Data Sampling and Analysis for the Improvement of Estimating Heating Loads in Alaska

Madelyn Gaumer

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington
2022

Reading Committee:
Nathan Kutz
Erin Trochim
David Beck

Program Authorized to Offer Degree:
Applied Mathematics

University of Washington

Abstract

Data Sampling and Analysis for the Improvement of Estimating Heating Loads in Alaska

Madelyn Gaumer

Co-chairs of the Supervisory Committee:

Nathan Kutz

Professor, Applied Mathematics, University of Washington

Erin Trochim

Research Assistant Professor, Alaska Center for Energy and Power, University of Alaska Fairbanks

Due to the accelerated effects of climate change over the past 10 years, Alaska and the larger Arctic region are in need of decarbonization far more than the rest of the world does. Over 75% of the energy utilized in the Arctic region is for heating houses and businesses. However, a key barrier to the switch to renewable energy is the absence of extensive and accurate heating load estimates in Alaska. This research builds upon previous work to establish a geospatial-first methodology using satellite data to estimate heating loads in Alaska. In this work, we analyze building data and climate data, including ERA5 and Daymet. We also use modern data sampling techniques to combat imbalanced data and show that random sampling performs well compared to other techniques.

Acknowledgements

I would like to thank Erin Trochim for all of her help and guidance during this work as well as my committee, Erin Trochim, Nathan Kutz, and David Beck, for their encouragement throughout this process. I would also like to thank Nick Bolten, Vidisha Chowdhury, Philippe Schicker, and Shamsi Soltani as well as the University of Washington Data Science for Social Good Program through the eScience Institute for instilling a passion in me for this work. Finally, I would like to thank my family and friends, whose unwavering support allows me to pursue my passions.

DEDICATION

To Pop

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Alaska and the larger Arctic region have been experiencing accelerated warming at a rate of almost four times as fast as the global average Rantanen et al. [2022]. Over 75% of the energy utilized in Alaska is for heating houses and businesses WHPacific [2012]. As such, heating is contributing significantly to global warming in areas with colder climates. Additionally, in 2020, Alaska had the second highest petroleum usage in the country, and overall, Alaska consumes more energy per person than the rest of the US U.S. Energy Information Administration. All of these statistics highlights the fact that in order for Alaska to make significant strides in reducing emissions statewide and progress towards decarbonization, heating usage in Alaska needs to be understood.

Currently, precise estimates for heating loads in Alaska do not exist. Researchers and officials at multiple levels of government have demanded accurate heating load estimates for Alaska at the neighborhood level. While the estimates produced through this work are at the building level, there are clear privacy concerns with releasing this data to the public. However, in future work, public participatory refinement of the heating loads is possible due to the estimates being at the building level. These heating load estimates are essential for modernizing Alaska's grid system and

informing the diversification of the state's energy supply and provide an interesting opportunity to advise state-sponsored retrofitting and energy programs.

1.2 Previous Work

Much of current data collection work for heating loads in Alaska involves on-the-ground approaches. These approaches are frequently invasive in nature and consist of entering people's houses and installing various fuel usage monitoring devices, like the PuMA pump Alaska Center for Energy and Power. The main limitation of current Alaskan heating load estimation work is its micro-level scale. These methods have many flaws, the main two being the high variation in fuel-use estimates and the micro-level scale of the approach. Fuel use monitoring apparatuses can have large variations in estimates even with identical stove models because of the variation in the apparatuses and sensitivity to apparatus placement. In addition, because these approaches are labor-intensive, they only occur on a small scale and cannot generate anywhere near the quantity of data needed for a large scale model to make use of. With over 230,000 buildings in Alaska's railbelt region, it would be impractical to visit each one and install a fuel use monitoring apparatus to generate a heating load estimate of each building. In the future, these micro-level approaches provide an interesting opportunity for cross validation of our methods.

The Alaska Housing Finance Corporation published two regression relationships relating to energy use for homes in 2014 Alaska Housing Finance Corporation. These regressions showed the relationship between the year a home was built and the average energy needed to heat it per square foot and were limited to two areas in Alaska: the Municipality of Anchorage and the Fairbanks Northstar Borough. These regression relationships use data from a comprehensive heating load database, created out of state retrofitting programs, and heating load estimates from AK Warm, the standard software in Alaska for modeling heat usage. These regressions were one of the first attempts at estimating Alaska's heating usage. However, they are flawed in that the regressions

and the synthetic data were not set up for predictive purposes and they are limited to two areas in Alaska and cannot be applied to the entire state.

In the summer of 2022, a group from the University of Washington's Data Science for Social Good Program Bolten et al., myself included undertook a project to model heating loads in Alaska using a geospatial first approach. This work is detailed further in Chapter 2.

1.3 Challenges

There are many challenges to the task of modeling heating loads in Alaska. These challenges can be broken down into several categories including data collection, access and export, geospatial approach, ethical considerations, and machine learning methods.

Alaska presents unique challenges when it comes to collecting relevant data for the task of modeling heating loads. In particular, due to a culture of independence, the remoteness of many buildings, and the harsh winters, data collection is difficult, and traditional data collection approaches that succeed in the contiguous 48 states don't always work in Alaska leading to erroneous and often missing data. Additionally, unlike many other states in the US, Alaska has a lack of consistent building codes for energy usage, which makes it hard to make accurate assumptions about energy being used. There is a comprehensive heating loads database containing information about retrofitted houses in Alaska. This database is in the process of becoming publicly available but is not used in this work yet. In addition, the model america dataset New et al. [2012], created by Oakridge National Laboratory, exists and contains relevant heating information. However, it currently only does so at a county level with building information anonymized. Using Google Earth Engine Gorelick et al. [2017] in this project also presented some challenges when it came to export due to the quanity and level of data being worked with.

The geospatial first, top down approach used in the previous 2022 Data Science for Social Good Work Bolten et al. and in this work also presents some challenges. Modeling heating loads

at this scale requires various assumptions and risks missing micro-level information. In this work, we try to make these assumptions clear. It is important to acknowledge that compiling heating load information at a building level and making various assumptions to do so does bring up data privacy concerns.

Finally, in using machine learning methods in this task, one of the main challenges and one that this work in particular aims to address is the fact that the data available for this task is very imbalanced in a number of ways and some of it is entirely missing.

1.4 Sampling

The preliminary sampling work done by the 2022 Data Science for Social Good group at UW Bolten et al. showed that it was possible to see lower error in the heating load estimations when the data was sampled to try to balance it in several ways. In this work, we do a more comprehensive exploration of various basic sampling strategies to improve the heating load estimations for the state of Alaska. In addition to the exploration of sampling strategies, we also do a deep dive into some additional building data and climate data.

Chapter 2

General Methods

This section details the prior work done during the summer 2022 Data Science For Social Good (DSSG) program at the University of Washington by the heating loads team Bolten et al.. Much of this preliminary work is used in this work.

2.1 Data Collection and Processing

In this study, we used the open source geospatial big data analysis tool Google Earth Engine Gorelick et al. [2017] to extract features from satellite data, which we then fed into our model. The seven datasets utilized in this research are all taken from the public archive of Google Earth Engine and include geometries of building outlines in Alaska as well as historical satellite pictures. We combined the geographical and temporal data to obtain building-level features and then fed these features into our model. We combat missing data and inaccurate data in the arctic by utilizing a variety of data sources.

We extracted information on the local climates of Alaska as well as building characteristics like height, base area. These characteristics were all picked because they relate to how much heat is required to heat a building.

In the preliminary work through DSSG, we compared the Open Street Maps Building Footprints dataset Open Street Maps and the Microsoft Buildings dataset Roy. In this work, we also consider the USA Structures dataset in Alaska Esri US Federal Data [2022]. The USA Structures dataset has the most up to date and accurate building outlines for the state of Alaska out of the three options, so we used the USA Structures dataset in all analysis here. Using these building footprints ensures there are no duplicates in our data processing.

In the preliminary work through DSSG, we used three independent datasets—World Settlements Footprint Evolution (1985-2015) Marconcini et al. [2021], World Settlements Footprint 2019 (2019) Marconcini et al. [2021], and Dynamic World (June 2015–present) Brown et al. [2022] to determine the age of buildings in Alaska. By obtaining the mean age value in each building outline, we first reduced the World Settlements Footprint Evolution dataset over the Open Street Maps Building Footprints dataset. We then repeated this using the Word Settlements Footprint 2019 dataset, giving age values for any buildings that did not yet have an age. The Dynamic World dataset was then used to repeat this process one last time, giving any buildings without age values but included in Dynamic World a 2020 designation. At this point in the process, we assigned an age of 1984 to any remaining buildings that lacked an age value. In this work, using the USA Structures dataset allowed us to use their provided height characteristic at a building level. To calculate building base area, we used the USA Structures dataset and calculated the area of each building geometry.

2.2 Training Procedure

We assigned a heating load estimate to each building in the Fairbanks Northstar Borough and the Municipality of Anchorage, the two regions included in the 2014 Alaska Housing Finance Corporation regression relationships Alaska Housing Finance Corporation.

After that, we developed a 70/30 train/test split and applied the regression models to our data.

We then predict on buildings in the rest of the railbelt, not including those in the Fairbanks Northstar Borough or in the Municipality of Anchorage. It is worth noting here that due to finite number of heating load estimates generated from on-the-ground micro-level approaches, traditional validation methods are not possible. Instead, we validate with the regression relationships from the Alaska Housing Finance Corporation Alaska Housing Finance Corporation.

2.3 Regression Models

Given that this project has a continuous output variable, btus-the heating demand of a building, regression was selected as the general model type. We compared the performance of five regression models: linear regression, ridge regression, ridge regression using polynomial features, decision tree regression, and random forest regression.

Chapter 3

Data

3.1 Climate Data

3.1.1 Alaska's Climate

Alaska's climate is very diverse, and the state is made up of 13 different climate divisions. These natural climate divisions can be seen in Figure 3.1 Connor et al. [2020]. All of the heating load estimates we have come from Fairbanks and Anchorage, which are in the Central Interior and the Cook Inlet divisions respectively. While we only have data from two of the thirteen climate divisions, we want our model to be able to generalize to all thirteen of Alaska's climate divisions.

Fairbanks is located in the rain shadow of the coastal mountains and the delta range, whereas Anchorage is on the rainy side of the coastal mountains. Figure 3.2 Connor et al. [2020] shows how this rain shadow affect in Fairbanks and rainy affect in Anchorage looks in practice.

Downtown Fairbanks lies in a valley and as you move out from downtown, there are buildings that lie in the surrounding hills. Due to this unique geography, Fairbanks experiences temperature inversions, shown in Figure 3.3 Connor et al. [2020]. Cool air slides off the surrounding mountains and sinks into the valley in downtown Fairbanks, forcing warm air upwards. This warm air becomes trapped in between cool air. This is a unique phenomenon that will most likely be difficult

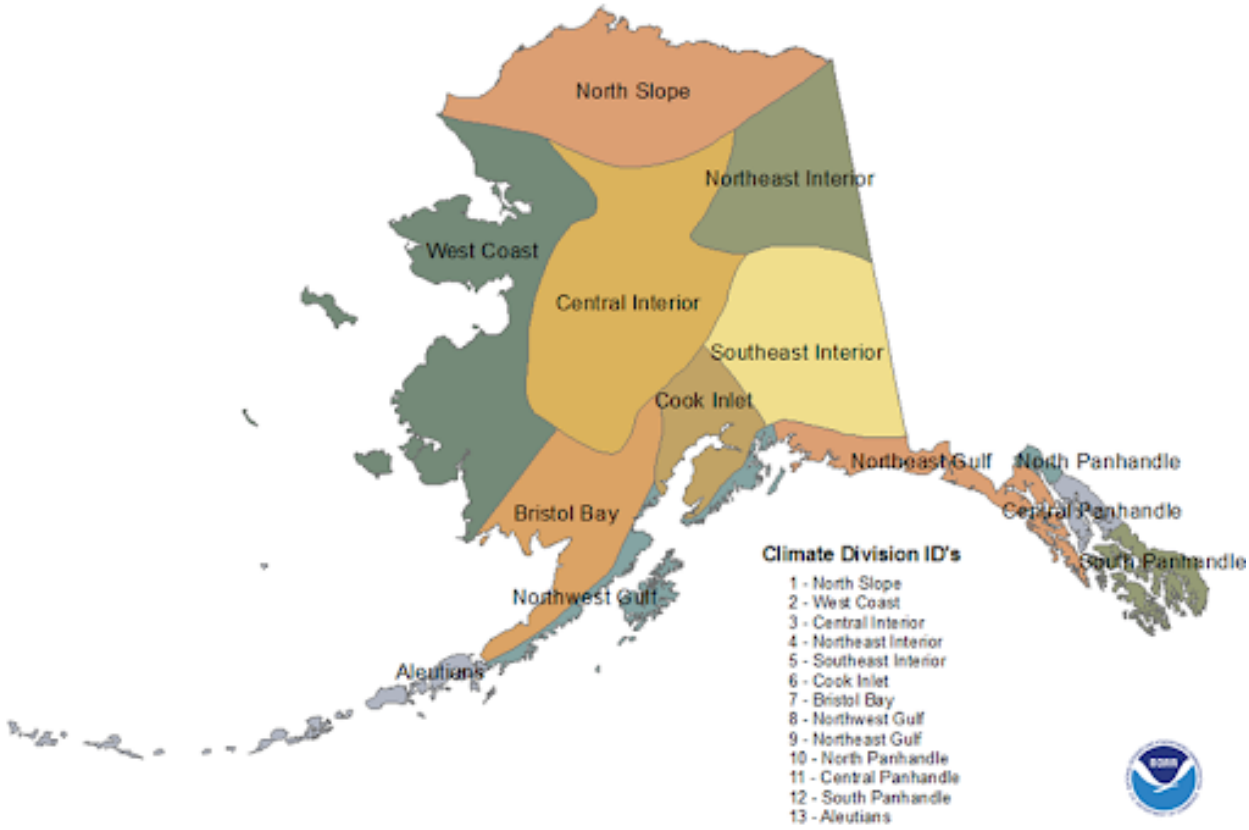

Figure 3.1: Current Alaska climate divisions. More information is available on the NOAA Alaska Climate Division FAQ page **?**.

for our models to capture.

3.1.2 ERA5 and Daymet V4

ERA5 and Daymet V4 are two climate satellite datasets we have extracted freezing and thawing degree days from, which are measurements of how often and how much the temperature is below and above freezing. The ERA5 dataset Copernicus Climate Change Service (C3S) (2017) is a historical reanalysis dataset, whereas the Daymet dataset Thornton et al. is a historical gridded dataset, as detailed in Figure 3.4 Connor et al. [2020].

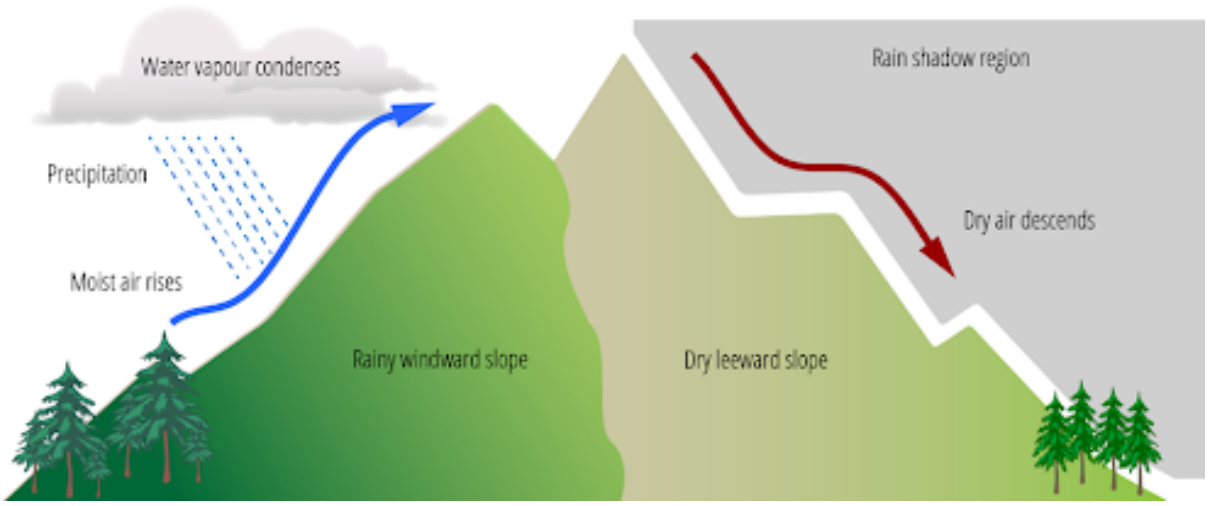

Figure 3.2: Example of orographic precipitation affect where there are rainy windward and dry leeward slopes, which creates a rain shadow region Connor et al. [2020].

ERA5 Copernicus Climate Change Service (C3S) (2017)

The ERA5 Daily Aggregate data contains climate information including air temperature, dewpoint temperature, precipitation, surface pressure, sea level pressure, and wind components at a spatial resolution of 27830 meters. This dataset is available starting from 1979 and goes up to present day. The spatial resolution of this dataset is quite large for our purposes, but the data is still useful as it contains accurate trends.

Daymet V4 Thornton et al.

The Daymet V4: Daily Surface Weather and Climatological Summaries dataset contains climate information including daylight, precipitation, incident shortwave radiation flux density, snow water equivalent, air temperature, and partial pressure of water vapor at a spatial resolution of 1000 meters. This spatial resolution is much finer than the spatial resolution of ERA5. As a result, we expect Daymet to be able to pick up on more local microclimates in the data, and we also expect to see the temperature inversions in Fairbanks we mentioned earlier. Thus, even though this data is more detailed, it could very well be harder for our models.

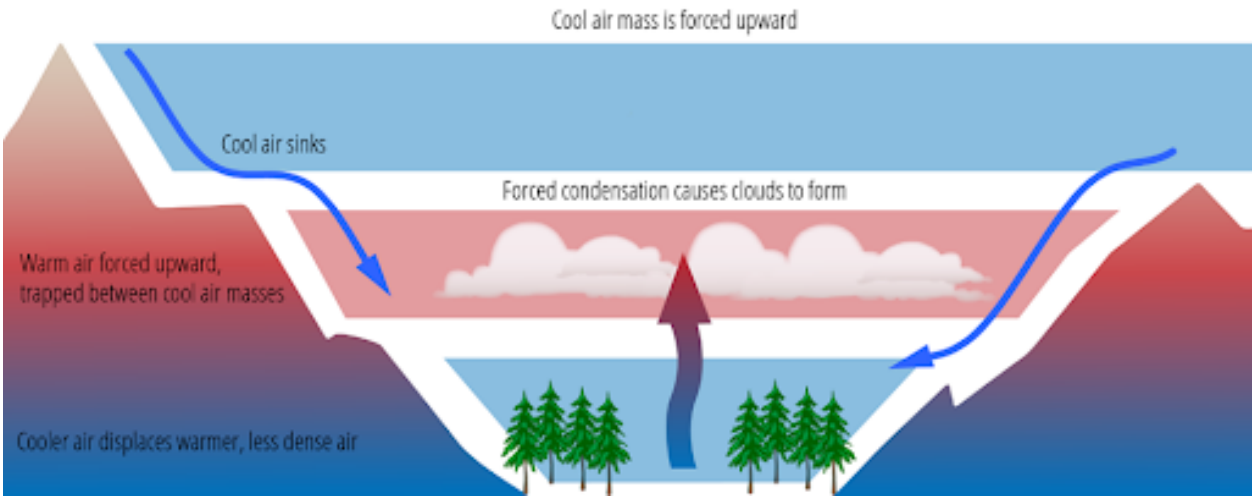

Figure 3.3: Formation of temperature inversions which can strongly affect the distribution of permafrost in areas like the Fairbanks North Star Borough Connor et al. [2020].

It is worth noting that there were 279 buildings total in Alaska's railbelt that are not contained in the Daymet V4 dataset due to the gridding they do. Thus, when using the Daymet V4 data, we did not include these buildings in our analyses. An example of the buildings being excluded from the dataset can be found in Figure 3.5.

3.2 Buildings

3.2.1 Railbelt Data Comparison

Over the course of this project, we considered four different datasets containing building outlines. Out of the Open Street Maps Building Footprints Open Street Maps, Microsoft Buildings Roy, and USA Structures Esri US Federal Data [2022], we found that for the state of Alaska, the USA Structures dataset has the most accurate set of buildings. The Model America New et al. [2012] dataset was also considered at this state, but building outlines are not available to the public at this time. Table 3.1 shows the building counts of each of the three datasets across Alaska's Railbelt region. From this, we can see that by far, USA Structures has the most building outlines in the

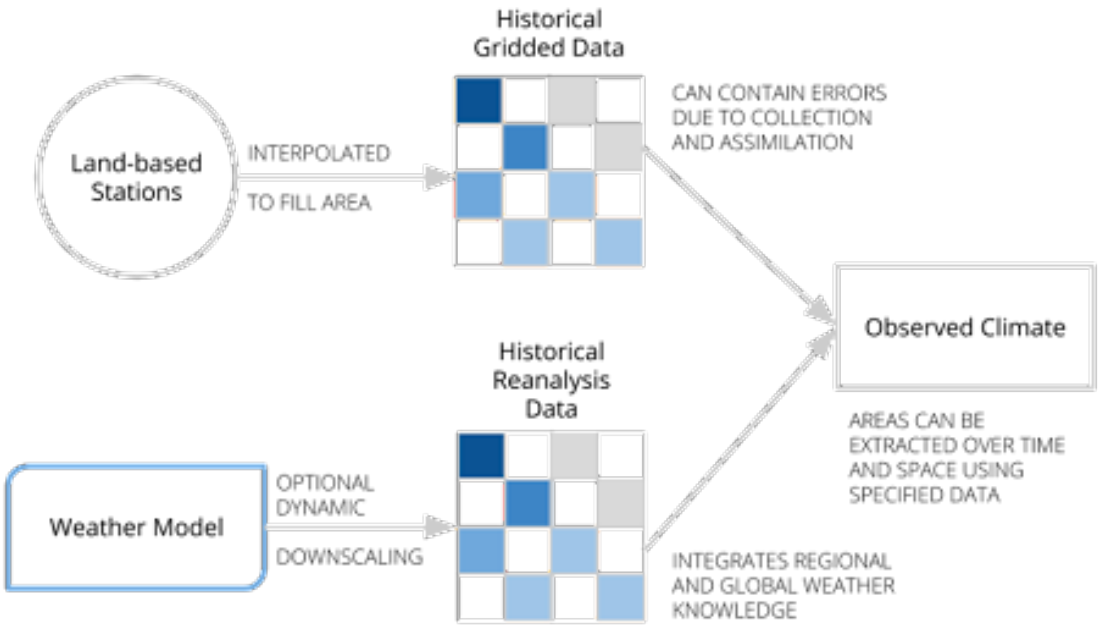

Figure 3.4: the two methods of creating gridded observed climate. Historical gridded data is primarily based on information interpolated from land-based station. Historical reanalysis data is based on weather models running over a specified domain Connor et al. [2020].

Railbelt region. While general building outlines for Alaska's Railbelt are not provided through Model America, we do have building counts for Fairbanks and Anchorage. In Table 3.2, we show the building counts for Fairbanks and Anchorage for the Model America Dataset, the USA Structures, and the borough government data for Fairbanks Fairbanks Northstar Borough. It is worth noting that the Fairbanks North Star Borough contains wood sheds and outbuildings, which may not all require full heating, as well as buildings that actually require heating. This is due to the fact that this data does not distinguish between building types. Thus, it is likely that the true building count, for heated buildings of any kind, in Fairbanks is somewhere in between the USA Structures estimate and the Borough Data estimate.

We can also compare building height data between the Model America dataset and the USA Structures dataset. Figure 3.6 contains a histogram of building heights from the Model America dataset for Fairbanks buildings, and 3.7 contains a histogram of building heights from the USA Structures dataset for Fairbanks buildings. The two histograms look very different. The Model

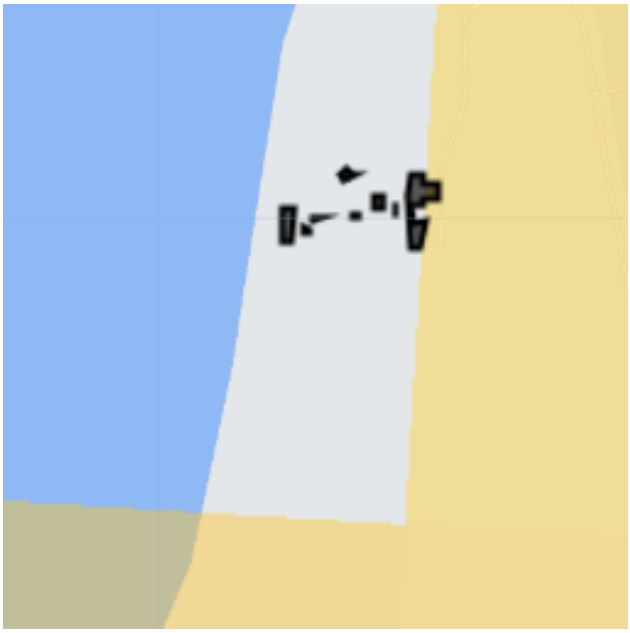

Figure 3.5: The yellow region is the region covered by Daymet V4. The black outlines are a few buildings that the Daymet Climate Data did not include. .

Dataset	Alaska Railbelt Building Count
Open Street Maps	109,296
Microsoft Buildings	101,474
USA Structures	231,029

Table 3.1: Building Counts in Alaska's Railbelt

Dataset	Fairbanks Building Count	Anchorage Building Count
Model America	23023	75440
USA Structures	34931	87104
Borough Data	76871	not available

Table 3.2: Various Building Dataset Count and Area information.

America Histogram shows that the maximum height building in Fairbanks is approximately eight meters tall, while the USA Structures dataset shows that all the buildings in Fairbanks have a height less than 0.2 meters. Clearly the USA Structures dataset is incorrect here, and these two figures show how much datasets in Alaska can vary across features.

We can do a similar comparison with building footprint area for the Model America dataset

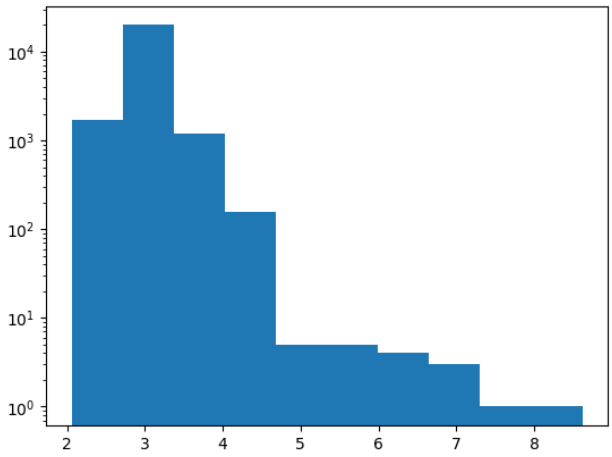

Figure 3.6: Model America Fairbanks Building Heights in meters (y-axis log scale) .

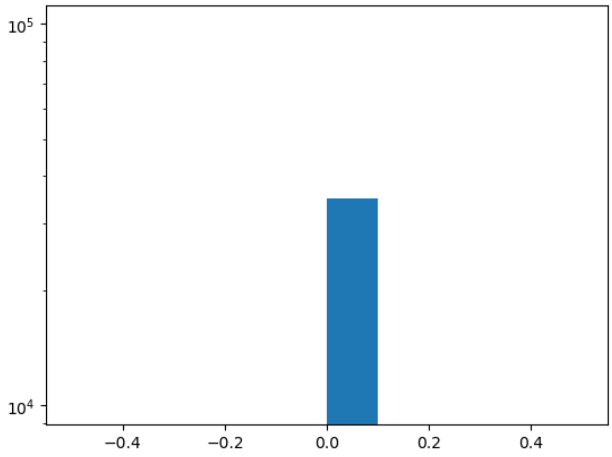

Figure 3.7: USA Structures Fairbanks Building Heights in meters (y-axis log scale) .

and the USA Structures dataset for buildings in Fairbanks. Figure 3.8 shows a histogram for the building footprint areas from Model America for Fairbanks buildings, and Figure 3.9 shows a histogram for the building footprint areas from USA Structures for Fairbanks buildings. These histograms have a much more similar general trend, but the x-axis differs between both images. This could have to do with how each dataset is defining building footprint area. For example, perhaps outdoor spaces that are part of the building are in the Model America dataset, making all of the areas bigger.

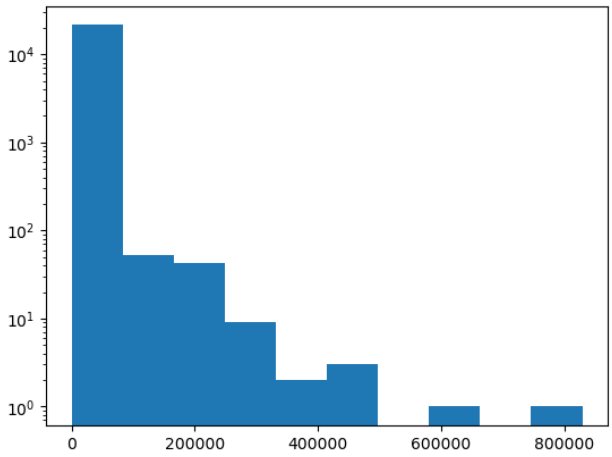

Figure 3.8: Model America Fairbanks Building Footprint Areas in square meters (y-axis log scale)
.

3.3 Feature Analysis and Imbalanced Data

3.3.1 Feature Analysis

In this work, we do a more extensive feature analysis than was performed in the summer of 2022 Data Science for Social Good team's work Bolten et al.. In Figure 3.10, we analyze the correlation between all the features that we extracted using Google Earth Engine and using the ERA5 Daily data where feature names are explained in Table 3.3. From our feature correlation matrix, we see that the features most highly correlated with 'combined_heating_load_per_sq_ft', our heating load estimation, are building age, 10 year average of Thawing Degree Days from 2011 to 2020, 10 year average of Thawing Degree Days from 2001 to 2010, 10 year average of Thawing Degree Days from 1991 to 2000, 30 year average of Thawing Degree Days from 1981 to 2010, and 10 year average of Thawing Degree Days from 1981 to 1990, and building height in that order.

In Figure 3.11, we analyze the same correlation but this time using the Daymet Climate data instead of the ERA5 data. From this feature correlation matrix, we see that the features most highly correlated are the same as previously, but in a different order. The order, from most to least correlated with 'combined_heating_load_per_sq_ft', our heating load estimation, is building age,

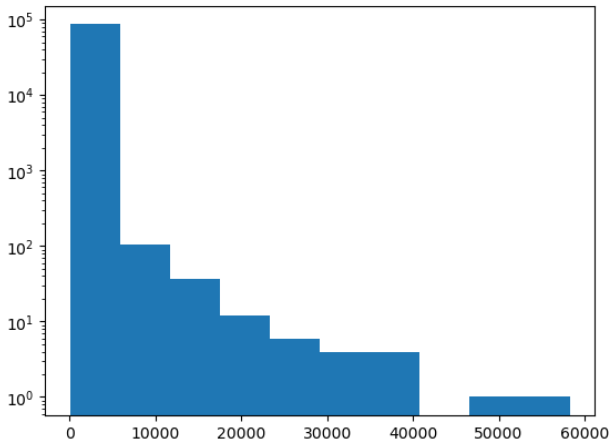

Figure 3.9: USA Structures Fairbanks Building Footprint Areas in square meters (y-axis log scale)
.

10 year average of Thawing Degree Days from 1991 to 2000, 10 year average of Thawing Degree Days from 1981 to 1990, 30 year average of Thawing Degree Days from 1981 to 2010, 10 year average of Thawing Degree Days from 2001 to 2010, 10 year average of Thawing Degree Days from 2011 to 2020, and building height. The different ordering of the feature correlations between using the two climate datasets suggests that for the Daymet dataset, the older climate data is more correlated with the heating load estimates whereas for the ERA5 dataset, the more current climate data is more correlated with the heating load estimates. This is an interesting finding that needs to be explored further.

However, the four climate features that are highly correlated with our heating load estimate are also highly correlated with each other. To eliminate repetitive information, we perform PCA on our normalized climate features. From Figure 3.12, we see that more than 96% of the variation in the five climate features coming from the ERA5 dataset is contained within the first principal component.

When we perform the same analysis using the Daymet data, we see that to surpass capturing 90% of the variation in the climate features, we need to use the first two principal components, which capture 98.7% of the variation. The plot of the components is located in Figure 3.13. For

Feature Abbreviation	Meaning
building_age	building age
30y_FD	30 year average of Freezing Degree Days from 1981 to 2010
HEIGHT	building height
LATITUDE	building centroid latitude
LONGITUDE	building centroid longitude
SQFEET	building footprint area
30y_TD	30 year average of Thawing Degree Days from 1981 to 2010
80_FD	10 year average of Freezing Degree Days from 1981 to 1990
80_TD	10 year average of Thawing Degree Days from 1981 to 1990
90_FD	10 year average of Freezing Degree Days from 1991 to 2000
90_TD	10 year average of Thawing Degree Days from 1991 to 2000
00_FD	10 year average of Freezing Degree Days from 2001 to 2010
00_TD	10 year average of Thawing Degree Days from 2001 to 2010
10_FD	10 year average of Freezing Degree Days from 2011 to 2020
10_TD	10 year average of Thawing Degree Days from 2011 to 2020
combined_heating_load_per_sq_ft	heating load estimation in btus per sqft

Table 3.3: Feature Abbreviations and their Meanings.

the remainder of the paper, unless otherwise stated, we use the PCA climate data features for any remaining analysis in an attempt to reduce overfitting in our models. We do understand that this interferes with the interpretability of our models, and so we include versions of select analyses using the non PCA climate features in the appendix.

3.4 Spatial Analysis

'Spatial heterogeneity', terms related to spatial data. We can define spatial dependence, by quoting Tobler's First Law of Geography, as the property that "near things are more related than distant things" GIS Geography [2022]. Spatial heterogeneity is the notion that the distribution of a spatially dependent variable is not constant over space.

We can measure the spatial dependence of our data using spatial autocorrelation statistics. Here we use Queen contiguity, a method for determing neighbors of each datapoint, to determine the

spatial weights matrix. The spatial weight matrix quantifies the relationship between a building's btu estimate and the btu estimate of its neighbors. From the spatial weight matrix, we can compute the spatial lag. The spatial lag for a building in this case is a weighted average of it's neighboring building's btu estimates.

From Figure 3.14, we can see that it is very likely that are data is spatially dependent and spatially heterogeneous, which is what we would expect given the problem.

3.4.1 Error Analysis of Unsampled Data

The data with heating load estimates we have is imbalanced across a number of different factors including by location and by building age. In practice, by location, this means that we have more heating load estimates to train our model coming from Anchorage than from Fairbanks since there are more buildings in Anchorage. This means that our model is seeing and learning more about buildings in Anchorage than it is in Fairbanks. In Table's 3.4 and 3.5, we see that training on unsampled imbalanced data results in worse test performance for buildings in Fairbanks compared with buildings in Anchorage. To try to rectify this issue and see more balanced performance by our models, we need to sample our data.

Model	Fairbanks MSE	Anchorage MSE
Linear	0.017617	8.701327e-03
Ridge	0.017614	8.702362e-03
Ridge w/ poly features	0.000703	5.358743e-04
Decision Tree	0.000015	3.768957e-08
Random Forest	0.000001	1.029798e-07

Table 3.4: Error Split between Fairbanks and Anchorage of Unsampled Data of PCA ERA5 Data.

Model	Fairbanks MSE	Anchorage MSE
Linear	0.019504	0.006878
Ridge	0.019501	0.006880
Ridge w/ poly features	0.004245	0.002656
Decision Tree	0.000028	0.000004
Random Forest	0.000064	0.000006

Table 3.5: Error Split between Fairbanks and Anchorage of Unsampled Data of PCA Daymet Data.

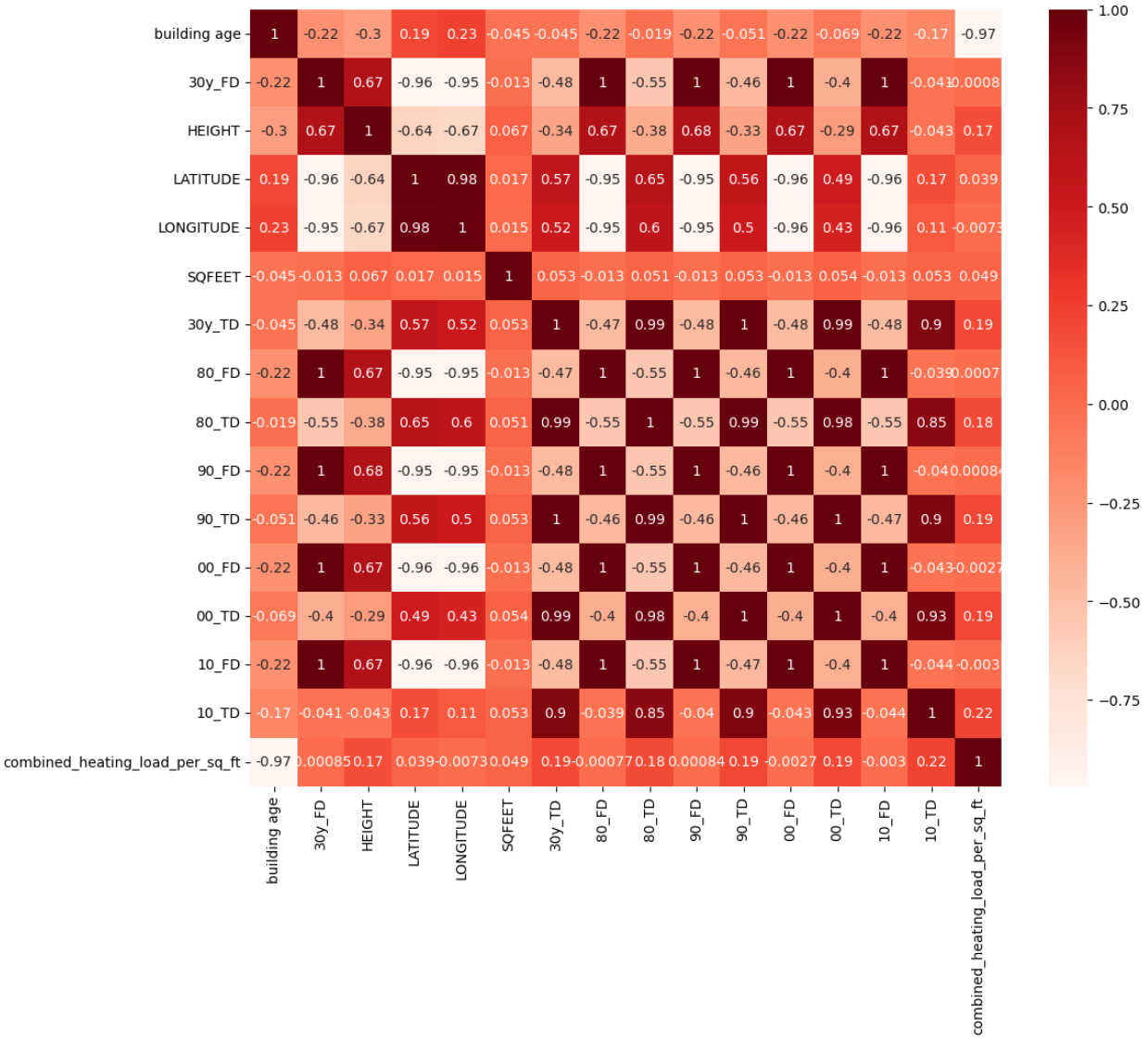

Figure 3.10: Correlation Matrix of Building and Climate Attributes using Pearson Correlation using ERA5

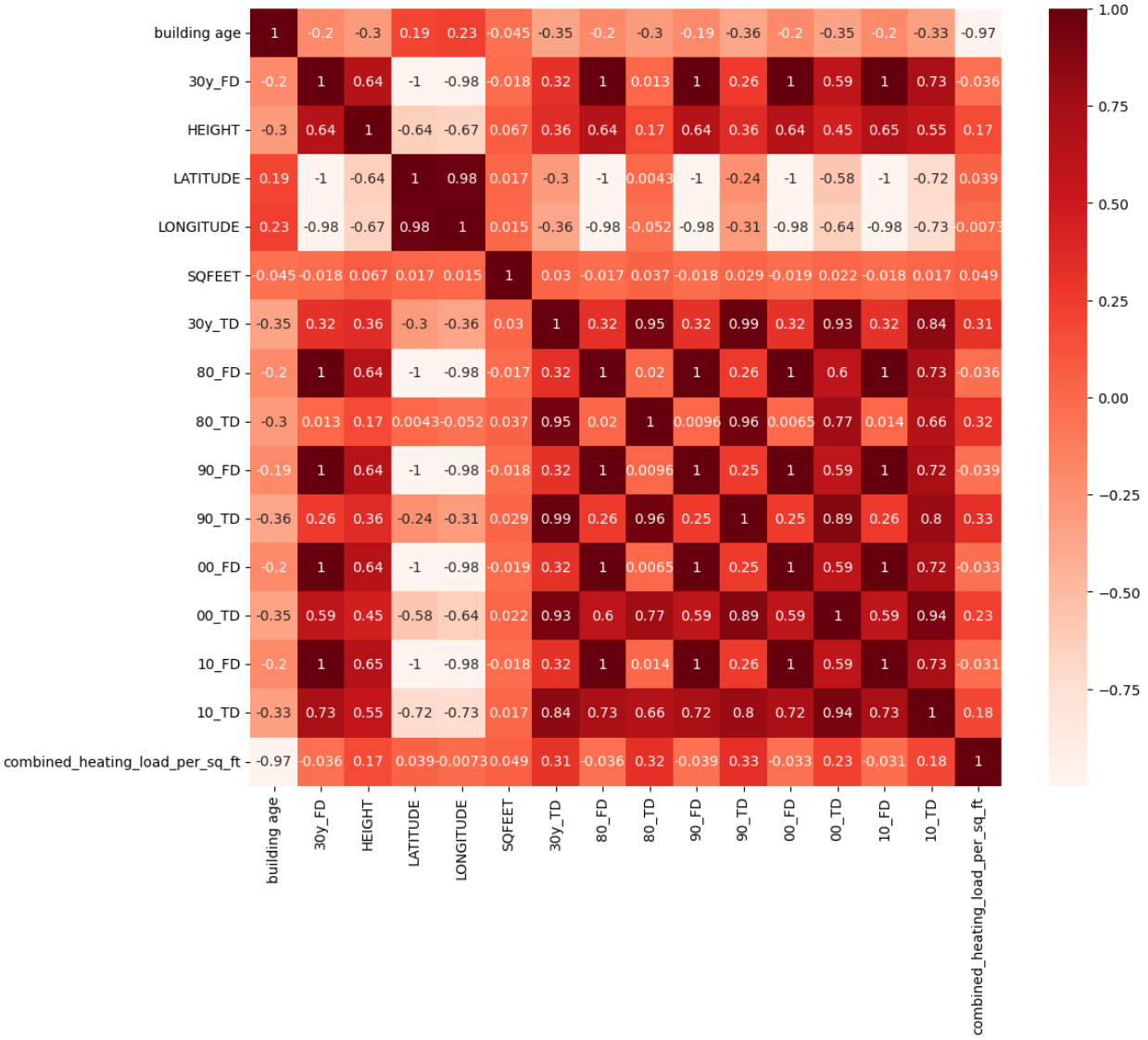

Figure 3.11: Correlation Matrix of Building and Climate Attributes using Pearson Correlation using Daymet

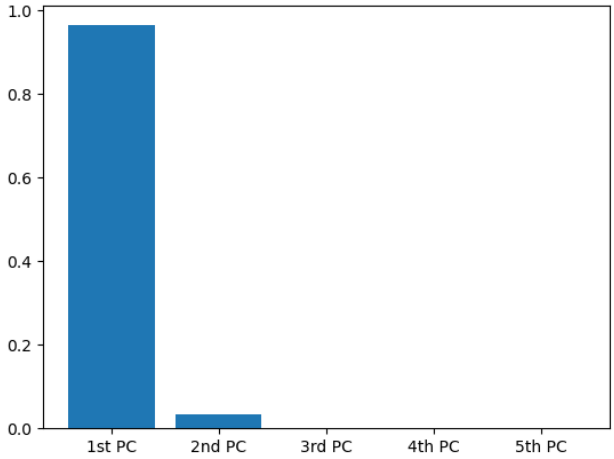

Figure 3.12: Principal Components from PCA on highly correlated Climate Features from the ERA5 Daily Data

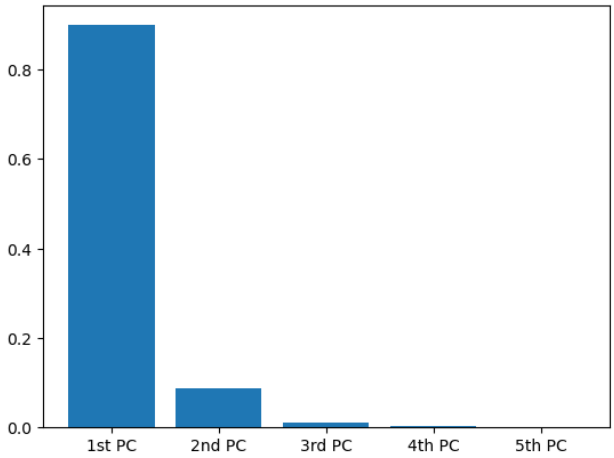

Figure 3.13: Principal Components from PCA on highly correlated Climate Features from the Daymet Data

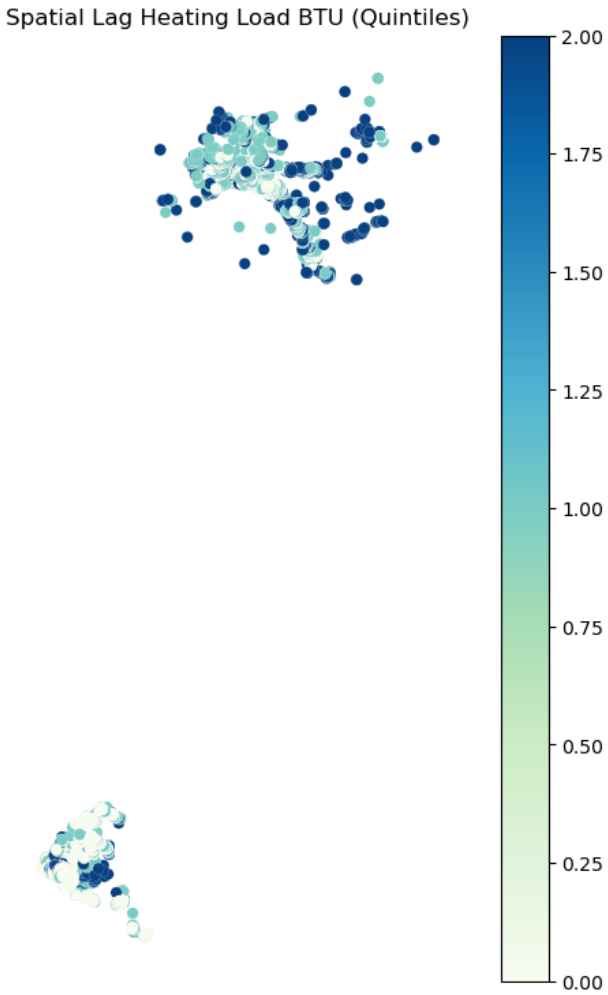

Figure 3.14: Spatial Lag of Heating Load Estimates (in terms of BTUs) using the Daymet climate data and building features .

Chapter 4

Basic Sampling Strategies

The total labeled dataset using the USA Structures building outlines has 34931 buildings from Fairbanks and 87104 buildings from Anchorage, making for a total of 122035 in Fairbanks and Anchorage combined. Thus, only 28.6% of the buildings in our dataset come from Fairbanks whereas 71.4% of the buildings come from Anchorage. The data is clearly not distributed equally over Fairbanks and Anchorage, the two locations we have btu estimatses for. If, we want our model to be able to perform equally well across all the areas in Alaska's Railbelt, we need to train it on balanced data.

Here we cover a variety of commonly used sampling strategies to combat imbalanced data through general approaches including oversampling and undersampling. These strategies consist "of the modification of an imbalanced dataset by some mechanism in order to provide a balanced distribution" He and Ma [2013].

4.1 Oversampling

The basic idea behind oversampling is to increase the number of elements in the minority class by adding samples to the minority class.

4.1.1 Random Oversampling

Random oversampling consists of adding samples chosen randomly with replacement from the minority class to the minority class. This is typically done until the majority and minority class are approximately the same size. Random oversampling is outlined in Algorithm 1.

Algorithm 1 Random Upsampling

Given a feature with a majority class X and a minority class Y

1 while $|X| > |Y|$:
2 Choose y from Y with replacement
3 Add y to Y
4 Return the increased Y

4.1.2 SMOTE

SMOTE stands for Synthetic Minority Over-Sampling Technique Chawla et al. [2002]. Here instead of oversampling via replacement, synthetic examples are created from the feature space. The basic idea is that for each minority class sample, you consider it's k nearest neighbors within the minority class and the lines from the sample to its nearest neighbors. Between the sample and each of its nearest neighbors, take the difference between them, and then multiply this by a randomly chosen number between zero and one. Then add this to the sample. This gives you k synthetic samples per original sample that are its new nearest neighbors. SMOTE is outlined in Algorithm 2.

4.1.3 Borderline Methods

Borderline SMOTE

Instead of creating synthetic samples that are closely related to any of the given samples in the original set, Borderline SMOTE aims to only oversample 'borderline' minority examples Han

Algorithm 2 SMOTE: Synthetic Minority Over-sampling Technique

Given data with a majority class X and a minority class Y
Let $N = |X| - |Y|$ be the number of synthetic samples needed, where $z = int(N/|Y|) => 1$

```
n = 0
synthetic_samples = []
while n < N:
    for y in Y:
        Compute y's k nearest neighbors from within Y, nn
        From nn, choose z nearest neighbors to use to create synthetic samples
        for i in range(z):
            new_sample = empty array of size |y|
            for each feature, j in y:
                Compute diff = nn[i][j] − y[j]
                Compute gap = rand(0, 1)
                new_sample[j] = y[j] + gap*diff
            synthetic_samples.append(new_sample)
            n+=1
return Y together with synthetic_samples
```

et al. [2005]. The general algorithm is similar to SMOTE, but the key difference lies the k nearest neighbors step of SMOTE.

Borderline Oversampling with SVM

Borderline oversampling with SVM Nguyen et al. [2009] is essentially the same as Borderline SMOTE except instead of using a k-nearest neighbors approach to determine which samples are 'borderline', a support vector machine (SVM) is used instead.

4.1.4 ADASYN

ADASYN stands for Adaptive Synthetic Sampling Approach for Imbalanced Learning He et al. [2008]. ADASYN is fairly similar to SMOTE. However, instead of evenly distributing the synthetic samples across the original samples of the minority class, the number of synthethic samples coming from each individual sample from the minority class is determined by weighting the examples from

Algorithm 3 Borderline SMOTE

Given data with a majority class X and a minority class Y such that $X \cup Y = D$ where D is the entire dataset.
Let $N = |X| - |Y|$ be the number of synthetic samples needed, where $z = int(N/|Y|) => 1$

$n = 0$
synthetic_samples = []
while $n < N$:
    for y in Y:
        Compute y's k nearest neighbors from within D, nn
        If all of y's nearest neighbors are from X:
            y is grouped as Noise, and we do not oversample using y.
        elif most of y's nearest neighbors are from X:
            y is grouped as Danger
        else:
            y is grouped as Safe, and we do not oversample using y.
                If y is grouped as Danger, we consider y a borderline sample.
    for y in Danger:
        Calculate y's k nearest neighbors from within Y, nn_y
        From nn_y, choose s nearest neighbors to use to create synthetic samples
        for i in range(s):
            new_sample = empty array of size $|y|$
            for each feature, j in y:
                Compute diff = $nn[i][j] - y[j]$
                Compute gap = $rand(0, 1)$
                new_sample[j] = y[j] + gap*diff
            synthetic_samples.append(new_sample)
            n+=1
return Y along with synthetic_samples

Note: In Borderline SMOTE 2, gap = $rand(0, 0.5)$ so that the synthetic examples are closer to the minority class.

the minority class by the number of their k-nearest neighbors that came from the majority class. ADASYN is described in further detail in Algorithm 4.

Algorithm 4 ADASYN

Given data with a majority class X and a minority class Y such that $X \cup Y = D$ where D is the entire dataset.
Given d_{th}, an acceptable threshold of the ratio of minority to majority class

Calculate the ratio of the minority class to the majority class $d = |Y|/|X|$
If $d < d_{th}$:
    Determine $G = (|X| - |Y|)\beta$ where $\beta = 1$ means a fully balanced dataset
    for y in Y:
        Compute y's k nearest neighbors from within D, nn
        Calculate r_y as the number of the k nearest neighbors that are from X divided by k
    Normalize the r_y's to get a set of $\hat{r}_y$'s
    Calculate $g_y = \hat{r}_y x G$ for each y
    For y in Y:
        # generate g_y synthetic samples
        For i in range(g_y):
            Choose one of y's nearest neighbors from nn that is in Y, y_{nn}
            new_sample $= y + (y_{nn} - y)\lambda$ where $\lambda = rand(0, 1)$
            Add this new_sample to Y
Return the increased Y

4.1.5 Upsampling Fairbanks Results

Here we give tabular results on upsampling based on various features in the data using the upsampling methods described previously in this section.

Table 4.1 shows the results of upsampling data coming from Fairbanks using the PCA ERA5 data. In this table, we can see that overall upsampling does increase the error slightly here for the linear regression, ridge regression, and ridge regression with polynomial features. However, for the decision tree and random forest models, there are some upsampling methods which decrease the overall error.

Table 4.2 shows the results of upsampling data coming from Fairbanks using the PCA Daymet data. In this table, we can see that overall upsampling does increase the error slightly here for the linear regression, ridge regression, and ridge regression with polynomial features. However, for the decision tree and random forest models, there are some upsampling methods which decrease the overall error. It is interesting when comparing this table with the previous table, how similar they are. Using different climate data does not seem to change our results as much as we thought it might.

In Table 4.3, we see a breakdown of the error between Fairbanks and Anchorage for upsampling data coming from Fairbanks using the PCA Daymet data. Here a negative value means the error for Fairbanks was larger than the error for Anchorage, and a positive value means the reverse. We can see that without doing any upsampling, the Fairbanks error is always larger. We also see that compared to the rest of the table, for the most part, the column without sampling shows a larger overall error difference. When we consider the upsampled columns, we see that the difference between the two errors is smaller, and depending on the model the error is sometimes bigger for Fairbanks and sometimes bigger for Anchorage. This is exactly what we were hoping to achieve by sampling our data. Our model is performing more equally with upsampled data across Fairbanks and Anchorage.

	None	Random	SMOTE	Borderline SMOTE	Borderline SVM SMOTE	ADASYN
Linear	0.01125	0.01268	0.01267	0.01293	0.01316	0.0132
Ridge	0.01125	0.01268	0.01267	0.01293	0.01316	0.0132
Ridge w/ poly features	0.0005838	0.0005988	0.0005981	0.0005994	0.0007239	0.0005992
Decision Tree	4.241e-06	4.241e-06	1.335e-05	3.01e-08	3.01e-08	3.007e-08
Random Forest	4.871e-07	1.691e-06	5.906e-06	1.483e-08	1.495e-08	1.577e-08

Table 4.1: Test MSE for Various Upsampling Methods for Upsampling Data from Fairbanks using PCA climate features from ERA5.

	None	Random	SMOTE	Borderline SMOTE	Borderline SVM SMOTE	ADASYN
Linear	0.01049	0.01198	0.01195	0.01231	0.01347	0.01135
Ridge	0.01049	0.01198	0.01195	0.01231	0.01347	0.01135
Ridge w/ poly features	0.003111	0.003457	0.003451	0.004043	0.004866	0.003925
Decision Tree	3.906e-05	0.0001412	9.725e-05	5.595e-06	2.866e-06	2.089e-05
Random Forest	2.445e-05	9.914e-05	0.0001027	1.105e-05	7.412e-06	2.516e-05

Table 4.2: Test MSE for Various Upsampling Methods for Upsampling Data from Fairbanks using PCA climate features from Daymet.

	None	Random	SMOTE	Borderline SMOTE	Borderline SVM SMOTE	ADASYN
Linear	-0.012626	0.002012	0.002205	-0.000114	0.000141	-0.006188
Ridge	-0.012621	0.002014	0.002207	-0.000111	0.000145	-0.006184
Ridge w/ poly features	-0.001589	0.001874	0.001883	0.002606	0.004495	0.002325
Decision Tree	-0.000024	-0.000112	-0.000144	-0.000019	-0.000009	-0.000072
Random Forest	-0.00006	-0.000139	-0.000147	-0.000025	-0.000009	-0.000069

Table 4.3: Error Difference between Anchorage and Fairbanks of Upsampled Data using PCA Daymet Data.

4.2 Undersampling

The basic idea behind undersampling is to decrease the number of elements in your class by removing elements in the majority class.

4.2.1 Random Undersampling

Random undersampling consists of removing samples chosen randomly without replacement from the majority class. This is typically done until the majority and minority class are approximately the same size. Random oversampling is outlined in Algorithm 5.

Algorithm 5 Random Undersampling

Given a feature with a majority class X and a minority class Y

```
while |X| > |Y|:
    Choose x from X without replacement
    Remove x from X
Return the reduced X
```

4.2.2 Neighborhood Methods

Condensed Nearest Neighbor (CNN)

The condensed nearest neighbor algorithm Hart [1968] removes instances from the majority class such that the resulting undersampled majority class is training set consistent, meaning that the training set would be classified the same way trained on the undersampled set of points as trained on the full set of points. In this way, condensed nearest neighbors maximizes the amount of information retained in the undersampled set. Condensed nearest neighbors is described more fully in Algorithm 6.

Algorithm 6 Condensed Nearest Neighbor

Given data with a majority class X and a minority class Y such that $X \cup Y = D$ where D is the entire dataset.

```
Choose k points from D randomly and put them in S
Find x in X such that, x's nearest neighbor in S is different from x's nearest neighbor in X.
If this x exists:
    Add x to S
else:
    You are finished modifying S
return S
```

Edited Nearest Neighbor (ENN)

Edited nearest neighbor Wilson [1972] is very similar to condensed nearest neighbor except that majority class elements and are being removed if the majority of the nearest neighbors lie in the minority class and majority class elements are also being removed if they make of the majority of the nearest neighbors of a minority class element. The edited nearest neighbor algorithm is detailed further in Algorithm 7.

Algorithm 7 Edited Nearest Neighbor

Given data with a majority class X and a minority class Y such that $X \cup Y = D$ where D is the entire dataset.

Choose a number k to be the number of nearest neighbors
Find x in X such that, the majority of x's k nearest neighbors in D are not in X.
If this x exists:
    Remove x from X
Now find neighbors x in X such that, for an element y in Y, the majority of y's k nearest neighbors are in X
If these x exist:
    Remove these x from X
If neither of the above two conditions cannot be satisfied further with the reduced X:
    return the reduced X

Neighborhood Cleaning

Neighborhood cleaning takes advantage of the Edited Nearest Neighbors algorithm by running it on the data first to remove any noisy datapoints. From there, it uses a k nearest neighbors approach to remove elements in the majority class that could be easily misclassified as the minority class. Neighborhood cleaning is detailed further in Algorithm 8.

4.2.3 Near Miss Undersampling

There are three different versions of near miss undersampling Zhang and Mani, near miss 1, near miss 2, and near miss 3. We use near miss 1 here as it performs best out of the three different

Algorithm 8 Neighborhood Cleaning

Given data with a majority class X and a minority class Y such that $X \cup Y = D$ where D is the entire dataset.

```
Remove all 'noisy' data in X using Edited Nearest Neighbors
If the remaining data in X > Y/2:
    For remaining x in X:
        If the majority of the 3 nearest neighbors of x belong to Y:
            Remove x from X
return the reduced X
```

versions on our data. Near miss 1 differs from 2 and 3 in that it chooses samples to keep from the majority class that have the smallest average distance to the three examples that are closest and are from the minority class. Near miss 1 is further outlined in Algorithm 9.

Algorithm 9 Near Miss 1

Given data with a majority class X and a minority class Y such that $X \cup Y = D$ where D is the entire dataset.

```
For x in X:
    Calculate x's three nearest neighbors from Y, y_1, y_2, y_3, and record the average
    distance out of the distance from x to y_1, x to y_2, and x to y_3
Sort the x's from shortest to largest in terms of the average distance previously calculated
return the first |Y| of the x's
```

Tomek Links Undersampling

Tomek links Tomek [1976] is an undersampling algorithm that can be used to remove all samples who's nearest neighbor is not of the same class. The Tomek links algorithm is more fully explained in Algorithm 10.

Algorithm 10 Tomek Links

Given data with a majority class X and a minority class Y such that $X \cup Y = D$ where D is the entire dataset.
A pair of data samples (x, y) form a Tomek Link if x and y are each other's nearest neighbors, and x is in X and y is in Y.

1 Find all tomek links in D, (x, y).
2 For each tomek link:
3 Remove the x component of the tomek link from X
4 return the reduced X

4.2.4 Undersampling Anchorage Results

Here we give tabular results on downsampling based on various features in the data using the downsampling methods described previously in this section.

Table A.5 shows the results of downsampling data coming from Anchorage using the PCA ERA5 data. In this table, we can see that overall downsampling does increase the error slightly here for the linear regression, ridge regression, ridge regression with polynomial features, and random forest. However, for the decision tree, there are some downsampling methods which decrease the overall error.

Table 4.5 shows the results of downsampling data coming from Anchorage using the PCA Daymet data. In this table, we can see that overall downsampling does increase the error slightly here for the linear regression, ridge regression, ridge regression with polynomial features, and random forest. However, for the decision tree, there are some downsampling methods which decrease the overall error. It is interesting when comparing this table with the previous table, how similar they are. Using different climate data does not seem to change our results as much as we thought it might, showing some similarity to the upsampling results.

In Table 4.6, we see a breakdown of the error between Fairbanks and Anchorage for downsampling data coming from Anchroage using the PCA Daymet data. Here a negative value means the error for Fairbanks was larger than the error for Anchorage, and a positive value means the reverse.

We can see that without doing any downsampling, the Fairbanks error is always larger. We also see that compared to the rest of the table, for the most part, the column without sampling shows a smaller overall error difference, meaning downsampling is not improving the models performance breakdown. When we consider the downsampled columns, we do see that depending on the model the error is sometimes bigger for Fairbanks and sometimes bigger for Anchorage. Overall, these results make sense due to the fact that downsampling is removing information.

	None	Random	Condensed Nearest Neighbor	Near Miss	Tomek Links	Edited Nearest Neighbor	One Sided Selection	Neighborhood Cleaning Rule
Linear	0.03532	0.03966	0.09914	0.04709	0.03532	0.03532	0.03784	0.03532
Ridge	0.03532	0.03966	0.09435	0.04708	0.03532	0.03532	0.03784	0.03532
Ridge w/ poly features	0.02443	0.02813	0.1969	0.1325	0.02443	0.02443	0.02917	0.02443
Decision Tree	0.0002773	0.0002492	0.07038	0.003898	0.0002579	0.000259	0.0002715	0.0002579
Random Forest	0.0001884	0.0002007	0.07725	0.003783	0.0001907	0.0001925	0.00028	0.000189

Table 4.4: Test MSE for Various Downsampling Methods for Downsampling Data from Anchorage using PCA climate features from ERA5

	None	Random	Condensed Nearest Neighbor	Near Miss	Tomek Links	Edited Nearest Neighbor	One Sided Selection	Neighborhood Cleaning Rule
Linear	0.01049	0.01194	0.08668	0.01157	0.01049	0.01049	0.01942	0.01049
Ridge	0.01049	0.01194	0.08594	0.01157	0.01049	0.01049	0.01942	0.01049
Ridge w/ poly features	0.003111	0.003443	0.1096	0.01931	0.003111	0.003111	0.005269	0.003111
Decision Tree	5.662e-06	0.0002539	0.09594	2.1e-05	1.109e-05	1.129e-05	1.118e-05	1.112e-05
Random Forest	2.672e-05	0.0001782	0.09597	3.782e-05	2.139e-05	2.546e-05	1.724e-05	2.322e-05

Table 4.5: Test MSE for Various Downsampling Methods for Downsampling Data from Anchorage using PCA climate features from Daymet

	None	Random	Condensed Nearest Neighbor	Near Miss	Tomek Links	Edited Nearest Neighbor	One Sided Selection	Neighborhood Cleaning Rule
Linear	-0.012626	0.001985	0.160178	-0.000853	-0.012626	-0.012638	0.033805	-0.012635
Ridge	-0.012621	0.001992	0.1485	-0.000847	-0.012621	-0.012633	0.033818	-0.01263
Ridge w/ poly features	-0.001589	0.002046	0.443736	0.022455	-0.001589	-0.00158	0.002141	-0.001577
Decision Tree	-0.000122	0.000012	0.134746	0.000016	-0.000015	-0.000011	-0.000018	-0.000011
Random Forest	-0.000049	-0.000089	0.136455	-0.000016	-0.000055	-0.000042	-0.000043	-0.000046

Table 4.6: Error Difference between Anchorage and Fairbanks of Downsampled Data using PCA Daymet Data.

Chapter 5

Conclusion

From this work, we reach several interesting conclusions about our data. The systemic approach outlined in this work is summarized in Figure 5.1.

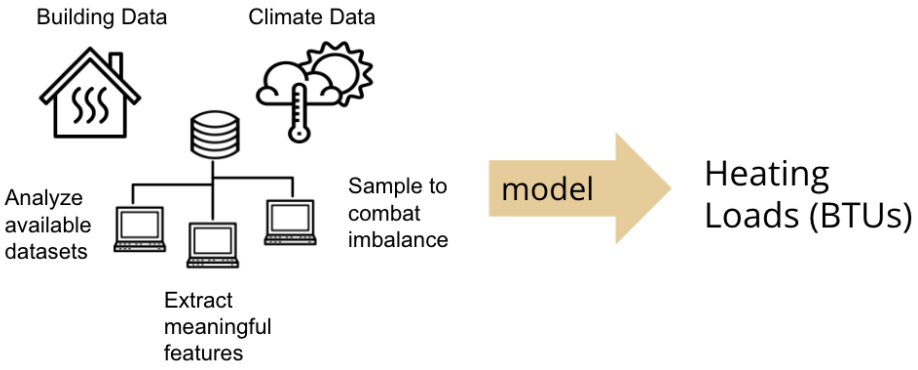

Figure 5.1: Geospatial First Data Exploration Workflow .

By comparing both the ERA5 climate data and the Daymet climate data and our models' performances using features extracted from these datasets respectively, we see there is less variability between the two datasets than we had previously hoped. Even though the Daymet data has a much finer spatial resolution than the ERA5 data, performance on the data was similar, and features extracted from the climate data could be represented using a similar amount of information.

From comparing multiple publicly available building datasets, we see that it is most likely there

is not a dataset available that has all of the buildings in Alaska's Railbelt, notably in Fairbanks and Anchorage. The variance between these building datasets is quite large in terms of building count, but also in terms of building height and area. Data quality in these building datasets also varies within the features of the dataset.

From analyzing all of the data in this problem, we found that our data is most likely spatially dependent and not spatially homogenous.

We try to combat the fact that the data available for modeling heating loads in Alaska is imbalanced, especially between Anchorage and Fairbanks. Without sampling our data at all, all of the models perform worse on Fairbanks than they do on Anchorage. When we upsample the data coming from Fairbanks, we see that the models' performance on Fairbanks and Anchorage is more even than when not sampled. When we downsample the data coming from Anchorage, we see that the model's performance on Fairbanks and Anchorage is also more even than when not sampled. However, we see that the model performs worse in general when downsampling the data whereas when upsampling, there are some instances where the model perform better overall. It is also worth noting that random sampling performs quite well overall, and perhaps, in future work estimating heating loads in Alaska, random sampling is sufficient over some of the more advanced sampling techniques covered here.

Bibliography

Alaska Center for Energy and Power. Puma fuel meters capture fuel use in fairbanks homes.

Alaska Housing Finance Corporation. 2014 alaska housing assessment.

Nick Bolten, Vidisha Chowdhury, Madelyn Gaumer, Philippe Schicker, Shamsi Soltani, and Erin Trochim.

C.F. Brown, S.P. Brumby, B. Guzder-Williams, et al. 2022. Dynamic world, near real-time global 10m land use land cover mapping. *Nature Scientific Data*, 9:251.

Nitesh V. Chawla, Lawrence O. Bowyer, Kevin W. an Hall, and W. Philip. Kegeelmeyer. 2002. Smote: Synthetic minorityu over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Billy Connor, Douglas J Goering, Mikhail Kanevskiy, Erin Trochim, Kevin L Bjella, and Robert L McHattie. 2020. Roads and airfields construction on permafrost: A synthesis of practice. *AK-DOTPF Report*, 000S927.

Copernicus Climate Change Service (C3S) (2017). Era5: Fifth generation of ecmwf atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store (CDS).

Esri US Federal Data. 2022. Usa structures.

Fairbanks Northstar Borough. Get fnsb gis datas.

GIS Geography. 2022. What is tobler's first law of geography?

Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. 2017. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*.

Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. 2005. Borderline-smote: A new over-sampling method in imbalanced data sets learning. *International Conference on Intelligent Computing*, Part 1:878–887.

Peter E. Hart. 1968. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 18:515–516.

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. 2008. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328.

Haibo He and Yuanqian Ma. 2013. *Imbalanced Learning: Foundations, Algorithms, and Applications*.

M. Marconcini, A. Metz-Marconcini, T. Esch, and N. Gorelick. 2021. Understanding current trends in global urbanisation – the world settlement footprint suite. *GI_Forum*, 1:33–38.

Joshua New, Mark Adams, Anne Berres, Brett Bass, and Nicholas Clinton. 2012. Model america-data and models of every u.s. building. Support for DOI 10.13139/ORNLNCCS/1774134 dataset is provided by the U.S. Department of Energy, project Automatic Building Energy Modeling (AutoBEM) under Contract DE-AC05-00OR22725. Project Automatic Building Energy Modeling (AutoBEM) used resources of the Oak Ridge Leadership Computing Facility at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

Hien M. Nguyen, Eric W. Cooper, and Katsuari Kamei. 2009. Borderline over-sampling for imbalanced data classification. *Proceedings : Fifth International Workshop on Computational Intelligence Applications*, 2009:24–29.

Open Street Maps. [link].

M. Rantanen, A.Y. Karpechko, A. Lipponen, et al. 2022. The arctic has warmed nearly four times faster than the globe since 1979. *Communications Earth Environment*, 322(168):1–10.

Samapriya Roy. Global ml building footprints. Created by Microsoft.

M.M. Thornton, R. Shrestha, P.E. Wei, S. Kao Thornton, and B.E. Wilson. Daymet: Daily surface weather data on a 1-km grid for north america, version 4. Oak Ridge, Tennessee, USA.

Ivan Tomek. 1976. Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(11):769–772.

U.S. Energy Information Administration. Alaska state profile and energy estimates.

WHPacific. 2012. Alaska energy authority end use study: 2012.

Dennis L. Wilson. 1972. "asymptotic properties of nearest neighbor rules using edited data.". *IEEE Trans. Syst. Man Cybern.*, 2:408–421.

Jianping Zhang and Inderjeet Mani. knn approach to unbalanced data distributions: A case study involving information extraction. *ICML: Workshop on Learning from Imbalanced Datasets II*.

Chapter A

Appendix One

Model	Fairbanks MSE	Anchorage MSE
Linear	0.015723	0.003886
Ridge	0.016216	0.004078
Ridge w/ poly features	0.000072	0.000016
Decision Tree	0.000090	0.000088
Random Forest	0.000087	0.000029

Table A.1: Error Split between Fairbanks and Anchorage of Unsampled Data of ERA5 Data (No PCA).

Model	Fairbanks MSE	Anchorage MSE
Linear	0.016703	0.004064
Ridge	0.016803	0.004268
Ridge w/ poly features	0.000910	0.000245
Decision Tree	0.000539	0.000029
Random Forest	0.000323	0.000031

Table A.2: Error Split between Fairbanks and Anchorage of Unsampled Data of Daymet Data (No PCA).

	None	Random	SMOTE	Borderline SMOTE	Borderline SVM SMOTE	ADASYN
Linear	0.007503	0.009024	0.009018	0.009111	0.009255	0.009208
Ridge	0.007552	0.009019	0.009022	0.009108	0.009261	0.009206
Ridge w/ poly features	0.007552	0.009019	0.009022	0.009108	0.009261	0.009206
Decision Tree	8.872e-05	3.904e-05	8.677e-05	0.000121	0.0001341	0.0001274
Random Forest	4.37e-05	3.304e-05	4.632e-05	7.456e-05	7.062e-05	6.897e-05

Table A.3: Test MSE for Various Upsampling Methods for Upsampling Data from Fairbanks using climate features from ERA5 (no PCA).

	None	Random	SMOTE	Borderline SMOTE	Borderline SVM SMOTE	ADASYN
Linear	0.007307	0.009014	0.009005	0.0091	0.00928	0.009212
Ridge	0.007552	0.009024	0.009018	0.009097	0.009298	0.009209
Ridge w/ poly features	3.199e-05	2.702e-05	2.7e-05	4.666e-05	4.061e-05	4.823e-05
Decision Tree	8.867e-05	6.217e-05	9.643e-05	0.0001316	0.0001277	0.0001316
Random Forest	4.427e-05	3.63e-05	4.399e-05	6.909e-05	6.63e-05	7.072e-05

Table A.4: Test MSE for Various Upsampling Methods for Upsampling Data from Fairbanks using climate features from Daymet (no PCA).

	None	Random	Condensed Nearest Neighbor	Near Miss	Tomek Links	Edited Nearest Neighbor	One Sided Selection	Neighborhood Cleaning Rule
Linear	0.007277	0.008992	0.03636	0.02036	0.007388	0.007283	0.007447	0.007279
Ridge	0.007552	0.008986	0.04053	0.02038	0.007552	0.007552	0.007596	0.007552
Ridge w/ poly features	3.199e-05	3.424e-05	0.01842	9.83e-05	3.199e-05	3.199e-05	7.751e-05	3.198e-05
Decision Tree	8.869e-05	0.0001146	0.08215	0.003707	8.868e-05	8.866e-05	8.326e-05	9.354e-05
Random Forest	4.534e-05	7.856e-05	0.08327	0.003629	4.618e-05	4.441e-05	4.631e-05	4.62e-05

Table A.5: Test MSE for Various Downsampling Methods for Downsampling Data from Anchorage using climate features from ERA5 (no PCA).

	None	Random	Condensed Nearest Neighbor	Near Miss	Tomek Links	Edited Nearest Neighbor	One Sided Selection	Neighborhood Cleaning Rule
Linear	0.007682	0.009055	0.06411	0.009062	0.007682	0.007683	0.01474	0.007683
Ridge	0.007857	0.009472	0.07087	0.00909	0.007857	0.007857	0.0158	0.007856
Ridge w/ poly features	0.0004355	0.0005828	0.114	0.001467	0.0004355	0.0004355	0.0006219	0.0004353
Decision Tree	0.0001575	0.000336	0.08094	0.0003289	0.000149	0.0001615	0.0001747	0.0001657
Random Forest	2.672e-05	0.0001782	0.07935	3.782e-05	2.139e-05	2.546e-05	1.724e-05	2.322e-05

Table A.6: Test MSE for Various Downsampling Methods for Downsampling Data from Anchorage using climate features from Daymet (no PCA).