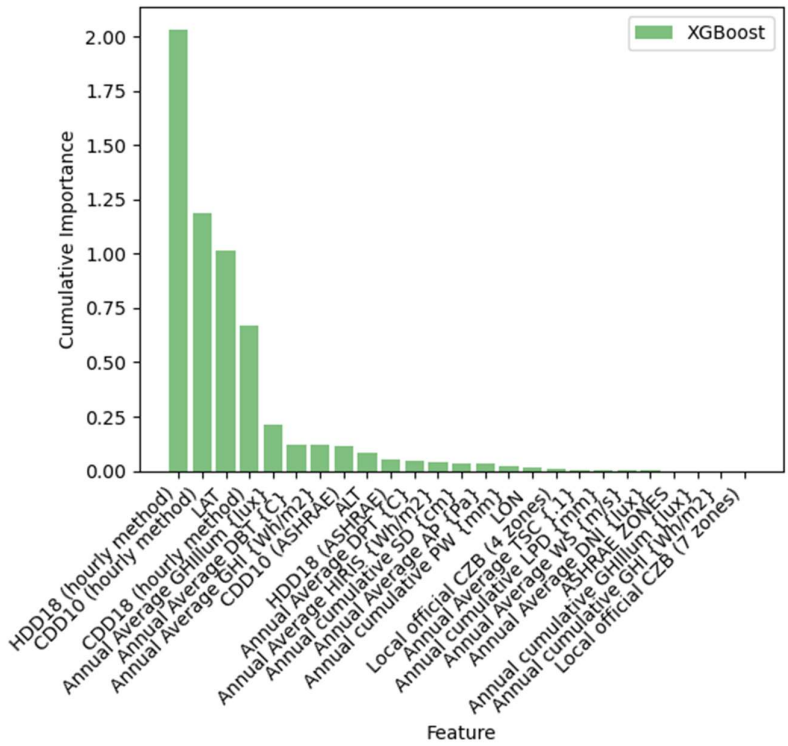


Figure 4.11: Cumulative bar chart of the most important variables for buildings' energy needs based on XGBoost.

4.4.5. **Summary**

The goal of this stage was not just a precise determination of several variables most strongly related to energy consumption, but a comprehensive search, based on various methods, for a set of such variables, which can form the basis for subsequent multivariate clustering. As a result, the analysis identified a set of key variables with substantial importance for building energy consumption across both archetypes.

- For annual space cooling, CDDs and annual average DBT were found to be consistently important;
- In the case of annual space heating, HDDs, LAT, annual average DBT, and annual average GHI exhibited the most significant importance;
- When considering the overall energy needs, including both heating and cooling, LAT, CDDs, HDDs, annual average DBT, and annual average GHI emerged as consistently impactful variables.

4.5. Phase 1 (Climate-based CZB)

Phase 1 of this research is a thorough investigation of CZB using a conventional approach based on climatic variables. It aims to understand the complexities of climate-based CZB development. The clustering process commenced by focusing on variables closely linked to

energy consumption, beginning with the primary use of the most significant variable, HDD, followed by iterative augmentations of variables based on their importance, including sequential additions like CDD and GHI, progressively expanding the range of factors in the dataset and encompassing both latitude-inclusive and latitude-exclusive configurations to discern the impact of the spatial variable on resultant classifications.

Phase 1 seeks to get a thorough grasp of the conventional CZB for a further comparison evaluation with the modern, performance-based methodology presented in Phase 2, and to answer the main research question: How do conventional climate-based and contemporary performance-based techniques for CZB compare in terms of outcomes, or can a reliable CZB be produced using a climate-based approach? In addition, the results of the developed CZB validation with building performance data will be presented.

4.5.1. **Optimal Number of climate zones**

The process of generating datasets was complete with a particular emphasis on variables most closely linked to buildings' energy needs. Six unique datasets were formed at the initial stage of Phase 1, see Table 4.3. The creation of these datasets began with the usage of one HDD18 (hourly method) variable. Increasingly, the following data sets included the gradual incorporation of other variables, including CDD18 (hourly technique), and GHI, based on their relative importance. The progressive expansion of the dataset is intended to investigate the need for using a wide range of variables for proper classification, examining if a concise set of variables may be enough. Also, the same sets of variables were used with and without LAT to track the influence of the spatial constraint on the ONCZ and classification results.

Table 4.3: Phase 1 datasets and used variables.

Name	Variable 1	Variable 2	Variable 3	Variable 4
Set 1	HDD18 (hourly method)			
Set 2	HDD18 (hourly method)	CDD18 (hourly method)		
Set 3	HDD18 (hourly method)	CDD18 (hourly method)	Annual Average Global Horizontal Irradiation	
Set 4	HDD18 (hourly method)			Latitude
Set 5	HDD18 (hourly method)	CDD18 (hourly method)		Latitude
Set 6	HDD18 (hourly method)	CDD18 (hourly method)	Annual Average Global Horizontal Irradiation	Latitude

The range of the ONCZ for each specific set of variables was determined using the Elbow technique. Figure 4.12 shows the Elbow graphs for all 6 datasets. The influence of LAT as an additional variable causing a change in the ONCZ determination (Figure 4.12 (b, d, f, h)) is noticeable.

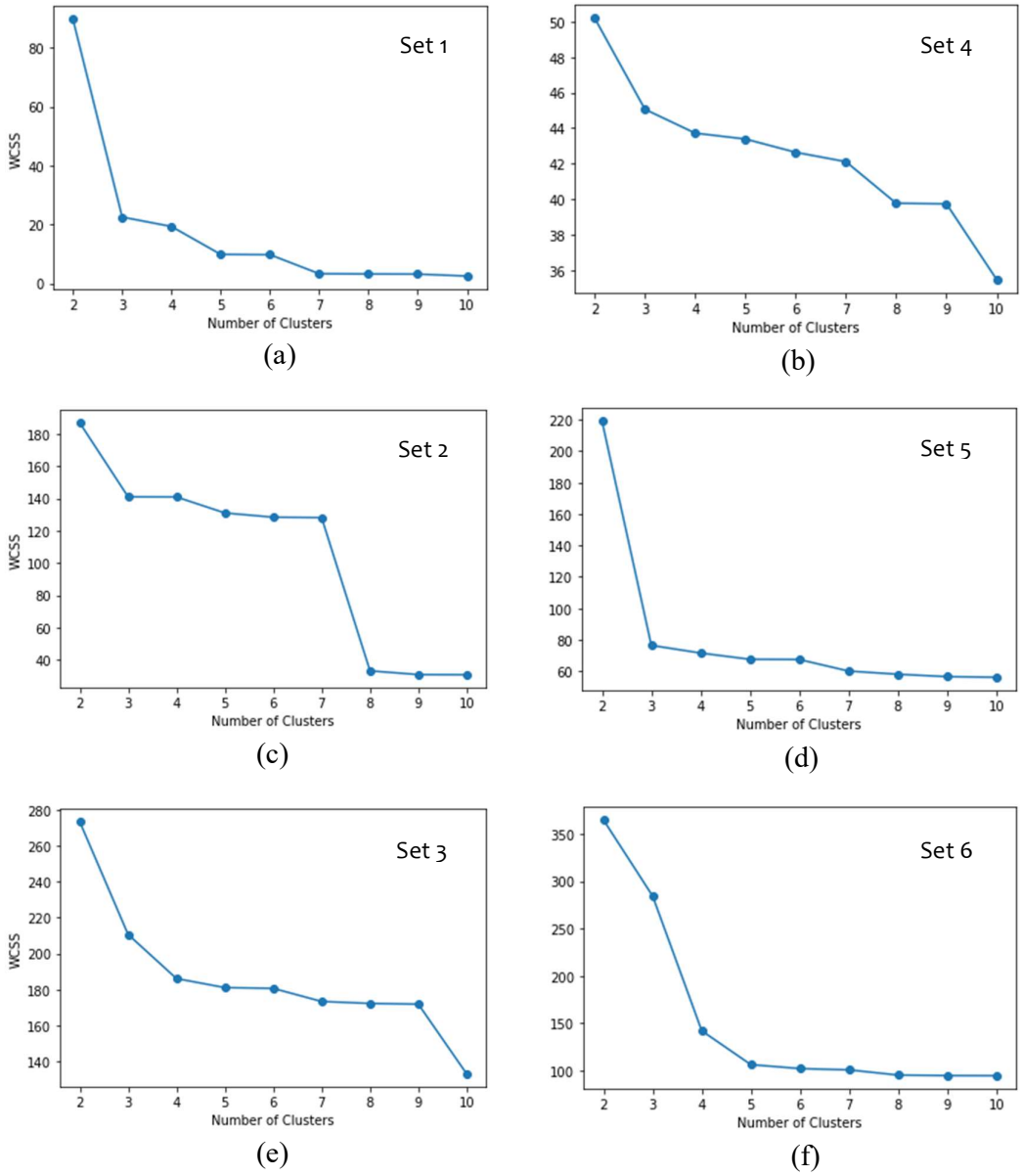


Figure 4.12: The ONCZ determination of Phase 1 methods. The Elbow graphs of set 1 (a), set 4 (b), set 2 (c), set 5 (d), set 3 (e), set 6 (f).

For each dataset, the ONCZ varied across different datasets, ranging from 3 to 9. ONCZ results were collected in Table 4.4. The findings obtained using the Elbow technique were established, indicating a preference for simplifying the CZB by selecting a smaller ONCZ, due to the rather small sample size. Nevertheless, it is vital to note that the choice of ONCZ within a specific range is rather arbitrary and subject to different influencing variables, with the stringency of energy efficiency criteria for buildings being a critical concern. In addition, it is worth mentioning that having a greater number of CZ, such as 8 or even 9, might offer more detailed information in particular situations, enabling a more precise adaptation to localized differences in climate conditions.

Table 4.4: ONCZs for Phase 1 datasets.

Name	ONCZ
Set 1	3
Set 2	8
Set 3	4
Set 4	8
Set 5	3
Set 6	5

When comparing the findings with the official CZB map of Kazakhstan, which includes 4 major CZ and 7 subzones, and the ASHRAE map, which defines 5 CZ, it is clear that the proposed ONCZ is different from the existing maps.

4.5.2. **Clustering Results**

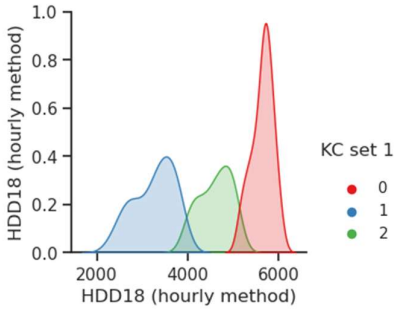
Initially, the data was loaded from a CSV file and normalized to ensure consistent scaling across the selected features. Two distinct clustering algorithms, KC and HC were applied to a climate data set using the Python script (Appendix B) both utilizing the user-specified cluster counts. The complete linkage method was chosen for HC due to its highest cophenetic coefficient in the pre-processing assessment, confirming its superiority for the analysis. 12 distinct runs were conducted that way, employing KC, HC, spatially constrained k-mean clustering (SCKC), and spatially constrained hierarchical clustering (SCHC).

Next, the scatterplot matrices were created to visually represent the results on a dataset with multiple dimensions. In the scatterplot matrix, every individual cell in the diagram represents a scatter plot, where the horizontal and vertical axes represent used variables, and the data points are color-coded according to their cluster assignments. Each set of plots in Figure 4.13 corresponds to KC (a, c, e), HC (b, d, f), SCKC (g, i, k), and SCHC (h, j, l), respectively. Comparing k-means (KC (a, c, e) and SCKC (g, i, k)) with its hierarchical counterparts (HC (b, d, f) and SCHC (h, j, l)) results, reveals distinct characteristics attributable to each method:

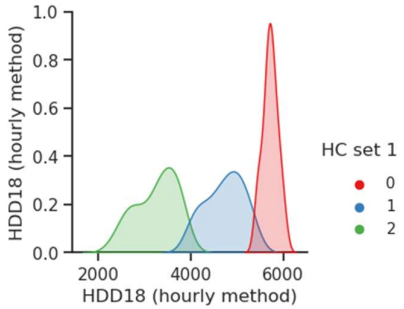
- KC and SCKC: These plots generally exhibit clusters that are more evenly distributed in space, with each cluster's points tending to form around a central value, indicative of K-means' centroid-based approach. This method optimizes for within-cluster variance, which often leads to relatively circular clusters in the feature space.
- HC and SCHC: The hierarchical approach is evident in the chaining of data points, forming clusters that may be elongated or strung out rather than tight groups. This method does not force clusters into a predefined shape, allowing for a more natural

grouping based on the actual distances between data points, which can lead to varied cluster shapes and sizes.

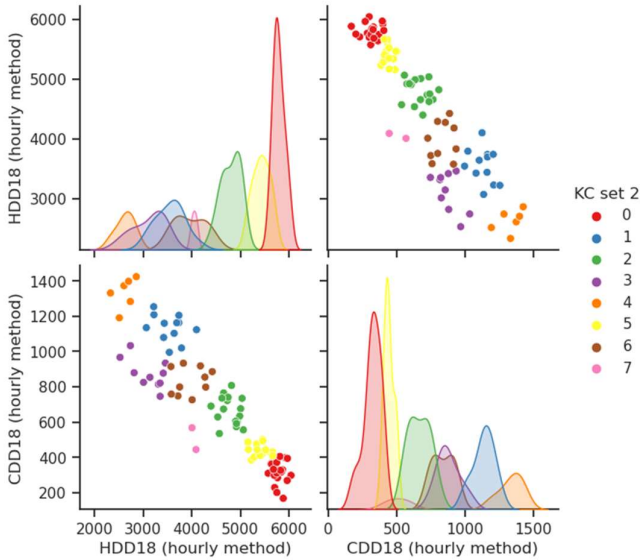
In essence, KC produced more defined and compact clusters, while the HC was capable of revealing complex structures within the data, potentially uncovering deeper insights into the inherent relationships among the data points. For further visual analysis, the clustering results were mapped using the folium library in Python and shown in Figure 4.14. The maps can be used to illustrate noticeable differences in geographical distribution between KC and HC and between regular and spatially constrained variants.


HDD18 (hourly method)
HDD18 (hourly method)
KC set 1
0
1
2

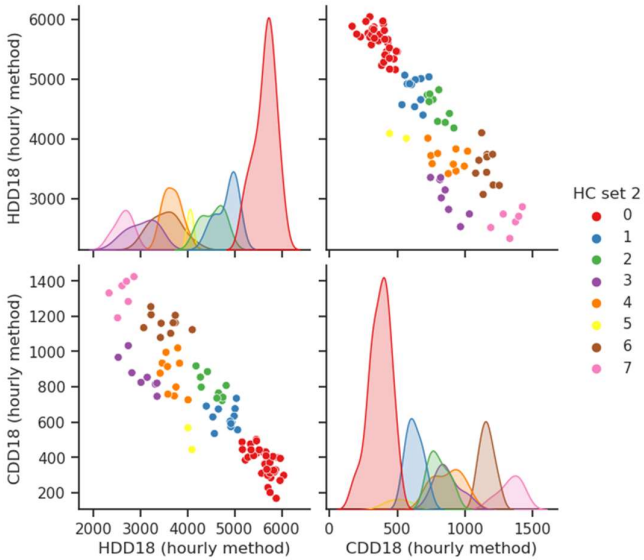
(a)


HDD18 (hourly method)
HDD18 (hourly method)
HC set 1
0
1
2

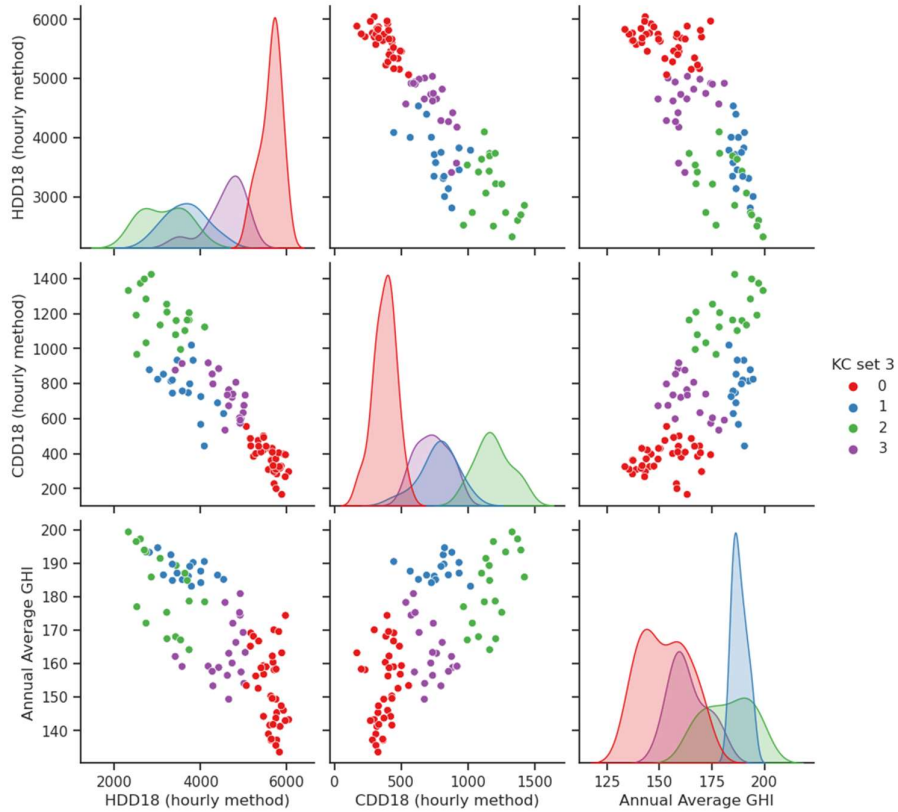
(b)


HDD18 (hourly method)
CDD18 (hourly method)
HDD18 (hourly method)
CDD18 (hourly method)
KC set 2
0
1
2
3
4
5
6
7

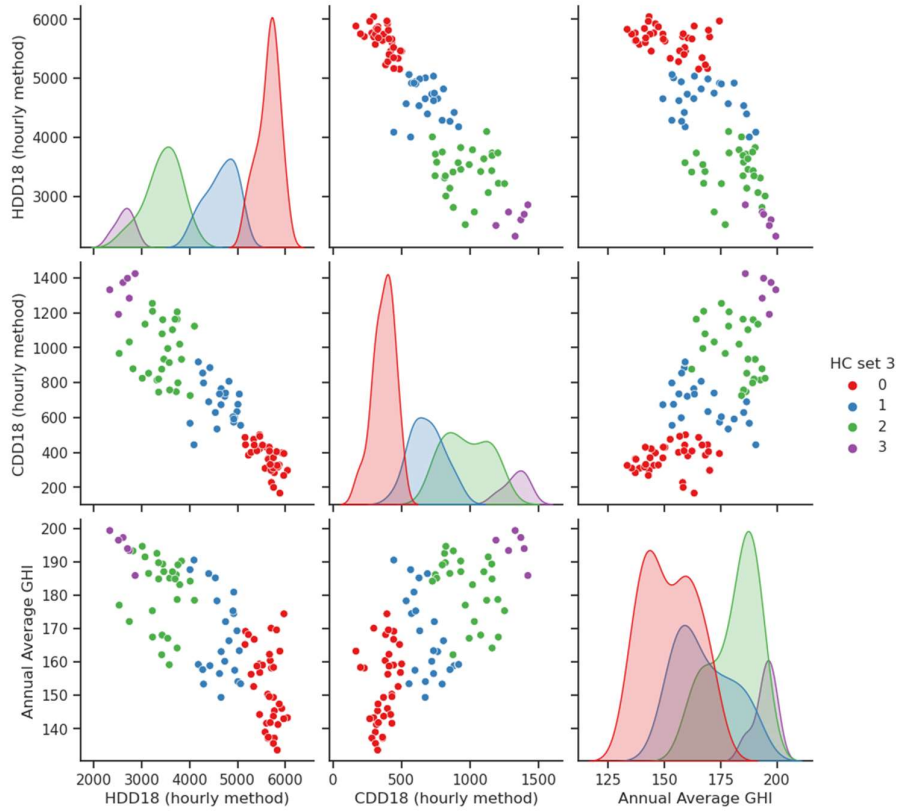
(c)


HDD18 (hourly method)
CDD18 (hourly method)
HDD18 (hourly method)
CDD18 (hourly method)
HC set 2
0
1
2
3
4
5
6
7

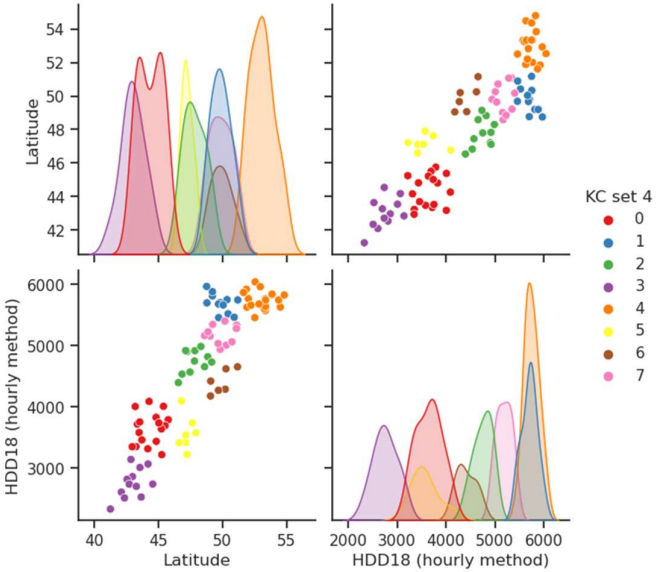
(d)


HDD18 (hourly method)
CDD18 (hourly method)
Annual Average GHI
KC set 3
0
1
2
3

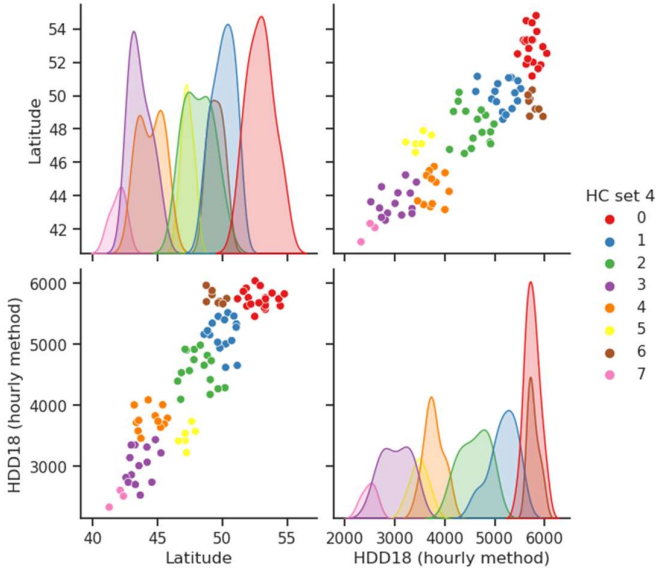
(e)


HDD18 (hourly method)
CDD18 (hourly method)
Annual Average GHI
HC set 3
0
1
2
3

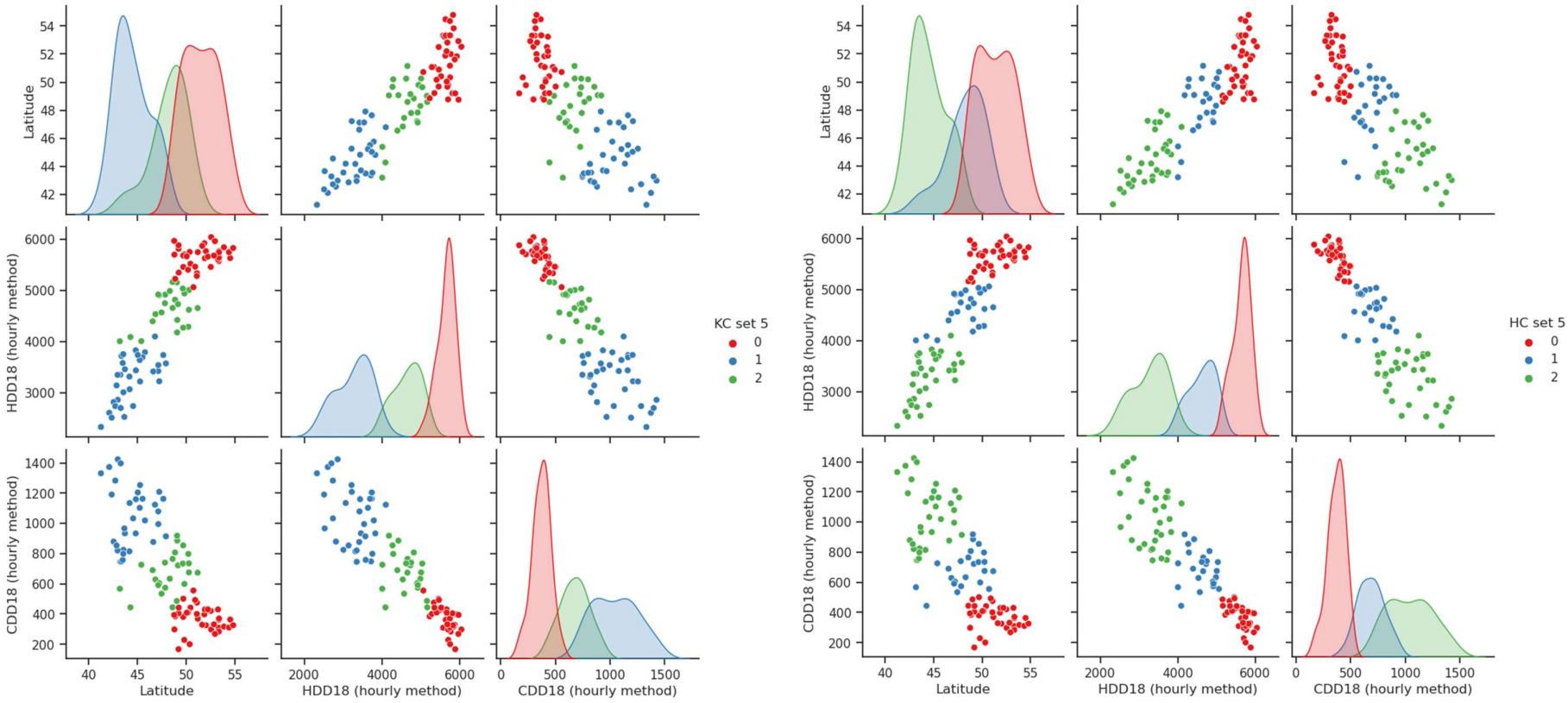
(f)


Latitude
HDD18 (hourly method)
KC set 4
0
1
2
3
4
5
6
7

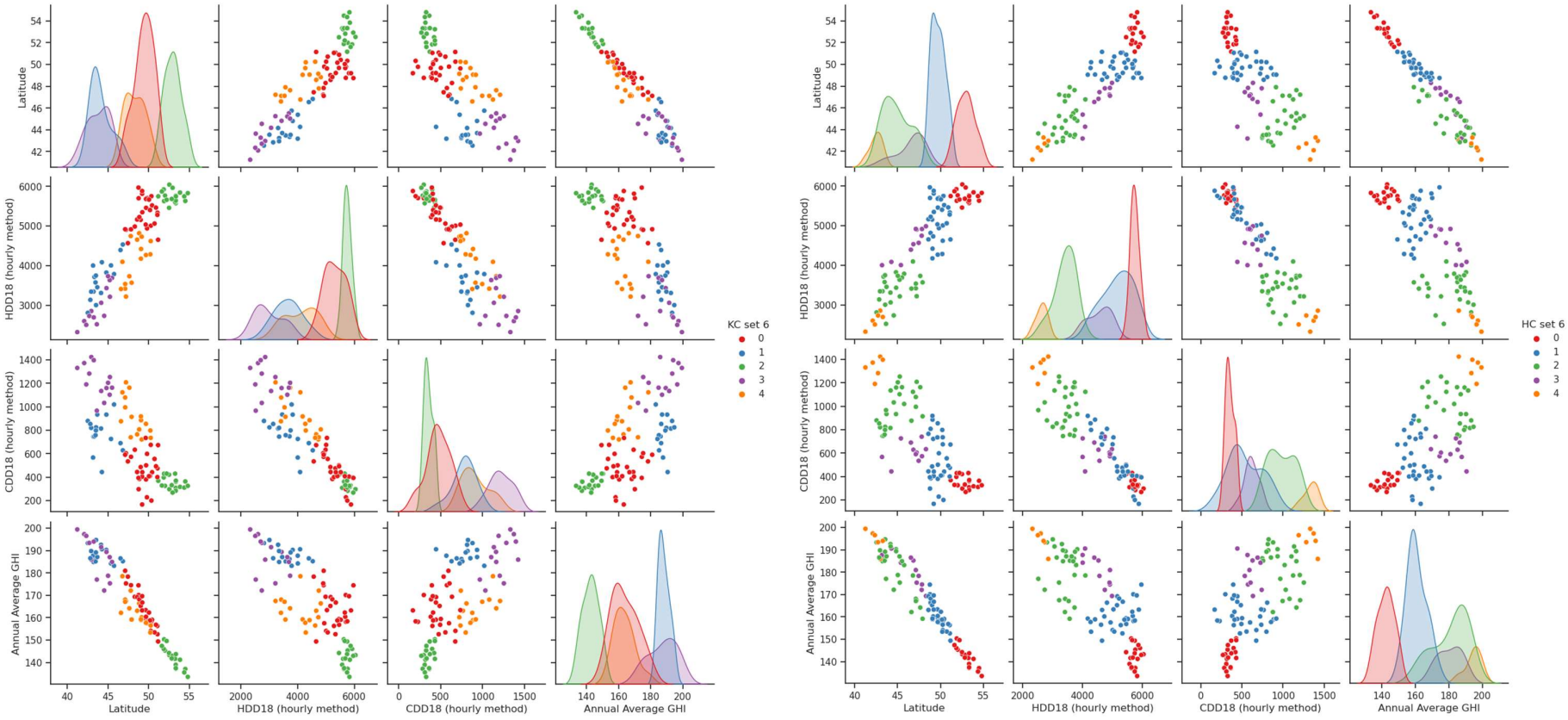
(g)


Latitude
HDD18 (hourly method)
HC set 4
0
1
2
3
4
5
6
7

(h)

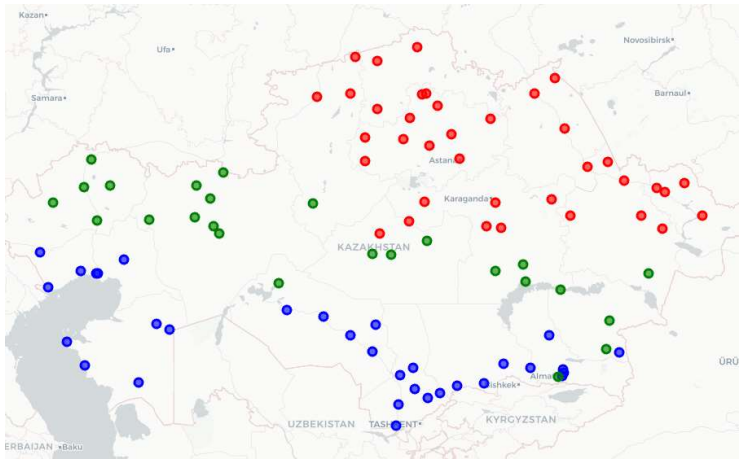

Latitude
HDD18 (hourly method)
CDD18 (hourly method)
KC set 5
0
1
2
HC set 5
0
1
2

(i) (j)

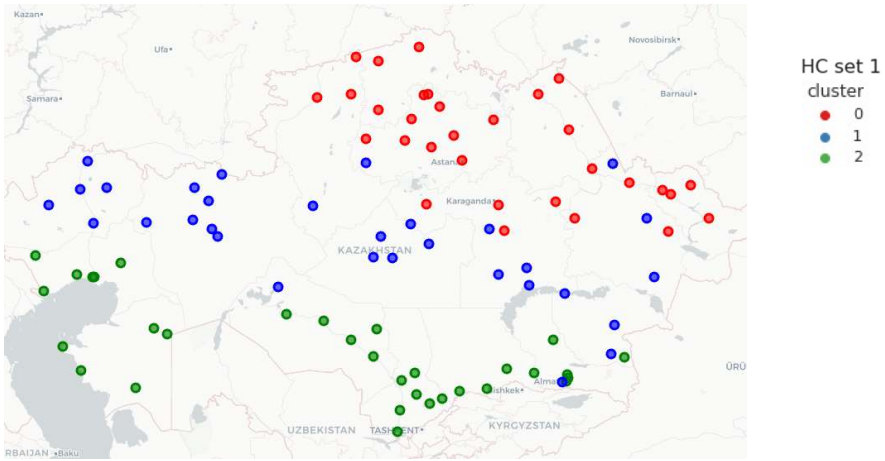


(k) (l)

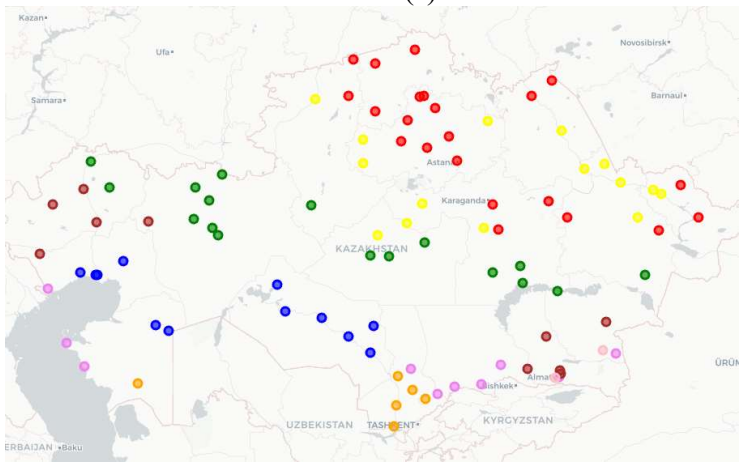
Figure 4.13: The CZB clustering scatterplot matrices of Phase 1. KC set 1 (a), HC set 1 (b), KC set 2 (c), HC set 2 (d), KC set 3 (e), HC set 3 (f), KC set 4 (SCKC) (g), HC set 4 (SCHC) (h), KC set 5 (SCKC) (i), HC set 5 (SCHC) (j), KC set 6 (SCKC) (k), HC set 6 (SCHC) (l).


Kazan
Ufa
Novosibirsk
Samara
Barnaul
Astana
Karaganda
KAZAKHSTAN
Almaty
Bishkek
UZBEKISTAN
TASHKENT
KYRGYZSTAN
KC set 1
cluster
0
1
2

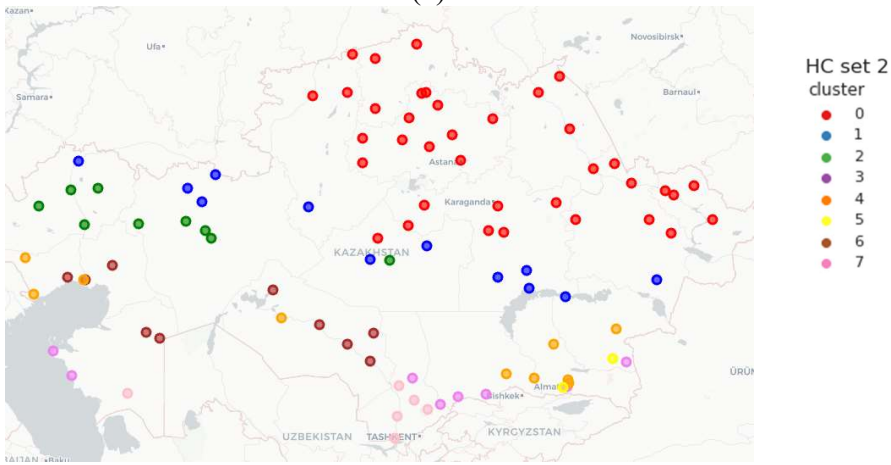
(a)


Kazan
Ufa
Novosibirsk
Samara
Barnaul
Astana
Karaganda
KAZAKHSTAN
Almaty
Bishkek
UZBEKISTAN
TASHKENT
KYRGYZSTAN
HC set 1
cluster
0
1
2

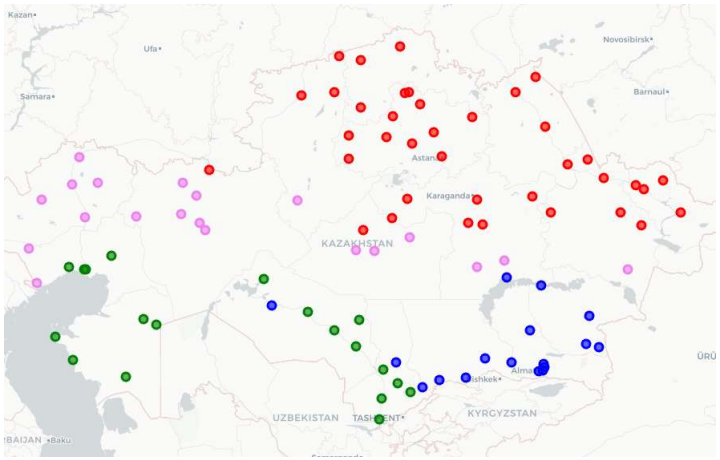
(b)


Kazan
Ufa
Novosibirsk
Samara
Barnaul
Astana
Karaganda
KAZAKHSTAN
Almaty
Bishkek
UZBEKISTAN
TASHKENT
KYRGYZSTAN
KC set 2
cluster
0
1
2
3
4
5
6
7

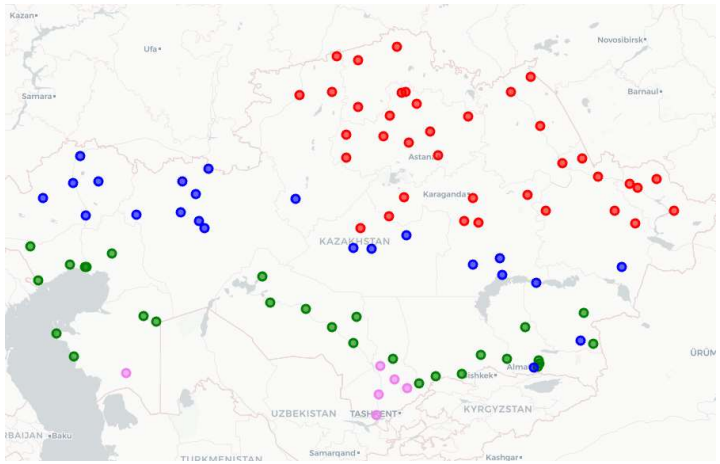
(c)


Kazan
Ufa
Novosibirsk
Samara
Barnaul
Astana
Karaganda
KAZAKHSTAN
Almaty
Bishkek
UZBEKISTAN
TASHKENT
KYRGYZSTAN
HC set 2
cluster
0
1
2
3
4
5
6
7

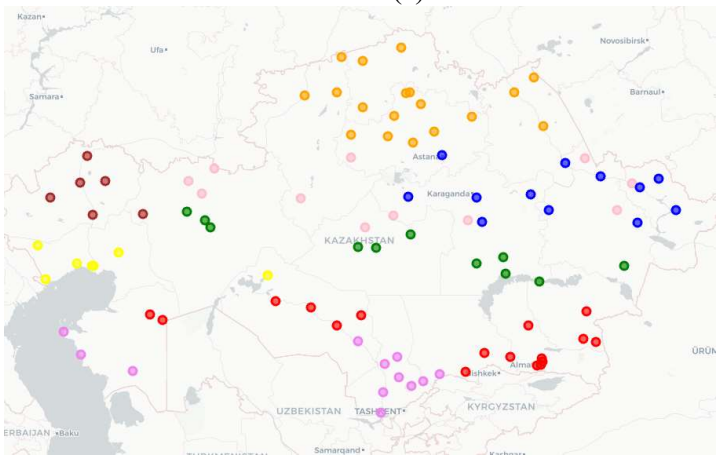
(d)


KC set 3
cluster
0
1
2
3
KAZAKHSTAN
UZBEKISTAN
KYRGYZSTAN

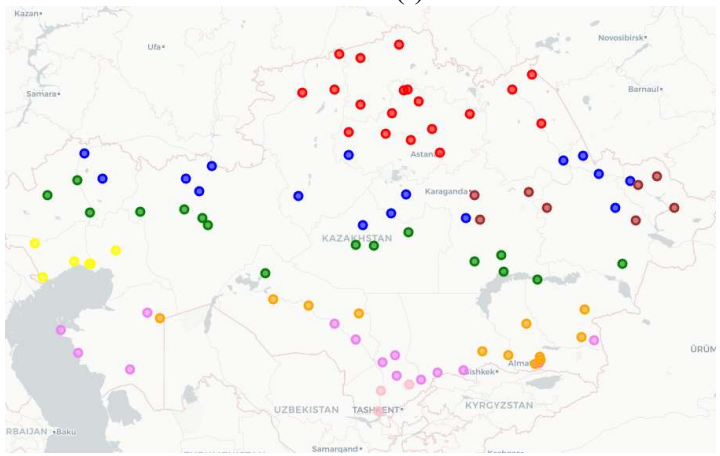
(e)


HC set 3
cluster
0
1
2
3
KAZAKHSTAN
UZBEKISTAN
KYRGYZSTAN

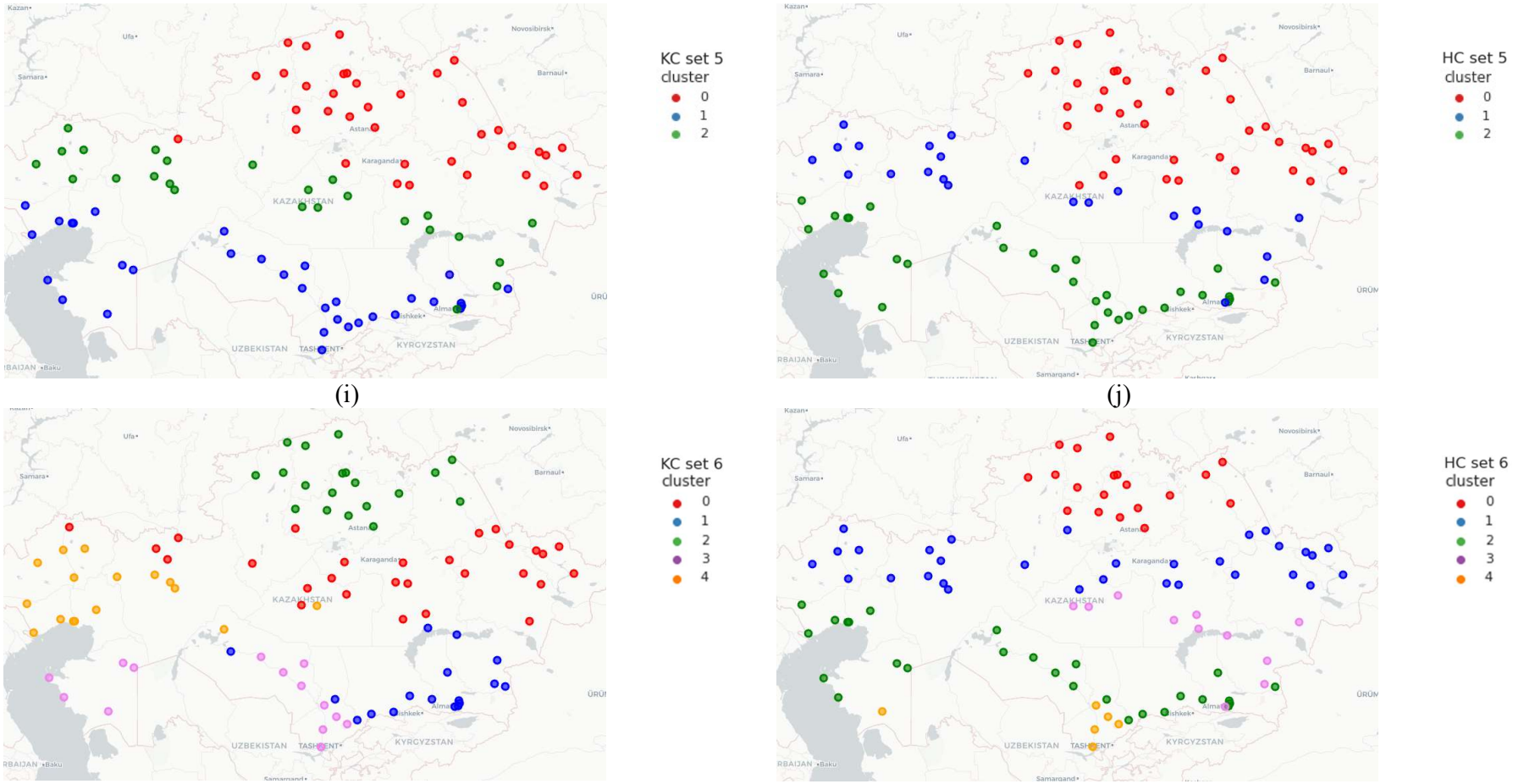
(f)


KC set 4
cluster
0
1
2
3
4
5
6
7
KAZAKHSTAN
UZBEKISTAN
KYRGYZSTAN

(g)


HC set 4
cluster
0
1
2
3
4
5
6
7
KAZAKHSTAN
UZBEKISTAN
KYRGYZSTAN

(h)



(i) (j)

(k) (l)

Figure 4.14: The maps of Phase 1 CZB clustering results. KC set 1 (a), HC set 1 (b), KC set 2 (c), HC set 2 (d), KC set 3 (e), HC set 3 (f), KC set 4 (SCKC) (g), HC set 4 (SCHC) (h), KC set 5 (SCKC) (i), HC set 5 (SCHC) (j), KC set 6 (SCKC) (k), HC set 6 (SCHC) (l).

The maps in Figure 4.13 show the clustering results of different methods and correspond to KC (a, c, e), HC (b, d, f), SCKC (g, i, k), and SCHC (h, j, l), respectively. It is seen how various methods produce different results. In the comparison between KC and HC for the same datasets across different regions of the country, notable differences emerge. In the northern region, both methods generally align, suggesting clear and distinct data characteristics easily captured by both algorithms. However, moving to the middle and southern regions, the differences become more pronounced. Here, HC often reveals more intricate and granular groupings, while KC tends to form broader, more uniform clusters. It's observed that the geographical shapes of the clusters also differ notably between the two methods. KC clusters tend to form more compact and geographically uniform groups, often centered around specific geographical locations, while HC clusters are often more dispersed and can follow more complex geographical shapes. Both clustering methods have a varied distribution of data points across clusters. In some sets (HC set 1), clusters are more evenly distributed (Figure 4.14 (b)), while in others (KC set 2, HC set 2), there is a wide range in the number of points per cluster (Figure 4.14 (c, d). This variation reflects the different ways each method groups the data, with KC tending to create more uniformly sized clusters and HC potentially capturing more diverse group sizes.

The maps also illustrate the geographical distribution of clusters for SCKC and SCHC. The distinction between SCKC and SCHC is applicable in this context as well, where SCHC typically uncovers more detailed and fine-grained groupings, in contrast to SCKC, which usually generates larger, more homogenous clusters. Clusters created by SCKC are generally more cohesive and geographically consistent, whereas SCHC-generated clusters tend to be more scattered, adhering to more intricate geographical patterns.

It is quite difficult to directly visually compare the SCKC and SCHC with their non-spatially constrained relatives due to the different number of clusters in the final classifications, however, a more detailed analysis of the efficiency of spatially constrained methods will be presented in section 4.5.4 where building performance-based validation will be discussed.

4.5.3. **Clustering results quality assessment**

This section aims to assess the quality of the clustering results achieved through various clustering methods for climate-based climate classification of Phase 1. This evaluation is conducted using a carefully chosen set of measures. Specifically, through three key dimensions: Uniqueness, as described in subsection 4.4.3.1, evaluates the distinctiveness of each cluster; Compactness, outlined in subsection 4.4.3.2, examines the tightness of the

clustering; and SS, explored in subsection 4.4.3.3, provides a comprehensive measure of both cohesion and separation within the clusters.

4.5.3.1. **Uniqueness**

In evaluating the climate-based clustering of Phase 1, a detailed analysis focusing on the mean and standard deviation of uniqueness values is conducted. The method yielding the highest mean uniqueness value is indicative of its superior performance in differentiating clusters, while a lower standard deviation reflects consistency across various CZs.

Figure 4.15 represents the heatmap of uniqueness percentage for each clustering method across different CZ, with the colour gradient indicating the level of uniqueness. Among all the methods, KC set 2, HC set 2, KC set 4, and HC set 4 emerge as the best performers with the highest mean uniqueness values of 87.50%, indicating their effectiveness in creating distinct clusters. Conversely, KC set 1, HC set 1, KC set 5, and HC set 5 display the lowest mean uniqueness values at 66.67%, suggesting less differentiation within their respective clustering outcomes.

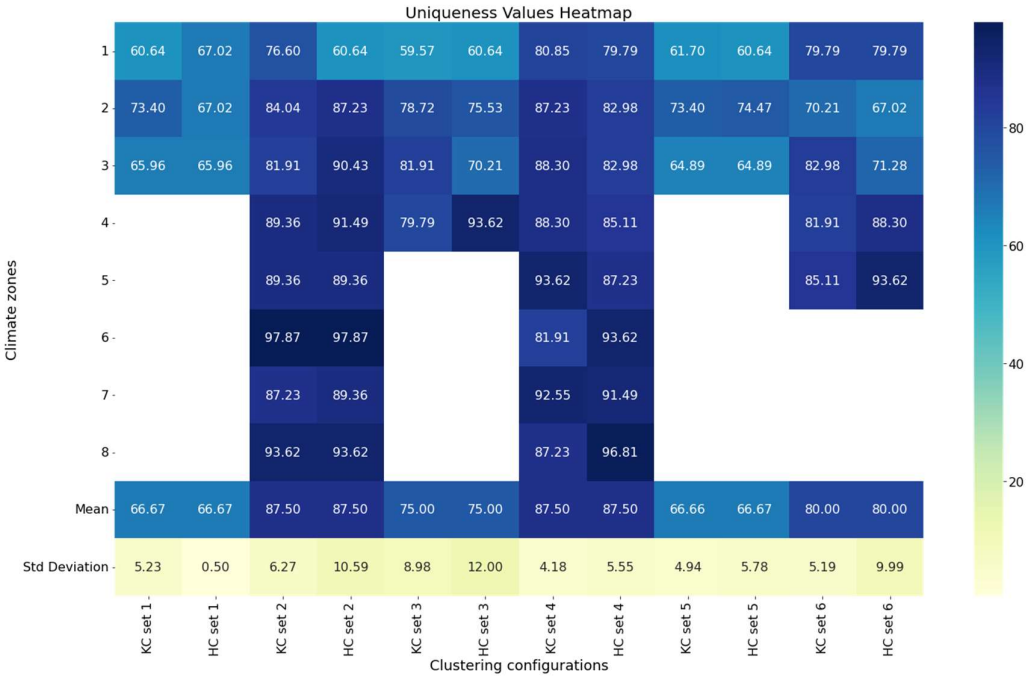


Figure 4.15: The uniqueness heatmap of Phase 1 clustering results.

A comparative analysis between KC and HC methods reveals that KC set 4 and HC set 4 are also the leading methods within their categories. However, KC set 4 demonstrates a slightly lower standard deviation (4.18%) compared to HC set 4 (5.55%), indicating lower variability in the uniqueness across CZs for the KC method. The same methods show the

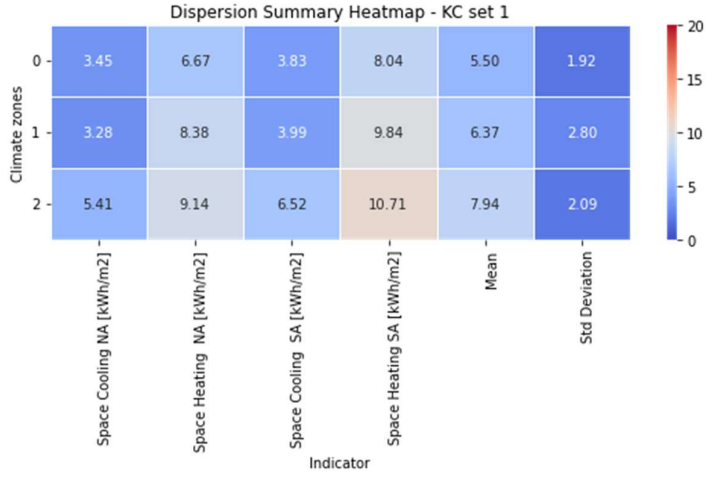
highest mean uniqueness values alongside a moderate standard deviation in the context of spatially constrained clustering.

Besides, all clustering methods display the same mean uniqueness values for the same datasets. This phenomenon is attributed to the limitations of uniqueness itself in capturing the nuances between different clustering approaches, and small variations in how clusters are generated may not be accurately measured.

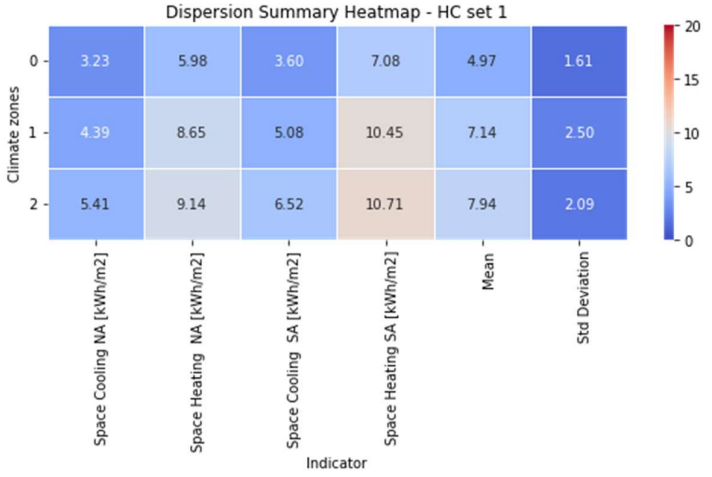
4.5.3.2. **Compactness**

Compactness as a quality measure provides a clear and quantitative measure of the degree to which data points within a cluster deviate from their cluster mean, thus reflecting the compactness of the clusters formed by different clustering algorithms. Based on the MAE of each building performance indicator within individual clusters, this metric is particularly valuable in discerning the effectiveness of clustering techniques in creating tightly grouped data points in a building performance dimension.

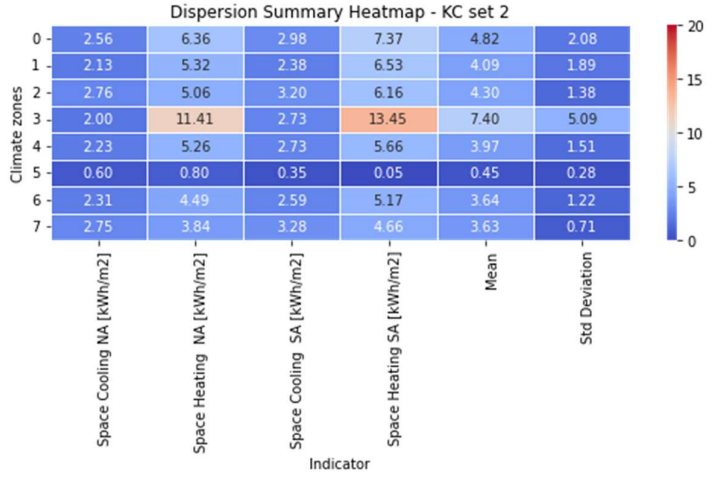
Figure 4.16 visually displays the dispersion values calculated for each clustering method of Phase 1, providing insights into their performance. The classification with the lowest mean dispersion value and low standard deviation values, such as HC set 2 (Figure 4.16 (d)), KC set 4 (Figure 4.16 (g)), and HC set 4 (Figure 4.16 (h)) have been considered the best in terms of cluster compactness. This indicates a high consistency within the clusters formed by these methods. In contrast, the methods with the highest mean dispersion, and therefore the least effective in creating compact clusters, are KC set 3 (Figure 4.16 (e)), KC set 5 (Figure 4.16 (i)), HC set 5 (Figure 4.16 (j)).


Dispersion Summary Heatmap - KC set 1
Climate zones
0
1
2
3.45
6.67
3.83
8.04
5.50
1.92
3.28
8.38
3.99
9.84
6.37
2.80
5.41
9.14
6.52
10.71
7.94
2.09
Space Cooling NA [kWh/m2]
Space Heating NA [kWh/m2]
Space Cooling SA [kWh/m2]
Space Heating SA [kWh/m2]
Mean
Std Deviation
Indicator
20
15
10
5
0

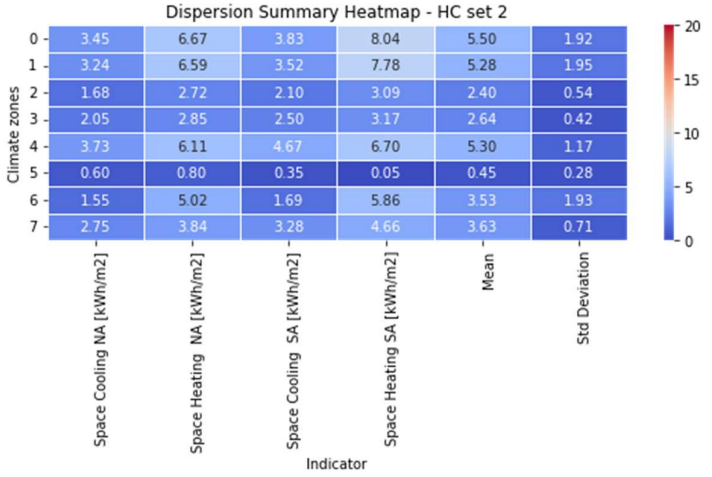
(a)


Dispersion Summary Heatmap - HC set 1
Climate zones
0
1
2
3.23
5.98
3.60
7.08
4.97
1.61
4.39
8.65
5.08
10.45
7.14
2.50
5.41
9.14
6.52
10.71
7.94
2.09
Space Cooling NA [kWh/m2]
Space Heating NA [kWh/m2]
Space Cooling SA [kWh/m2]
Space Heating SA [kWh/m2]
Mean
Std Deviation
Indicator
20
15
10
5
0

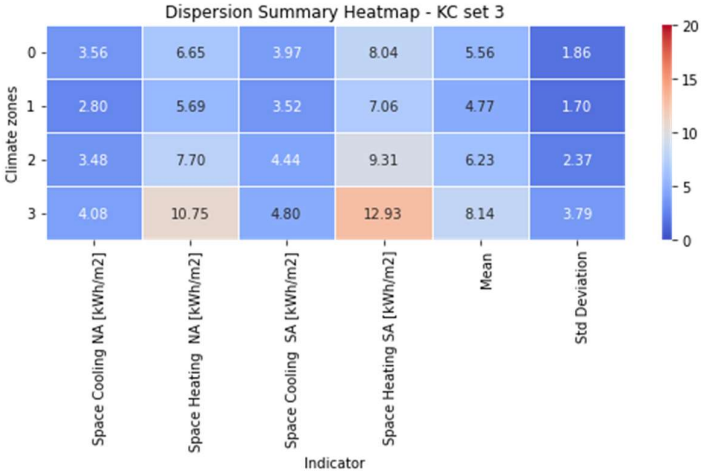
(b)


Dispersion Summary Heatmap - KC set 2
Climate zones
0
1
2
3
4
5
6
7
2.56
6.36
2.98
7.37
4.82
2.08
2.13
5.32
2.38
6.53
4.09
1.89
2.76
5.06
3.20
6.16
4.30
1.38
2.00
11.41
2.73
13.45
7.40
5.09
2.23
5.26
2.73
5.66
3.97
1.51
0.60
0.80
0.35
0.05
0.45
0.28
2.31
4.49
2.59
5.17
3.64
1.22
2.75
3.84
3.28
4.66
3.63
0.71
Space Cooling NA [kWh/m2]
Space Heating NA [kWh/m2]
Space Cooling SA [kWh/m2]
Space Heating SA [kWh/m2]
Mean
Std Deviation
Indicator
20
15
10
5
0

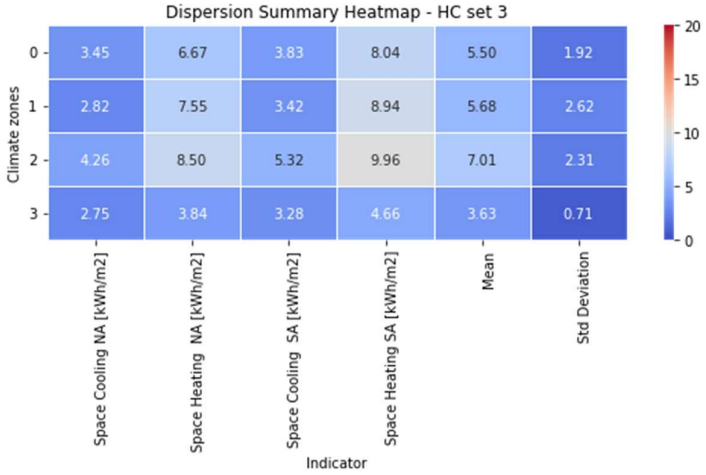
(c)


Dispersion Summary Heatmap - HC set 2
Climate zones
0
1
2
3
4
5
6
7
3.45
6.67
3.83
8.04
5.50
1.92
3.24
6.59
3.52
7.78
5.28
1.95
1.68
2.72
2.10
3.09
2.40
0.54
2.05
2.85
2.50
3.17
2.64
0.42
3.73
6.11
4.67
6.70
5.30
1.17
0.60
0.80
0.35
0.05
0.45
0.28
1.55
5.02
1.69
5.86
3.53
1.93
2.75
3.84
3.28
4.66
3.63
0.71
Space Cooling NA [kWh/m2]
Space Heating NA [kWh/m2]
Space Cooling SA [kWh/m2]
Space Heating SA [kWh/m2]
Mean
Std Deviation
Indicator
20
15
10
5
0

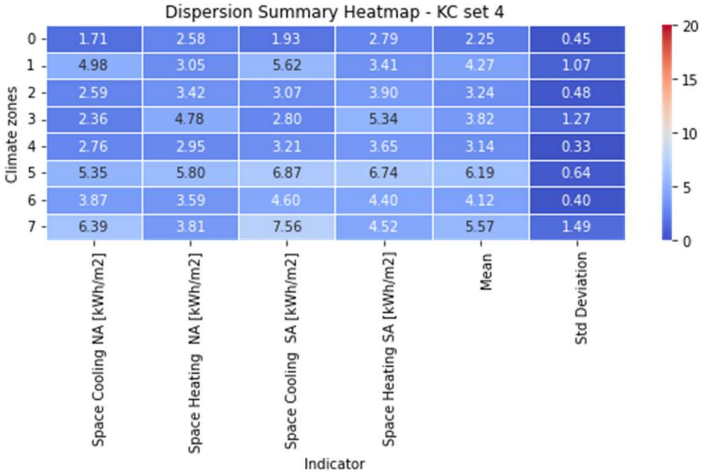
(d)


Dispersion Summary Heatmap - KC set 3
Climate zones
0 3.56 6.65 3.97 8.04 5.56 1.86
1 2.80 5.69 3.52 7.06 4.77 1.70
2 3.48 7.70 4.44 9.31 6.23 2.37
3 4.08 10.75 4.80 12.93 8.14 3.79
Space Cooling NA [kWh/m2]
Space Heating NA [kWh/m2]
Space Cooling SA [kWh/m2]
Space Heating SA [kWh/m2]
Mean
Std Deviation
Indicator
20
15
10
5
0

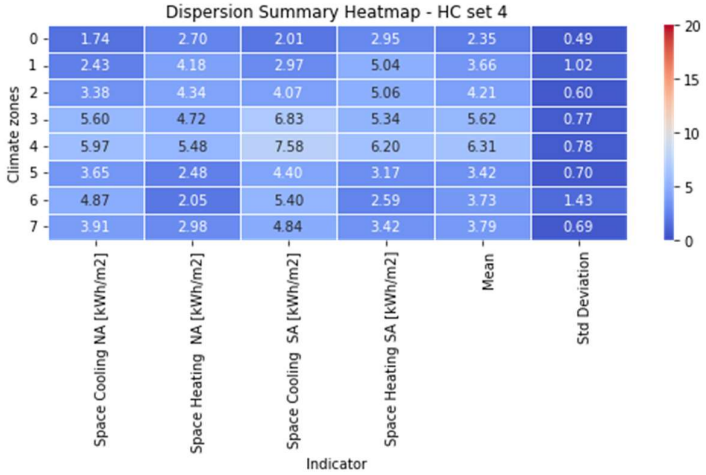
(e)


Dispersion Summary Heatmap - HC set 3
Climate zones
0 3.45 6.67 3.83 8.04 5.50 1.92
1 2.82 7.55 3.42 8.94 5.68 2.62
2 4.26 8.50 5.32 9.96 7.01 2.31
3 2.75 3.84 3.28 4.66 3.63 0.71
Space Cooling NA [kWh/m2]
Space Heating NA [kWh/m2]
Space Cooling SA [kWh/m2]
Space Heating SA [kWh/m2]
Mean
Std Deviation
Indicator
20
15
10
5
0

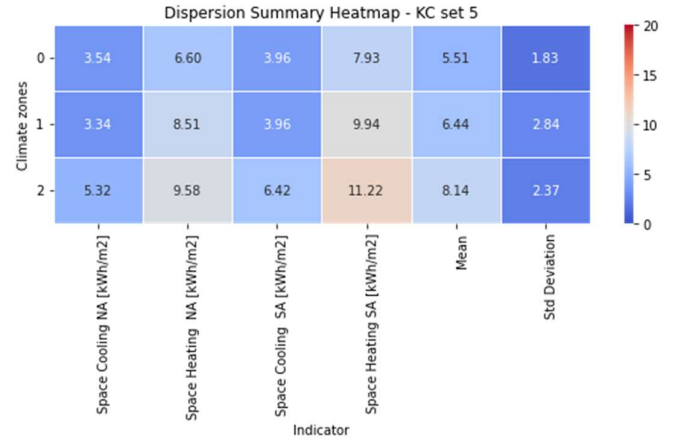
(f)


Dispersion Summary Heatmap - KC set 4
Climate zones
0 1.71 2.58 1.93 2.79 2.25 0.45
1 4.98 3.05 5.62 3.41 4.27 1.07
2 2.59 3.42 3.07 3.90 3.24 0.48
3 2.36 4.78 2.80 5.34 3.82 1.27
4 2.76 2.95 3.21 3.65 3.14 0.33
5 5.35 5.80 6.87 6.74 6.19 0.64
6 3.87 3.59 4.60 4.40 4.12 0.40
7 6.39 3.81 7.56 4.52 5.57 1.49
Space Cooling NA [kWh/m2]
Space Heating NA [kWh/m2]
Space Cooling SA [kWh/m2]
Space Heating SA [kWh/m2]
Mean
Std Deviation
Indicator
20
15
10
5
0

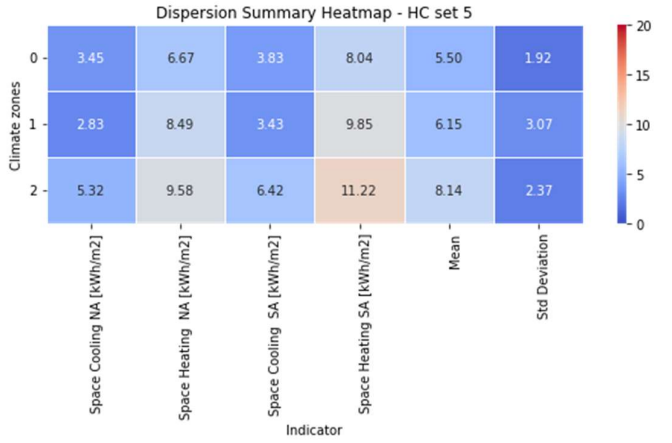
(g)


Dispersion Summary Heatmap - HC set 4
Climate zones
0 1.74 2.70 2.01 2.95 2.35 0.49
1 2.43 4.18 2.97 5.04 3.66 1.02
2 3.38 4.34 4.07 5.06 4.21 0.60
3 5.60 4.72 6.83 5.34 5.62 0.77
4 5.97 5.48 7.58 6.20 6.31 0.78
5 3.65 2.48 4.40 3.17 3.42 0.70
6 4.87 2.05 5.40 2.59 3.73 1.43
7 3.91 2.98 4.84 3.42 3.79 0.69
Space Cooling NA [kWh/m2]
Space Heating NA [kWh/m2]
Space Cooling SA [kWh/m2]
Space Heating SA [kWh/m2]
Mean
Std Deviation
Indicator
20
15
10
5
0

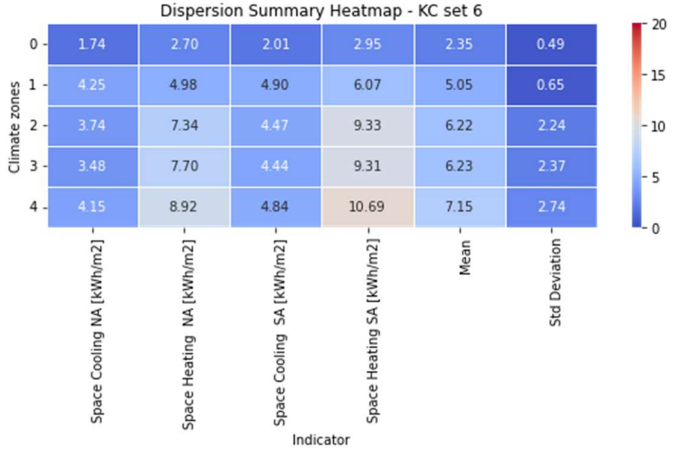
(h)



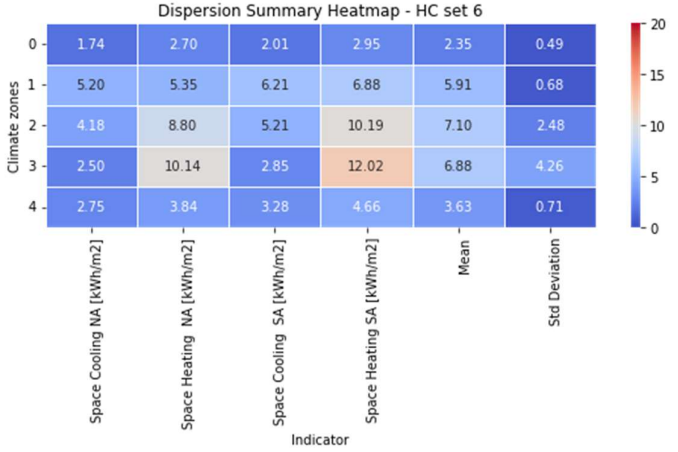
(i)



(j)



(k)



(l)

Figure 4.16: The dispersion values for each Phase 1 clustering method. KC set 1 (a), HC set 1 (b), KC set 2 (c), HC set 2 (d), KC set 3 (e), HC set 3 (f), KC set 4 (SCKC) (g), HC set 4 (SCHC) (h), KC set 5 (SCKC) (i), HC set 5 (SCHC) (j), KC set 6 (SCKC) (k), HC set 6 (SCHC) (l).

Comparing the overall performance, HC methods (HC set 2, HC set 4) demonstrate lower mean dispersion values compared to KC methods, indicating tighter and more consistent clusters. In the spatially constrained category, the KC set 4 and HC set 4 emerge as the most efficient, reflecting their capability to form more compact clusters. Also, no significant difference was found between SCKC and SCHC.

4.5.3.3. **The Silhouette Score**

In assessing the clustering quality across 12 different clustering results of Phase 1, the SS analysis utilized four principal features: space cooling NA, space heating NA, space cooling SA, and space heating SA. The analysis uncovered significant differences in the quality of clustering, identifying certain configurations as highly effective in delineating separate clusters, while others were less successful. The SS results are shown in Figure 4.17, where the KC set 1, HC set 5, KC set 5, HC set 1, and HC set 3 configurations emerged as the best overall performers with SS of 0.44, 0.44, 0.43, 0.42, and 0.39 respectively, exemplifying optimal clustering efficiency. The extremely small difference in the SS among top-performing methods suggests that they can also demonstrate similar quality. In contrast, KC set 2 and KC set 4 recorded the lowest SS of 0.17 and 0.19 revealing considerable cluster overlap and potential misclassification of data points.

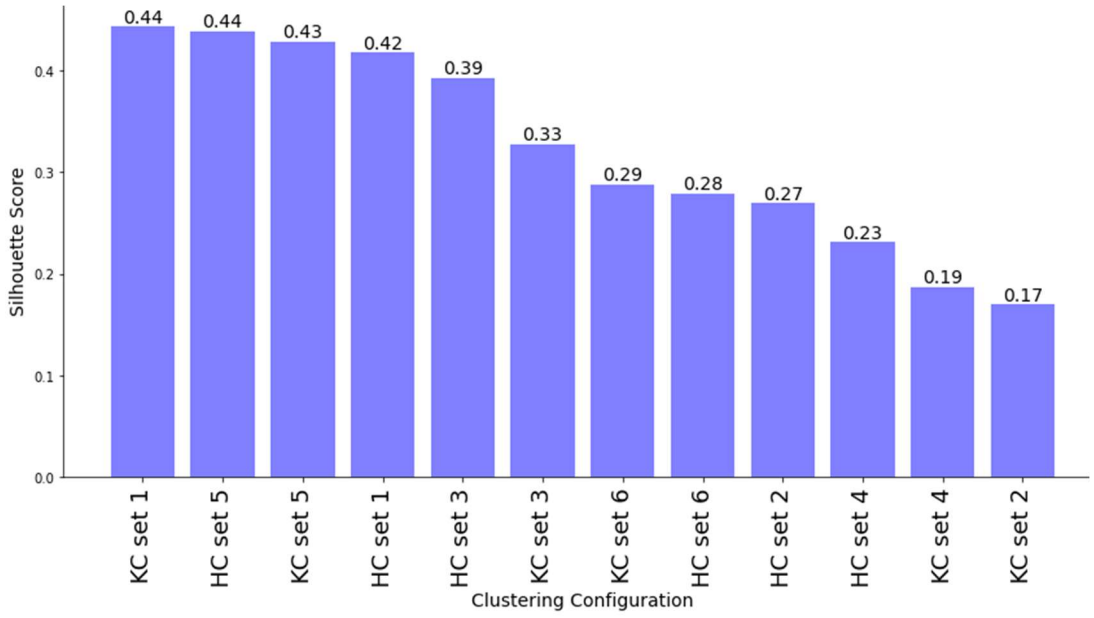


Figure 4.17: The SS results of each clustering method of Phase 1.

Within the specific clustering methods, the top performers were: for non-spatial variants - KC set 1 with a score of 0.44, and HC set 1 with a score of 0.42; for spatially constrained - HC set 5 with a score of 0.44, and KC set 5 with 0.43 SS score. No significant difference was found between KC and its HC variant except in set 2, where HC outperformed the KC method significantly. Also, no stable dominance of spatial methods over non-spatial ones has been found.

4.5.4. **Building performance-based validation**

The performance-based validation focuses on evaluating the distinctiveness of clusters within different building archetypes, and performance indicators. This is achieved by calculating the overlap in heating and cooling energy needs between clusters for each clustering method within each archetype, providing insights into the effectiveness of various clustering methods.

Using Python script, 48 overlap graphs for each clustering, building performance indicator, and archetype were obtained, visually representing the amount of overlap for different clustering methods. For the sake of conserving space, only 4 overlap graphs are depicted in Figure 4.18. These graphs once again give an idea of the essence of the proposed CZMI, where the average value of cluster overlaps is taken as a quantitative indicator of the quality of the clustering results. When analyzing the results of the CZMI, it is important to avoid being misled by its values. The index is not a direct percentage of the overlap but rather an adjusted value that not only takes the overlap into account but also considers the degree of intra-cluster separation.

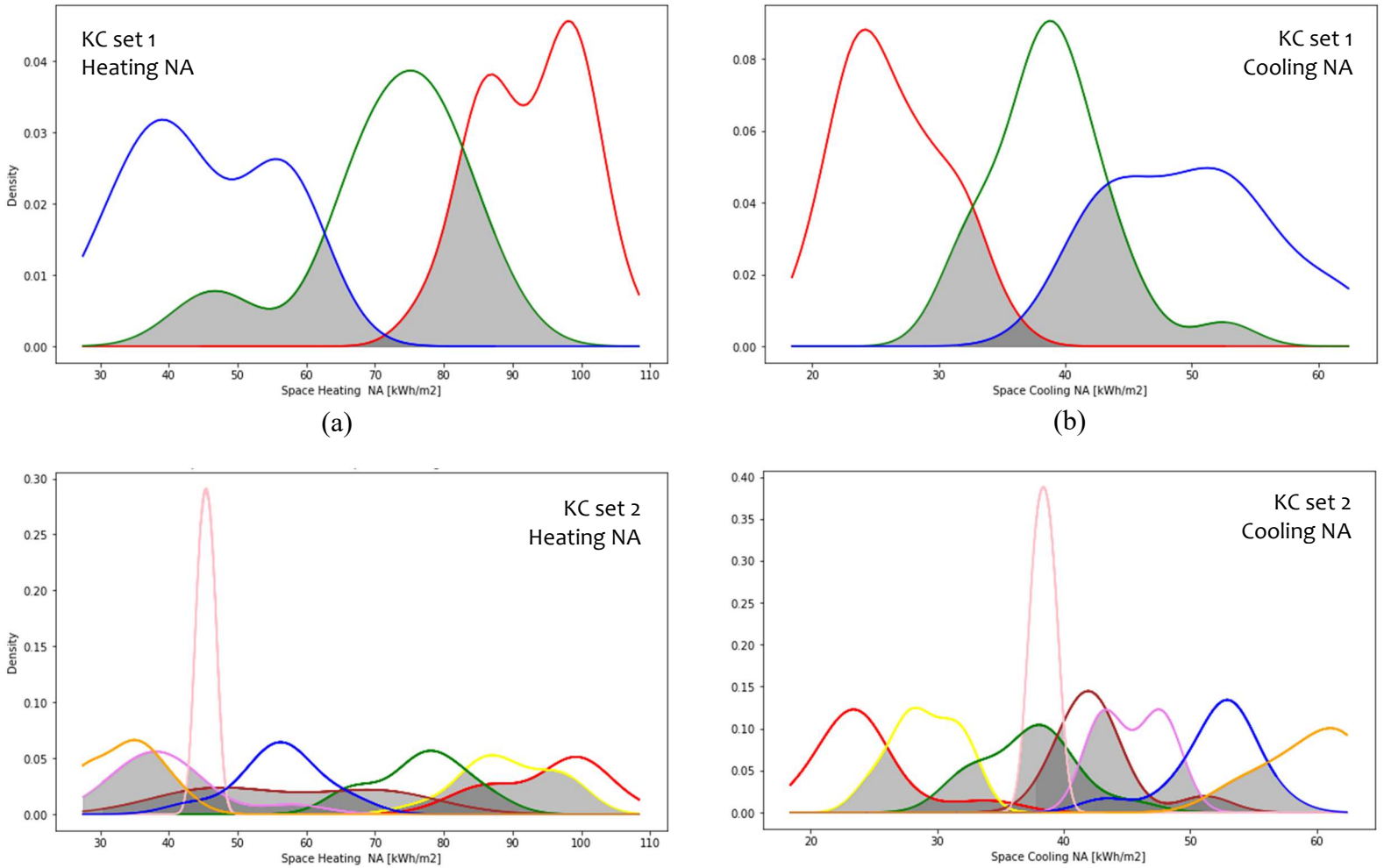


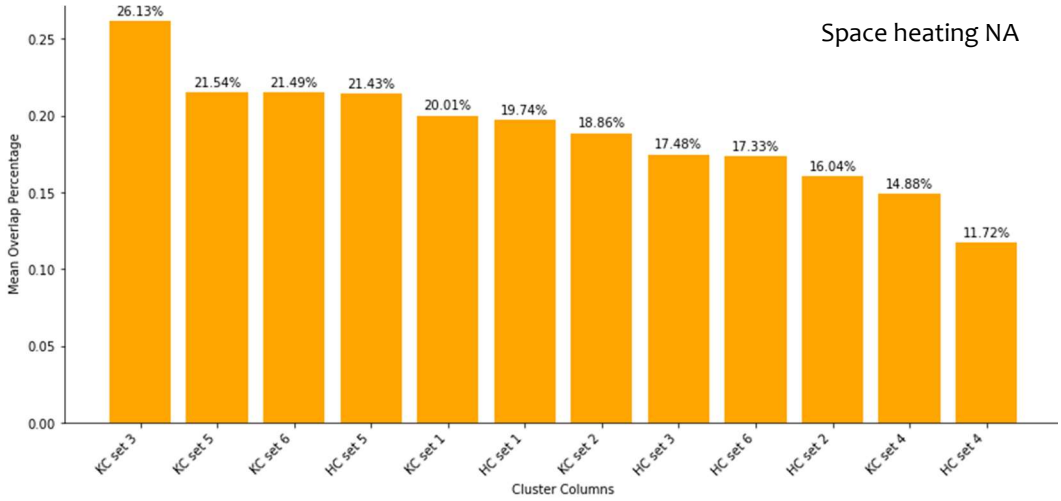
Figure 4.18: KDE overlap between clusters for KC set 1 based on space heating NA (a), and space cooling NA (b), KC set 2 based on space heating NA (c), and space cooling NA (d).

A specialized Python script was developed to calculate the CZMI (Appendix C). The computational technique implemented in the script can be outlined as follows:

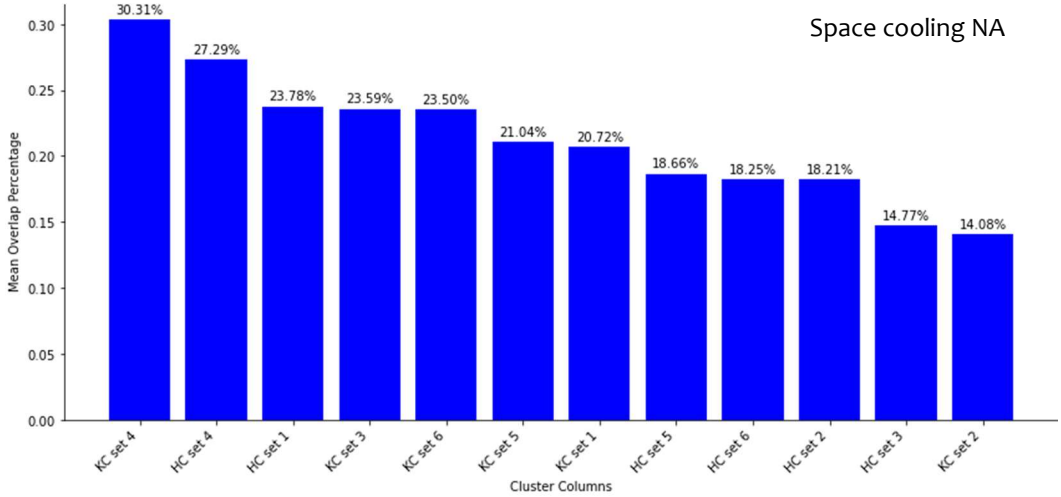
- The code initiates by importing a dataset that includes clustering results for each method (cluster labels) and energy consumption data. The “calculate_overlap_area” function calculates the overlap area between two KDE curves for pairs of cluster labels for performance indicators for each archetype. After that, the total KDE for the entire performance range is calculated. Overlap percentages are defined as the proportion of overlap area and total KDE. At the end of that stage, the mean overlap percentages are calculated (Appendix C (a)).
- Euclidean distances between centroids of each pair of clusters are calculated. These distances are normalized within each cluster method so that they can be compared on a common scale. The overlap percentages are adjusted using the normalized distances. (Appendix C (b)).

The results of this computational analysis are visually represented through a series of bar charts in Figure 4.19, each serving a distinct purpose in elucidating the findings. For both NA and SA the range of mean overlap percentage values for each performance indicator is from 10.90 to 26.13% for heating and from 13.10 to 30.55 % for cooling performance indicator. Also, the overlap results are almost identical for NA and SA for the same performance indicators, the difference on average does not exceed 2 % for the same clustering results.

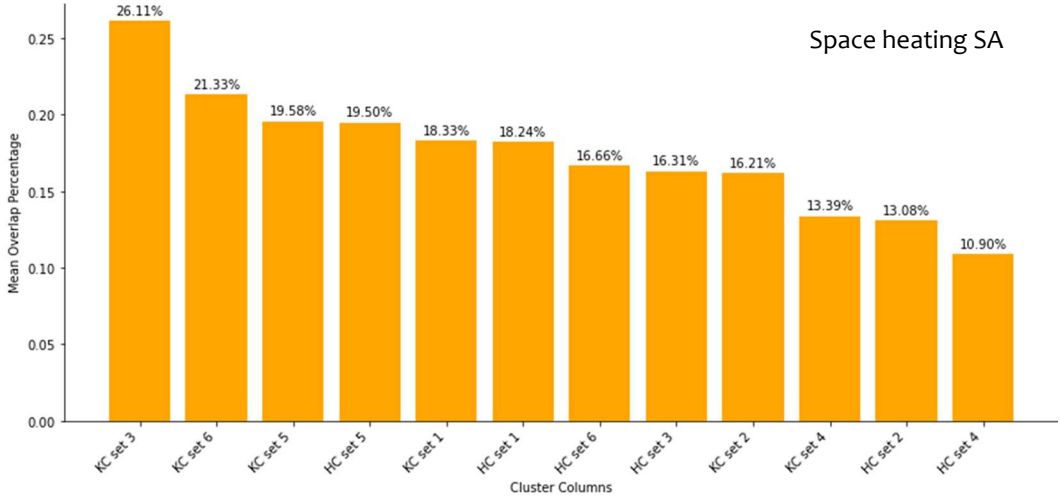
Considering each indicator separately, within space heating (Figure 4.19 (a, c)), HC set 4 emerged as the most effective method, achieving the lowest mean overlap percentage values (11.72% for NA and 10.90% for SA). Conversely, KC set 3 exhibits the highest mean overlap values (26.13 and 26.11% for NA and SA respectively), thus signaling less efficacy in distinguishing clusters. For space cooling, the method with the lowest mean overlap is KC set 2 (14.08 for NA and 13.10% for SA). It is also evident that the clustering methods that yield optimal outcomes for heating are, as anticipated, the least effective for cooling.


Space heating NA
Mean Overlap Percentage
0.00
0.05
0.10
0.15
0.20
0.25
26.13%
21.54%
21.49%
21.43%
20.01%
19.74%
18.86%
17.48%
17.33%
16.04%
14.88%
11.72%
KC set 3
KC set 5
KC set 6
HC set 5
KC set 1
HC set 1
KC set 2
HC set 3
HC set 6
HC set 2
KC set 4
HC set 4
Cluster Columns

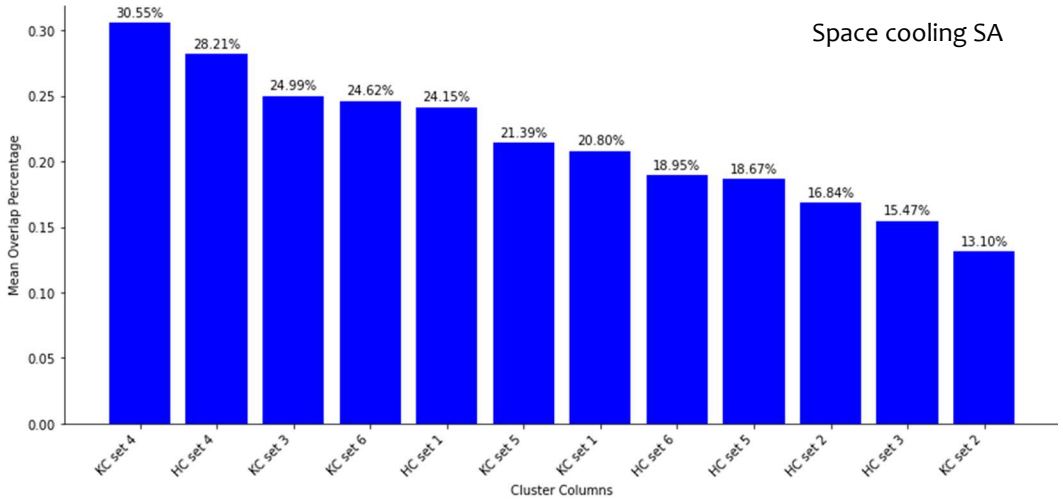
(a)


Space cooling NA
Mean Overlap Percentage
0.00
0.05
0.10
0.15
0.20
0.25
0.30
30.31%
27.29%
23.78%
23.59%
23.50%
21.04%
20.72%
18.66%
18.25%
18.21%
14.77%
14.08%
KC set 4
HC set 4
HC set 1
KC set 3
KC set 6
KC set 5
KC set 1
HC set 5
HC set 6
HC set 2
HC set 3
KC set 2
Cluster Columns

(b)


Space heating SA
Mean Overlap Percentage
0.00
0.05
0.10
0.15
0.20
0.25
26.11%
21.33%
19.58%
19.50%
18.33%
18.24%
16.66%
16.31%
16.21%
13.39%
13.08%
10.90%
KC set 3
KC set 6
KC set 5
HC set 5
KC set 1
HC set 1
HC set 6
HC set 3
KC set 2
KC set 4
HC set 2
HC set 4
Cluster Columns

(c)



(d)

Fig 4.19: Mean overlap percentage values of Phase 1 clustering results for space heating NA (a), space cooling NA (b), space heating SA (c), and space cooling SA (d).

In the domain of CZMI analysis, the evaluation of clustering methods reveals that HC set 5 (based on HDD18 (hourly method), CDD18 (hourly method) and LAT with 3 clusters) yielded the most favourable results with CZMI of 9.42% (Figure 4.20). The second best option is KC set 1 (based on HDD18 (hourly method) only with 3 clusters). Overall, the gradually increasing values of the CZMI make a range from 9.42 to 17.07%, with KC set 4 emerging as the worst option, exhibiting the highest CZMI (17.07%) among all considered clustering methods, and suggesting less distinct classification and higher misclassification.

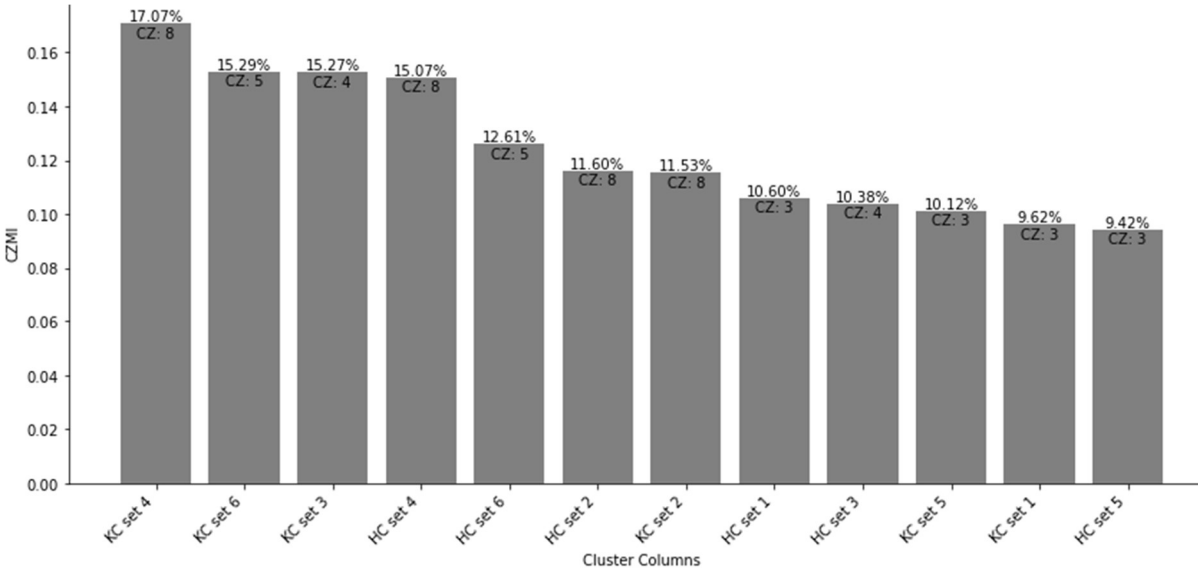


Figure 4.20: CZMI values of Phase 1 clustering methods.

Comparing KC with HC methods, among HC methods, HC set 5 has the lowest CZMI (9.42) underscoring its effectiveness in creating well-separated clusters. Among KC methods, KC set 1 emerged as the most effective, indicating its ability to minimize overlaps between clusters. Overall HC methods have a lower average CZMI (11.95) compared to KC clustering (12.82). This suggests that HC on average is generally more effective than KC, also out of the four poorest results, three were acquired using KC.

It's not possible to claim confidently that spatially constrained techniques always have lower misclassification than non-constrained ones. In some instances, non-spatial clustering approaches tend to yield lower CZMI values (KC set 1, HC set 3), suggesting a potential preference for non-spatial approaches when aiming to minimize CZMI. Additionally, there is a substantial variance in outcomes for both approaches. For example, HC set 4 has a relatively high CZMI of 15.07, but HC set 5 (also spatially constrained), has the lowest scores of 9.42. It seems that the outcome is instead influenced by the number of clusters and particular variables in the dataset, rather than the method itself. It is noteworthy that within spatially constrained approaches HC consistently outperforms KC in terms of CZMI.

The average CZMI values increase with the number of CZs, with 3 zones having the lowest average CZMI (9.94), followed by 4 zones (12.83), 5 zones (13.96), and 8 zones (13.82). This pattern suggests that a lower number of zones might contribute to achieving lower CZMI values, indicating an inverse relationship between the number of zones and CZMI.

It is observed that sets employing a single variable exhibit the lowest mean CZMI, recorded at 10.11. This suggests that a minimalist approach in variable selection could be advantageous in achieving lower CZMI values. For example, sets with just HDD have lower CZMI scores (9.62, 10.6) compared to sets that include HDD, CDD, GHI, and Latitude (e.g., KC set 6 with CZMI of 15.29). This indicates that an increase in the number of variables may not linearly correlate with the optimization of climate zoning classifications. This finding implies a nuanced relationship between the number of variables and the resultant CZMI, suggesting that a balanced selection of variables, specifically three, may offer a more optimized approach toward climate zoning.

Visually the overlaps of best and worst methods can be seen in Figure 4.21. The low CZMI of these methods suggests a high degree of precision in segregating clusters, implying that it can accurately differentiate between various performance indicators prevalent in both archetypes.

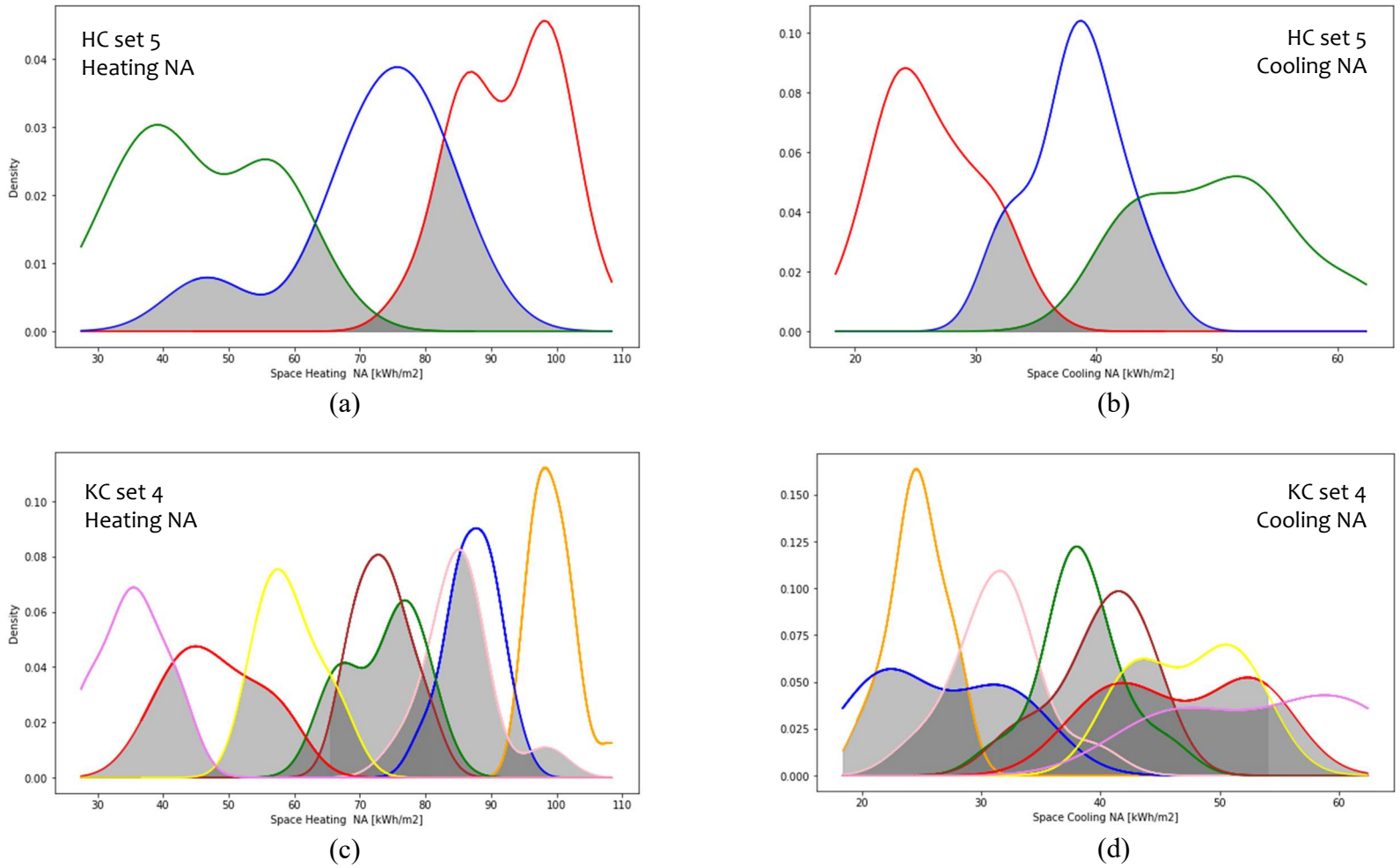


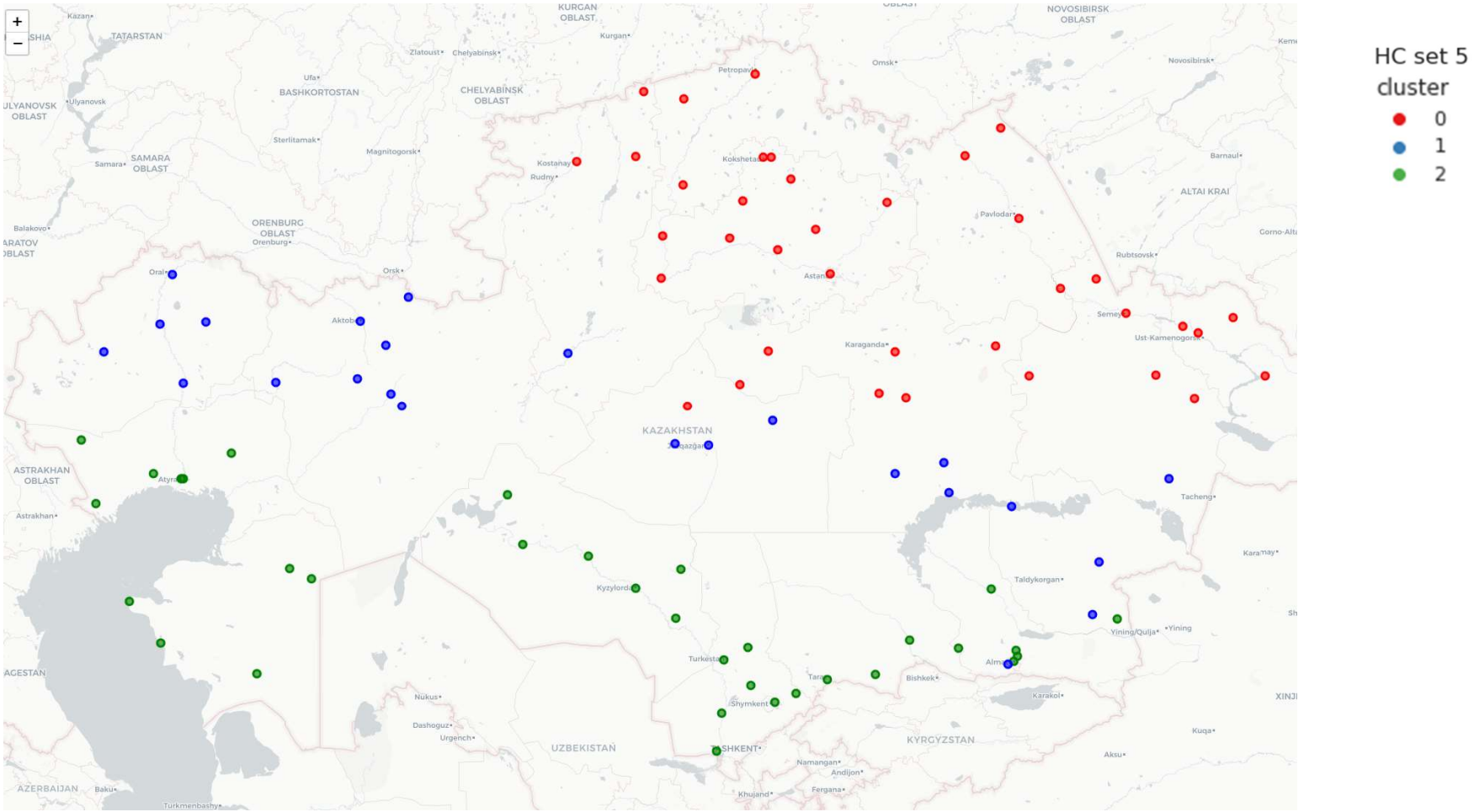
Figure 4.21: Overlap graphs of Phase 1 clustering methods with highest (a, b) and lowest (c, d) CZMI. HC set 5 for space heating NA (a), HC set 5 for space cooling NA (b), KC set 4 for space heating NA (c), and KC set 4 for space cooling NA (d).

4.5.5. **Summary of Phase 1 findings**

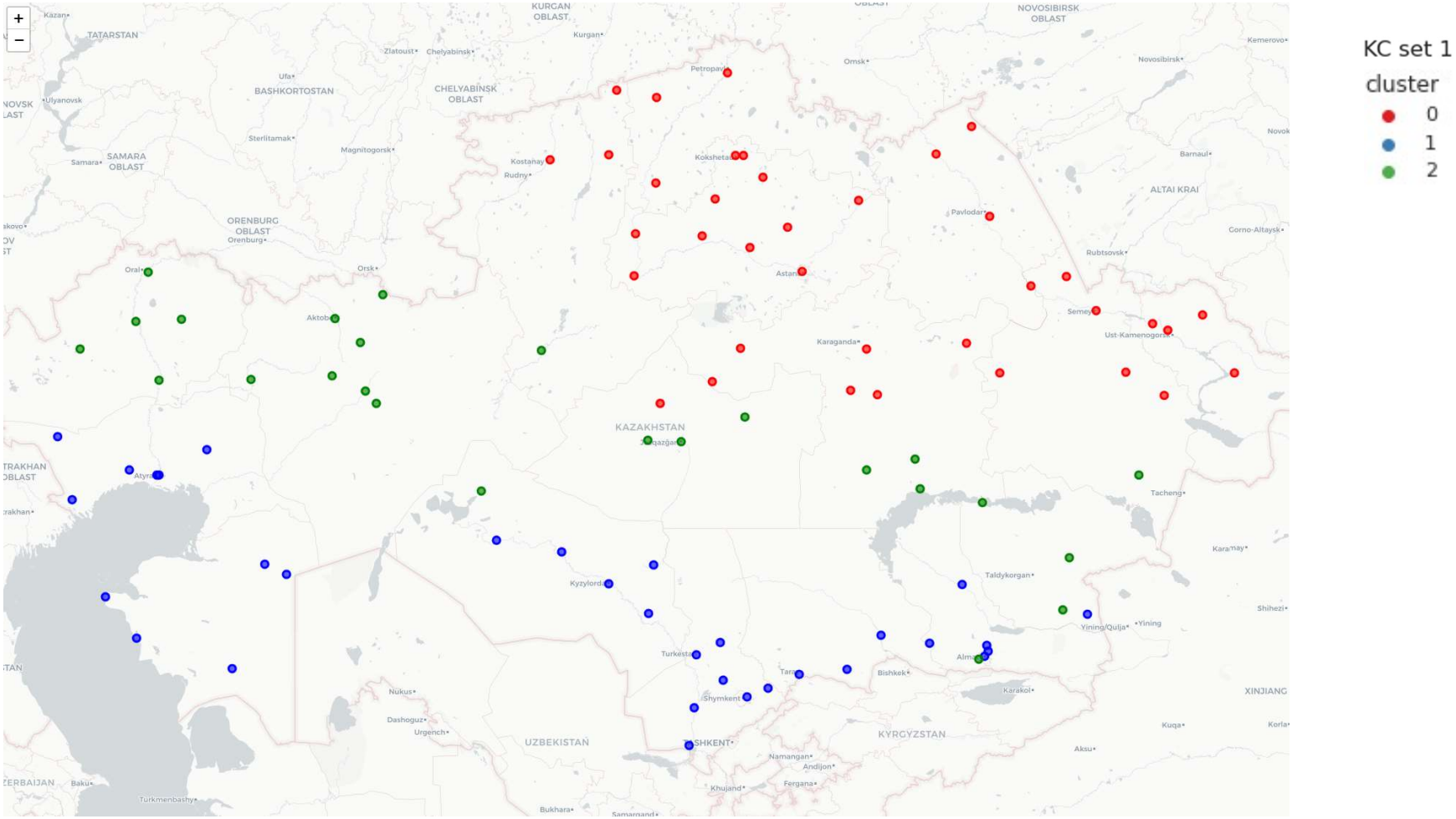
Phase 1 of the research delves into climate-based CZB. By systematically incorporating variables directly tied to energy consumption, such as HDD, CDD, and GHI, and analyzing these with and without latitude to gauge the impact of spatial constraints, phase 1 lays the groundwork for a comparative evaluation with the subsequent, performance-based methodology outlined in Phase 2. Key findings from Phase 1 include:

- The Elbow technique identified a range for the ONCZ from 3 to 8 for different sets of variables.
- Clustering results demonstrated distinct characteristics between KC and HC methods, with KC showing more evenly distributed clusters and HC revealing complex structures within the data.
- Determining the "best" clustering method based on the quality assessment results is challenging due to the variation in which different methods excelled across distinct quality metrics.
- Uniqueness favoured KC set 4 and HC set 4, suggesting these methods are particularly effective at generating distinct clusters that are easily distinguishable from one another.
- Compactness was best achieved by HC set 2, KC set 4, and HC set 4 indicating these methods excel at creating tightly grouped clusters, where data points within a cluster are closely packed together.
- As a primary statistical clustering quality metric, the SS, which evaluates both cohesion within clusters and separation between them, identified KC set 1, HC set 5, KC set 5, HC set 1, and HC set 3 configurations as superior, indicating a well-balanced clustering structure.
- The performance-based validation using KDE overlap analysis revealed that HC set 5 and KC set 1 yielded the most favourable results with the lowest CZMI under 10% and 3 CZs each, indicating a high degree of connection between clusters and energy consumption patterns.
- There is no direct connection between the use of the specific clustering methods and the expected quality of the resulting CZB. The proposed CZMI is more clearly correlated with the number of clusters in classification, the number of variables used, and the SS.
- A lower number of zones is advantageous for achieving reduced CZMI values.
- Typically, the utilization of a single variable or a moderate number of variables, with three variables emerging as optimal within the analyzed conditions, significantly influences

quality of climate classification both in terms of SS and CZMI outcomes. This suggests that a nuanced approach to variable selection is essential for optimizing climate zoning classifications.


HC set 5
duster
0
1
2
KAZAKHSTAN
UZBEKISTAN
KYRGYZSTAN
TATARSTAN
BASHKORTOSTAN
CHELYABINSK OBLAST
SAMARA OBLAST
ORENBURG OBLAST
KURGAN OBLAST
NOVOSIBIRSK OBLAST
ALTAI KRAI
ASTRAKHAN OBLAST
AZERBAIJAN
Kazan
Ufa
Ulyanovsk
Sterlitamak
Magnitogorsk
Samara
Balakovo
Orenburg
Orsk
Aktobe
Atyrau
Astrakhan
Baku
Turkmenbashy
Nukus
Dashoguz
Urgench
Kyzylorda
Turkestan
Shymkent
Tashkent
Namangan
Andijon
Fergana
Khujand
Bishkek
Taraz
Almaty
Taldykorgan
Yining/Qulja
Karakol
Aksu
Kuqa
Tacheng
Karamay
Pavlodar
Ust-Kamenogorsk
Semey
Rubtsovsk
Barnaul
Novosibirsk
Omsk
Petropavl
Kostanay
Rudny
Kokshetau
Karaganda
Astana
Zlatoust
Chelyabinsk
Kurgan

(a)



(b)

Figure 4.22: CZB maps which CZMI not exceeding 10%. HC set 5 (a), and KC set 1 (b).

4.6. Phase 2 (Performance-based CZB)

To effectively classify the climate for buildings, it is essential to incorporate the buildings' aspect (energy needs) as a fundamental component [1, 3, 14, 56]. The application of BES has shown significant potential in optimizing the CZB. It suggests reorienting the focus of CZB from climate to performance-based criteria, which is characterized by its simplicity and reliability. While previous works often used BES to validate climate-based CZB results [3, 4, 54], Phase 2 of this research directly incorporates BES data into the classification process.

Phase 2 of this research introduces a novel method that utilizes the same traditional and spatially constrained clustering techniques (HC, KC, SCKC, and SCKC) as in Phase 1 to categorize the climate using building energy performance variables. This method aims to bridge the current disparity between CZB and building energy usage in Kazakhstan. In contrast to alternative methodologies that use both climatic factors and building performance indicators in the classification process, phase 2 of this study employs energy demands data and omits traditional climate variables. Phase 2 is intended to address two main research questions. Is it more effective to propose a CZB directly utilizing building energy data, while eliminating climatic variables? Which classification (clustering) approach produces the most optimal outcomes?

4.6.1. **The optimal number of climate zones**

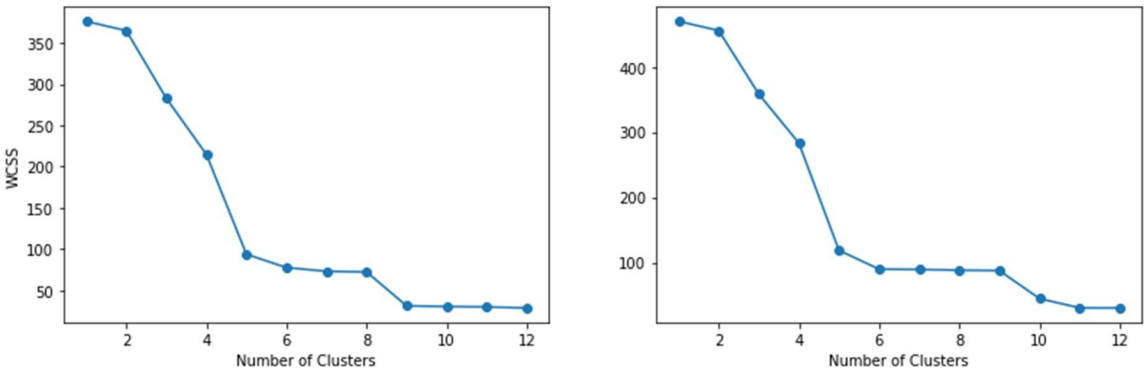
Four unique datasets, each including four building performance indicators were formed at the initial stage of Phase 2, see Table 4.5. The same sets of variables were used with and without LAT to track the influence of the spatial constraint on the ONCZ and classification results.

Table 4.5: Phase 2 datasets and used variables.

Name	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
HC	annual space heating NA	annual space cooling NA	annual space heating SA	annual space cooling SA	
KC	annual space heating NA	annual space cooling NA	annual space heating SA	annual space cooling SA	
SCHC	annual space heating NA	annual space cooling NA	annual space heating SA	annual space cooling SA	Latitude
SCKC	annual space heating NA	annual space cooling NA	annual space heating SA	annual space cooling SA	Latitude

The Elbow method, unaffected by the spatial factor, demonstrated a decline in the WCSS until K=6, reaching a plateau from K=6 to K=9, with a slight subsequent decline. The identified ONCZs were 6, 9, and 10 (Figure 4.23 (a, b)). The results indicated that the impact of latitude

as a spatially constrained variable on the determination of ONCZ was negligible. However, it will be highlighted later that the significance of this variable becomes more pronounced in the direct clustering of the climate data.



(a) (b)

Figure 4.23: The ONCZ determination of Phase 2. Elbow graph based on spatially constrained data (a) and Elbow graph based on non-spatially constrained data (b).

Given the limited sample size, opting for a smaller number of ONCZ could simplify the clustering procedure and lead to the identification of 6 clusters as the most suitable choice. Moreover, like in Phase 1 ONCZ additional clusters, like 9 and 10, might provide greater detail in some situations, allowing for more particular adaptations to localized differences in building energy efficiency trends. Furthermore, similar to Phase 1, on comparing the results with the official CZB map of Kazakhstan, it is evident that the proposed ONCZ differs from the official maps. The findings of ONCs are displayed in Table 4.6.

Table 4.6: ONCZs for Phase 2 datasets.

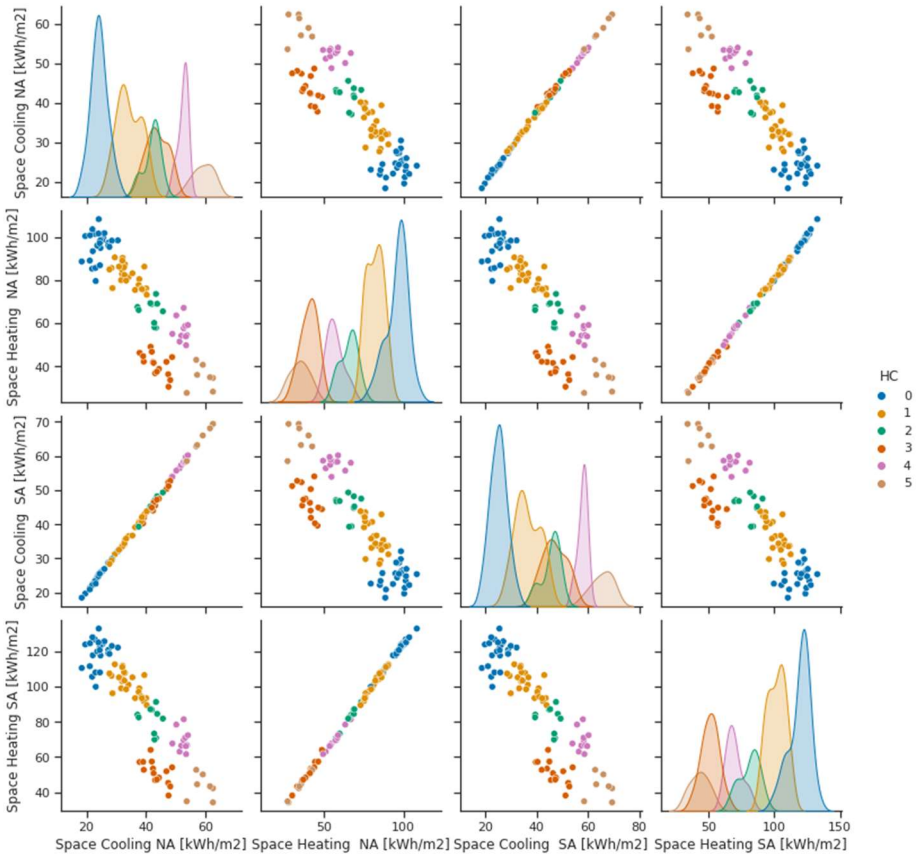
Name	ONCZ
HC	6
KC	6
SCHC	6
SCKC	6

4.6.2. **Clustering results**

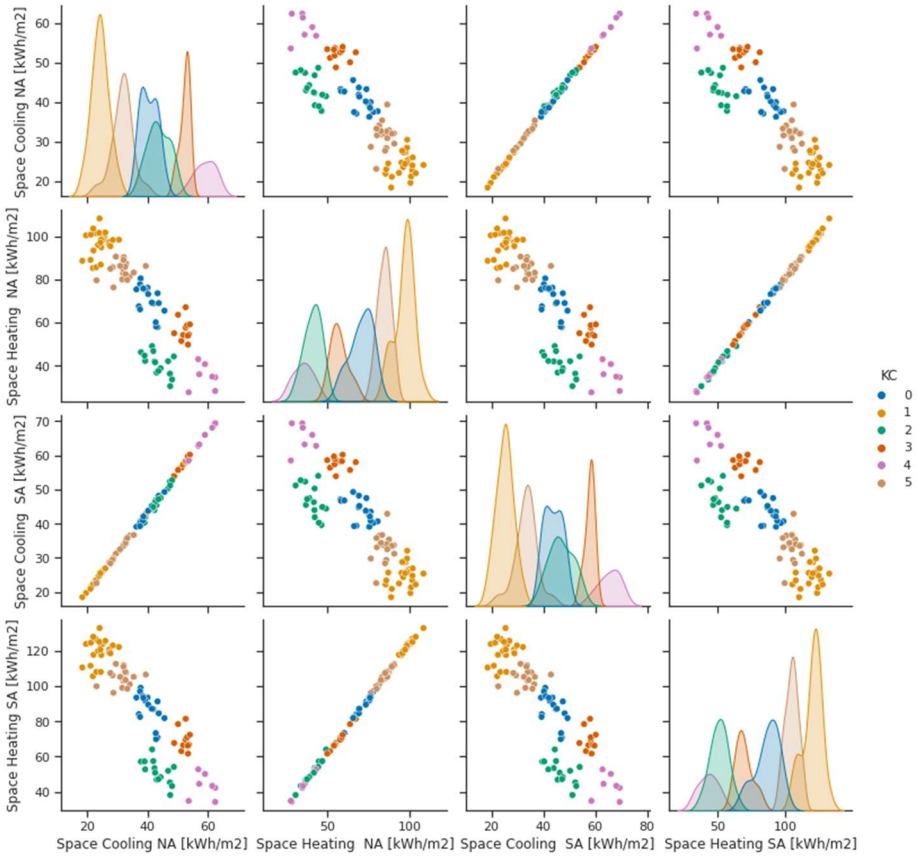
Four separate runs (HC, KC, SCHC, and SCKC) were conducted to execute clustering techniques, both with and without spatial constraints. After, the scatterplot matrix was created to display the results of clustering done on Phase 2 datasets. Figure 4.24 facilitates a thorough

analysis of the clustering findings by graphically representing the performance metrics of each pair of buildings in a matrix configuration.

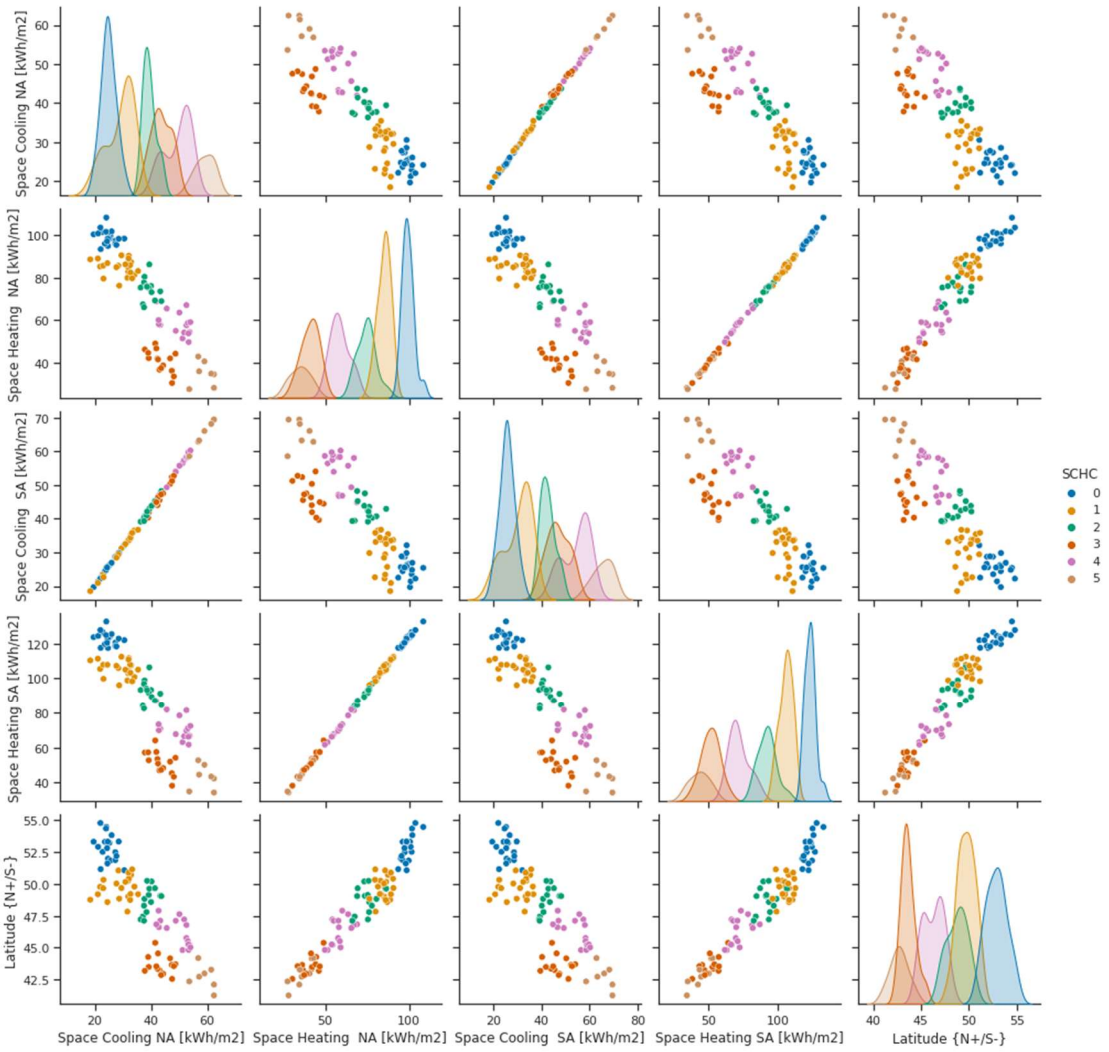
Although the clusters are named differently in various approaches, it is clear that the outcomes of HC and KC are comparable. The discrepancy is seen when comparing HC and KC methods with their spatially constrained counterparts. To analyze in-depth and visually represent the borders and spatial distribution of each CZ, the clustering results were shown using ArcGIS Online (Figure 4.25). Both the HC and KC methods produced comparable cluster allocations for the furthest northern and southern areas. However, significant differences are seen in the center regions of the country, where 8 out of 97 data points showed misclassifications between the two techniques (Figure 4.25 (a, b)). The latitude influence varied among clusters in the southern, northern, and central areas of the country. In SCHC and SCKC, similar to HC and KC, the geographical component had limited impact on the outcomes in the furthest northern and southern areas. The spatial constraint had a notable effect on clusters 0, 1, 2, and 4 (SCHC) in the central areas, as seen in Figure 4.25 (c, d). Utilizing spatial analysis improved the consistency of classifying data points near the CZs borders, ensuring alignment with their geographic context. KC showed a smaller change in comparison to HC when latitude was used in the clustering procedure.


Space Cooling NA [kWh/m2]
Space Heating NA [kWh/m2]
Space Cooling SA [kWh/m2]
Space Heating SA [kWh/m2]
HC
0
1
2
3
4
5

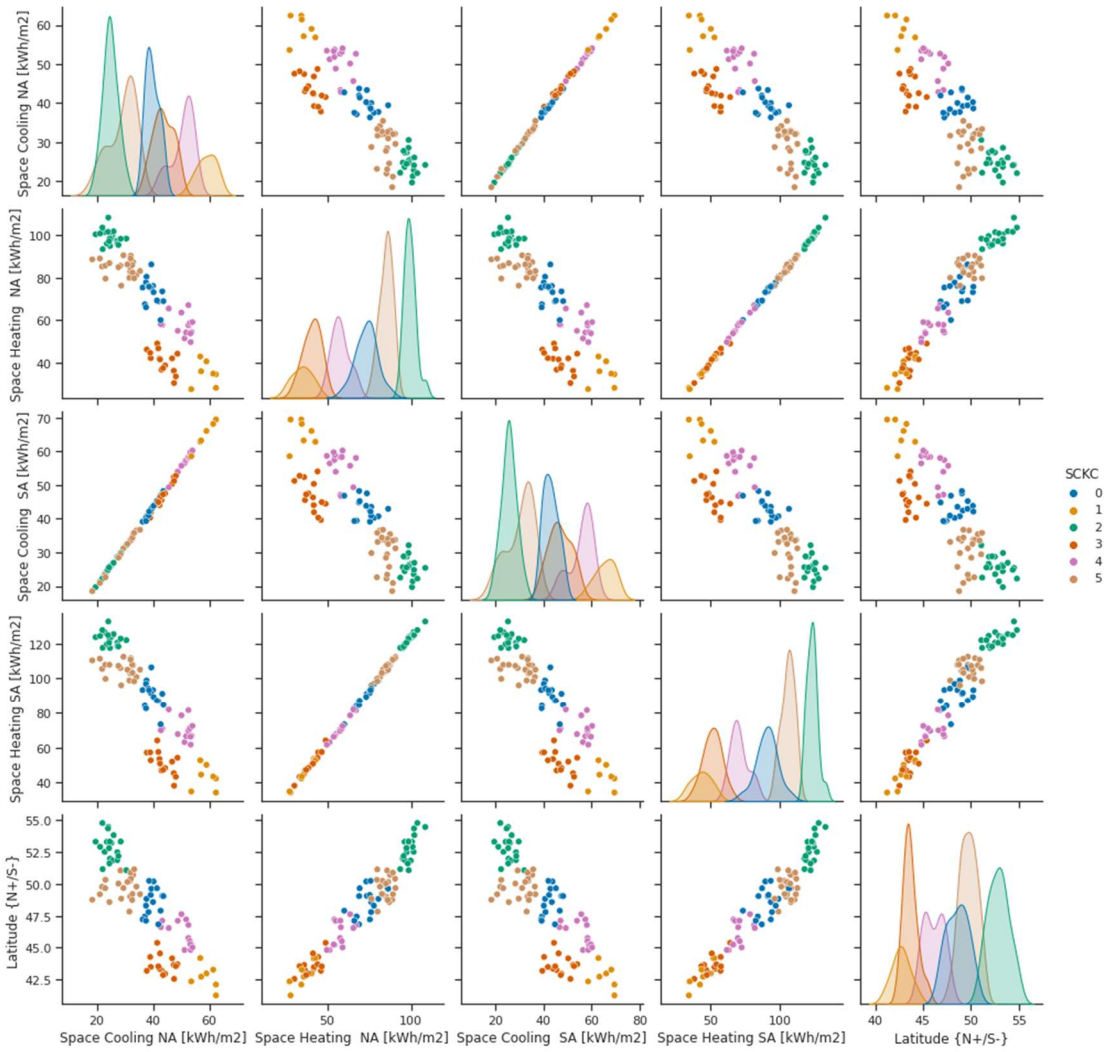
(a)


Space Cooling NA [kWh/m2]
Space Heating NA [kWh/m2]
Space Cooling SA [kWh/m2]
Space Heating SA [kWh/m2]
KC
0
1
2
3
4
5

(b)

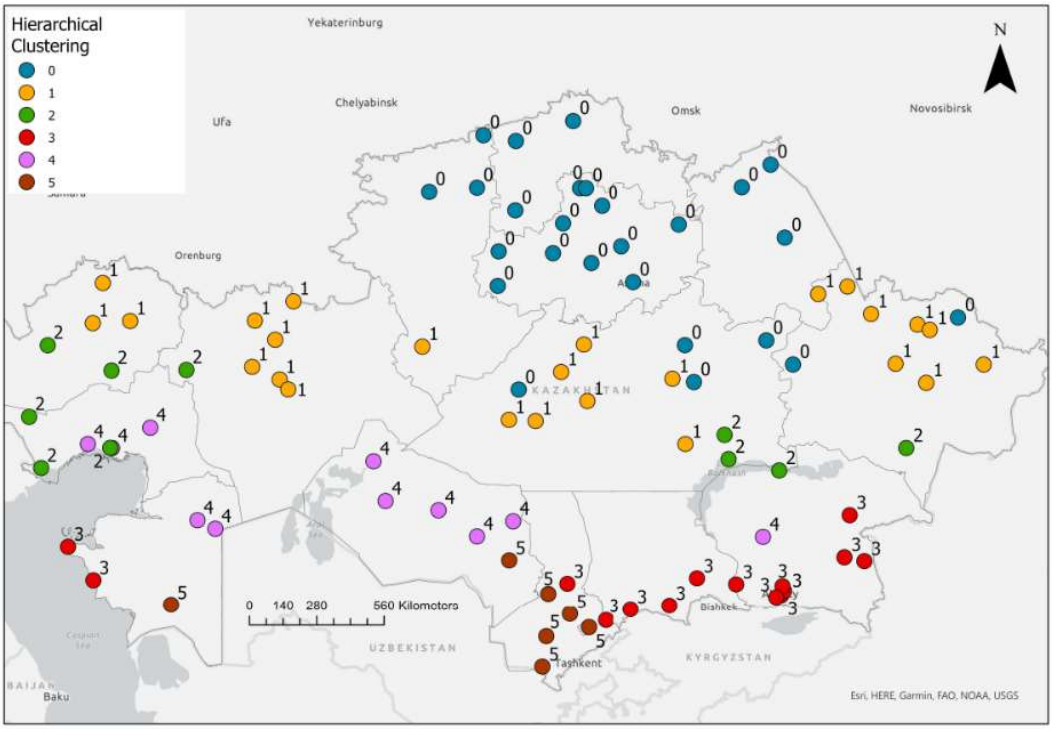

Space Cooling NA [kWh/m2]
Space Heating NA [kWh/m2]
Space Cooling SA [kWh/m2]
Space Heating SA [kWh/m2]
Latitude {N+/S-}
SCHC
0
1
2
3
4
5

(c)

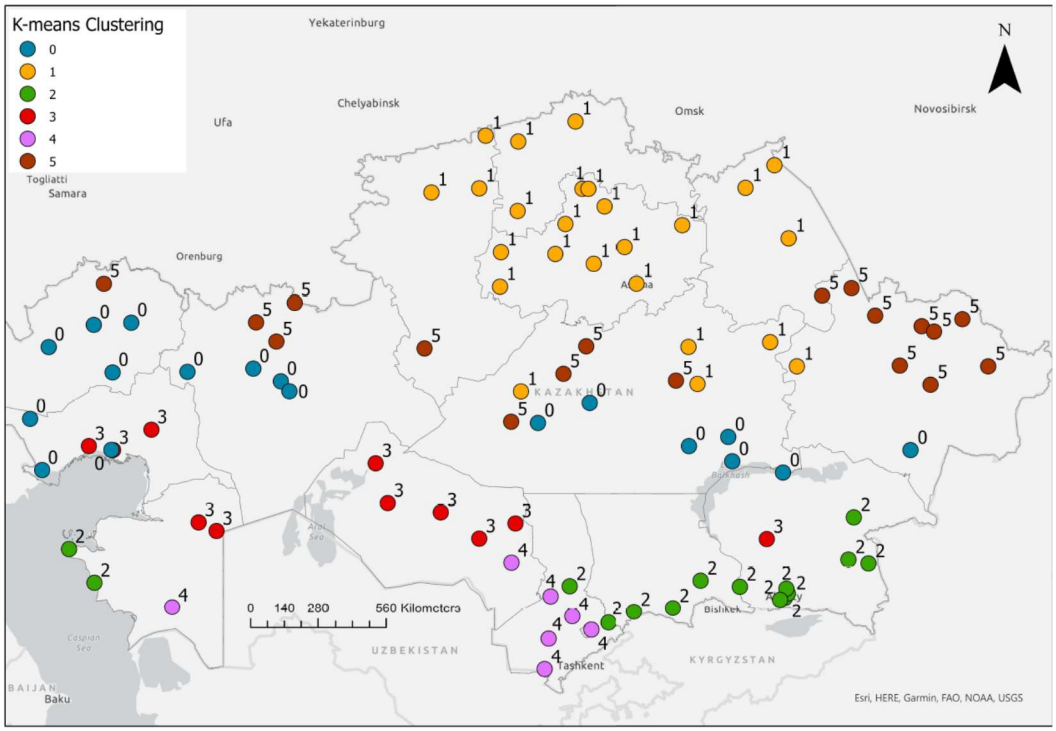


(d)

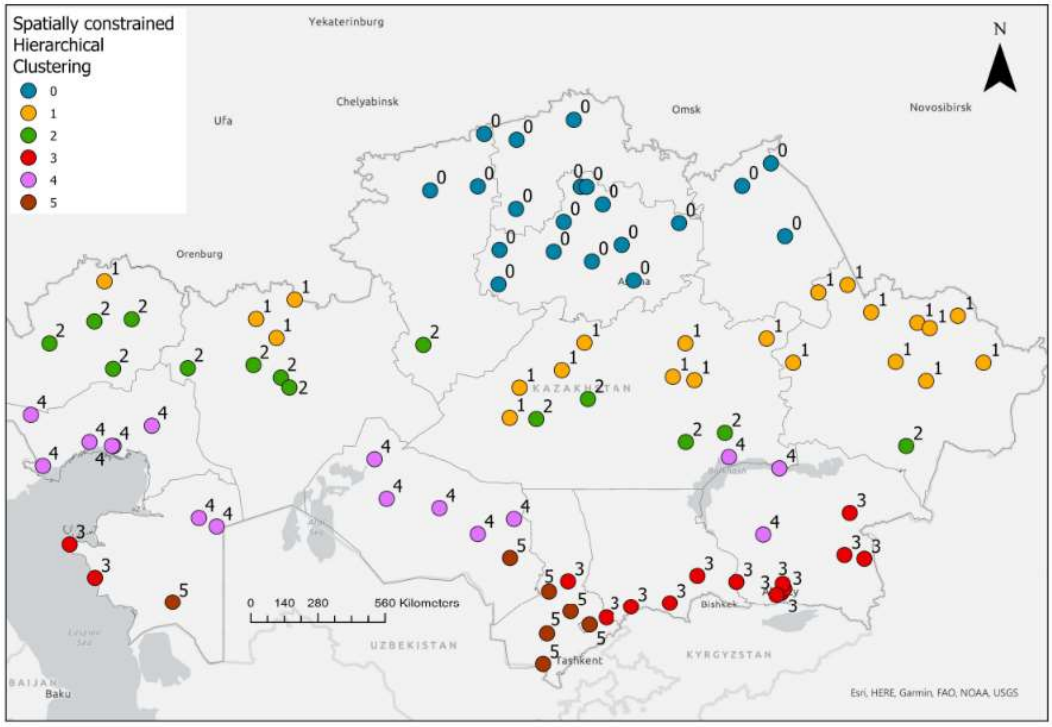
Figure 4.24: The CZ clustering scatterplot matrices. HC (a), KC (b), SCHC (c), SCKC (d).


Hierarchical Clustering
0
1
2
3
4
5
Yekaterinburg
Chelyabinsk
Ufa
Omsk
Novosibirsk
Orenburg
KAZAKHSTAN
UZBEKISTAN
KYRGYZSTAN
Tashkent
Bishkek
Baku
0 140 280 560 Kilometers
N
Esri, HERE, Garmin, FAO, NOAA, USGS

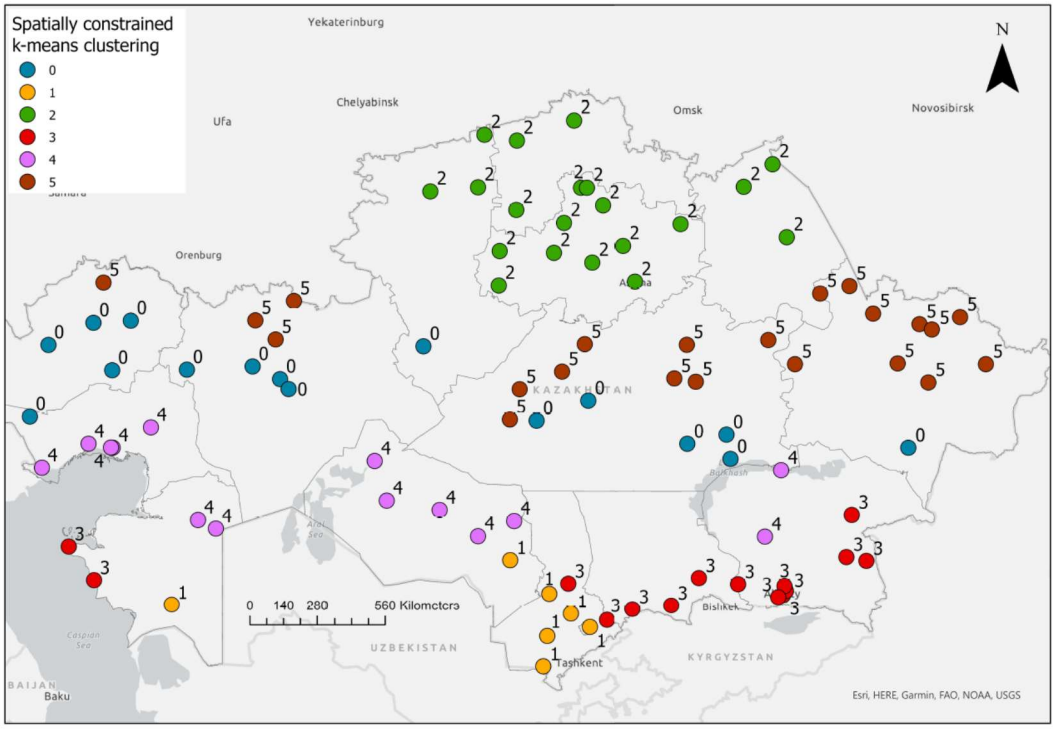
(a)


K-means Clustering
0
1
2
3
4
5
Yekaterinburg
Chelyabinsk
Ufa
Omsk
Novosibirsk
Togliatti
Samara
Orenburg
KAZAKHSTAN
UZBEKISTAN
KYRGYZSTAN
Tashkent
Bishkek
Baku
0 140 280 560 Kilometers
N
Esri, HERE, Garmin, FAO, NOAA, USGS

(b)



(c)



(d)

Figure 4.25: Performance-based maps. HC (a), KC (b), SCHC (c), SCKC (d).

4.6.3. **Clustering Quality Assessment**

This section aims to assess the quality of the clustering results achieved through various methods for performance-based climate classification of Phase 2. This evaluation is conducted through three key dimensions: Uniqueness, as described in subsection 4.5.3.1, evaluates the distinctiveness of each cluster; Compactness, outlined in subsection 4.5.3.2, examines the tightness of the clustering; and SS, explored in subsection 4.5.3.3, provides a comprehensive measure of both cohesion and separation within the clusters.

4.6.3.1. **Uniqueness**

Upon examination, it was found that all clustering approaches of Phase 2 displayed exceptional distinctiveness, with spatially constrained methods (SCHC, SCKC) achieving the lowest standard deviation values (5.09%). Figure 4.26 displays the uniqueness outcomes of every method as % values for each CZ along with mean and standard deviation values.

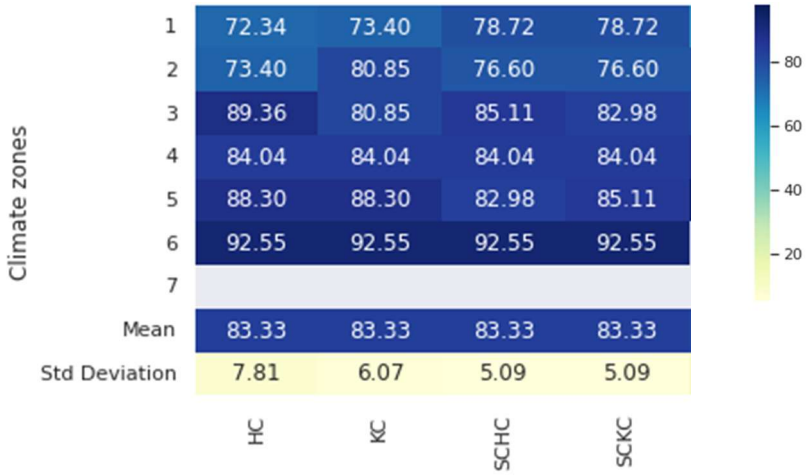
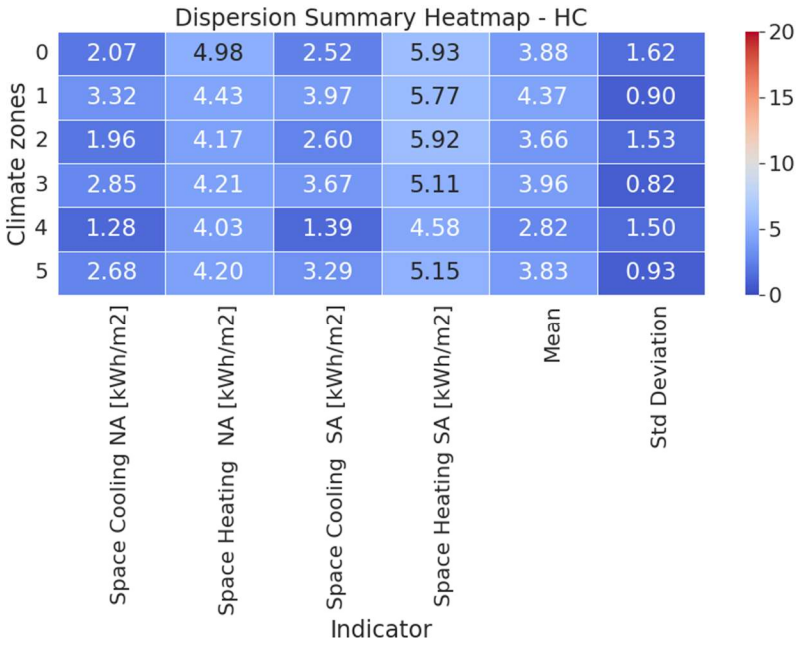


Figure 4.26: The uniqueness heatmap of Phase 2 clustering results.

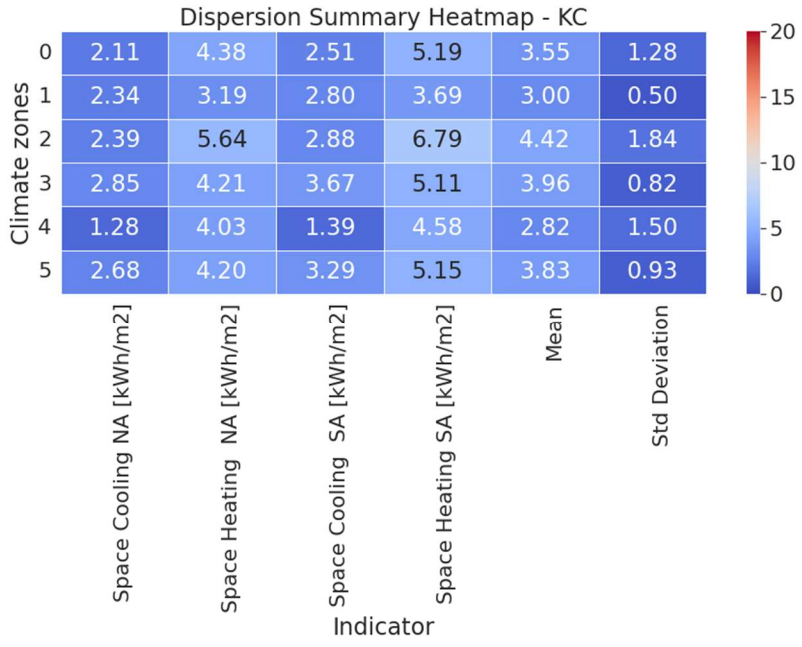
4.6.3.2. **Compactness**

The dispersion values produced for each clustering technique are depicted in Figure 4.27, providing a deeper understanding of the performance of the various Phase 2 clustering methods. Despite the variations observed, it is noteworthy that the results across all methods and CZs are relatively close to each other. This proximity in values underscores a certain level of consistency and reliability across the clustering methods, although the spatially constrained ones still show a slight edge in performance and exhibit the lowest mean and standard deviation values of the dispersion. However, the SCHC method's standard deviation ranges from approximately 0.33

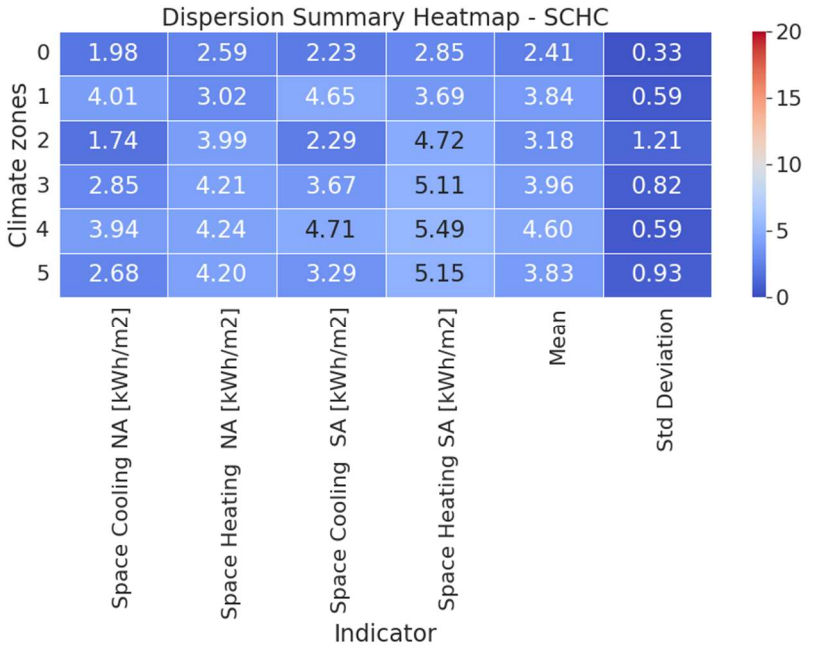
to 1.21, while the SCKC method shows a similar range from 0.33 to 1.57. In contrast, the HC method displays a broader standard deviation range, from about 0.82 to 1.62, and the KC method's range is from 0.50 to 1.84, reflecting a higher variability in cluster compactness. It is also evident that SCHC and SCKC show a more balanced dispersion between space heating and cooling across the CZ. This suggests that SCHC and SCKC lead to more homogenous clusters concerning heating and cooling needs.



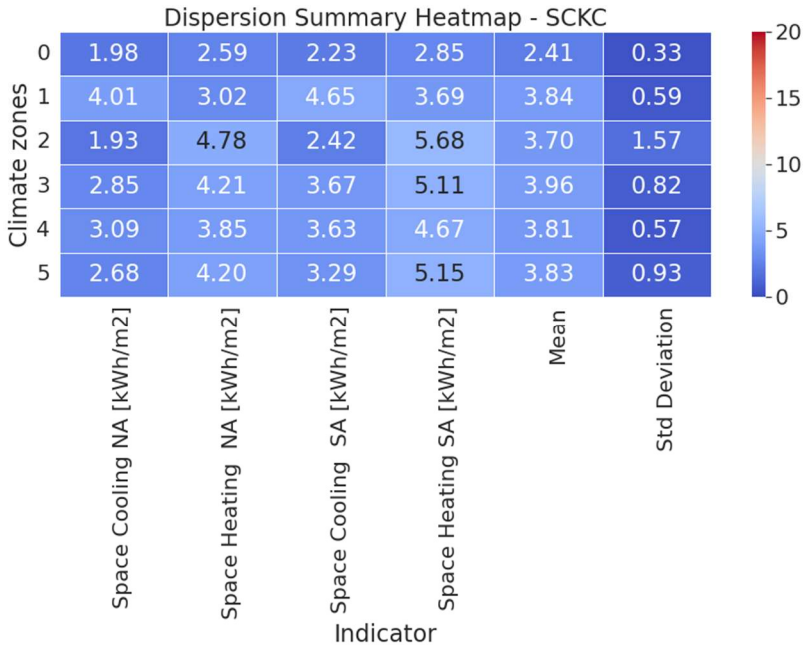
(a)



(b)



(c)



(d)

Figure 4.27: Heatmap displaying the dispersion summary for all CZ techniques. HC (a), KC (b), SCHC (c), SCKC (d)

4.6.3.3. **The Silhouette Score**

In assessing the clustering quality across 4 different clustering results of Phase 2, the SS analysis utilized four principal features: space cooling NA, space heating NA, space cooling SA, and space heating SA. The SS for SCKC clusters is the highest, at approximately 0.49, indicating a

relatively better-defined clustering structure. Following closely is SCHC with a score of around 0.48, suggesting that cooling and heating clustering configurations manifest comparatively coherent and separated clusters. The KC and HC configurations show slightly lower scores, around 0.46 and 0.45 respectively, implying a less pronounced separation between clusters. Notably, the results across all configurations are quite close, highlighting that SCKC and SCHC configurations demonstrate slightly better performance in grouping similar entities based on space cooling and heating metrics. The SS results are shown in Figure 4.28.

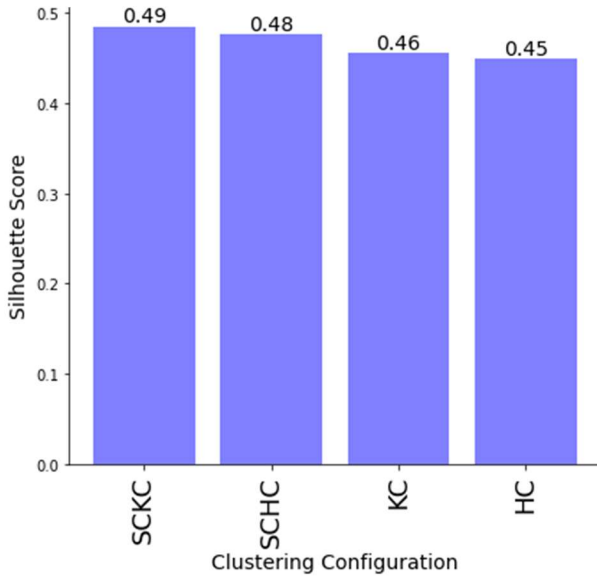
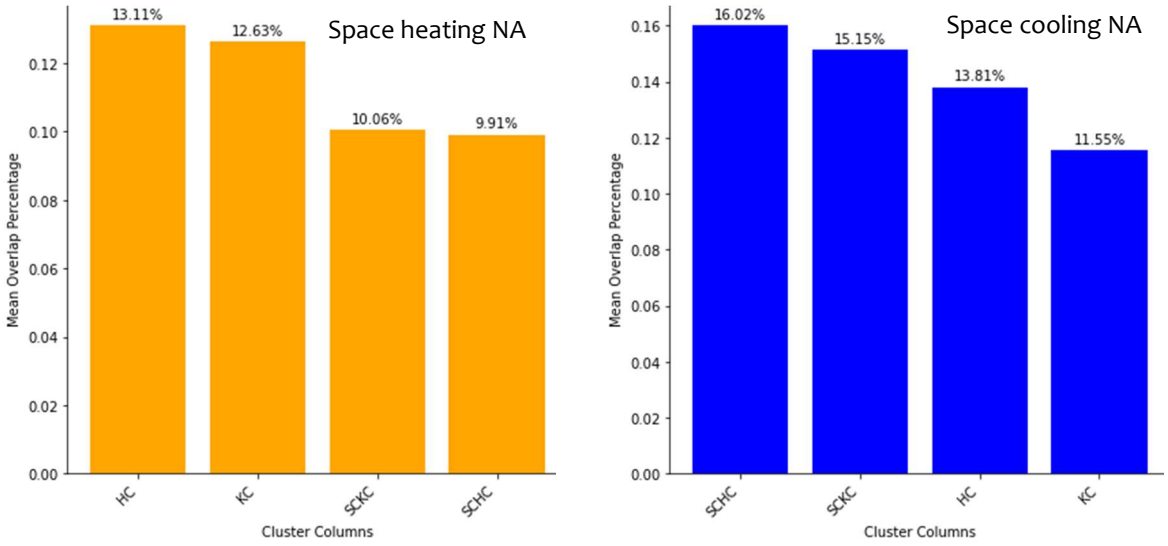


Figure 4.28: The SS results of each Phase 2 clustering method.

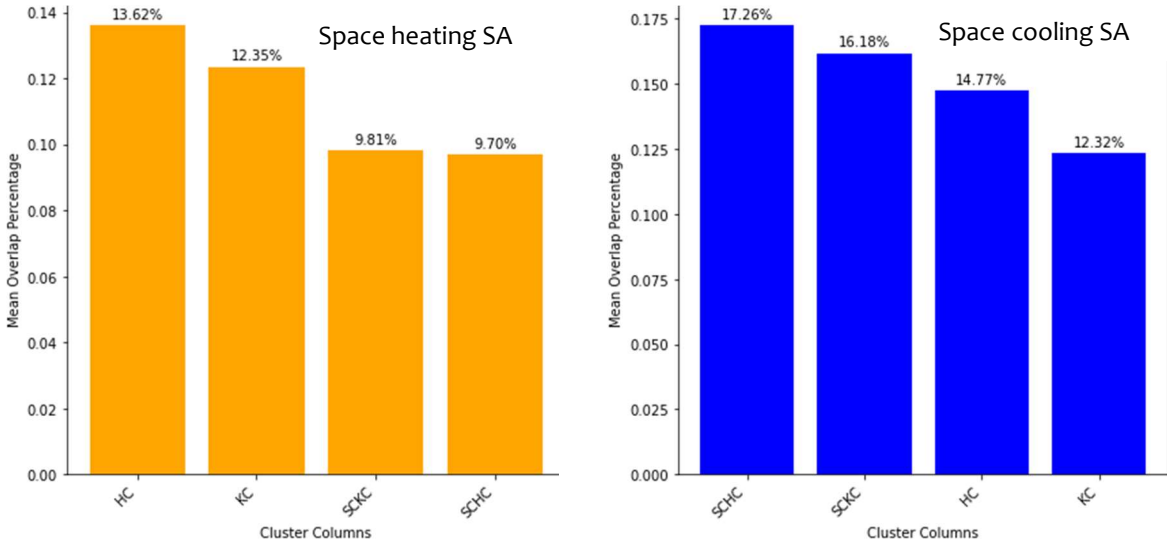
4.6.4. **Overlap calculation**

In addressing the research questions comparing conventional climate-based CZB and contemporary performance-based techniques, Phase 2 employs KDE solely for reference purposes. It is imperative to note that the principal aim of overlap calculation in Phase 2 is not validation but rather the establishment of reference values for subsequent comparative analyses, offering a comprehensive evaluation of climate-based versus performance-based clustering methods. Comparing the overlap observed in both phases allows for an assessment of whether the climate-based approach in Phase 1 creates clusters with similar or potentially less overlap compared to the performance-based CZB. With that, it seeks to answer whether proposing a CZB directly using building energy consumption data, yields more efficient outcomes compared to the conventional approach.

The mean overlap percentages for various energy metrics revealed a range of intersection degrees, specifically, overlaps in cooling energy needs showed a range from approximately 11.55% (KC) up to 17.26% (SCHC) (Figure 4.29 (b, d)). In contrast, the heating energy needs exhibited overlaps within a narrower band, from around 9.70% to 13.62% (Figure 4.29 (a, c)), suggesting a closer alignment in heating consumption patterns across all clustering methods. CZMI are shown in Figure 4.30 and ranged from 8.59% for the KC method to 9.69% for HC, making KC the best option from cluster overlapping perspectives, also indicating that, despite initial overlaps, clusters are generally well-separated when both energy consumption behaviors are considered. As the overlap results are almost identical (difference less than 2%) for NA and SA for the same performance indicators, Figure 4.31 shows the overlap graphs of NA to visually represent the amount of intersection of all used clustering methods. As expected, all clustering approaches exhibit a high degree of intra-cluster separation and a small amount of overlap.



(a) (b)



(c) (d)

Figure 4.29: Mean overlap percentage values of Phase 2 clustering results for space heating NA (a), space cooling NA (b), space heating SA (c), and space cooling SA (d).

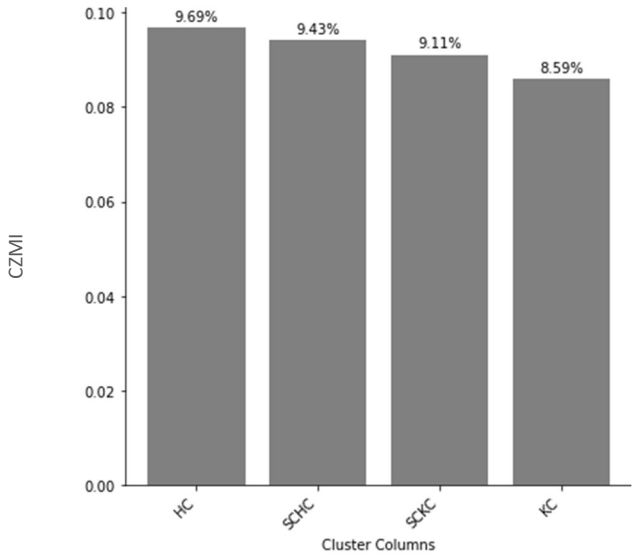
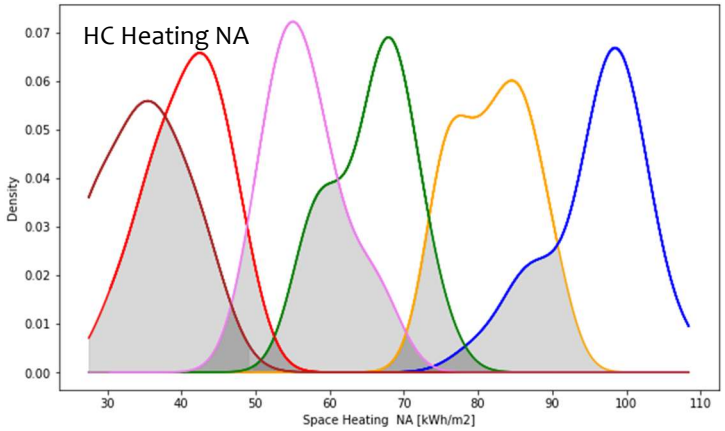
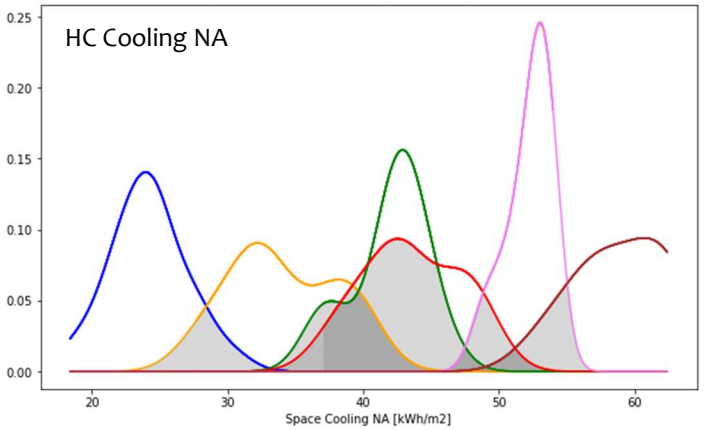


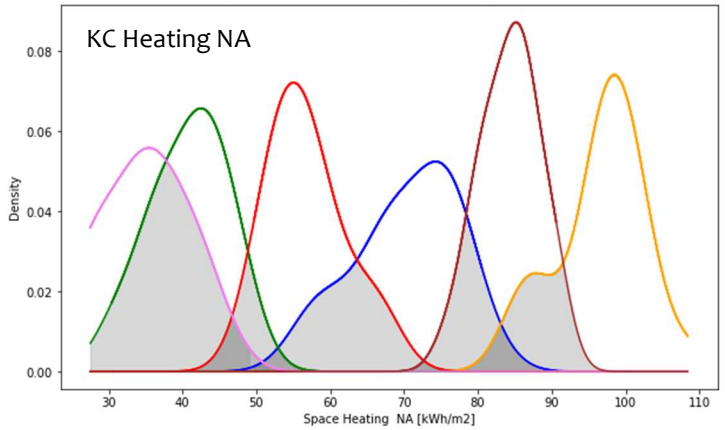
Figure 4.30: CZMI values of Phase 2 clustering methods.


HC Heating NA
Density
Space Heating NA [kWh/m2]

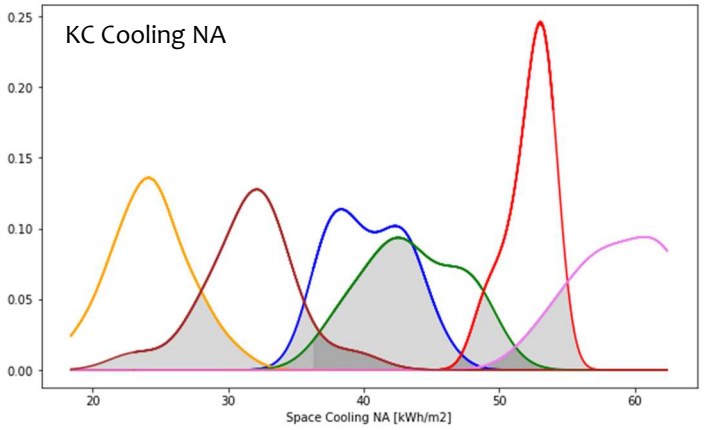
(a)


HC Cooling NA
Space Cooling NA [kWh/m2]

(b)


KC Heating NA
Density
Space Heating NA [kWh/m2]

(c)


KC Cooling NA
Space Cooling NA [kWh/m2]

(d)

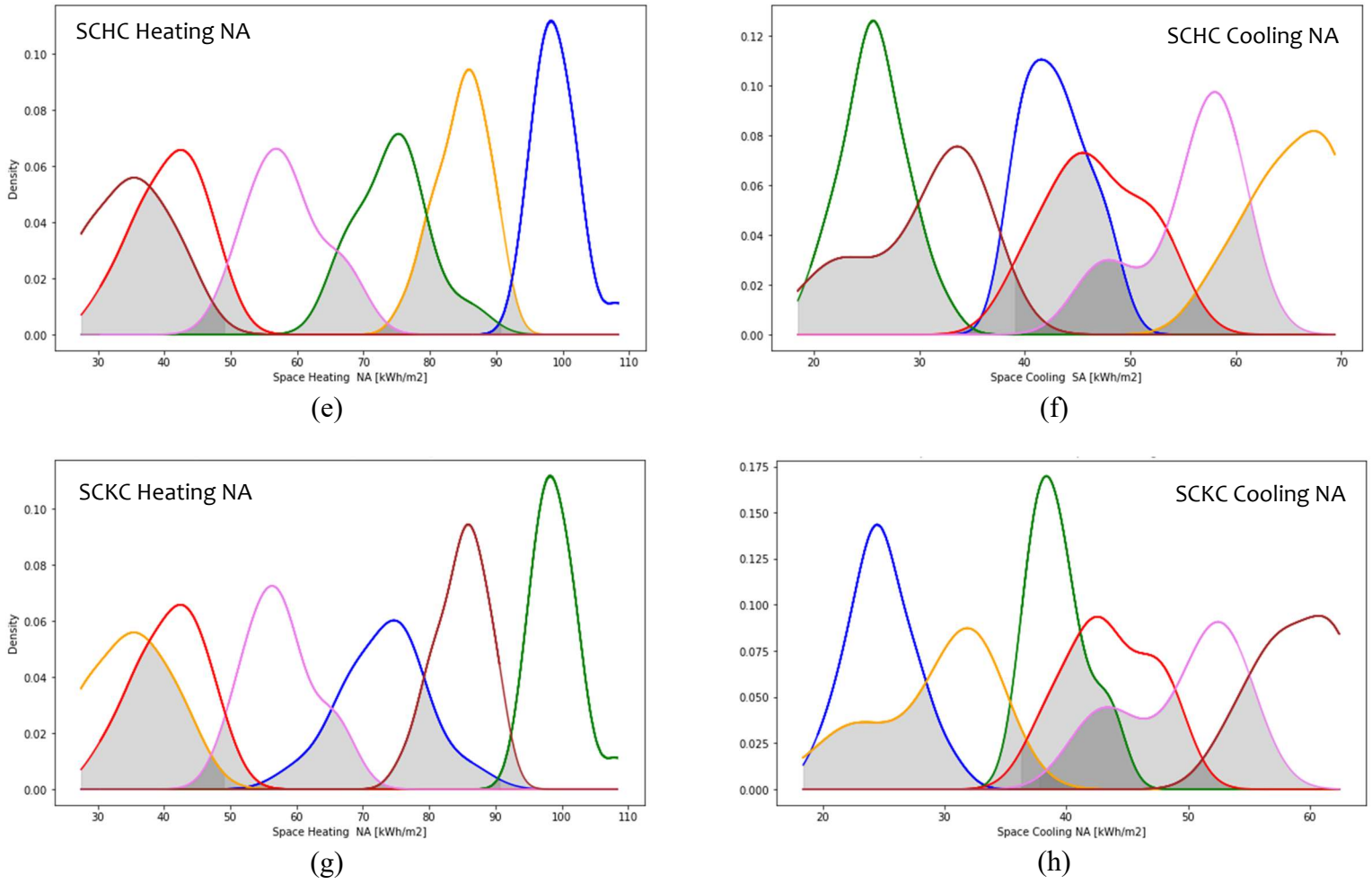


Figure 4.31: Overlap graphs of phase 2 clustering methods. HC for heating NA (a), HC for cooling NA (b), KC for heating NA (c), KC for cooling NA (d), SCHC for heating NA (e), SCHC for cooling NA (f), SCKC for heating NA (g), and SCKC for cooling NA (h).

4.6.5. **Summary of Phase 2 findings**

Reliable CZB guarantees that buildings with comparable energy consumption levels are classified together within the same climate zone. By focusing on energy needs data only and excluding conventional climate variables, phase 2 aimed to establish a direct link between CZ and actual building energy usage in Kazakhstan. It underscores the significant potential of a performance-based classification approach to CZB, leveraging BES data to achieve a more nuanced and practical understanding of CZB. Key findings from this approach are summarized as follows:

- The ONCZ identified through the Elbow method suggested 6 (optionally extending up to 10) as the preferred number, proposing a more granular CZ pattern compared to existing ASHRAE and official local maps, with 4 main CZs.
- Spatial constraints, particularly latitude, introduced in SCHC and SCKC methods, played a minimal role in altering the ONCZ.
- SCHC and SCKC demonstrated increased clustering quality across three key dimensions: uniqueness, compactness, and SS, indicating their effectiveness in creating more coherent and distinct clusters. The SCHC and SCKC algorithms had higher cluster compactness and cluster separation, as indicated by their high mean values (83.33%). Additionally, these algorithms exhibited the lowest standard deviation of uniqueness (5.09%). SCKC had the greatest SS of 0.49, and SCHC scored 0.48, following closely. The KC and HC showed slightly lower scores, around 0.46 and 0.45 respectively.
- The spatial aspect showed mixed impacts on various clusters, with no impact in the northern and southern areas but substantial influence in the central regions of the country.
- The KDE overlap analysis, serves as an additional quantitative measure. All CZMI values were under 10% (ranging from 8.59% to 9.69%), illustrating well-separated clusters. The KC method performs the best in terms of minimizing overlaps.

4.7. Comparative analysis and synthesis

The comparative analysis part of this study conducts a thorough assessment and comparison of the findings from two separate phases, each emphasizing a distinct approach to the construction of CZB. The analysis commences by contrasting the climate-based methodology employed in Phase 1 with the performance-based approach presented in Phase 2. This two-step approach has been devised to comprehensively evaluate and compare the effectiveness of these different approaches in producing CZB, to determine the most efficient method for CZB generation. It also examines whether using building energy consumption data directly for CZB creation is a more efficient alternative to existing approaches that depend on climate variables.

The comparative analysis also includes a careful evaluation of the current conditions of the local official CZ map, to assess how well it aligns with the research findings and identify the possible amount of misclassification. This qualitative analysis is crucial for finding any inconsistencies in the traditional CZB of Kazakhstan that could guide future changes.

4.7.1. **Evaluation of discrepancies and misclassifications**

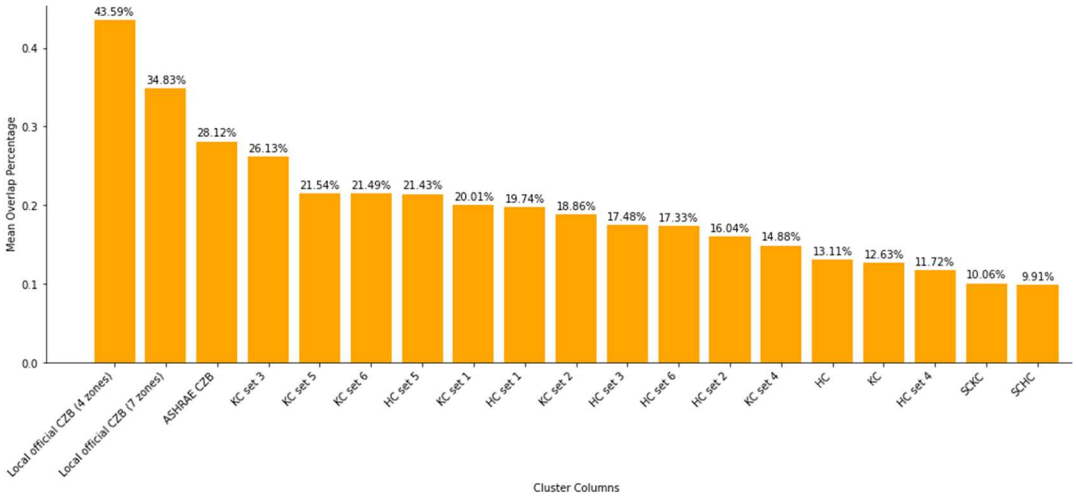
The evaluation of discrepancies and misclassifications is segmented into two key subsections using the proposed misclassification metrics: Mean Overlap Percentages and CZMI (4.6.1.1), and Adjusted Rand Index (4.6.1.2).

4.7.1.1. **Mean Overlap Percentages and CZMI**

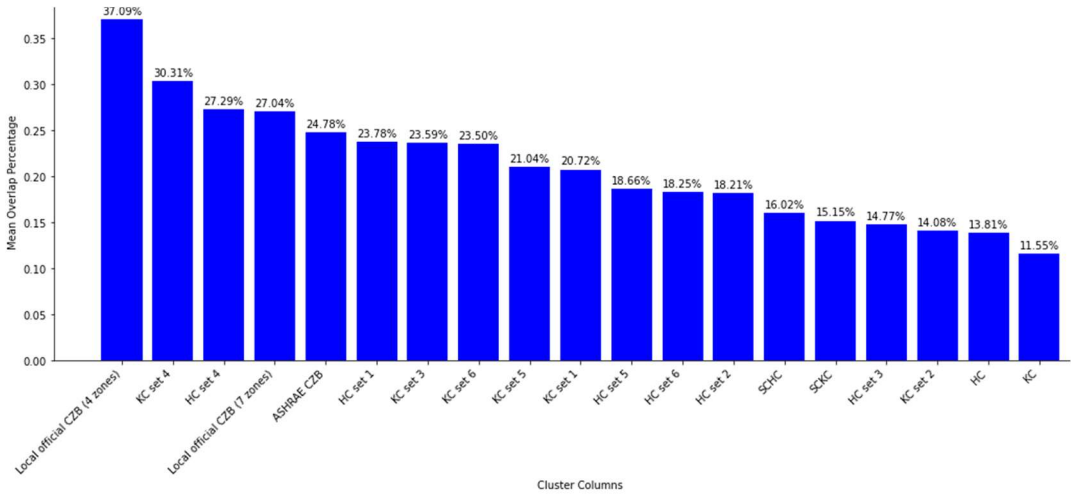
Calculating the mean overlap percentages based on heating energy needs (Figure 4.32 (a, c)), the performance-based clustering methods generally show lower overlaps, implying they are more accurate in classifying CZs for heating needs compared to climate-based methods. The SCHC method consistently appears among the lower overlap percentages (9.70-9.91%), which suggests it has a smaller misclassification rate for heating. However, results similar to performance-based methods can also be demonstrated by some climate-based methods (HC set 4). Considering cooling energy needs overlapping (Figure 4.32 (b, d)) similarly shows that performance-based clustering methods (SCHC, HC, SCKC, KC) result in lower percentages (11.55-17.26%), indicating better classification for cooling needs. SCHC and HC methods in particular seem to demonstrate better performance for cooling, with lower rates (11.55-12.32% and 13.81-14.77% respectively) of misclassification compared to the other methods. Among climate-based methods, the best performance was shown by KC set 2 and HC set 3 (13.10-14.08% and 14.77-15.47% respectively).

Official CZB map of Kazakhstan and ASHRAE map, have a broader range of performance (26.84-43.59%) typically placed on the higher end of the overlap percentage scale, indicating more significant misclassification rates for heating applications. However, for heating the ASHRAE map has an overlap of 26.84-28.12%, lower than the official CZB map of Kazakhstan. Among all explored methods, the official CZB map and the ASHRAE map indicate a higher overlap percentage, showing their poor reliability and performance-based classification accuracy.

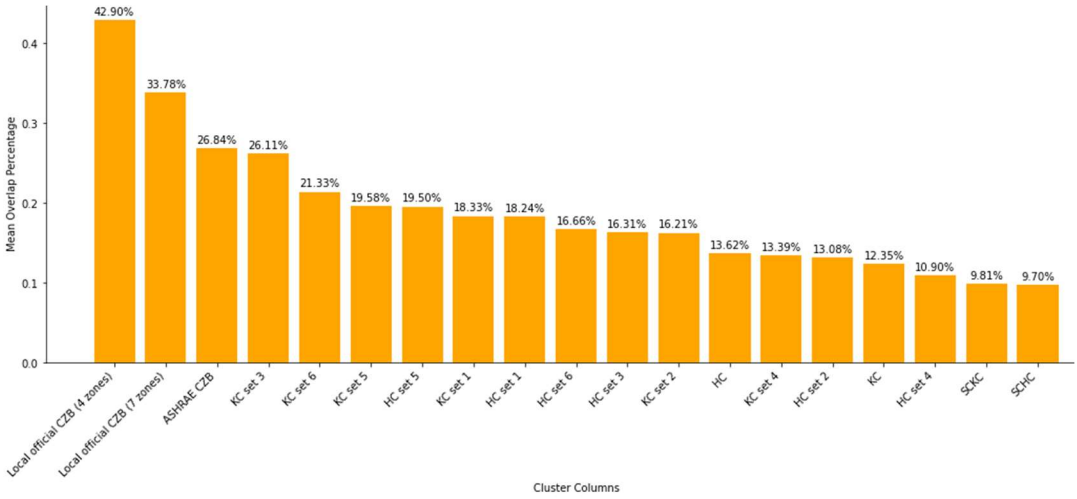
For both heating and cooling, the results suggest that performance-based clustering methods offer a more precise classification of CZB, which is critical for optimizing energy use and system design. These methods present consistent and high-quality results with a minimal spread. However, it is important to mention that specific climate-based approaches exhibit similarly low overlaps, matching the effectiveness of performance-based clustering.



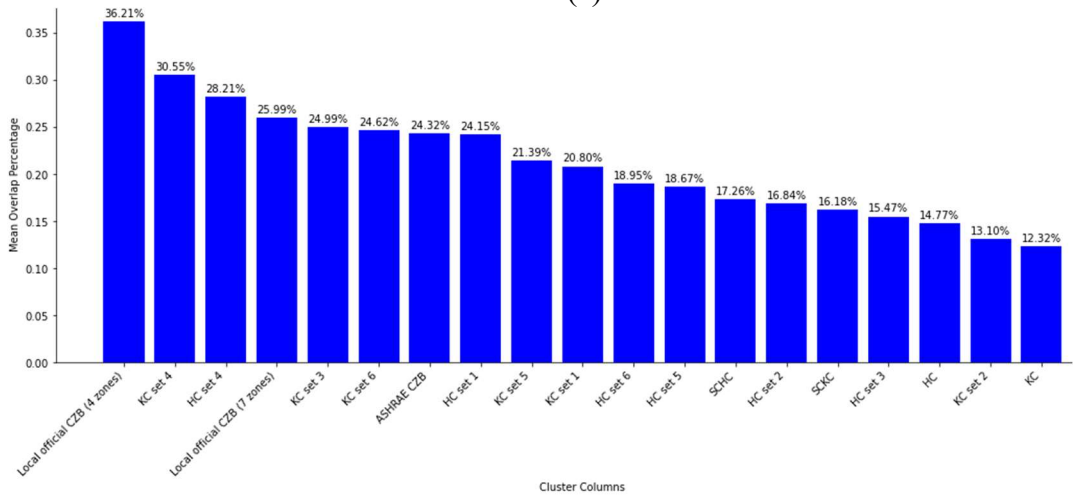
(a)



(b)



(c)



(d)

Figure 4.32: Mean overlap percentage values of all used clustering methods for space heating NA (a), space cooling NA (b), space heating SA (c), and space cooling SA (d).

After a thorough analysis of the CZMI for all involved CZB methods, it is clear that the performance-based classification approaches, specifically represented by KC and SCHC, consistently achieve higher accuracy with the lowest CZMI percentages (8.59% and 9.11% respectively), indicating a reduced percentage of misclassification. Figure 4.33 represents the CZMI values of all clustering methods, with traditional CZB maps marked green, climate-based classification methods in gray, and performance-based methods marked red. It is important to mention that specific climate-based approaches (HC set 5 and KC set 1) exhibit similar overlap outcomes (9.42% and 9.62% respectively) as performance-based methods, matching their effectiveness.

This finding suggests that climate-based classification, when tailored to specific regional climates, can serve as a viable alternative to performance-based methods, achieving comparable levels of accuracy. Given its more straightforward implementation (requiring only a set of climatic variables), this approach can deliver high-quality outcomes. Emphasizing the range of results generated by both methods is crucial. Although performance-based methods consistently produce outputs of high quality, the results obtained from climate-based methods can vary significantly. As a result, the probability of inaccuracy in climatic approaches is considerably higher. However, careful selection of variables (focusing on those most closely linked to building energy consumption) and the determination of an optimal number of climate zones are critical for error minimization in climate-based CZB. Suggestions here are to employ a concise dataset and a moderate number of zones. Preference to climate-based methods over performance-based alternatives should be given in conditions where the demands for final classification quality are moderate.

Traditional approaches such as the ASHRAE and official local CZB maps are notably ineffective, resulting in the highest CZMI percentages and hence, the most significant misclassification. Also, in this study, no significant evidence was obtained to support the superiority of spatially constrained methods over non-spatially constrained methods in overlap percentage reduction. Figure 4.34 shows the overlap graphs of the best-performed method (b, d) in comparison with the official local CZB (a, c). The proposed method has, in comparison with the official one, much smaller overlaps between climatic zones and more distinct CZ separation.

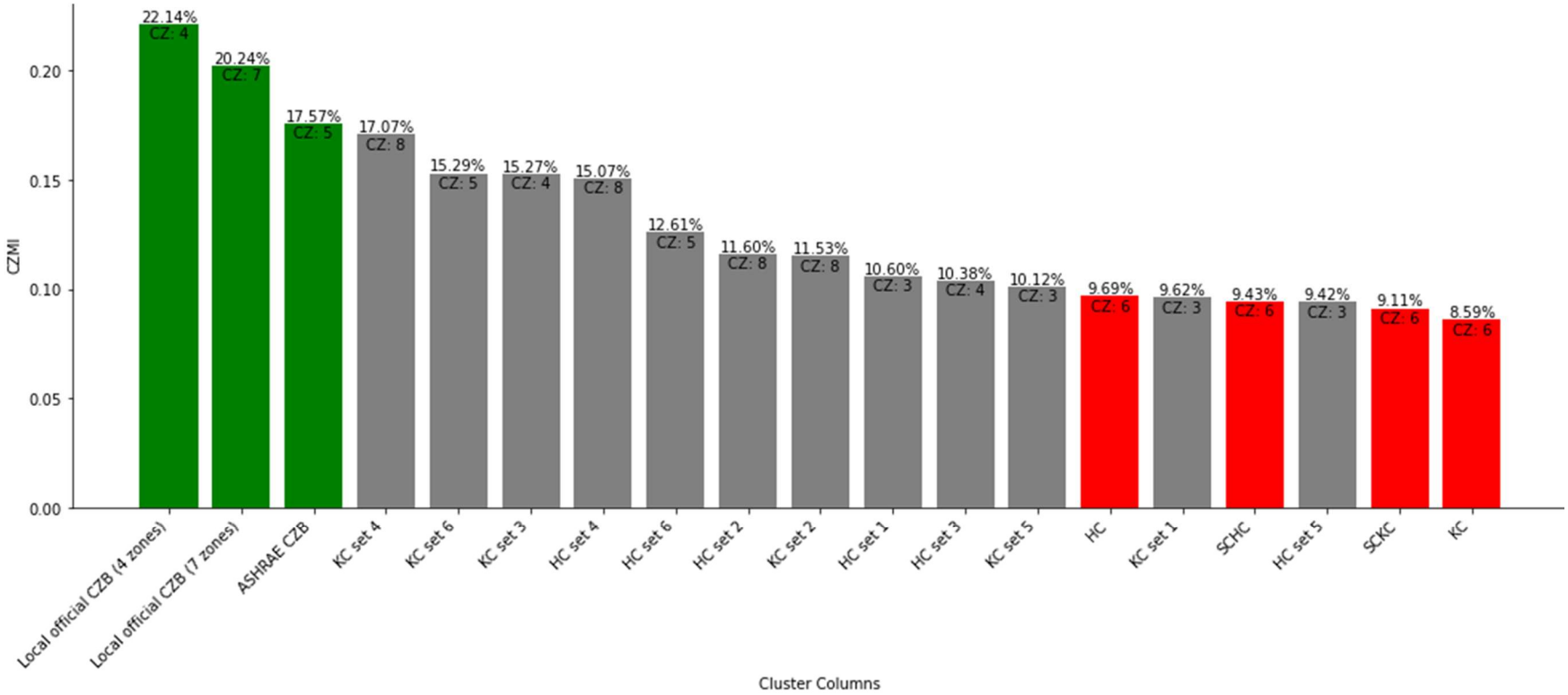


Figure 4.33: CZMI values of all used clustering methods.

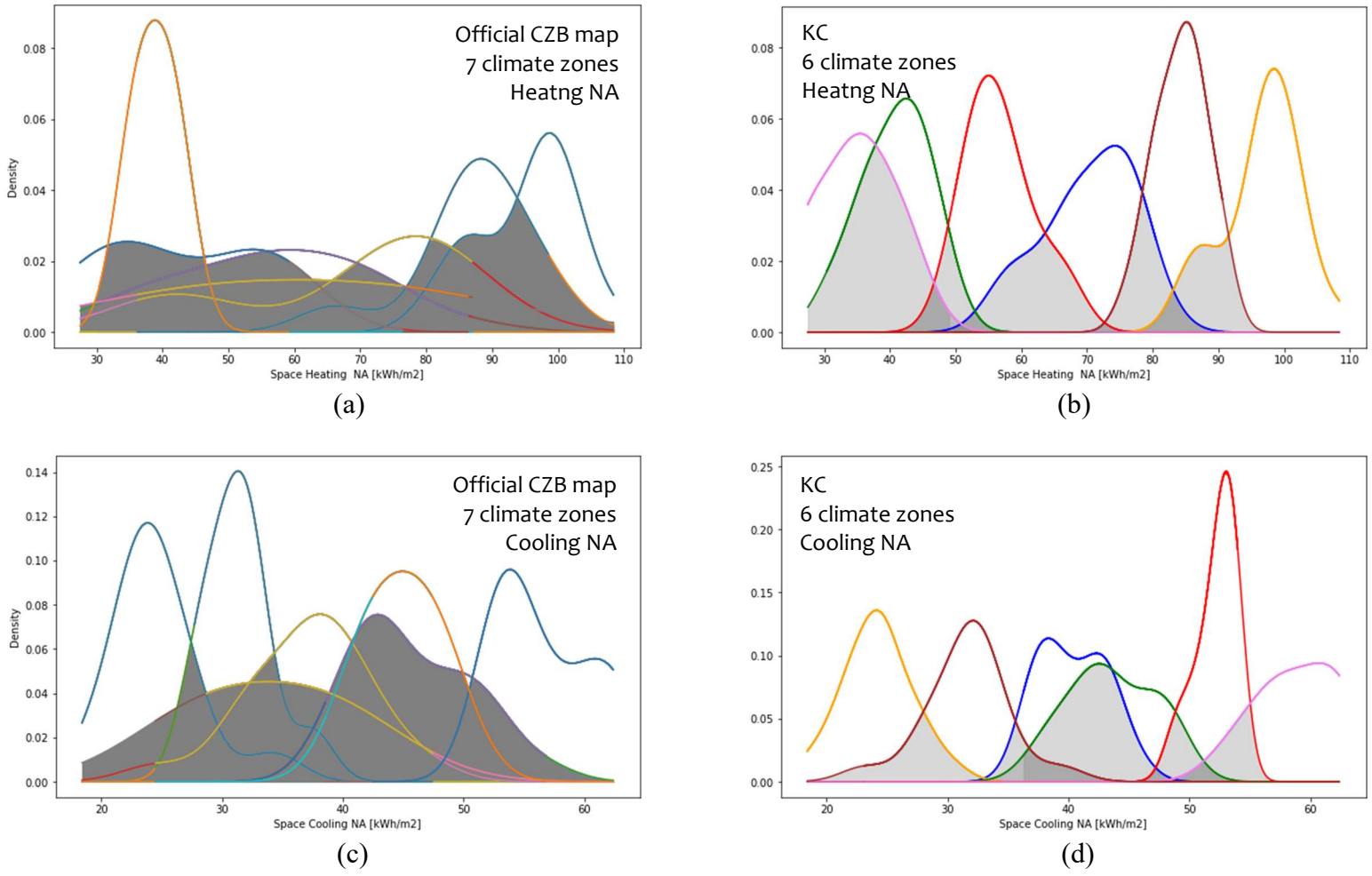


Figure 4.34: Overlap graphs of the local official CZB map with 7 zones based on heating NA (a), KC method based on heating NA (b), the local official CZB map with 7 zones based on space cooling NA (b), and KC method based on cooling NA (d).

Based on the comparison of CZMI for KC and HC in Figure 4.35 it can be seen that none of the methods consistently exhibits better outcomes. However, while KC works better for performance-based clustering, HC performs with much smaller CZMI for climate-based clustering. Some datasets exhibit substantial improvements in overlap percentages for HC, while others show only marginal differences between the two methods. The degree of improvement varies across different datasets and can reach around 30% of the difference (KC set 3 and HC set 3).

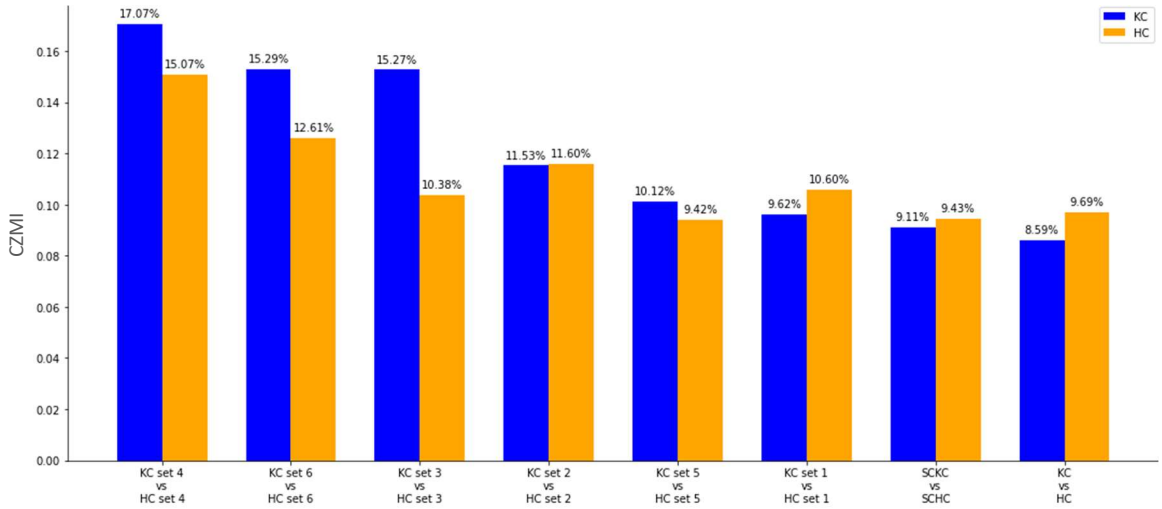


Figure 4.35: Comparison of CZMI between KC and HC methods.

Upon a side-by-side comparison of CZMI percentages in Figure 4.36 between traditional and spatially constrained methods, some instances are seen where the traditional method (blue) has a lower CZMI than the spatially constrained method (green), indicating a lower misclassification. This suggests that in some cases the non-spatial method outperforms the spatial method, and there is no uniform superiority of one method over the other, which is contrary to the typical expectation that spatial constraints improve clustering performance.

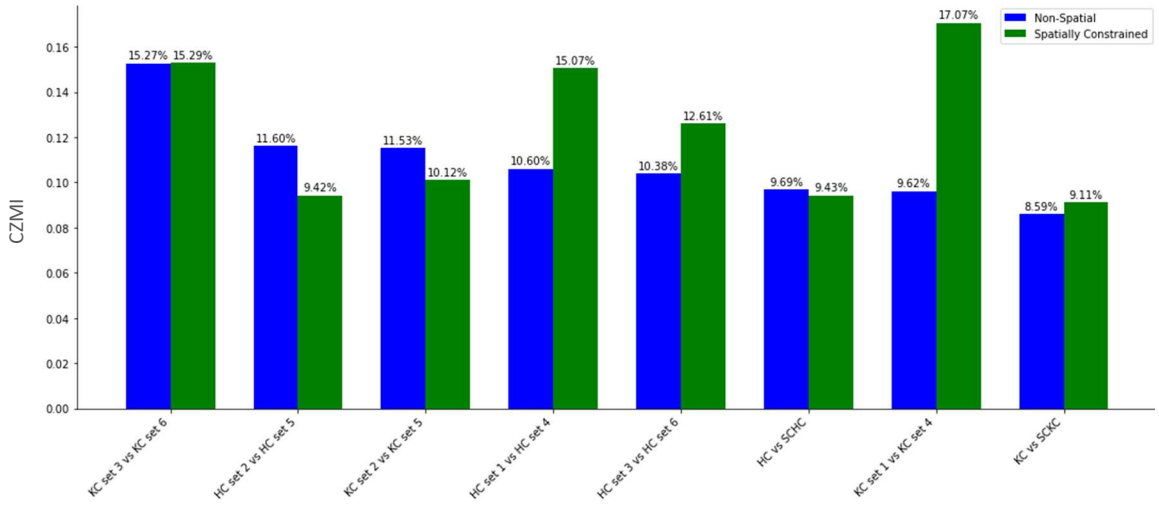


Figure 4.36: Comparison of CZMI between non-spatially and spatially constrained clustering methods.

4.7.1.2. **The Adjusted Rand Index**

The main purpose of employing the ARI in this context is to quantitatively evaluate the extent to which traditional CZB methods diverge from the best-performing proposed CZB methods. ARI values have been calculated for traditional classification methods compared with the best-recognized methods. This computational analysis was conducted using Pythons "sklearn.metrics" library (Appendix D).

The ARI scores can range from -1 to 1, with negative scores indicating disagreement, while positive scores show agreement.

The Local official CZB methods, with 4 and 7 CZs, yield moderate ARI values when compared with more sophisticated clustering algorithms, showing a modest congruence (Figure 4.37). Specifically, the local official CZB with 4 zones and local official CZB with 7 zones manifest ARIs that predominantly hover around the 0.30 mark in contrast to best-proposed methods, denoting that while there is some commonality in clustering outcomes, significant divergences in the identification of CZs persist. It is within this context that the ASHRAE CZB method emerges as a traditional technique with unexpectedly potent scores to its contemporary counterparts. Its ARI values—particularly 0.75 with KC set 1 and 0.73 with HC set 5 — revealed a connection, suggesting that the foundational principles of the ASHRAE map may capture essential CZ characteristics resonant with those deduced by more complex, data-driven approaches. This revelatory insight underscores the method's relevance and its potential to

remain a core concept in climate zoning analysis, bridging the gap between historical consistency and modern precision.

Furthermore, when delving into the best-recognized methods, the inter-comparisons reveal a compelling narrative of consistency of all performance-based methods (SCHC, HC, SCKC, and KC). However, these methods demonstrate significant differences in classification (ARI ranging from 0.40 to 0.43) against KC set 1 and HC set 5. The disparities in ARI values between SCHC, HC, SCKC, KC, and KC set 1, HC set 5 might be mostly attributed to the differing number of clusters among these methods. While all these methods are recognized for their quality clustering capabilities, their comparison reveals nuances in how each method approaches the data segmentation task.

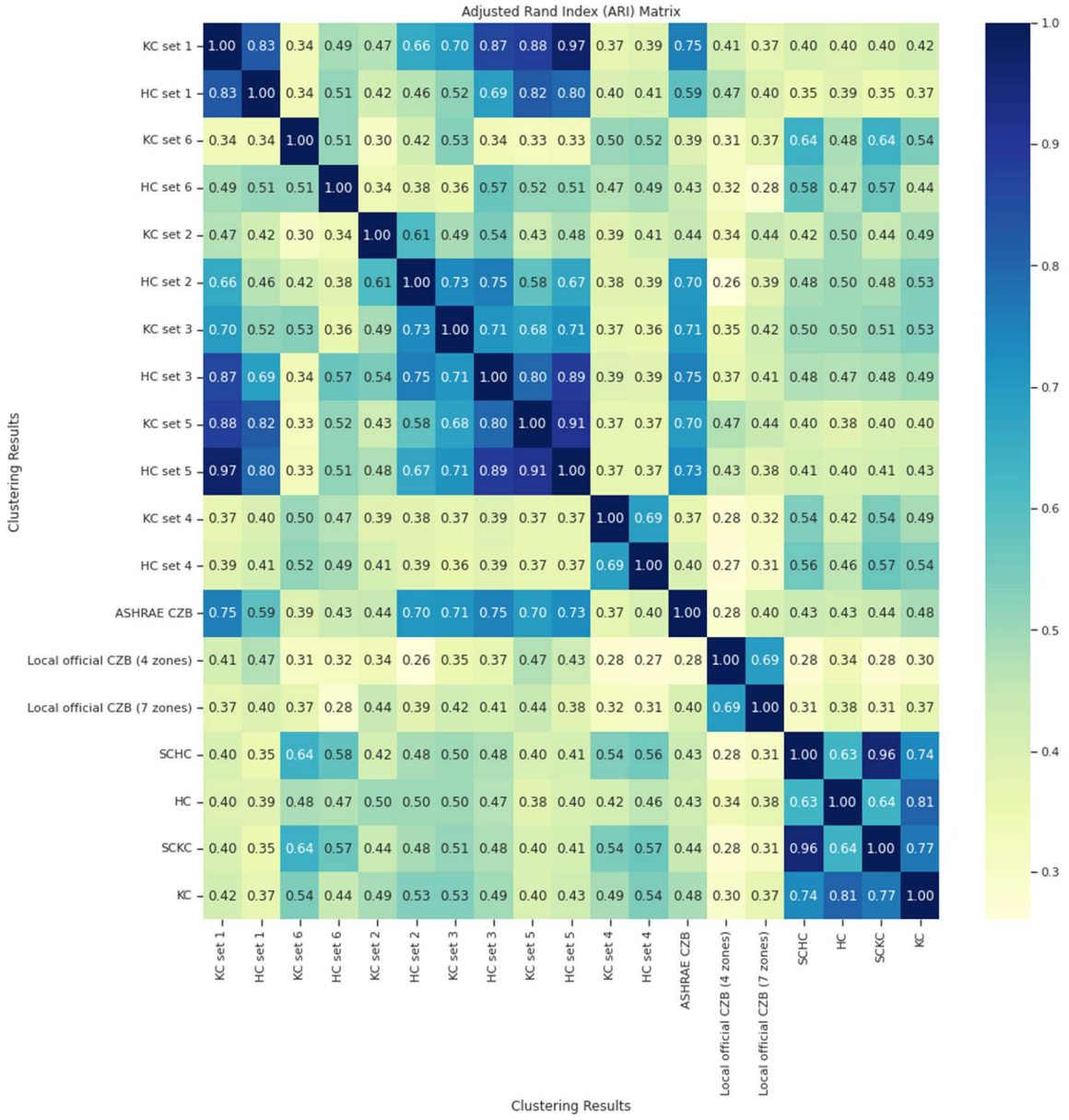


Figure 4.37: ARI matrix for all clustering methods.

This leads to the hypothesis that a clustering technique with a higher number of clusters and comparatively less intra-cluster overlaps is inherently a "better" choice. A higher number of clusters can provide more detailed data segmentation, potentially capturing subtler distinctions within the dataset. With that in mind, all performance-based (SCHC, HC, SCKC, and KC) methods should be considered the best for CZB classification in Kazakhstan. However, the optimality of a clustering result depends on the specific goals of classification, including the desired granularity of segmentation and the interpretability of the results. In contexts where a broader overview is preferred, or where the focus is on larger, more generalized patterns, a method producing fewer clusters might be considered more effective. The key lies in aligning the clustering approach with the analytical objectives and the inherent characteristics of the data.

The clustering results of the best-performing method (KC) were subjected to spatial interpolation within the ArcGIS Online software platform (Figure 4.38) to mark the transition from point data display to regional designations. The map presents a discernible north-south oriented pattern of CZB. In the northern area, there is a large climate zone 1, with a mean energy need of NA 96.7 kWh/m², and a range extending from 85.3 to 108.4 kWh/m², indicating a high demand for heating energy. For cooling in zone 1, NA buildings have a mean of 24.2 kWh/m², with a range of 12.1 kWh/m². The center of the country has a mix of narrow climatic zones (5, 0, and 3), extended along latitude. Southern regions have bigger and more uniform CZs 2 and 4, where in zone 4 NA exhibits the lowest heating demands among the zones, with a mean requirement of 34.9 kWh/m², while SA has a slightly higher mean heating demand of 43.3 kWh/m². However, both NA and SA buildings in Zone 4 display the highest cooling needs within their respective categories, with means of 59.0 kWh/m² and 65.3 kWh/m², respectively. Table 4.7 provides detailed information on the energy requirements for heating and cooling in different CZs.

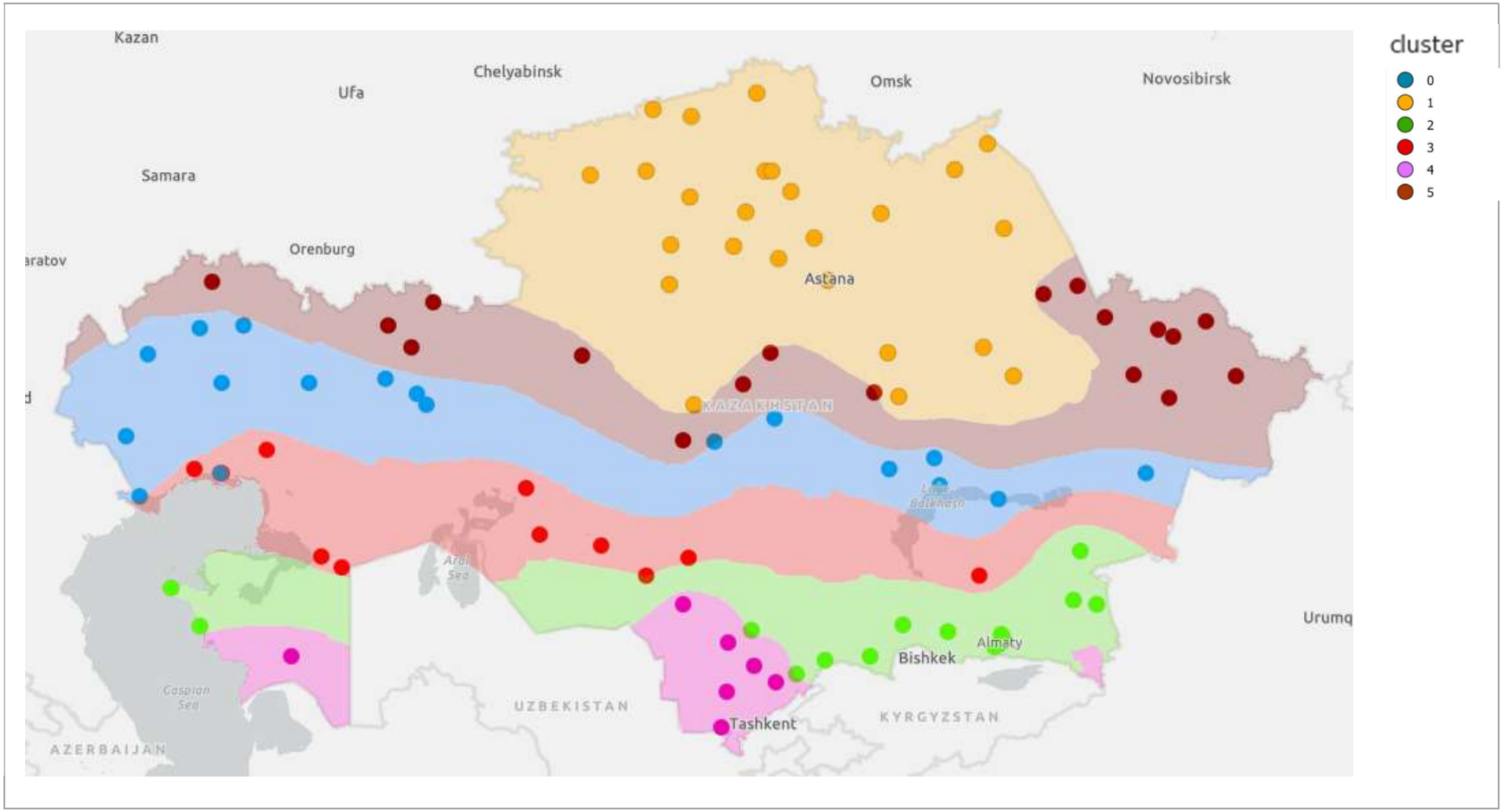


Figure 4.38: Final CZB maps of Kazakhstan based on the best-performed proposed methods.

Table 4.7: The mean values and range of heating and cooling energy consumption for each building type in the most effective KC approach.

Climate zone	Space Cooling NA [kWh/m^2] mean	Space Cooling NA [kWh/m^2] min	Space Cooling NA [kWh/m^2] max	Space Cooling NA [kWh/m^2] range	Space Heating NA [kWh/m^2] mean	Space Heating NA [kWh/m^2] min	Space Heating NA [kWh/m^2] max	Space Heating NA [kWh/m^2] range
0	40.4	36.3	45.6	9.3	70.3	57.8	80.5	22.7
1	24.2	18.4	30.5	12.1	96.7	85.3	108.4	23.1
2	43.6	37.8	48.7	10.9	40.7	30.4	49.0	18.6
3	52.2	48.8	54.0	5.2	56.8	49.7	67.1	17.4
4	59.0	53.6	62.4	8.8	34.9	27.5	42.9	15.4
5	31.6	23.1	39.4	16.3	84.3	76.3	90.6	14.3

	Space Cooling SA [kWh/m^2] mean	Space Cooling SA [kWh/m^2] min	Space Cooling SA [kWh/m^2] max	Space Cooling SA [kWh/m^2] range	Space Heating SA [kWh/m2] mean	Space Heating SA [kWh/m^2] min	Space Heating SA [kWh/m^2] max	Space Heating SA [kWh/m^2] range
0	43.7	39.1	49.3	10.2	86.8	69.9	98.9	29.0
1	25.0	18.5	32.1	13.6	120.0	105.5	132.7	27.2
2	46.8	39.6	54.0	14.4	51.3	38.3	64.1	25.8
3	57.8	53.9	60.2	6.3	69.5	61.7	81.4	19.7
4	65.3	58.5	69.4	10.9	43.3	34.4	52.8	18.4
5	33.2	22.6	42.9	20.3	104.7	96.1	112.5	16.4

4.8. Chapter Summary

This chapter delves into the comprehensive analysis of building performance simulations and energy performance spatial patterns in Kazakhstan. It investigates the significance of various variables for climate classification and examines two distinct phases of CZB. Phase 1 focuses on climate-based classification, validating results through building performance analysis, while Phase 2 shifts to performance-based classification, utilizing overlap analysis for deeper insights. A comparative analysis and synthesis of the findings highlight discrepancies and misclassifications, offering valuable insights into the effectiveness of different classification methodologies in optimizing CZB in Kazakhstan. The main findings from the chapter are listed below:

- The most buildings' energy-dependent climate variables for Kazakhstan are HDDs, CDDs, LAT, annual average DBT, and annual average GHI.
- The optimal number of CZ for Kazakhstan is 3 (in climate-based classification (KC set 1 and HC set 5)) and 6 (based on performance-based classification ((SCKC, KC)).
- Performance-based classification approaches, notably KC and SCHC, achieve higher accuracy with the lowest CZMI percentages (8.59% and 9.11% respectively), indicating reduced misclassification compared to all other methods and traditional CZB maps.
- Specific climate-based approaches (HC set 5 and KC set 1) show similarly low overlaps (9.42% and 9.62% respectively), matching the effectiveness of performance-based clustering.
- Although KC is more effective for performance-based clustering, HC exhibits much lower CZMI for climate-based clustering. Certain datasets provide significant enhancements in overlap percentages using HC, but others display very minimal disparities between the two techniques. The extent of enhancement varies across various datasets and can reach around 30% of the disparity (KC set 3 and HC set 3).
- The dataset and number of CZs influence the performance of specific clustering methods. Overall, increasing the number of variables in a data set and growing the number of clusters leads to decreased performance and higher CZMI values
- No significant evidence supports the superiority of spatially constrained methods over non-spatially constrained methods in reducing overlap percentage. Some instances show non-spatial methods with lower CZMI than spatially constrained methods, contrary to the typical expectation that spatial constraints improve clustering performance.

- The Local official CZB methods, with 4 and 7 CZs, yield moderate ARI values compared to the best-proposed methods, indicating little agreement. The ARIs for the Local official CZB (4 zones) and Local official CZB (7 zones) are primarily around 0.30, far departing from the best-proposed methods, suggesting notable differences in CZ identification and their connection with building energy performance.

Chapter 5: Conclusions

This thesis tackled the challenges presented in Chapter 1 by providing a framework for the development of building performance-based climate maps of Kazakhstan. This chapter addresses the conclusions, limitations, and future research recommendations.

5.1. Conclusions

This thesis proposed a novel building energy performance-based CZB map of Kazakhstan, which included a few steps:

- Typical building archetypes selection, where SFB was revealed as the predominant residential archetype in Kazakhstan with a noteworthy majority (71.8%) falling within the range of two to four rooms and an area of 52 to 95 m^2. R-values for external walls range from 3.20 for northern regions to 2.40 m^2K/W for southern regions, while roofs R-value vary from 4.00 to 2.40 m^2K/W. The process ensured that the proposed CZB was highly targeted and effective for the majority of the building.
- Selection of building performance indicators and performing simulations, as a result of which annual space heating (kWh/m^2) and annual space cooling (kWh/m^2) needs were chosen as the two primary energy performance components. The selection of performance indicators was influenced by the juxtaposing cold and warm seasons in Kazakhstan, which result in significant heating and cooling needs that affect the total yearly energy usage of buildings.
- Acquiring and analyzing data to discern spatial patterns in energy consumption levels among the dominant building types in Kazakhstan revealed that the annual total energy needs range from 140.5 to 174.2 kWh/m^2. The average annual total energy needs for NA are 125.3 kWh/m^2, whereas SA has a slightly higher average of 143.4 kWh/m^2. The regional climate is primarily characterized by a higher demand for heating, with heating needs representing around 60% of the overall energy usage. In general, the energy consumption pattern throughout the country is strongly linked to changes in latitude.
- Identifying the climate variables exerting the most substantial influence on building energy consumption within the Kazakh context using correlation analysis, random forest regression, gradient boosting, and extreme gradient boosting techniques revealed that HDDs, CDDs, LAT, annual average DBT, annual average GHI emerged as consistently impactful variables. In total, 47 variables were analyzed.

- Proposing CZ for local building archetypes by using multivariate cluster analysis ended up proposing 16 climate maps (12 climate-based and 4 performance-based), with a range of CZs (established by the Elbow method) from 3 to 8. To find the best CZB solution among 16 proposed maps the clustering quality assessment and verification were implemented. The main conclusions are:
- Cluster quality assessment in terms of uniqueness, dispersion indicators, and SS showed that determining the "best" clustering method is challenging due to the variation in which different methods excelled across distinct quality metrics. For climate-based maps, uniqueness favored KC set 2, HC set 2, KC set 4, and HC set 4. Compactness was best achieved by HC set 2, KC set 4, and HC set 4. The SS, which evaluates both cohesion within clusters and separation between them, identified KC set 5, HC set 5, KC set 5, and HC set 5 as superior, indicating a well-balanced clustering structure. For performance-based methods, SCHC and SCKC demonstrated increased clustering quality across three key dimensions: uniqueness, compactness, and SS, indicating their effectiveness in creating more coherent and distinct clusters. However, the results of SS across all configurations are quite close, highlighting that SCKC and SCHC configurations demonstrate insignificantly better performance.
- Accessing the accuracy and reliability of maps novel CZMI, which calculates the mean overlap percentages of KDE based on performance indicators was proposed. It showed that performance-based classification approaches, notably KC and SCHC, achieve higher accuracy with the lowest CZMI percentages (8.59% and 9.11% respectively), indicating reduced misclassification compared to all other methods. However, some climate-based approaches (HC set 5 (based on HDD18 (hourly method), CDD18 (hourly method), and LAT with 3 clusters) and KC set 1 (based on HDD18 (hourly method) with 3 clusters) show similarly low overlaps (9.42% and 9.62% respectively), matching the effectiveness of performance-based clustering.
- Climate-based classification, customized for specific regions, can match the accuracy of performance-based methods.
- For climate-based classification, a lower number of CZs is advantageous for achieving reduced CZMI values. Also the utilization of a single variable or a moderate number of variables, with three variables emerging as optimal within the analyzed conditions, significantly influences CZMI outcomes. This suggests a nuanced approach to variable selection is essential for optimizing climate zoning classifications.

- While performance methods are consistently accurate, climate-based results can greatly vary, raising the risk of errors.
- No significant evidence supports the superiority of spatially constrained methods over non-spatially constrained methods in reducing CZMI. Some instances show non-spatial methods with lower CZMI than spatially constrained methods, contrary to the typical expectation that spatial constraints improve clustering performance.
- The official CZB map showed the worst results of CZMI achieving the highest corrected CZMI percentage (20.24% and 22.14% respectively) among all methods. However, the ASHRAE map has a slightly higher accuracy (17.27%).
- As the final result, the KC CZB map was chosen as the best option, with the optimal number of CZ of 6.

Following the detailed steps undertaken for the development of the new CZB map, its effectiveness in refining Kazakhstan's existing standards and its possible practical applicability for policymakers and industry professionals should be examined. The existing CZB maps, although foundational, have shown limitations in accurately reflecting the diverse climatic impacts on building energy consumption across different regions. The new map, developed through advanced spatial analysis and multivariate clustering techniques, offers a more granular and accurate representation of climate zones, tailored to enhance energy efficiency standards. This refined zoning not only addresses the previous inaccuracies but also provides a tool with the potential to significantly influence policy and industry practices.

For policymakers, the updated map delivers a robust framework to guide the development of regional energy regulations and building codes that are better aligned with the specific climatic characteristics and building energy needs of each zone. By adopting this map, policymakers can enforce more precise standards that encourage sustainable building practices and enhance energy conservation measures, ultimately leading to reduced energy costs and lower carbon emissions. For the industry, particularly in the realms of architecture, construction, and urban planning, the new map serves as a critical reference that can inform smarter decisions regarding building design, material selection, and HVAC system implementation, tailored to the unique demands of Kazakhstan's varied climates. The introduction of this map promises to bridge the gap between theoretical zoning and practical, actionable insights that can drive energy efficiency in building practices across the nation.

5.2. Limitations of the current research

Although significant progress has been achieved in this study, it is crucial to recognize certain limitations that might affect the applicability and strength of the results. The findings may have limited application and generalizability due to the relatively small climatic data sample size and the special emphasis on a single case study in Kazakhstan. Future research should aim to increase the sample size in order to confirm the effectiveness of the suggested technique in more detail. Furthermore, the exclusion of severe weather occurrences in energy simulations as a result of using Typical Meteorological Year (TMY) data is a constraint. The research primarily aims to develop a CZB method that improves the energy consumption patterns of buildings. In the future, the method could be expanded to include considerations for extreme weather conditions, making it more comprehensive and applicable. In addition, the dependence on energy simulations that utilize historical weather data may fail to include the dynamic characteristics of climate change and its possible impacts on the future energy efficiency of buildings. This constraint emphasizes the importance of being careful when extending conclusions to future climate scenarios since they cannot completely encompass the whole spectrum of possible climatic fluctuations and extremes. In addition, this work uses a simple limitation based on latitude to evaluate the impact of spatial phenomena on CZB. Future studies might explore advanced spatial analysis methods such as Spatial Autocorrelation, SKATER, and DBSCAN to gain a deeper understanding of spatial patterns and interrelationships within CZB. Another constraint arises from the possibility of biases being created due to the choice of a solitary building archetype and the underlying assumptions that govern its portrayal. The validity of the findings relies on the accuracy and representativeness of a broader group of archetypes.

5.3. Recommendations for Future Research

Although the proposed method for CZB offers improvements in methodology, other potential areas require more investigation and development. The aforementioned research directions have the potential to improve and refine CZ classification approaches:

- Future work could focus on validating the CZB with a diverse set of buildings to refine its accuracy and enhance its generalizability. This would not only corroborate the initial findings but also enable the development of tailored energy-efficiency strategies across the broader architectural spectrum, thereby reinforcing the utility of climate zoning as a pivotal tool in sustainable building practices.

- The present study focuses on a particular geographic area, and extending the suggested approach to a broader or worldwide scope might be a promising direction for future studies. Creating an accurate and universally applicable classification system would increase its usefulness for international building design, energy policy creation, and comparative analysis.
- Exploring the potential integration of advanced machine learning techniques and data analytics methods could unlock new perspectives for CZB research, facilitating more sophisticated pattern recognition and classification algorithms. This could lead to more accurate and adaptive CZB frameworks capable of capturing complex interactions between climate, building characteristics, and energy consumption data.
- The proposed method relies on building performance indicators and meteorological data as its main inputs. To improve the accuracy of classification and obtain a more comprehensive understanding of the impact of energy consumption patterns and climate on buildings, future research could investigate the incorporation of supplementary data sources, such as socio-economic data, land use patterns, and urban morphology.
- The efficacy of the suggested approach in capturing long-term performance patterns and its adaptation to future climate scenarios should be investigated, considering its inclusion in building energy simulations. To achieve this, future weather files (F-TMY), generated based on regional climate models (RCMs) or global climate models (GCMs) that provide future climate data based on different Representative Concentration Pathways (RCPs), could be integrated into the CZB process [200, 201]. This approach would allow for the evaluation of the long-term efficacy and resilience of the proposed CZB framework under anticipated climatic shifts. Furthermore, the integration of future weather files into CZB analysis would facilitate the development of dynamic climate zoning maps that can adapt over time, reflecting the evolving nature of climate impacts on building energy performance. This dynamic modeling would be critical in ensuring that building codes and energy efficiency strategies remain relevant and effective as the climate continues to change.
- It is essential to comprehend the usability and user satisfaction of the proposed approach in order to effectively apply it. Additional research might involve user-centric evaluations, such as questionnaires, interviews, and usability testing, to get input from architects, engineers, policymakers, and other parties. This feedback can offer valuable perspectives

and enhance the continuous improvement of the system, guaranteeing its usefulness and applicability in real-life scenarios.

Bibliography

1. Walsh, A., D. Cóstola, and L.C. Labaki, *Review of methods for climatic zoning for building energy efficiency programs.* Building and Environment, 2017. **112**: p. 337-350.
2. Albatayneh, A., et al., *The Significance of Building Design for the Climate.* Environmental and Climate Technologies, 2018. **22**: p. 165-178.
3. Mazzaferro, L., et al., *Do we need building performance data to propose a climatic zoning for building energy efficiency regulations?* Energy and Buildings, 2020. **225**.
4. Walsh, A., D. Cóstola, and L.C. Labaki, *Performance-based validation of climatic zoning for building energy efficiency applications.* Applied Energy, 2018. **212**: p. 416-427.
5. Walsh, A., D. Cóstola, and L.C. Labaki, *Comparison of three climatic zoning methodologies for building energy efficiency applications.* Energy and Buildings, 2017. **146**: p. 111-121.
6. Cory, S., et al. *Formulating a building climate classification method*. in *12th Conference of International Building Performance Simulation Association Building Simulation 2011, BS 2011*. 2011. Sydney, NSW: International Building Performance Simulation Association.
7. Jain, K., et al., *Climatic Classification of India for Building Design Using Data Analytics.* National Academy Science Letters, 2022. **45**(3): p. 235-239.
8. Benevides, M.N., D.B.D.S. Teixeira, and J.C. Carlo, *Climatic zoning for energy efficiency applications in buildings based on multivariate statistics: The case of the Brazilian semiarid region.* Frontiers of Architectural Research, 2022. **11**(1): p. 161-177.
9. Bienvenido-Huertas, D., et al., *Climate classification for new and restored buildings in Andalusia: Analysing the current regulation and a new approach based on k-means.* Journal of Building Engineering, 2021. **43**.
10. Deng, X., et al., *A clustering-based climatic zoning method for office buildings in China.* Journal of Building Engineering, 2021. **42**.
11. Tükel, M., et al., *Reclassification of climatic zones for building thermal regulations based on thermoeconomic analysis: A case study of Turkey.* Energy and Buildings, 2021. **246**.
12. Roshan, G., M. Farrokhzad, and S. Attia, *Climatic clustering analysis for novel atlas mapping and bioclimatic design recommendations.* Indoor and Built Environment, 2021. **30**(3): p. 313-333.
13. Verichev, K., et al., *Analysis of Climate-Oriented Researches in Building.* Applied Sciences, 2021. **11**: p. 3251.
14. Wang, R. and S. Lu, *A novel method of building climate subdivision oriented by reducing building energy demand.* Energy and Buildings, 2020. **216**.

15. Architecture", J.S.C.K.R.a.D.I.o.C.a., *SP RK 2.04-01-2017 - Building Climatology*. 2017, Construction and Housing-Communal Services Affairs Committee of the Ministry of Industry and Infrastructural Development of the Republic of Kazakhstan.
16. Structures, R.I.o.B.P.a.F., *SNiP 2.01.01-82 CONSTRUCTION CLIMATOLOGY AND GEOPHYSICS*. 1983, USSR State Committee Of Construction.
17. Xiong, J., et al., *A hierarchical climatic zoning method for energy efficient building design applied in the region with diverse climate characteristics.* Energy and Buildings, 2019. **186**: p. 355-367.
18. Verichev, K., M. Zamorano, and M. Carpio, *Assessing the applicability of various climatic zoning methods for building construction: Case study from the extreme southern part of Chile.* Building and Environment, 2019. **160**.
19. Committee for Construction, H.a.C.S.o.t.M.o.I.a.I.D.o.t.R.o.K., *Thermal Protection of Buildings*. 2022, Committee for Construction, Housing and Communal Services of the Ministry of Industry and Infrastructure Development of the Republic of Kazakhstan: Astana.
20. Kazakhstan, M.o.I.a.I.D.o.t.R.o., *On approval of the Concept for the development of energy saving and increasing energy efficiency of the Republic of Kazakhstan for 2023 – 2029*. 2023: Astana.
21. Tokayev, K.-J., *MESSAGE FROM THE HEAD OF STATE TO THE PEOPLE OF KAZAKHSTAN "Economic course of a Fair Kazakhstan"*. 2023.
22. Uyzbayeva, A., T. Valeriya, and S. Artem, *A Case Study of Energy Modeling of a School Building in Astana City (Kazakhstan): Applications*. 2018. p. 967-984.
23. Tukhtamisheva, A., et al. *Optimization of the Thermal Insulation Level of Residential Buildings in the Almaty Region of Kazakhstan*. Energies, 2020. **13**, DOI: 10.3390/en13184692.
24. Kim, Y. and C. Sun, *The Energy-Efficient Adaptation Scheme for Residential Buildings in Kazakhstan.* Energy Procedia, 2017. **118**: p. 28-34.
25. Tokbolat, S., R. Tokpatayeva, and S.N. Al-Zubaidy, *The Effects of Orientation on Energy Consumption in Buildings in Kazakhstan.* Journal of Solar Energy Engineering, 2013. **135**(4).
26. IEA, *Kazakhstan energy profile*. 2020: Paris
27. Bank, W., *Unlocking Energy Efficiency Potentials in*

Cities in Kazakhstan. 2018.
28. Sanderson, M., *The Classification of Climates from Pythagoras to Koeppen.* Bulletin of the American Meteorological Society, 1999. **80**(4): p. 669-673.
29. Oliver, J., *The history, status and future of climatic classification.* Physical Geography, 2013. **12**: p. 231-251.

30. Robinson, A.H. and H.M. Wallis, *Humboldt's Map of Isothermal Lines: A Milestone in Thematic Cartography.* Cartographic Journal, 1967. **4**: p. 119-123.
31. Kottek, M., et al., *World Map of the Köppen-Geiger Climate Classification Updated.* Meteorologische Zeitschrift, 2006. **15**: p. 259-263.
32. Beck, H.E., et al., *Present and future köppen-geiger climate classification maps at 1-km resolution.* Scientific Data, 2018. **5**.
33. Rubel, F., et al., *The climate of the European Alps: Shift of very high resolution Köppen-Geiger climate zones 1800-2100.* Meteorologische Zeitschrift, 2017. **26**(2): p. 115-125.
34. Rubel, F. and M. Kottek, *Observed and projected climate shifts 1901-2100 depicted by world maps of the Köppen-Geiger climate classification.* Meteorologische Zeitschrift, 2010. **19**(2): p. 135-141.
35. Rubel, F. and M. Kottek, *Comments on: "The thermal zones of the Earth" by Wladimir Köppen (1884).* Meteorologische Zeitschrift, 2011. **20**(3): p. 361-365.
36. Zscheischler, J., M.D. Mahecha, and S. Harmeling. *Climate classifications: The value of unsupervised clustering*. in *12th Annual International Conference on Computational Science, ICCS 2012*. 2012. Omaha, NB: Elsevier B.V.
37. Republic, C.C.o.t.R.S.F.S., *Rules and regulations for the development of populated areas, design and construction of buildings and structures*. 1930, State technical publishing house: Moscow.
38. Kupriyanov, V.N., *Construction Climatology And Environmental Physics*. 2007, Kazan Federal Agency for Education Kazan State University of Architecture and Civil Engineering.
39. *National Building Code: 1941*. 1941, National Research Council of Canada.
40. American Society of Heating, R.A.-C., Engineers, *ASHRAE Standard 90-75: Energy Conservation in New Building Design*. 1975: American Society of Heating, Refrigerating, and Air-Conditioning Engineers, Incorporated.
41. LAUSTSEN, J., *Energy Efficiency Requirements in Building Codes: Policies for New Buildings*. 2008, International Energy Agency (IEA).
42. Bohne, R.A., L. Huang, and J. Lohne, *A global overview of residential building energy consumption in eight climate zones.* International Journal of Sustainable Building Technology and Urban Development, 2016. **7**(1): p. 38-51.
43. Díaz-López, C., et al., *Evolution of climate zones for building in Spain in the face of climate change.* Sustainable Cities and Society, 2021. **74**.
44. Omer, A.M., *Renewable building energy systems and passive human comfort solutions.* Renewable and Sustainable Energy Reviews, 2008. **12**(6): p. 1562-1587.

45. Abebe, S. and T. Assefa, *Development of climatic zoning and energy demand prediction for Ethiopian cities in degree days.* Energy and Buildings, 2022. **260**: p. 111935.
46. Alrashed, F. and M. Asif. *Climatic Classifications of Saudi Arabia for Building Energy Modelling*. in *7th International Conference on Applied Energy, ICAE 2015*. 2015. Elsevier Ltd.
47. Bai, L. and S. Wang, *Definition of new thermal climate zones for building energy efficiency response to the climate change during the past decades in China.* Energy, 2019. **170**: p. 709-719.
48. Bai, L., et al., *A new approach to develop a climate classification for building energy efficiency addressing Chinese climate characteristics.* Energy, 2020. **195**.
49. Carpio, M., et al., *A proposed method based on approximation and interpolation for determining climatic zones and its effect on energy demand and CO2 emissions from buildings.* Energy and Buildings, 2015. **87**: p. 253-264.
50. Chen, Y., et al., *Effect of climate zone change on energy consumption of office and residential buildings in China.* Theoretical and Applied Climatology, 2021. **144**(1-2): p. 353-361.
51. Mazzaferro, L., et al. *Climatic zoning methodology based on data-driven approach*. in *16th International Conference of the International Building Performance Simulation Association, Building Simulation 2019*. 2019. International Building Performance Simulation Association.
52. Naveen Kishore, K. and J. Rekha, *A bioclimatic approach to develop spatial zoning maps for comfort, passive heating and cooling strategies within a composite zone of India.* Building and Environment, 2018. **128**: p. 190-215.
53. Praene, J.P., et al., *GIS-based approach to identify climatic zoning: A hierarchical clustering on principal component analysis.* Building and Environment, 2019. **164**.
54. Walsh, A., D. Cóstola, and L.C. Labaki, *Validation of the climatic zoning defined by ASHRAE standard 169-2013.* Energy Policy, 2019. **135**.
55. Yang, L., et al., *Building climate zoning in China using supervised classification-based machine learning.* Building and Environment, 2020. **171**.
56. Zeleke, B., M. Kumar, and E. Rajasekar, *A Novel Building Performance Based Climate Zoning for Ethiopia.* Frontiers in Sustainable Cities, 2022. **4**.
57. Day, T., *TM 41 Degree-days: theory and application*, in *TM 41 Degree-days: theory and application*. 2006, Chartered Institution of Building Services Engineers (CIBSE): London, UK.
58. Quayle, R.G. and H.F. Diaz, *Heating Degree Day Data Applied to Residential Heating Energy Consumption.* Journal of Applied Meteorology (1962-1982), 1980. **19**(3): p. 241-246.

59. Le Comte, D.M. and H.E. Warren, *Modeling the Impact of Summer Temperatures on National Electricity Consumption.* Journal of Applied Meteorology, 1981. **20**: p. 1415-1419.
60. Lehman, R.L. and H.E. Warren, *Residential Natural Gas Consumption: Evidence That Conservation Efforts to Date Have Failed.* Science, 1978. **199**(4331): p. 879-882.
61. Warren, H.E. and S.K. LeDuc, *Impact of Climate on Energy Sector in Economic Analysis.* Journal of Applied Meteorology (1962-1982), 1981. **20**(12): p. 1431-1439.
62. Pusat, S. and I. Ekmekci, *A study on degree-day regions of Turkey.* Energy Efficiency, 2016. **9**(2): p. 525-532.
63. Noh, B., J. Choi, and D. Seo, *A Study on the Classification Criteria of Climatic Zones in Korean Building Code Based on Heating Degree-Days.* Korean Journal of Air-Conditioning and Refrigeration Engineering, 2015. **27**: p. 574-580.
64. Ghedamsi, R., et al., *Modeling and forecasting energy consumption for residential buildings in Algeria using bottom-up approach.* Energy and Buildings, 2016. **121**: p. 309-317.
65. Wan, K.K.W., et al., *Climate classifications and building energy use implications in China.* Energy and Buildings, 2010. **42**(9): p. 1463-1471.
66. Bishop, C.M. and N.M. Nasrabadi, *Pattern recognition and machine learning*. Vol. 4. 2006: Springer.
67. Mohammed, M., M. Khan, and E. Bashier, *Machine Learning: Algorithms and Applications*. 2016.
68. Sarker, I.H., *Machine Learning: Algorithms, Real-World Applications and Research Directions.* SN Computer Science, 2021. **2**(3): p. 160.
69. Jolliffe, I.T., *Principal component analysis for special types of data*. 2002: Springer.
70. Fovell, R.G. and M.Y.C. Fovell, *Climate zones of the conterminous United States defined using cluster analysis.* Journal of Climate, 1993. **6**(11): p. 2103-2135.
71. Lau, C.C.S., J.C. Lam, and L. Yang, *Climate classification and passive solar design implications in China.* Energy Conversion and Management, 2007. **48**(7): p. 2006-2015.
72. Shi, J. and L. Yang, *A climate classification of China through k-nearest-neighbor and sparse subspace representation.* Journal of Climate, 2020. **33**(1): p. 243-262.
73. Erell, E., B. Portnov, and Y. Etzion, *Mapping the potential for climate-conscious design of buildings.* Building and Environment, 2003. **38**(2): p. 271-281.

74. van Schijndel, A.W.M. and H.L. Schellen, *The simulation and mapping of building performance indicators based on European weather stations.* Frontiers of Architectural Research, 2013. **2**(2): p. 121-133.
75. Board, A.B.C., *NatHERS heating and cooling load limits* 2019, Australian Building Codes Board Standard
76. (ADEREE), N.A.f.t.D.o.R.E.a.E.E., *THEMIC CONSTRUCTION REGULATIONS IN MAROC (RtCm)*. 2013.
77. Semahi, S., et al., *Development of spatial distribution maps for energy demand and thermal comfort estimation in Algeria.* Sustainability (Switzerland), 2020. **12**(15).
78. China, t.M.o.C.o., *GB 50176-2016*. 2016.
79. 175, A.s.K., *ÖNORM B 8110-2 Wärmeschutz im Hochbau Teil 2: Wasserdampfdiffusion und Kondensationsschutz. Thermal insulation in building construction Part 2: Water vapor diffusion and protection against condensation.* 2003.
80. norm, M.O.C.A.U.O.T.R.O.L.c., *Construction climatology. RSN 156-94*. 2002.
81. Michał Strzeszewski, P.W., *"Norma PN–EN 12831Nowa metoda obliczania projektowego obciążenia cieplnego (A new method for calculating the design heat load)".* 2009.
82. COUNCIL, C., *ORDIN nr. 386 din 28 martie 2016*. 2016.
83. Ministry of Environment and Sustainable Development Vice Ministry of Environment and Sustainable Development Directorate of Environmental, S.a.U.A.R.o.C., *Environmental criteria for the design and construction of urban housing*. 2012.
84. Sciences", F.S.B.I.R.I.o.B.P.o.t.R.A.o.A.a.B., *SP 131.13330.2018 "BUILDING CLIMATOLOGY".* 2019.
85. Li, M., et al., *Climate Impacts on Extreme Energy Consumption of Different Types of Buildings.* PloS one, 2015. **10**: p. e0124413.
86. Arens, E.A. and P.B. Williams, *The effect of wind on energy consumption in buildings.* Energy and Buildings, 1977. **1**(1): p. 77-84.
87. Jovanović, S., et al., *The impact of the mean daily air temperature change on electricity consumption.* Energy, 2015. **88**: p. 604-609.
88. M. N. K. De Silva, Y.G.S., *Building energy consumption factors: a literature review and future research agenda* in *World Construction Conference 2012 – Global Challenges in Construction Industry*

2012: Colombo, Sri Lanka.
89. Lam, J.C., L. Yang, and J. Liu, *Development of passive design zones in China using bioclimatic approach.* Energy Conversion and Management, 2006. **47**(6): p. 746-762.
90. Singh, M.K., S. Mahapatra, and S.K. Atreya, *Development of bio-climatic zones in north-east India.* Energy and Buildings, 2007. **39**(12): p. 1250-1257.

91. Pajek, L. and M. Košir, *Implications of present and upcoming changes in bioclimatic potential for energy performance of residential buildings.* Building and Environment, 2018. **127**: p. 157–172.
92. (IRAM), E.I.A.d.N.y.C., *ESQUEMA 1. Acondicionamiento térmico de edificios. Clasificación bioambiental de la República Argentina*. 2011.
93. Olgyay, V., *Design with Climate: A Bioclimatic Approach to Architectural Regionalism*. 1992, New York: Van Nostrand Reinhold.
94. M. Milne, B.G., *Architectural design based on climate.* Energy Conservation through Building Design, 1979: p. 96-113.
95. Givoni, B., *Man, climate, and architecture,*. Elsevier architectural science series. 1969, Amsterdam; London; New York: Elsevier Publishing Company Limited.
96. Rakoto-Joseph, O., et al., *Development of climatic zones and passive solar design in Madagascar.* Energy Conversion and Management, 2009. **50**(4): p. 1004-1010.
97. Ali, S.I.A. and Z. Szalay, *Overview and analysis of the overheating effect in modern sudanese buildings.* Pollack Periodica, 2020. **15**(3): p. 208-219.
98. Hobaica, M., F. Allard, and R. Belarbi, *Passive Cooling Systems for Buildings. Potential of use within Venezuela's Climatic Zones.* 2002.
99. Sarricolea, P., M. Herrera-Ossandon, and Ó. Meseguer-Ruiz, *Climatic regionalisation of continental Chile.* Journal of Maps, 2017. **13**(2): p. 66-73.
100. Peel, M.C., B.L. Finlayson, and T.A. McMahon, *Updated world map of the Köppen-Geiger climate classification.* Hydrology and Earth System Sciences, 2007. **11**(5): p. 1633-1644.
101. Butera, F., N. Aste, and R. Adhikari, *Sustainable Building Design for Tropical Climates*. 2015.
102. Netzel, P. and T. Stepinski, *On using a clustering approach for global climate classification.* Journal of Climate, 2016. **29**(9): p. 3387-3401.
103. Falquina, R. and C. Gallardo, *Development and application of a technique for projecting novel and disappearing climates using cluster analysis.* Atmospheric Research, 2017. **197**: p. 224-231.
104. Attia, S., et al., *Analysis tool for bioclimatic design strategies in hot humid climates.* Sustainable Cities and Society, 2019. **45**: p. 8-24.
105. da Casa Martín, F., E. Echeverría Valiente, and F. Celis D'Amico, *Climate zoning for its application to bioclimatic design. Application in Galicia (Spain).* Informes de la Construccion, 2017. **69**(547).
106. Federici, A., et al., *Climatic Severity Index: definition of summer climatic zones in Italy through the assessment of air conditioning energy need in buildings*. 2013.
107. Moral, F.J., et al., *Climatic zoning for the calculation of the thermal demand of buildings in Extremadura (Spain).* Theoretical and Applied Climatology, 2017. **129**(3-4): p. 881-889.

108. Verichev, K. and M. Carpio, *Climatic zoning for building construction in a temperate climate of Chile.* Sustainable Cities and Society, 2018. **40**: p. 352-364.
109. Roriz, M., E. Ghisi, and R. Lamberts, *Bioclimatic zoning of Brazil: a proposal based on the Givoni and Mahoney methods.* Proceedings volume 1, 1999.
110. Selek, B., I. Kaan Tuncok, and Z. Selek, *Changes in climate zones across Turkey.* Journal of Water and Climate Change, 2018. **9**(1): p. 178-195.
111. Izzo, M., et al., *A new climatic map of the dominican republic based on the thornthwaite classification.* Physical Geography, 2010. **31**(5): p. 455-472.
112. Vietnam, M.o.C.o., *Vietnam Building Code 2009 - QCVN 02: 2021/BXD - Natural Physical & Climatic Data for Construction*. 2009.
113. Vondráková, A., A. Vávra, and V. Voženílek, *Climatic regions of the Czech Republic.* Journal of Maps, 2013. **9**(3): p. 425-430.
114. Gangolells, M., et al., *Energy mapping of existing building stock in Spain.* Journal of Cleaner Production, 2016. **112**: p. 3895-3904.
115. Hjortling, C., et al., *Energy mapping of existing building stock in Sweden – Analysis of data from Energy Performance Certificates.* Energy and Buildings, 2017. **153**: p. 341-355.
116. Joan Felix, L.D.P., Raykenler Izquierdo, *Análisis comparativo de las diferentes zonas climáticas de la república dominicana.* Proceedings of the 1st Iberic Conference on Theoretical and Experimental Mechanics and Materials/11th National Congress on Experimental Mechanics., 2018: p. 865-876.
117. Khedari, J., A. Sangprajak, and J. Hirunlabh, *Thailand climatic zones.* Renewable Energy, 2002. **25**(2): p. 267-280.
118. Díaz-López, C., et al., *Dynamics of changes in climate zones and building energy demand. A case study in spain.* Applied Sciences (Switzerland), 2021. **11**(9).
119. Mayes Boustead, B., et al., *The Accumulated Winter Season Severity Index (AWSSI).* Journal of Applied Meteorology and Climatology, 2015. **54**: p. 150326122910004.
120. Walker, C.L., et al., *Developing a winter severity index: A critical review.* Cold Regions Science and Technology, 2019. **160**: p. 139-149.
121. Verichev, K., A. Salimova, and M. Carpio, *Thermal and climatic zoning for construction in the southern part of Chile.* Advances in Science and Research, 2018. **15**: p. 63-69.
122. Ogunsote, O. and B. Prucnal-Ogunsote, *Defining climatic zones for architectural design in Nigeria: a systematic delineation.* J Environ Technol, 2002. **1**: p. 1-14.
123. Thornthwaite, C.W., *An Approach toward a Rational Classification of Climate.* Geographical Review, 1948. **38**(1): p. 55-94.

124. Tsikaloudaki, K., K. Laskos, and D. Bikas, *On the establishment of climatic zones in Europe with regard to the energy performance of buildings.* Energies, 2012. **5**(1): p. 32-44.
125. Woods, J. and C. Fuller, *Estimating base temperatures in econometric models that include degree days.* Energy Economics, 2014. **45**: p. 166-171.
126. Li, L., et al., *Impact of natural and social environmental factors on building energy consumption: Based on bibliometrics.* Journal of Building Engineering, 2021. **37**: p. 102136.
127. Thornton, B.A., et al. *Technical Support Document: 50% Energy Savings for Small Office Buildings*. 2010.
128. Chen, Y., Z. Wang, and P. Wei, *Climatic zoning for the building thermal design in China's rural areas.* Building Services Engineering Research & Technology, 2021. **42**: p. 567 - 581.
129. Ascione, F., et al., *Resilience of robust cost-optimal energy retrofit of buildings to global warming: A multi-stage, multi-objective approach.* Energy and Buildings, 2017. **153**: p. 150-167.
130. Gillingham, K.T., et al., *The climate and health benefits from intensive building energy efficiency improvements.* Science Advances, 2021. **7**(34).
131. Guttman, N.B. and R.L. Lehman, *Estimation of Daily Degree-hours.* Journal of Applied Meteorology, 1992. **31**: p. 797-797.
132. Omarov, B., S.A. Memon, and J. Kim, *A novel approach to develop climate classification based on degree days and building energy performance.* Energy, 2023. **267**: p. 126514.
133. *Construction statistics*. 2022 [cited 2023 1.02.2023]; Available from: https://stat.gov.kz/en/industries/business-statistics/stat-inno-build/publications/.
134. Rousseeuw, P.J., *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.* Journal of Computational and Applied Mathematics, 1987. **20**: p. 53-65.
135. Syakur, M.A., et al., *Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster.* IOP Conference Series: Materials Science and Engineering, 2018. **336**(1): p. 012017.
136. Breiman, L., et al., *Classification and regression trees*. Classification and Regression Trees. 2017. 1-358.
137. Gupta, A., K. Gusain, and B. Popli. *Verifying the value and veracity of extreme gradient boosted decision trees on a variety of datasets*. in *2016 11th International Conference on Industrial and Information Systems (ICIIS)*. 2016.
138. Wei, Y., et al., *A review of data-driven approaches for prediction and classification of building energy consumption.* Renewable and Sustainable Energy Reviews, 2018. **82**: p. 1027-1047.

139. Anselin, L., *Local Indicators of Spatial Association—LISA.* 1995. **27**: p. 93-115.
140. Dowd, P., et al., *Constrained Spatial Clustering of Climate Variables for Geostatistical Reconstruction of Optimal Time Series and Spatial Fields*. 2017. p. 879-891.
141. Kazakhstan, N.h.s.o. *Climate Of Kazakhstan*. 2023 [cited 2023 july 4]; Available from: https://www.kazhydromet.kz/en/klimat/klimat-kazahstana.
142. Dru Crawley, L.L. *Development of Global Typical Meteorological Years (TMYx)*. 2022; Available from: https://climate.onebuilding.org/.
143. 89, T.E.C.f.S.C.T.C.C.T., *ISO 15927-4:2005*, in *Hygrothermal performance of buildings — Calculation and presentation of climatic data — Part 4: Hourly data for assessing the annual energy use for heating and cooling*, 2022.
144. S. Wilcox, W.M., *Users Manual for TMY3 Data Sets*. 2008, National Renewable Energy Laboratory: Golden, Colorado.
145. Bre, F., et al., *Assessment of solar radiation data quality in typical meteorological years and its influence on the building performance simulation.* Energy and Buildings, 2021. **250**.
146. Crawley, D. and L. Lawrie, *Should We Be Using Just 'Typical' Weather Data in Building Performance Simulation?* 2019.
147. Huld, T., et al. *Assembling Typical Meteorological Year Data Sets for Building Energy Performance Using Reanalysis and Satellite-Based Data*. Atmosphere, 2018. **9**, DOI: 10.3390/atmos9020053.
148. Chan, A.L.S., et al., *Generation of a typical meteorological year for Hong Kong.* Energy Conversion and Management, 2006. **47**(1): p. 87-96.
149. Sun, J., Z. Li, and F. Xiao, *Analysis of Typical Meteorological Year selection for energy simulation of building with daylight utilization.* Procedia Engineering, 2017. **205**: p. 3080-3087.
150. Honglian, L., et al., *Compare several methods of select typical meteorological year for building energy simulation in China.* Energy, 2020. **209**: p. 118465.
151. Lechner, N., *Heating, Cooling, Lighting: Sustainable Design Methods for Architects*. 2014: Wiley.
152. Eto, J.H., *On using degree-days to account for the effects of weather on annual energy use in office buildings.* Energy and Buildings, 1988. **12**(2): p. 113-127.
153. Fathi, A., et al., *The Effect of Outdoor Air Temperature on the Thermal Performance of a Residential Building.* Journal of Multidisciplinary Engineering Science and Technology, 2015. **2**: p. 3159-40.
154. Fumo, N., P. Mago, and R. Luck, *Methodology to estimate building energy consumption using EnergyPlus Benchmark Models.* Energy and Buildings, 2010. **42**(12): p. 2331-2337.

155. Echenagucia, T., et al., *The early design stage of a building envelope: Multi-objective search through heating, cooling and lighting energy performance analysis.* Applied Energy, 2015. **154**: p. 577-591.
156. American Society of Heating, R. and E. Air-Conditioning, *2017 ASHRAE handbook. Fundamentals*. Inch-pound edition ed. 2017, Atlanta, GA: ASHRAE, [2017] Atlanta, GA.
157. Willmott, C.J. and K. Matsuura, *Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance.* Climate Research, 2005. **30**(1): p. 79-82.
158. Willmott, C.J., *On the validation of models.* Physical Geography, 1981. **2**(2): p. 184-194.
159. Hodson, T.O., T.M. Over, and S.S. Foks, *Mean Squared Error, Deconstructed.* Journal of Advances in Modeling Earth Systems, 2021. **13**(12): p. e2021MS002681.
160. Biau, G. and E. Scornet, *A random forest guided tour.* Test, 2016. **25**(2): p. 197-227.
161. Breiman, L., *Random forests.* Machine Learning, 2001. **45**(1): p. 5-32.
162. Sen, P.K., *Gini diversity index, Hamming distance, and curse of dimensionality.* Metron, 2005. **63**(3): p. 329-349.
163. Strobl, C., A.L. Boulesteix, and T. Augustin, *Unbiased split selection for classification trees based on the Gini Index.* Computational Statistics and Data Analysis, 2007. **52**(1): p. 483-501.
164. Svetnik, V., et al., *Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling.* Journal of Chemical Information and Computer Sciences, 2003. **43**(6): p. 1947-1958.
165. Natekin, A. and A. Knoll, *Gradient boosting machines, a tutorial.* Frontiers in Neurorobotics, 2013. **7**(DEC).
166. Bühlmann, P. and B. Yu, *Boosting with the L2 loss: Regression and classification.* Journal of the American Statistical Association, 2003. **98**(462): p. 324-339.
167. Friedman, J.H., *Greedy function approximation: A gradient boosting machine.* Annals of Statistics, 2001. **29**(5): p. 1189-1232.
168. Chen, T. and C. Guestrin. *XGBoost: A scalable tree boosting system*. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016.
169. Liu, W., Z. Chen, and Y. Hu, *XGBoost algorithm-based prediction of safety assessment for pipelines.* International Journal of Pressure Vessels and Piping, 2022. **197**: p. 104655.
170. Wang, Z., T. Hong, and M.A. Piette, *Building thermal load prediction through shallow machine learning and deep learning.* Applied Energy, 2020. **263**.

171. Legendre, P. and L. Legendre, *Cluster analysis*, in *Developments in Environmental Modelling*. 2012. p. 337-424.
172. Murtagh, F. and P. Contreras, *Algorithms for hierarchical clustering: An overview.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2012. **2**(1): p. 86-97.
173. Anas, H., et al., *Novel climate classification based on the information of solar radiation intensity: An application to the climatic zoning of Morocco.* Energy Conversion and Management, 2021. **247**.
174. Jain, A.K., *Data clustering: 50 years beyond K-means.* Pattern Recognition Letters, 2010. **31**(8): p. 651-666.
175. Wu, X., et al., *Top 10 algorithms in data mining.* Knowledge and Information Systems, 2008. **14**(1): p. 1-37.
176. Johnson, S.C., *Hierarchical clustering schemes.* Psychometrika, 1967. **32**(3): p. 241-254.
177. Fraley, C. and A.E. Raftery, *How many clusters? Which clustering method? Answers via model-based cluster analysis.* Computer Journal, 1998. **41**(8): p. 586-588.
178. Sokal, R.R. and F.J. Rohlf, *THE COMPARISON OF DENDROGRAMS BY OBJECTIVE METHODS.* TAXON, 1962. **11**(2): p. 33-40.
179. Farris, J.S., *On the Cophenetic Correlation Coefficient.* Systematic Biology, 1969. **18**(3): p. 279-285.
180. Unal, Y., T. Kindap, and M. Karaca, *Redefining the climate zones of Turkey using cluster analysis.* International Journal of Climatology, 2003. **23**(9): p. 1045-1055.
181. Daly, C., *Guidelines for assessing the stability of spatial climate data sets.* International Journal of Climatology, 2006. **26**: p. 707-721.
182. Raymundo, C.E., et al., *Spatial analysis of COVID-19 incidence and the sociodemographic context in Brazil.* PLOS ONE, 2021. **16**(3): p. e0247794.
183. HALL, A. and G.V. JONES, *Spatial analysis of climate in winegrape-growing regions in Australia.* Australian Journal of Grape and Wine Research, 2010. **16**(3): p. 389-404.
184. Hammer, R.B., et al., *Characterizing dynamic spatial and temporal residential density patterns from 1940–1990 across the North Central United States.* Landscape and Urban Planning, 2004. **69**(2): p. 183-199.
185. Tobler, W., *On the First Law of Geography: A Reply.* Annals of the Association of American Geographers, 2004. **94**(2): p. 304-310.
186. Jain, A.K., M.N. Murty, and P.J. Flynn. *Data clustering: A review*. in *ACM Computing Surveys*. 1999.
187. Dubes, R.C., *CLUSTER ANALYSIS AND RELATED ISSUES*, in *Handbook of Pattern Recognition and Computer Vision*. 1993, WORLD SCIENTIFIC. p. 3-32.

188. Gordon, A.D., *Identifying genuine clusters in a classification.* Computational Statistics and Data Analysis, 1994. **18**(5): p. 561-581.
189. Bezdek, J.C. and N.R. Pal, *Some new indexes of cluster validity.* IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 1998. **28**(3): p. 301-315.
190. Tran, L.T., *Kernel density estimation on random fields.* Journal of Multivariate Analysis, 1990. **34**(1): p. 37-53.
191. Parzen, E., *On Estimation of a Probability Density Function and Mode.* The Annals of Mathematical Statistics, 1962. **33**(3): p. 1065-1076.
192. Rand, W.M., *Objective criteria for the evaluation of clustering methods.* Journal of the American Statistical Association, 1971. **66**(336): p. 846-850.
193. Hubert, L. and P. Arabie, *Comparing partitions.* Journal of Classification, 1985. **2**(1): p. 193-218.
194. Zhang, S., H.S. Wong, and Y. Shen, *Generalized adjusted rand indices for cluster ensembles.* Pattern Recognition, 2012. **45**(6): p. 2214-2226.
195. Steinley, D., M.J. Brusco, and L. Hubert, *The variance of the adjusted Rand index.* Psychological Methods, 2016. **21**(2): p. 261-272.
196. Steinley, D., *Properties of the Hubert-Arable Adjusted Rand Index.* Psychological Methods, 2004. **9**: p. 386-396.
197. Inc, N. *pandas.DataFrame.corr*. 2023 [cited 2024 6 jan]; Available from: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html.
198. *Pearson's Correlation Coefficient*, in *Encyclopedia of Public Health*, W. Kirch, Editor. 2008, Springer Netherlands: Dordrecht. p. 1090-1091.
199. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* Journal of Machine Learning Research, 2011. **12**: p. 2825-2830.
200. Herrera, M., et al., *A review of current and future weather data for building simulation.* Building Services Engineering Research & Technology, 2017. **38**: p. 602 - 627.
201. Jentsch, M.F., et al., *Transforming existing weather data for worldwide locations to enable energy and building performance simulation under future climates.* Renewable Energy, 2013. **55**: p. 514-524.

Appendices

Appendix A

The Python script for the EnergyPlus verification procedure

```
# Area of the house
floor_area = 92.30  # in square meters

# Ceiling height
ceiling_height = 3  # in meters

# Volume of the house
house_volume = floor_area * ceiling_height  # in cubic meters

ACH = 5  # Air Changes per Hour
specific_heat_air = 0.0175  # Specific heat of air in kWh/m³·K

# Setpoints
heating_setpoint = 20  # in degrees Celsius
cooling_setpoint = 26  # in degrees Celsius

# Window Area
wall_area = (10 * ceiling_height * 4)
window_area = wall_area * 0.25

# Temperature difference for heating and cooling from the base temperature of 18 degrees Celsius
deltaT_heating = 53
deltaT_cooling = 10

# Archetypes
archetypes = {
    "NA": {"R_wall": 3.20, "R_ceiling": 4.00},
    "SA": {"R_wall": 2.40, "R_ceiling": 2.40}
}

for archetype, values in archetypes.items():
    # R-values
    U_wall = 1 / values["R_wall"]
    U_ceiling = 1 / values["R_ceiling"]

    # Heating Load Calculations
    Q_walls_heating = U_wall * wall_area * deltaT_heating
    Q_ceiling_heating = U_ceiling * floor_area * deltaT_heating
    Q_windows_heating = U_window * window_area * deltaT_heating
    Q_infiltration_heating = ACH * house_volume * specific_heat_air * deltaT_heating

    Subtotal_total_heating_load = (Q_walls_heating + Q_ceiling_heating + Q_windows_heating + Q_infiltration_heating) * 1.1
    Heating_Distribution_loss = Subtotal_total_heating_load * 0.13

    total_heating_load = Subtotal_total_heating_load + Heating_Distribution_loss
    total_heating_load_per_sqm = total_heating_load / floor_area

    # Cooling Load Calculations
    Q_walls_cooling = U_wall * wall_area * deltaT_cooling
    Q_ceiling_cooling = U_ceiling * floor_area * deltaT_cooling
    Q_windows_cooling = U_window * window_area * deltaT_cooling + SHGC * window_area * deltaT_cooling
    Q_infiltration_cooling = ACH * house_volume * specific_heat_air * deltaT_cooling

    Subtotal_total_cooling_load = (Q_windows_cooling + Q_walls_cooling + Q_ceiling_cooling) * 1.1
    Cooling_Distribution_loss = Subtotal_total_cooling_load * 0.27

    total_cooling_load = Subtotal_total_cooling_load + Cooling_Distribution_loss
    total_cooling_load_per_sqm = total_cooling_load / floor_area

    print(f"Archetype: {archetype}")
    print(f"Heating Load (kWh): {total_heating_load}")
    print(f"Cooling Load (kWh): {total_cooling_load}")
    print(f"Heating Load per sqm (kWh/m²): {total_heating_load_per_sqm}")
    print(f"Cooling Load per sqm (kWh/m²): {total_cooling_load_per_sqm}")
    print()
```

```python
import pandas as pd
import numpy as np
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score

# Read the CSV file into a pandas DataFrame
df = pd.read_csv("All data DD and climate and sim OFFICIAL ZONES for pyhton all clustering results (kw sq m) (validation.csv")

# Extract main dataset values and validation dataset values for NA and SA categories
space_cooling_na_main = df['Space Cooling NA [kWh/m2]']
space_heating_na_main = df['Space Heating  NA [kWh/m2]']
space_cooling_sa_main = df['Space Cooling  SA [kWh/m2]']
space_heating_sa_main = df['Space Heating SA [kWh/m2]']
space_cooling_na_validation = df['Space Cooling NA [kWh/m2] validation']
space_heating_na_validation = df['Space Heating NA [kWh/m2] validation']
space_cooling_sa_validation = df['Space Cooling SA [kWh/m2] validation']
space_heating_sa_validation = df['Space Heating SA [kWh/m2] validation']

# Function to calculate metrics
def calculate_metrics(main, validation):
    rmse = np.sqrt(mean_squared_error(main, validation)) # The RMSE provides a easure of the differences between the validation values and the m
    cv_rmse = (rmse / np.mean(main)) * 100 # CV RMSE: CV RMSE is the RMSE expressed as a percentage of the mean of the main dataset values, prov
    mae = mean_absolute_error(main, validation) # Mean Absolute Error (MAE): MAE measures the average absolute differences between the validatio
    mse = mean_squared_error(main, validation) # Mean Squared Error (MSE): MSE measures the average of the squared differences between the valid
    mape = np.mean(np.abs((main - validation) / main)) * 100 # Mean Absolute Percentage Error (MAPE): MAPE is similar to MPE but calculates the
    return cv_rmse, mae, mse, mape

# Calculate metrics for Space Cooling NA
cv_rmse_space_cooling_na, mae_space_cooling_na, mse_space_cooling_na, mape_space_cooling_na = calculate_metrics(space_cooling_na_main, space_coo
# Calculate metrics for Space Heating NA
cv_rmse_space_heating_na, mae_space_heating_na, mse_space_heating_na, mape_space_heating_na = calculate_metrics(space_heating_na_main, space_hea
# Calculate metrics for Space Cooling SA
cv_rmse_space_cooling_sa, mae_space_cooling_sa, mse_space_cooling_sa, mape_space_cooling_sa = calculate_metrics(space_cooling_sa_main, space_coo
# Calculate metrics for Space Heating SA
cv_rmse_space_heating_sa, mae_space_heating_sa, mse_space_heating_sa, mape_space_heating_sa= calculate_metrics(space_heating_sa_main, space_heat

# Print results
print("Space Cooling NA:")
print("CV RMSE:", cv_rmse_space_cooling_na)
print("MAE:", mae_space_cooling_na)
print("MSE:", mse_space_cooling_na)
print("MAPE:", mape_space_cooling_na)

print("\nSpace Heating NA:")
print("CV RMSE:", cv_rmse_space_heating_na)
print("MAE:", mae_space_heating_na)
print("MSE:", mse_space_heating_na)
print("MAPE:", mape_space_heating_na)

print("\nSpace Cooling SA:")
print("CV RMSE:", cv_rmse_space_cooling_sa)
print("MAE:", mae_space_cooling_sa)
print("MSE:", mse_space_cooling_sa)
print("MAPE:", mape_space_cooling_sa)

print("\nSpace Heating SA:")
print("CV RMSE:", cv_rmse_space_heating_sa)
print("MAE:", mae_space_heating_sa)
print("MSE:", mse_space_heating_sa)
print("MAPE:", mape_space_heating_sa)
```

Appendix B

The Python script for the clustering procedure

```python
import pandas as pd
from sklearn.cluster import KMeans, AgglomerativeClustering
from sklearn.preprocessing import MinMaxScaler

# Read CSV file
input_file = "climate data file.csv"
df = pd.read_csv(input_file)

# Select columns to use for clustering
columns = ['Latitude {N+/S-}', 'HDD18 (hourly method)']

# Normalize the data
scaler = MinMaxScaler()
df_norm = scaler.fit_transform(df[columns])

# K-means Clustering Block
n_clusters_kmeans = int(input("Enter the number of clusters for K-means: "))
kmeans = KMeans(n_clusters=n_clusters_kmeans, random_state=0).fit(df_norm)
kmeans_column_name = input("Enter the name of the new column for K-means clustering labels: ")
df[kmeans_column_name] = kmeans.labels_

# Hierarchical Clustering Block with Complete Linkage Method
n_clusters_hierarchical = int(input("Enter the number of clusters for Hierarchical clustering: "))
hierarchical = AgglomerativeClustering(n_clusters=n_clusters_hierarchical, linkage='complete').fit(df_norm)
hierarchical_column_name = input("Enter the name of the new column for Hierarchical clustering labels: ")
df[hierarchical_column_name] = hierarchical.labels_

# Save the updated dataframe to the same CSV file
df.to_csv(input_file, index=False)

# Print message
print(f"K-means cluster labels added as a new column '{kmeans_column_name}' and Hierarchical clustering labels as
```

Appendix C

Python script for CZMI calculations. Mean overlap percentages calculation (a), intra-cluster distance corrected CZMI calculations (b)

```python
import pandas as pd
import numpy as np
from scipy.stats import gaussian_kde
from itertools import combinations
from scipy.spatial.distance import euclidean
import matplotlib.pyplot as plt

# Load the dataset
# Load the data
data = pd.read_csv("All data DD and climate and sim OFFICIAL ZONES for pyhton all clustering results (kw sq m) (final) (2).csv

# Define cluster label columns and energy consumption columns
cluster_columns = ['SCKC_SET5', 'SCKC_SET6', 'SCKC_SET7', 'SCKC_SET8',
                   'KC_SET1', 'KC_SET2', 'KC_SET3', 'KC_SET4',
                   'SCHC_SET5', 'HC_SET1', 'HC_SET2', 'HC_SET3',
                   'HC_SET4', 'SCHC_SET7', 'SCHC_SET6', 'SCHC_SET8']
energy_columns = ['Space Cooling NA [kWh/m2]', 'Space Heating  NA [kWh/m2]',
                  'Space Cooling  SA [kWh/m2]', 'Space Heating SA [kWh/m2]']

# Function to calculate overlap area between two KDEs
def calculate_overlap_area(kde1, kde2, x_range):
    x_vals = np.linspace(x_range[0], x_range[1], 1000)
    overlap = np.minimum(kde1(x_vals), kde2(x_vals))
    return np.trapz(overlap, x_vals)

# Calculate total KDE for each energy column
total_kde = {}
for energy_col in energy_columns:
    energy_data = data[energy_col]
    kde = gaussian_kde(energy_data)
    x_range = (energy_data.min(), energy_data.max())
    x_vals = np.linspace(x_range[0], x_range[1], 1000)
    total_kde[energy_col] = np.trapz(kde(x_vals), x_vals)

# Recalculate the overlap percentages
overlap_percentages = {col: {energy: [] for energy in energy_columns} for col in cluster_columns}
for cluster_col in cluster_columns:
    for energy_col in energy_columns:
        for cluster_pair in combinations(data[cluster_col].unique(), 2):
            cluster_a, cluster_b = cluster_pair
            cluster_data_a = data[data[cluster_col] == cluster_a][energy_col]
            cluster_data_b = data[data[cluster_col] == cluster_b][energy_col]

            if len(cluster_data_a) > 1 and len(cluster_data_b) > 1:
                kde_a = gaussian_kde(cluster_data_a)
                kde_b = gaussian_kde(cluster_data_b)

                x_range = (min(cluster_data_a.min(), cluster_data_b.min()), max(cluster_data_a.max(), cluster_data_b.max()))
                overlap_area = calculate_overlap_area(kde_a, kde_b, x_range)

                if total_kde[energy_col] > 0:
                    overlap_percentage = overlap_area / total_kde[energy_col]
                    overlap_percentages[cluster_col][energy_col].append(overlap_percentage)

# Calculate mean overlap percentages
mean_overlap_percentages = {col: {energy: np.mean(overlap_percentages[col][energy]) if overlap_percentages[col][energy]
                                  else 0 for energy in energy_columns} for col in cluster_columns}

# Create bar charts for mean overlap percentages
```

(a)

```python
# Function to calculate the centroid of a cluster
def calculate_centroid(cluster_data):
    return np.mean(cluster_data, axis=0)

# Calculate centroids for each cluster in each cluster column
centroids = {col: {cluster: calculate_centroid(data[data[col] == cluster][energy_columns].values)
                   for cluster in data[col].unique()} for col in cluster_columns}

# Calculate inter-cluster distances
inter_cluster_distances = {col: {} for col in cluster_columns}
for col in centroids:
    for cluster_a, cluster_b in combinations(centroids[col].keys(), 2):
        distance = euclidean(centroids[col][cluster_a], centroids[col][cluster_b])
        inter_cluster_distances[col][(cluster_a, cluster_b)] = distance

# Normalize the distances within each cluster column
max_distances = {col: max(inter_cluster_distances[col].values()) for col in inter_cluster_distances}
normalized_distances = {col: {pair: distance / max_distances[col] for pair, distance in inter_cluster_distances[col].items()}
                        for col in inter_cluster_distances}

corrected_overlap_percentages = {col: {} for col in cluster_columns}
for col in cluster_columns:
    for energy_col in energy_columns:
        corrected_values = []
        cluster_pairs = list(combinations(data[col].unique(), 2))  # List of cluster pairs for current column
        for index, overlap_value in enumerate(overlap_percentages[col][energy_col]):
            cluster_pair = cluster_pairs[index]
            distance_correction = 1 - normalized_distances[col][cluster_pair]
            corrected_overlap = overlap_value * distance_correction
            corrected_values.append(corrected_overlap)
        corrected_overlap_percentages[col][energy_col] = np.mean(corrected_values) if corrected_values else 0

# Calculate the corrected overall average overlap percentage for each cluster column
corrected_overall_average_overlap = {cluster_col: np.mean([corrected_overlap_percentages[cluster_col][energy_col]
                                                           for energy_col in energy_columns]) for cluster_col in cluster_columns}

# Sort the corrected overall averages in descending order
sorted_corrected_overall_average_overlap = {k: v for k, v in sorted(corrected_overall_average_overlap.items(),
                                                                    key=lambda item: item[1], reverse=True)}

# Create a bar graph for the corrected overall average overlap percentages
```

(b)

Appendix D

Python script for ARI calculation.

```python
import pandas as pd
import numpy as np
from sklearn.metrics import adjusted_rand_score
import seaborn as sns
import matplotlib.pyplot as plt

# Load the clustering results data from the CSV file
data = pd.read_csv('All data DD and climate and sim OFFICIAL ZONES for pyhton all clustering results (kw sq m) (final) (2).csv')

# Define the column names for clustering results
clustering_columns = ['SCKC_SET5', 'SCKC_SET6', 'SCKC_SET7', 'SCKC_SET8',
                      'KC_SET1', 'KC_SET2', 'KC_SET3', 'KC_SET4',
                      'SCHC_SET5', 'HC_SET1', 'HC_SET2', 'HC_SET3',
                      'HC_SET4', 'SCHC_SET7', 'SCHC_SET6', 'SCHC_SET8',
                      'ASHRAE CZB','Local official CZB (4 zones)', 'Local official CZB (7 zones)',
                      'SCHC', 'HC', 'SCKC', 'KC']

# Create an empty matrix to store the ARI values
ari_matrix = np.zeros((len(clustering_columns), len(clustering_columns)))

# Calculate ARI for all combinations of clustering results
for i in range(len(clustering_columns)):
    for j in range(len(clustering_columns)):
        true_labels = data[clustering_columns[i]]
        predicted_labels = data[clustering_columns[j]]
        ari = adjusted_rand_score(true_labels, predicted_labels)
        ari_matrix[i, j] = ari

# Create a heatmap to visualize the ARI matrix
fig, ax = plt.subplots(figsize=(15, 15))
sns.heatmap(ari_matrix, annot=True, fmt=".2f", cmap='YlGnBu', xticklabels=clustering_columns, yticklabels=clustering_columns, ax=ax)
plt.xlabel('Clustering Results')
plt.ylabel('Clustering Results')
plt.title('Adjusted Rand Index (ARI) Matrix')
plt.show()

# Convert the ARI matrix to a pandas DataFrame for easier analysis
ari_df = pd.DataFrame(ari_matrix, index=clustering_columns, columns=clustering_columns)
print("ARI Matrix:")
print(ari_df)
```

Appendix E – Published articles

(Q1, impact factor 11.2)

All Export to Excel Save to source list

View metrics for year: 2022

	Source title	CiteScore	Highest percentile	Citations 2019-22	Documents 2019-22	% Cited
1	Applied Energy	21.1	99% 1/200 Building and Construction	138,931	6,584	91

https://doi.org/10.1016/j.apenergy.2023.122238

Applied Energy 355 (2024) 122238

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Novel building energy performance-based climate zoning enhanced with spatial constraint

Alexey Remizov, Shazim Ali Memon *, Jong R. Kim

Department of Civil and Environmental Engineering, School of Engineering and Digital Sciences, Nazarbayev University, Nur-Sultan, Kazakhstan

HIGHLIGHTS

- Proposed novel climate zoning using building energy performance & spatial constraint.
- Spatial constraints enhance buildings' climate zoning and reduce misclassification.
- Comparison using Adjusted Rand Index reveals disparities in existing climate maps.

ARTICLE INFO

Keywords:
Buildings' climate zoning
Building energy simulation
Cluster analysis
Spatial constraint

ABSTRACT

Existing buildings' climate zoning approaches often overlook the incorporation of building energy usage data, leading to discrepancies between climate zoning classifications and the ever-increasing demands for energy efficiency in modern construction practices. This research proposes a novel spatially constrained approach that incorporates multivariate clustering and building energy needs indicators to create a climate classification aligned with buildings' energy performance patterns. The study focuses on Kazakhstan as a case study, where buildings experience distinct climatic conditions across different regions. A spatial constraint was used to enhance hierarchical and k-means clustering methods, which were applied to the space heating and cooling energy needs of the most typical building archetypes. The quality of clustering results was evaluated using uniqueness and dispersion indicators. Furthermore, The Adjusted Rand Index was introduced to compare the proposed method with the ASHRAE and the official buildings' climate map of Kazakhstan. The existing climate maps failed to match the patterns of building energy performance. Overall, the proposed spatially constrained method exhibits promising results and offers optimized buildings' climate zoning supported by buildings' energy performance data and rigorous measures of clustering quality.

1. Introduction

Buildings represent a significant contributor to energy consumption and greenhouse gas emissions, thereby exerting a notable influence on global sustainability conditions [1–3]. About a third of the world's energy usage and a quarter of its CO2 emissions are caused by buildings [4]. Building operation's direct and indirect emissions rebounded after the "COVID-19" period to roughly 10 Gt in 2021, which is 2% more than in 2019 and 5% more than in 2020. The anticipated 2030 development forecasts a 20% increase in building floor space compared to 2022. More

(Q1, impact factor 3.8)

https://doi.org/10.1016/j.apenergy.2023.122238

buildings

MDPI

Review

Climate Zoning for Buildings: From Basic to Advanced Methods—A Review of the Scientific Literature

Alexey Remizov, Shazim Ali Memon * and Jong R. Kim

Department of Civil and Environmental Engineering, School of Engineering and Digital Sciences, Nazarbayev University, Astana 010000, Kazakhstan
* Correspondence: shazim.memon@nu.edu.kz

Abstract: Understanding the link between the energy-efficiency of buildings and climatic conditions can improve the design of energy-efficient housing. Due to global climate change and growing requirements for building energy-efficiency, the number of publications on climate zoning for buildings has grown over the last 20 years. This review attempted to give the reader an up-to-date assessment of the scientific literature in the field of climate mapping for buildings on a global and national scale, filling in the gaps of previous works and focusing on details that were not presented before. There were 105 scientific sources examined. The most dominant climate zoning variables were thoroughly analyzed. A clear categorization of climate zoning methods with specific criteria was shown. The most used methods were evaluated, emphasizing their similarities and differences, as well as their essential components and advantages. The main literature review was supported with bibliometric and bibliographic analysis. The existence of many climate zoning methods can be an indicator of the lack of agreement on the most effective strategy. A tendency has been established for the popularization among scientists of methods based on machine learning and building energy simulations, which are relatively easy to use and have proven to be the most reliable climate zoning methods. A transformation is emerging by shifting from a climate-based to a building performance-based climate zoning approach.

Keywords: building energy-efficiency; building energy simulation; climate zoning; climatic variables; cluster analysis; degree-days; machine learning

Citation: Remizov, A.; Memon, S.A.; Kim, J.R. Climate Zoning for Buildings: From Basic to Advanced Methods—A Review of the Scientific Literature. *Buildings* **2023**, *13*, 694. https://doi.org/10.3390/buildings13030694

Academic Editor: Adrian Pitts

Received: 29 January 2023
Revised: 22 February 2023
Accepted: 28 February 2023
Published: 6 March 2023

1. Introduction

People are becoming more conscious about the link between energy use and environmental impacts as global warming and climate change progress more significantly [1,2]. The present energy-related greenhouse gas (GHG) emissions are around 39 Gt CO2 equivalent, according to the International Energy Agency. The building industry was directly or indirectly responsible for nearly 50% of global energy consumption and 39% of total GHG emissions in 2018 [3]. While developed countries have taken significant progress to reduce their energy consumption, the energy demand for buildings rose by over 20% between 2000 and 2017 due to factors including the rapidly expanding floor area of dwellings, the relatively small reduction in energy intensity, and the rising energy requirements of the energy services [4]. Existing and future buildings will be largely responsible for determining global energy consumption [5–10]. Future growth in energy use and accompanying emissions is prominent. The increased access of billions of people in developing countries to decent housing, electricity, and improved cooking facilities is a significant trend. By 2040, buildings are expected to be the most significant source of GHG emissions [11]. In addition to the issue of climate change, there are important economic reasons why energy-efficient buildings are becoming increasingly attractive. There are between 100 and 150 million people in developed countries that are unable to afford the cost of energy due to low incomes [12]. In 2018, nearly 13% of Europeans said they live in homes that are too cold,