Building and occupant characteristics as predictors of temperature-related health hazards in American homes

Arfa N. Aijazia,\*, Stefano Schiavona, Duncan Callawayb

a Center for the Built Environment, University of California, Berkeley, CA 94720

b Energy and Resources Group, University of California, Berkeley, CA 94720

\* Corresponding author:

Email address: [arfa@berkeley.edu](mailto:arfa@berkeley.edu)

Address: Center for the Built Environment (CBE)

University of California, Berkeley

390 Wurster Hall #1839

Berkeley, CA 24720-1839

Keywords: Thermal resilience, Housing, Extreme heat, Extreme cold, Overheating, Public health, Climate change

Target publication: Building and Environment

# Highlights

# Graphical Abstract

# Abstract

1. Introduction
   1. Background

Though less visibly destructive than floods, hurricanes, and other hazards, prolonged periods of extreme temperatures are the leading cause of weather-related deaths in the United States (Berko et al. 2014). Globally, extreme ambient temperature (either or hot or cold) contributes to 6.5-10% of all deaths (Sera et al. 2019; Zhao et al. 2021). While the media focus is largely on extreme heat, in both studies, nearly 90% of global deaths attributed to temperature were cold-related, which is consistent with findings in the United States (Berko et al. 2014). It is hard to pinpoint the true public health impact of extreme temperature because it aggravates underlying conditions, but may not be noted as a contributing cause on hospital records or death certificates (Ostro et al. 2009; Lane 2018). At least 17 chronic conditions such as heart disease, diabetes, kidney disease, and respiratory infection show a J-shaped relationship with temperature (Burkart et al. 2021), meaning that the disease prevalence increases in both extreme high and low ambient temperatures. In addition to these serious health impacts, extreme temperature exposure impairs labor productivity (Lai et al. 2023), sleep quality (Obradovich et al. 2017), and cognitive performance (Laurent et al. 2018).

Several trends necessitate prioritizing emergency planning and disaster mitigation with regards to extreme temperatures, the first being anthropogenic climate change. Future projections show rising ambient temperatures as well as increasing frequency and intensity of heat waves (IPCC 2021). Individuals may also be exposed to hazardous temperatures in the event of a power outage coinciding with an extreme weather event, as was seen across the Northeastern United States following Hurricane Sandy in 2012 (Henry and Ramirez-Marquez 2016) the state of Texas in February 2021 (King, Rhodes, and Zarnikau 2021). Major electrical grid failures have increased by more than 60% in recent years (Stone et al. 2021). The final trend is global aging. In 2019, less than a tenth of the global population was over the age of 65 and by 2050 this number will increase to 1 in 6 (United Nations 2020). Age is a well-documented risk factor for temperature-related morbidity and mortality (Oudin Åström, Bertil, and Joacim 2011), so an older population has greater vulnerability.

* 1. Municipal planning for extreme temperatures

Most cities face substantial variation in intra-city vulnerability to heat0F[[1]](#footnote-1). Many cities and other jurisdictions use heat vulnerability indices (HVI) to better allocate resources on emergency response to heat, such as the location of cooling centers (Reid et al. 2009; Rinner et al. 2010; Uejio et al. 2011; Nayak et al. 2018). An HVI is typically a weighted sum of variables related to heat vulnerability such as income or poverty level, age, social isolation, and land cover characteristics calculated at a localized level like census tract.

HVI seldom include variables related to the local building stock even though building characteristics can exacerbate or mitigate occupant exposure to heat. Indoor exposure, particularly at home, accounts for a sizable portion, 38-85%, of heat-related deaths (Fouillet et al. 2006; Wheeler et al. 2013; Iverson et al. 2020). An individual will on average spend around 67% of their time in a residence, out of 87% of their total time indoors (Klepeis et al. 2001). This proportion is even higher for vulnerable populations such as infants and the elderly, who on average spend 89% and 78% of their time in a residence respectively (Matz et al. 2014).

A review by Samuelson et al. (Samuelson et al. 2020), found that out of 20 HVI’s from different cities and regions, eight included the year of construction, nine included central air-conditioning (AC) ownership, three included floor of residence, and one included rooftop albedo and thermal mass. Many building characteristics that could potentially impact indoor heat exposure such as orientation, envelope properties, and typology are not included. City-level tax assessor data typically records year of construction and the presence of central AC at the parcel level, so these variables are attractive proxies for the contribution of the built environment. However, there is no obvious link between these variables and heat vulnerability.

HVI for New York State (Nayak et al. 2018) and the cities of Toronto (Rinner et al. 2010) and Philadelphia (Uejio et al. 2011) all considered older homes to have a higher risk for heat exposure due to a presumption of lack of insulation, lower likelihood of AC, and correlation with other risk factors like poverty. However, several studies monitoring temperatures in European residences found that older buildings had significantly cooler temperatures than newer ones (Beizaee, Lomas, and Firth 2013; Maivel, Kurnitski, and Kalamees 2015; Pathan et al. 2017), perhaps because of the thermal mass properties of stone construction typical of older European homes. Temperature monitoring in American residential buildings did not find a strong correlation between construction age and measured indoor temperature during times the home was actively heated or cooled (Booten et al. 2017). A simulation-based comparison of housing archetypes in Boston found older typologies had lower maximum indoor temperatures (Samuelson et al. 2020). These studies demonstrate that construction age alone cannot capture indoor heat exposure.

HVI also typically consider AC prevalence, particularly that of central AC systems. In homes where AC is present, the cost of operating and maintaining AC systems may prohibit their use in a way that sufficiently protects residents from the adverse effects of heat. In recent investigations of indoor heat deaths, the Maricopa County Department of Public Health (MCDPH) found that in 91% of cases, AC was present (MCDPH 2019). In those cases, the AC was either broken (87%), disconnected from electricity (5%), or functioning but not turned on (8%). Clearly, the sheer presence of AC alone is not a protective factor against overheating.

The primary barrier to including additional building level characteristics is data availability at sufficient scale. However, new methods of data acquisition are rapidly becoming available such as smart thermostat data (Ecobee 2021) and satellite and street-level imagery (New et al. 2020). Understanding the role of the building and other household characteristics will enable public agencies to target emergency planning efforts and resources in the short-term like locations of cooling and warming centers and the long-term such as social programs and building weatherization.

* 1. Research gaps and objectives

Several research gaps relate to the role of building characteristics on temperature-related morbidity and mortality.

First, is the lack of empirical evidence to link building characteristics to temperature-related morbidity and mortality. Studies assessing the sensitivity of overheating risk to building characteristics often use building performance simulations. These studies use simulation outputs such as maximum daily room temperature (Samuelson, Baniassadi, and Gonzalez 2020; Mavrogianni et al. 2012), percent of time in different U.S. Occupational Safety and Health Administration (OSHA) heat index (HI) risk categories (Sun, Specian, and Hong 2020), or the degree-hours the wet-bulb globe temperature (WBGT) index exceeded a threshold value (Baniassadi, Heusinger, and Sailor 2018) to model the risk of overheating. Holmes et al. (Holmes, Phillips, and Wilson 2016) discuss the suitability of different thermal indices with building performance simulation outputs and recommend acceptable thresholds for residential spaces based on WBGT. While there are many thermal indices, as yet none of them are validated for personal exposure in residential environments, meaning the recommended thresholds are not based on empirical observations of temperature-related health hazards in this context (Kuras et al. 2017). Recently, the Occupational Safety and Health Review Commission criticized OSHA’s use of HI to assess risk for heat exposure in occupational settings due to the lack of scientific evidence for risk categories (“Secretary of Labor v. A.H. Sturgill Roofing, Inc.” 2019).

The second research gap is limited consideration for building characteristics in conjunction with other markers of socioeconomic vulnerability such as income and age. Very few studies using building performance simulations review the interaction of building and occupant characteristics. Baniassadi et al. (Baniassadi et al. 2019) account for some effects of occupant income by modeling AC non-functionality and occupant age by using a conservative value for their overheating threshold.

To overcome these research gaps, this study uses supervised machine learning to predict temperature-related morbidity based on a nationwide survey of building and household characteristics in American homes. This study revolves around two research questions:

1. Would a HVI with detailed information about the building be more accurate at predicting the risk of health hazards? If so, by how much?
2. Which building and occupant characteristics contribute most to predicting the risk of a health hazards?

More accurate predictions will allow public agencies to better identify at-risk households and strategize limited resources for short-term planning like locations of cooling and warming centers and long-term planning like building weatherization and social programs. Understanding the contributions of building and occupant characteristics can prioritize data collection efforts.

1. Materials and methods
   1. Residential energy consumption survey (RECS) data

The main source of data for this study is the Residential Energy Consumption Survey (RECS), which is administered by the U.S. Energy Information Administration (EIA) (EIA 2018; 2022). RECS is a periodic survey that has collected detailed energy characteristics, usage patterns, and demographics of American households since 1978. The primary objective of RECS is to estimate future energy demand and improve energy efficiency and building design.

Of relevance for this study, the three most recent cycles of RECS, 2009, 2015, 2020, ask respondents if in the last year, if anyone in their household needed medical attention because the home was too hot or too cold (EIA 2016; 2020). This study treats an affirmative response to either question as a temperature-related morbidity. While the questions are vague regarding duration and severity of extreme temperatures and who in the household needed medical assistance, it provides a source of empirical evidence that the household experienced a hazardous interior thermal environment. We only analyzed data from the 2015 and 2020 RECS. Responses to our questions of interest are not available in the public data file for the 2009 RECS due to infrequent responses risking disclosure of sensitive and confidential household information.[[2]](#footnote-2)

Each RECS is an independent cross-sectional study of residential energy use, so each iteration of the survey is slightly different. Theoretically, it’s possible to select the same home twice, but it is highly unlikely and occurs rarely.[[3]](#footnote-3) The EIA selects samples to statistically represent all U.S. households occupied as a primary residence at the time of the survey. The most significant difference between the 2015 and 2020 survey cycles is the mode of execution. The 2015 survey cycle collected data through a combination of computer-assisted personal interviews, internet, and mailings. The 2020 survey cycle relied entirely on self-administered web and paper questionaries. Because there were no in-person interviews, the 2020 survey did not use a clustered sampling method like in 2015. The impact of this change is a three-fold increase in sample size – 5,686 in 2015 to 18,496 in 2020. Sample size is inversely proportional to the standard error and, so larger samples generally result in narrower confidence intervals for population and subpopulation estimates. The two survey years also had minor differences in variable coding, which we describe in more detail in Appendix X.

Table 1 shows counts of heat-, cold-, or any temperature-related morbidity in the 2015 and 2020 RECS. For the predictive model, we treat each sample as an independent observation. However, for population estimate, we reviewed the results of each year separately due to differences in sampling methods. For each year, RECS calculates the sample weight, which indicates the number of households in the population that observation represents. Inclusion of replicate weights allows for calculation of the sampling error. We followed the EIA’s procedure for calculating population estimates, standard errors, and confidence intervals in R (EIA 2019; 2023).

Table 1. Observations of temperature-related morbidity in RECS

|  |  |  |  |
| --- | --- | --- | --- |
| **Morbidity** | **2015** | **2020** | **Total** |
| Heat-related | 39 | 76 | 115 |
| Cold-related | 54 | 120 | 174 |
| Any temperature | 81 | 171 | 252 |
| None | 5,605 | 18,496 | 24,101 |

To explore patterns in households that experienced temperature-related morbidity we narrowed the over 750 household characteristics described in the RECS dataset to approximately 25 related to either vulnerability or exposure to extreme temperature. These variables fall under 3 categories: climate, demographics, and buildings. We describe these building and household characteristics in the subsequent sections. Table 2 provides an overall summary of all input variables.

By default, the RECS dataset encodes all variables as numerical quantities. We retained the numerical values for truly numerical household characteristics like the construction age, categorical household characteristics describing ordinal data such as the level of insulation, or binary variables such as presence of back-up generator or on-site solar. We converted categorical variables into informed groups based on the literature on heat and cold-related vulnerability for example non-white versus individual racial and ethnic categories. We also derived new variables of interest such as poverty, which combines the number of household members with income level and thermal mass, which combines insulation level and exterior wall or roof material.

Climate

Ground surface temperature is a climatic variable often reported in HVI (Uejio et al. 2011), because it represents local exposure to extreme temperatures. Due to the EIA’s objective to forecast energy demand, climatic variables in RECS are oriented towards HVAC system operation, such as cooling and heating design-temperature.

Demographics

Many epidemiological studies have investigated the relationship between different demographic and socioeconomic variables on heat-related mortality. Several studies identify elderly age as a vulnerability factor, but there is some ambiguity around the cut-off for higher risk: 60, 65, 70, or 75 (Applegate et al. 1981; O’Neill, Zanobetti, and Schwartz 2003; Ballester et al. 1997; Centers for Disease Control and Prevention (CDC) 1995; Conti et al. 2005). The elderly may be more likely to have co-morbidities and take medication that affect thermoregulation. They also may have limited mobility to access cooling centers or limit AC usage due to fixed income. Several studies also suggest a link between income and heat-related mortality as measured by poverty (Naughton et al. 2002; Curriero et al. 2002), unemployment (Nayak et al. 2018), renter status (Uejio et al. 2011; Wright et al. 2020), and utility payment (Wright et al. 2020). Klinenberg’s sociological analysis of the 1995 Chicago heat wave found a higher risk of death in individuals with limited social connections such as those living alone (Klinenberg 2015). These individuals may be at higher risk of not being checked on regularly during a heat emergency. On the other hand large households (7+ members) may also have elevated heat mortality risk (Uejio et al. 2011). The impact of race and ethnicity on heat-related mortality is mixed, with some studies finding a higher risk for African Americans or non-white racial and ethnic groups (O’Neill, Zanobetti, and Schwartz 2005; Schwartz 2005).

Table 2. Summary of household characteristics derived from the RECS dataset relevant to the household’s vulnerability or exposure to extreme temperature.

| **Category** | **Variable** | **Variable description** | **Type a** |
| --- | --- | --- | --- |
| Climate | Cooling design temperature | Dry bulb design temperature (F) expected to be exceeded 1% of the time | N |
| Heating design temperature | Dry bulb design temperature (F) expected to be exceeded 99% of the time | N |
| Demographic | Non-white | Householder (respondent) race is non-white, or ethnicity is of Spanish descent. | B |
| Older than 65 | Respondent or household member age is > 65 | B |
| Lives alone | Number of household members = 1 | B |
| Large households (7+ members) | Number of household members > 7 | B |
| Poverty | Calculated from gross income and number of household members based on U.S. Census Bureau definition for poverty threshold for that year | B |
| Unemployed | Respondent is unemployed or retired | B |
| Education level | Highest education attained is high school or equivalent |  |
| Renting | Household pays rent | B |
| Pays utility or for other fuel | Household pays for electricity, natural gas, propane, and/or fuel oil | B |
| Buildings: construction | Construction age | Estimated year when housing unit was built (taken as the maximum of the range in RECS response coding) | N |
| Apartment | Type of housing unit is low-rise or high-rise apartment | B |
|  | Mobile | Type of housing unit is mobile home | B |
| Buildings: envelope | Thermally massive wall | Estimated thermal mass based on exterior wall material and presence of insulation | N |
| Thermally massive roof | Estimated thermal mass based on exterior roof material and presence of insulation | N |
| Insulation | Level of insulation | N |
| Infiltration | Frequency of draft | N |
| Window per room | Number of windows per room as an approximation for window-to-wall ratio | N |
| Glazing type | Type of glass in most windows | N |
| Buildings: HVAC | AC type | Air conditioning equipment used | N |
| Heating type | Space heating equipment used | N |
| Energy insecurity | Household reported difficulty paying energy bills; repairing or replacing broken heating or cooling equipment; or that they had kept their home at unsafe temperatures because of cost concerns | B |
| Fans | Number of ceiling, floor, window, and/or table fans used | N |
| Off-grid | Home has back-up generator or on-site solar electricity generation | B |

a Type includes numerical (N) and binary (B)

Buildings: construction

Building construction represents variables related to the building age and form. As mentioned in Section 1.2, several city and state-level HVI use construction age as a catch-all for other building characteristics (Rinner et al. 2010; Uejio et al. 2011; Nayak et al. 2018). A building performance simulation study of London dwellings found a significant impact of archetype, a combination of construction age and construction type on overheating risk (Mavrogianni et al. 2012). Samuelson et al. (Samuelson et al. 2020) suggests that detached buildings may be less vulnerable due to a greater potential for exposed walls to exchange heat and more opportunities for cross-ventilation. Similarly, Lomas (Lomas 2021) singles out flats or apartments because of more limited opportunities for natural ventilation. Mobile or manufactured homes may also increase heat or cold exposure due to poor energy efficiency (Harrison and Popke 2011), an issue common in even newer mobile homes (Hart et al. 2002).

Buildings: envelope

The building envelope represents the materials that separate the interior and exterior of the building. Building performance simulations find wall insulation reduces the overheating risk when applied to the exterior, but may increase overheating risk when applied to the interior (Mavrogianni et al. 2012; Porritt et al. 2012). Porritt et al. (Porritt et al. 2012) also found a correlation between roof and wall surface reflectivity i.e., inverse of solar absorptivity and overheating risk. Samuelson et al. (Samuelson et al. 2020) infers that other building envelope characteristics like infiltration, and window-to-wall ratio may also be significant. One HVI considered houses with thermally massive materials to have greater adaptive capacity (Inostroza, Palme, and de la Barrera 2016). Thermal mass describes building materials with high heat capacity, such as brick, stone, and concrete, which can buffer temperature fluctuations. For this strategy to work, the material must be exposed to both the interior and exterior i.e. limited insulation.

Buildings: HVAC

Building HVAC characteristics describe the presence (Curriero et al. 2002), type (O’Neill, Zanobetti, and Schwartz 2005), and functionality (Naughton et al. 2002; MCDPH 2019) of HVAC systems. Samuelson et al. (Samuelson et al. 2020) also suggests consideration for electric fans. While fans cannot reduce air temperature, they can aid evaporative cooling in air temperatures as high as 50°C and 10% relative humidity (Jay et al. 2015). Finally, we also consider availability of alternate power sources such as a back-up generator or on-site solar panels as they may reduce interruptions to HVAC systems.

* 1. Machine learning

We used machine learning to predict the occurrence of a temperature-related morbidity event, which we treated as a binary classification problem since the RECS survey is coded as a yes or no response. The input features for the machine learning model are the variables described in Table 2. We focus on comparing the performance of models trained with and without building characteristics.

We know from Table 1 that there is an extreme imbalance in the RECS data – less than 1% of all households experienced temperature-related morbidity. This imbalance is problematic because a naïve model that always predicts the majority class, i.e. no temperature-related morbidity, will have a high accuracy, 99% in this case, but will fail to predict any observations in the minority class, i.e. occurrence of temperature-related morbidity. Imbalanced data is a common issue in other domains such as disease diagnosis, customer churn prediction, and fraud detection. As in our case, imbalanced data problems generally have a high cost associated with failure to predict the minority class. We employ several techniques in the machine learning model building process to address the imbalanced data (He and Garcia 2009; Kaur, Pannu, and Malhi 2019; Krawczyk 2016).

We first pre-processed the data set to remove variables with zero or near-zero variance. These variables can negatively impact model performance as they may become zero variance after the data is subdivided. This step removed the variable for large households and pays utility and/or fuel. We also checked for highly correlated variables (magnitude of Spearman’s correlation coefficient > 0.75) and linear combinations, but no variable met the threshold for removal. We then standardized input variables to have zero mean and unit variance. This step prevents variables with larger numerical quantities from having undue influence, particularly in regression-based modeling methods.

We then split the RECS dataset into training and test data, using 80% for training and holding 20% for testing, which prevents overfitting. We bootstrapped this process with 30 iterations to quantify the uncertainty in model performance due to the training data split. For each training and test split, we then used 5-fold cross validation repeated 5 times to further split the training data into training and validation sets for selecting machine learning model hyperparameters.

We compared performance from several machine learning algorithms, listed in Table 3. These algorithms vary in their underlying structure and assumptions about input features. We selected these algorithms because of their ability to accept class weights and availability in the R caret library. We applied an exhaustive grid search of 100 values to find the best performing hyperparameter settings for each machine learning algorithm.

Table 3. Summary of machine learning algorithms

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Hyperparameters (min, max)** | **R implementation** |
| Generalized linear model | None | glm (Kuhn et al. 2023) |
| Multivariate Adaptive Regression Spline | #Terms: (1, 100)  Product degree: 1 | earth (Milborrow 2023) |
| Penalized discriminant analysis | Shrinkage penalty coefficient: (0, 0.1) | pda (Hastie and Tibshirani 2023) |
| Penalized multinomial regression | Weight decay = (0, 0.1) | Multinom (Ripley and Venables 2023) |
| Bagged classification and regression tree | None | treebag (Peters et al. 2023; Wickham 2023; Meyer et al. 2023) |
| Stochastic gradient boosting | # Boosting iterations: (50, 500)  Max. tree depth: 1  Shrinkage: (510-3, 510-2)  Min. terminal node size: 10 | gbm (Greenwell et al. 2022; Wickham 2023) |
| Random forest | # Randomly selected predictors: (1, # of variables)  Splitting rule: Gini impurity, extremely randomized  Min. node size: (1, 5) | ranger (Meyer et al. 2023; Greenwell et al. 2022; Wickham et al. 2023) |
| Single layer neural network | # Hidden units: (1, # of variables)  Weight decay: (10-7, 10-1) | nnet (Ripley and Venables 2023) |

We employed the following strategies to address the inherent class imbalance in the RECS data set: 1) stratified sampling 2) fewer cross-validation folds 3) class weights 4) sub-sampling 5) appropriate performance metrics. Stratified sampling means that any time we created divisions in the data set such as splitting the training and test data or subdividing the training data into cross-validation folds, we partitioned the data based on occurrence of temperature-related morbidity. This way each subset maintained the same proportion of the dependent variable as the original data. In other words, we did not want any cross-validation folds to end up with no cases of temperature-related morbidity. We also set 5 folds versus the common practice of 10 folds for cross-validation. This allowed us to hold more observations of temperature-related morbidity for the validation set when tuning hyperparameters. Thirdly we test the effect of class weights on model performance. Class weights impose a heavier cost on errors in the minority class. Fourth, we considered the effect of several sub-sampling techniques during cross-validation. Up-sampling randomly replicates instances of the minority class. We also tried two hybrid methods, the synthetic minority oversampling technique (SMOTE) and random oversampling examples (ROSE), which down-sample the majority class and synthesize new data points in the minority class. SMOTE draws artificial samples by choosing points on the line connecting minority class observations to its nearest neighbors in the feature space (Fernandez et al. 2018). ROSE uses smoothed bootstrapping to draw artificial samples from the feature space neighborhood around the minority class (Menardi and Torelli 2014).

Finally, we considered the class imbalance in our choice of performance metric. As illustrated earlier, the model’s overall accuracy (ratio of correct classifications to total observations) can be biased for imbalanced classes. The routine choice for binary classification problems is the Receiver Operating Characteristic (ROC) curve. To understand this metric, we define a positive and negative class – the two outcomes of the predictive model. In our imbalanced data set, the positive class is the minority class and the negative class is the majority class. The ROC curve plots the true positive rate, also called the sensitivity (True Positives / (True Positives + False Negatives)), versus the false positive rate, (False Positives / (False Positives + True Negatives)) or 1 – specificity, the true negative rate , with different discrimination thresholds. The area under the receiver operator curve summarizes the ROC curve into a single metric that represents the prediction accuracy of the model. This metric can be misleading for imbalanced data because the false positive rate becomes very small when the number of negatives is very large. (Davis and Goadrich 2006; Fawcett 2006). The Precision-Recall (PR) curve, on the other hand, plots the precision, defined as the number of correct positive predictions divided by the total number of positive predictions (True Positives / (True Positives + False Positives)) by the recall, which quantifies that number of correct positive predictions by total number of positives (True Positives / (True Positives + False Negatives)) i.e. the same as the true positive rate in the ROC curve.

The PR curve is better suited for imbalanced data sets because it is not concerned with negative class predictions i.e. the majority class. As with the ROC curve, the area under the PR curve summarizes the curve into a single metric, which we use to select the best hyperparameter values during cross-validation. In the test set, we will evaluate the model along three performance metrics: 1) balanced accuracy, 2) recall, and 3) precision. Balanced accuracy is defined as the average accuracy on either class or in other words the arithmetic mean of the sensitivity and specificity. For a naïve model that always predicts the majority class the sensitivity is 0, the specificity is 1, and so the balanced accuracy is 0.5. This serves as a benchmark for good model performance. Recall and precision are of interest because of the high-cost of not only temperature-related health hazards but also preventive measures.

For statistical analysis, we will use a paired t-test by bootstrap iteration i.e. the same training and test data split to compare models trained with different groups of input features. For results with statistical significance, p < 0.05, we will use Cohen’s d to quantify the effect size. We will interpret Cohen’s d as follows: 0.4 < |d| < 1.15 for recommended minimum practical effect, 1.15 < |d| < 2.70 for moderate effect, and |d| > 2.70 for strong effect (Ferguson 2009).

We used the statistical software R (R Core Team 2022) and its associated integrated development environment RStudio (RStudio Team 2021) to build and analyze all machine learning models. In particular, we used the tidyverse package (Wickham and RStudio 2023) for reading, manipulating, and visualizing data and the caret package (Kuhn et al. 2023) as a wrapper to conduct data pre-processing, resampling, and cross-validation as well as interface with the different machine learning algorithms.

1. Results
   1. Prevalence of temperature-related morbidity in population

First, we estimated the prevalence of temperature-related morbidity in U.S. households. Figure 1 compares the estimated number of households affected by heat-related, cold-related, or any temperature-related morbidity in 2015 and 2020. Like the global and national trends discussed in Section 1.1, we find that cold-related hazards were more prevalent than heat-related ones. While overall the number of households with any temperature-related morbidity represents less than 1% of the total population, this still means that nearly 2 million households are impacted annually in the United States. A graph of numbers and a number of households affected by the number of households

Description automatically generated

Figure 1: Prevalence of temperature-related morbidity in U.S. households based on the 2015 and 2020 Residential Energy Consumption Survey conducted by the U.S. Energy Information Administration. We calculated population estimates and standard errors from sample weights and replicate weights as recommended by the EIA (EIA 2019; 2023). Error bars represent the 95% confidence interval.

* 1. Predicting temperature-related morbidity

We next constructed machine learning models to predict temperature-related morbidity. Figure 2a) shows the performance of all model iterations along three performance metrics: balanced accuracy, recall, and precision. Each bar represents machine learning models trained from the same set of input features, class imbalance scheme, and machine learning algorithm, a total of 80 models. The error bars represent the 95% confidence interval, which we calculated from 30 bootstrapped sample iterations, each with a different training and test data split. Generally, about half of the machine learning models performed better than a naïve model. Many of the poor performing models did not converge during model training. For well-performing models, the balanced accuracy and recall range from 70 to 85%. In comparison, the model precision is quite low, around 5%. This means that the models produce many false positives – households that we incorrectly predicted would have temperature-related morbidity.

Figure 2b) compares the best model performance from each input group. For the Climate + Demographics model the best machine learning algorithm was penalized discriminant analysis with SMOTE sub-sampling. For the + Buildings model, the best machine learning algorithm was stochastic gradient boosting with up-sampling. We find a 13% increase in balanced accuracy, 15% increase in recall, and 3% increase in precision. These results are statistically significant with a p < 0.001 and have a moderate to strong effect size.

Figure 2c) compares the magnitude of variable coefficients for the best regression model from each input group. Regression models allow for clearer interpretability of variable contributions, so even though this is not the best performing model for either input features group, its performance is within the 95% confidence interval. For both input feature groups, the best regression model happens to be penalized multinomial regression(Nibbering and Hastie 2022). This model type performs regularization, i.e. aims to reduce the number of input features by forcing coefficients of insignificant variables towards 0. For the Climate + Demographics input features group the best multinomial model used SMOTE sub-sampling and that of the best + Buildings input features group used ROSE sub-sampling. The error bars represent the 95% confidence interval, which we calculated from 30 bootstrapped sample iterations, each with a different training and test data split. We greyed out points where the 95% confidence interval included 0. Variable names that are crossed out were eliminated during pre-processing due to zero or near-zero variance. We find that in the Climate + Demographics model, the variables with the largest magnitude are (in decreasing order): poverty, non-white, renting, employment, and older than 65, and lives alone. For the + Buildings model, the variables with the largest magnitude are (in decreasing order): energy insecurity, non-white, poverty, infiltration, and renting, AC type, employment, and thermally massive roof. What is most striking is the high magnitude of the coefficient for energy insecurity, in comparison to that of any other variable. When comparing the input groups, we see that almost the model selects almost the same demographics variables, however the magnitude of the coefficient is higher for the same variable in the Climate + Demographics model.

* 1. Machine learning modeling

Finally, we review the impact of two key machine learning modeling parameters: the algorithm and the class imbalance handling scheme along two performance metrics: balanced accuracy as a proxy for model performance and computation time. Figure 3 plots the interaction between the two machine learning modeling parameters and both performance metrics.

A screenshot of a graph

Description automatically generated

Figure 2: a) Overall machine learning model performance across all iterations along three metrics: balanced accuracy, recall, and precision. Each bar represents a machine learning model trained with the same input features, class imbalance handling scheme, and algorithm. The error bars represent the 95% confidence interval calculated from 30 bootstrapped samples, each with a different training and test data split. b) Shows the performance for the best machine learning model from each input features group. We calculated statistical significance using a paired t-test by bootstrap iteration i.e. the same training and test data split, and the effect size from Cohen’s d. We interpreted Cohen’s d as follows: 0.4 < |d| < 1.15 for recommended minimum practical effect, 1.15 < |d| < 2.70 for moderate effect, and |d| > 2.70 for strong effect (Ferguson 2009). c) Shows the variable contribution from the best regression

A chart of different colors

Description automatically generated with medium confidence

Figure 3: Interaction between two machine learning modeling parameters: the algorithm (on y-axis) and the class imbalance handling scheme (on x-axis). Results are shown for one input features group (Climate + Demographics + Buildings), but the trends are applicable to all groups. We binned the computation time for clarity. Computation time is the time it takes for model training included cross-validation and is binned as follows for clarity: fast (< 30 s), medium (30 s to 2 min) and slow (> 2 min).

The color of the point represents balanced accuracy, and the shape of the point represents the computation time binned into three groups, fast (< 20 s), medium (20 s to 1.5 min) and slow (> 1.5 min). From the plot, we can see that across all machine learning algorithms, no class imbalance handling scheme and class-weights had balanced accuracies near 50%, like a naïve model. We can also see that two machine learning algorithms, classification and regression trees and random forest had balanced accuracies near 50%, like a naïve model, across all class imbalance handling schemes except ROSE. Some machine learning algorithms like multivariate adaptive regression spline and random forest had consistently slow computation times across all class imbalance handling schemes. In some machine learning algorithms like penalized discriminant analysis and neural network, we find that up-sampling and SMOTE increase computation time relative to the other class imbalance handling schemes. Generalized linear models have the fastest computation times and models with up-sampling and SMOTE have balanced accuracies near the maximum of 85%. Generalized linear models with ROSE have balanced accuracy significantly worse than a naïve model.

1. Discussion

RECS excludes vacant, seasonal or vacation homes, and group quarters such as prisons, military barracks, dormitories, and nursing homes.

# Conclusions

# CRediT authorship contribution statement

**Arfa Aijazi:** Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation Writing–Original Draft, Visualization. **Stefano Schiavon**: Supervision, Validation, Writing–Review and Editing. **Duncan Callaway:** Methodology, Writing–Review and Editing

# Declaration of competing interest

The Center for the Built Environment at the University of California, Berkeley – with which the authors are affiliated, is advised by and funded in part by many partners that represent a diversity of organizations from the building industry – including manufacturers, building owners, facility managers, contractors, architects, engineers, government agencies, and utilities.

# Acknowledgements

The Center for the Built Environment at the University of California, Berkeley funded this study.

# Appendix

# Reference

1. The discourse in public agencies and academic literature around thermal vulnerability focuses on extreme heat, even though the mortality rate from extreme cold is nearly double that of extreme heat (Berko et al. 2014). In principle, many of the socioeconomic vulnerabilities contributing to heat-related morbidity and mortality also apply to extreme cold.  [↑](#footnote-ref-1)
2. W. McNary (personal communication, July 14, 2021) [↑](#footnote-ref-2)
3. C. A. Hronis (personal communication, August 15, 2022) [↑](#footnote-ref-3)