Methods for Estimating Residential Evaporative Cooler Water Consumption and Prevalence using Account-Level Water and Energy Consumption Data

Prepared by: UC Davis Center for Water-Energy Efficiency
215 Sage Street Suite 200, Davis, CA 95616

Prepared for: California Department of Water Resources
Report Number: WUES-DWR-2021-05.T1

Submitted on: 01-06-2022

Amended on: 04-15-2022

Contents

EXECUTIVE SUMMARY

The California Department of Water Resources will be delivering recommendations on water use efficiency standards, variances, and performance measures to the State Water Resources Control Board for adoption under the legislative requirements of Senate Bill 606 and Assembly Bill 1668 of 2018. Variances are significant water uses that should be reasonably accounted for but not included in the considerations of water use efficiency standards. One such use case is a high prevalence of residential evaporative coolers (REC) in certain regions. The aim of this study was to develop data-driven methods from readily accessible data for calculating a single-family residential REC variance with minimal data requirements and easy implementations. However, the data available was proven inadequate to develop an approach based on use data and therefore, an engineering-based approach is recommended.

Two pieces of information are fundamental to calculating a variance for a given supply region: (1) the typical cooling-day water usage of an evaporative cooling unit in that region; and (2) the number of evaporative cooling units in that region. While this information could be gathered from resident surveys and unit-based metering, it would be resource intensive to do so.

This study investigated the possibility of using account-level water and energy data, publicly available weather and tax-assessor data, and statistical modeling to estimate the quantities needed to calculate an REC variance. However, none of the models yielded results consistent with other field and laboratory studies. The cause of the disagreement proved to be unreliable data: the data describing home cooling system, found in the tax assessor dataset, vastly under-counted the number of REC's in the region. In fact, regional specialists suggested that it would be rare for a home in the region not to have an REC. With no reliable sub-sample of homes without REC's to use as a comparison, it was not possible to derive an reliable estimate of typical daily REC water use.

INTRODUCTION

In 2018, Assembly Bill 1668 and Senate Bill 606 laid out a new framework for water conservation and drought planning in California. The California Department of Water Resources (DWR) and the State Water Resources Control Board have been collaborating with stakeholders across a range of use cases in the residential, commercial, industrial, and institutional sectors to establish new water use efficiency standards within this framework. These standards will be used to calculate annual water budgets for urban retail water suppliers. In addition to standards, the DWR will be recommending variances for certain unique uses that may materially impact an urban retail urban supplier's (URWS) ability to meet its budget limitations. One such unique use is a significant water use of residential evaporative coolers (REC) in an URWS service area.

DWR contracted the UC Davis Center for Water-Energy Efficiency (CWEE) to develop methods to (1) measure the water consumption of evaporative coolers in single family residences (SFR) and (2) estimate the prevalence of residences equipped with an REC unit, both under a variety of data-availability scenarios. For each scenario, methods are presented assuming a specific set of data availability constraints. DWR intended to use the methods corresponding to a URWS's data assets to help them calculate the appropriate water use volume for this specific use as a variance.

BACKGROUND AND LITERATURE REVIEW

This section provides background on evaporative cooling systems, their range of use, and their water and energy consumption.

EVAPORATIVE COOLERS

Evaporative cooling is the process by which thermal energy transfers from hot dry air to liquid water, causing some of that water to vaporize and create cool, moist air. REC's use this process to cool homes; however, there are variations in the technology used to do so. Direct and indirect cooling are two basic types of REC systems. Most REC's use direct technology (Spartz et al. 2004). Both types operate by passing hot, dry outdoor air through a sheet of moistened evaporative media, creating cool, moist supply air. Direct REC's circulate the moist, evaporatively cooled air throughout the residence with no intervening steps. Indirect REC's use the evaporatively cooled air to cool dry, outdoor air via a heat exchanger and then circulate the cooled, dry air throughout the residence (Spartz et al. 2004; Sahai et al. 2012).

In addition to variation in evaporative cooling technology, REC's also vary in size (denoted by cubic feet per minute or CFM). Online retailers recommend purchasing an REC with CFM value equal to ½ the cubic footage of the residence. That is, for a 1,000 square foot residence with 8-foot ceilings, a 4,000 CFM REC unit is recommended. At online retailers such as homedepot.com and lowes.com, CFM values range from less than 100 for single-room hand-refilled units, to 1,000-10,000 for window-mounted multi-room units, central-air roof-mounted units, and stand-alone units. Under fixed temperature and usage conditions, every cubic foot of evaporatively cooled air requires the same water consumption for cooling. Thus, under fixed conditions, the size of an REC is highly correlated with its total water consumption.

REC's can reduce indoor dry-bulb temperature to within a few degrees of outdoor wet-bulb temperature (Sahai et al. 2012). The actual temperature reduction is approximated by an evaporative efficiency value, usually between .8 and .95 (ASHRAE 2012). An REC with evaporative efficiency of .8 will, ideally, cool indoor air by 80% of the difference between outdoor dry-bulb and wet-bulb temperatures. Some REC systems, such as the Maisotsenko Cycle, can reduce supply air dry-bulb temperature below outdoor wet-bulb temperature (Sahai et al. 2012). However, these models are less common, and in general the outdoor wet-bulb temperature is a lower bound on the temperature-reduction potential of an REC.

EXPECTED RANGE OF USE

Due to the wet-bulb lower bound, the cooling potential of an REC depends on the regional climate. Figure 1 demonstrates the relationship between Climate Zone and average number of hours per day with ideal REC conditions (wet-bulb temperature less than 68 degree Fahrenheit (°F) and dry-bulb temperature greater than 78.8°F (Sahai et al. 2012.)) using weather data from summer 2017.

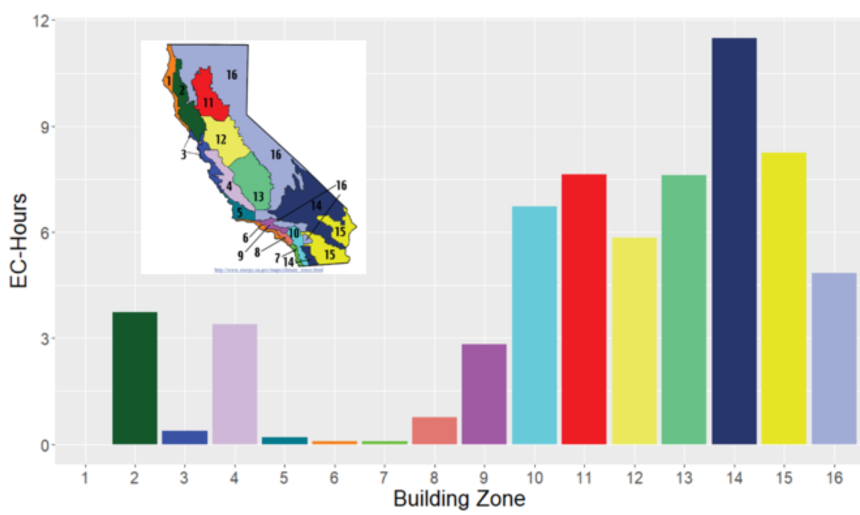


Figure 1. Average Daily Summer Hours Appropriate to Evaporative Cooling in the 16 California Building Climate Zones.

The plot gives no information about *actual* REC prevalence. Rather, it indicates that suppliers with significant REC use are mostly likely found in the Central Valley and in inland regions south and east of the Sierra Nevada.

WATER CONSUMPTION

Previous studies have quantified SFR- and unit-scale REC water use both experimentally and in-situ. Under experimental conditions, a mixed design indirect-direct REC was found to use between 15.9 and 70.3 liters/hour (WCEC et al. 2020). In-situ REC systems in central and southern California were found to use between 7.6 to 90.7 liters/hour during the cooling season (Spartz et al. 2004). REC systems in Phoenix were found to use on average 250 liters/day, or 12.4 to 14.4 liters/hour during the cooling season (Karpiscak et al. 1998). A model-based calculation of hourly cooling loads estimated REC water consumption to be between 300 and 500 liters per SFR per day (Sahai et al. 2012). Most recently, a 2020 collaboration between two URWS's (A and C) in Climate Zone 14 found an average daily REC water usage of 196.8 liters (52 gallons) per unit per day during the cooling season (see Supplemental Dataset in Appendix).

In addition to the water consumed by the process of evaporative cooling, some REC's require water for a maintenance process of flushing the system to remove mineral build up that has accrued with use. This process is known as bleed-off and it has been found to increase water consumption by 10 to 50 percent (Karpiscak et al. 1998).

FIRST PRINCIPLES WATER CONSUMPTION CALCULATION

An estimate for REC water consumption can be made using engineering calculations based on first principles of evaporative cooling (see Appendix for details). Assuming a cooling season of June through August, all-day cooling, and an evaporative efficiency of 80%, 2020 hourly weather data from Building Climate Zone 14 (see Figure 1) was used to estimate REC water consumption for various CFM values (Table 1).

Table 1. 2020 weather data and engineering principles of evaporative cooling were used to estimate the average liters per day consumed by REC systems of various sizes during the cooling season.

Unit Size (Cubic Feet per Minute; CFM)	Average Water Use (Liters Per Day)
500	96.3
1,000	192.6
2,000	385.2
3,000	577.8
4,000	770.4

ENERGY CONSUMPTION

REC systems are generally more energy efficient than comparable air conditioning (AC) systems. A common measure of energy efficiency is the coefficient of performance (COP), or the ratio of useful heating/cooling provided to energy required. Higher COP values indicate more useful heating/cooling per unit of energy input. Under experimental conditions, the COP of REC systems was between 4.9 and 23.3, whereas the COP of AC systems (some of which incorporated REC pre-cooling) was between 2.3 and 4.9 (Pistochini and Modera 2011).

Lab and field test have been used to determine specific energy consumption of REC systems over a range of conditions (WCEC et. al. 2020, Spartz et. al. 2004). However, without a building model and simulation of heating/cooling load, those studies do not easily translate into quantities on the scale of home and cooling season. Field monitoring of homes in Phoenix indicates that during the cooling season a 1,600 square-foot home with a central-air direct REC will use around 1,500 kWh, compared to 6,000 kWh of consumption in an equivalent home with AC (Karpiscak and Marion 2016).

METHODS

CWEE collaborated with two URWS's, referred to here as URWS A and URWS B and with the investor-owned energy utility, referred to as IOU A, that provided energy for the URWS A service region. Both URWS A and B were located in San Bernardino County (Building Climate Zone 14, see Figure 1) and were thus well suited to studying REC use.

CWEE used the data provided to create four data-availability scenarios, each modeling a different combination of data resources available to a supplier and the DWR. This section describes the datasets; broadly presents the regression and classification methods used; and defines each data scenario and how each method applies. For the technical details of the methods, see the Appendix.

DATASETS USED

Several datasets were used for the study and are summarized below.

Water Consumption Data

URWS A and B provided account-level, monthly water consumption data from 2015 through 2020. The original datasets included records for 4,561 (URWS A) and 5,916 (URWS B) accounts identified by address and assessor parcel number (APN). The datasets went through pre-processing before use, including geocoding and removing extreme data points. Any records with water consumption values equal to or exceeding the 99th percentile of overall monthly consumption were not included in the study as they were assumed to indicate some abnormal condition such as major pipe failure or erroneous recording. The final number of accounts (2,912 for URWS A and 4,057 for URWS B) is a result of these procedures.

Energy Consumption Data

IOU A provided account-level, hourly energy consumption data from 09/2017 – 09/2020 for all SFR's in the URWS A region. Not all account information provided was complete: some were missing hours and even days of records. Overall, this was rare, but an address record was only used when at most one day was missing from the specific service period, and when all hours of each day were accounted for in the record.

Property Data

CWEE obtained public property data for all San Bernardino County SFR's from the San Bernardino County's tax assessor website. This dataset included home size (in square feet), home vintage, lot size, number of rooms and bedrooms, and many other variables that affect property values. Most importantly, the dataset also included an indicator for the cooling system associated with the home, be it AC, REC, or some alternative or combination. The data distinguish between central and non-central cooling systems, with non-central being by far the majority. Table 2 presents the number of homes with each cooling system in each participating supplier.

Table 2. Summary counts of the cooling systems found in the San Bernardino County tax assessor property characteristic dataset for each data-provider.

Supplier	Accounts Geocoded	Number of AC, Non-Central	Number of AC, Central	Number of REC, Non-Central	Number of REC, Central
URWS A	2,912	21	1,095	1631	60
URWS B	4,057	20	3,186	831	20

There was no way to validate the accuracy of these data easily. Tax assessor's data are often (but not always) updated when a property changes ownership; when new construction is finished; and when valuation reviews and appeals are made. These data are assumed to be correct for the purposes of the regression and classification methods to be discussed below. However, there is good reason to believe these data are ultimately unreliable and require that the methods be interpreted differently if they are to be believed at all. This, too, will be discussed below.

Weather Data

Hourly weather data was acquired from DWR's California Irrigation Management Information System (CIMIS) weather stations closest to the addresses in each district. These data included dry-bulb air temperature, relative humidity, dew point, and vapor pressure. When making psychrometric calculations, atmospheric pressure was estimated using the elevation of the relevant location. Missing data were imputed by fitting a loess smoothing function of time between the previous and following non-missing values at the same location.

COMPUTATIONAL METHODS

Given the range of data resources available to URWS's across the state, the methods used for variance implementation needed to be applicable under a variety of data scenarios. In particular, an URWS may or may not have access to energy data or detailed information about which residences in their region have REC units.

In the following sections the regression and classification methods are described, with technical implementation details available in the Appendix.

Regression Method

The regression method was developed to compare the water consumption of SFR homes cooled exclusively with REC systems to water consumption in homes cooled exclusively with AC systems, and thereby distinguish the water consumption due to the REC systems. This method to address the study goal to the water consumption of REC systems for SFR in a service region.

The approach is based on the assumption that after controlling for the physical characteristics of a home (e.g., the square footage, lot size, vintage, number of bedrooms), water consumption will be higher on hot days in a home with an REC compared to a home with an AC. Using this assumption, the method applied an ordinary least squares model to estimate how much more rapidly water consumption increases with the outdoor temperature in homes with REC systems compared to homes with AC systems. The model formula is as follows:

$$Cons_{ij} = \alpha_i + (\beta_1 \cdot c_i + \boldsymbol{\gamma'\theta_i}) \cdot CD_j + (\beta_2 \cdot c_i + \boldsymbol{\rho'\theta_i}) \cdot HD_j + \varepsilon_{ij}$$

Where $Cons_{ij}$ is the average daily water consumption (in liters) at household i during billing period j, α_i is a premise-level fixed effect, c_i is an indicator variable that equals 1 if the premise has REC system and 0 if the premise has an AC system (regardless if that system is central or non-central), CD_j and HD_j represent the average daily cooling-degrees and heating degrees across the days included in the billing period, and θ_i is a vector of four binned premise-level characteristics (home size, home vintage, number of bedrooms, and home quality). The key coefficient of interest is β_1. If β_1 is positive, that implies that, after controlling for the observable premise-level characteristics, water consumption increases more on hot days in homes with REC systems relative to homes with AC systems.

To account for the possibility that the unobserved errors are correlated over time within observation from a single premise, as well as across premises during a given time-period, standard errors that are robust to heteroskedasticity and two-way cluster by premise and month-of-year were calculated.

Classification Method

The above regression method allows a supplier to estimate the expected water consumption of a typical REC system in its service region. However, to apply this method a supplier must first have a sample of residences in its service region containing both homes with REC systems and homes with AC systems. Furthermore, to use the water-use estimate to calculate the total water use volume, the supplier must also have an estimate of the total number of residences in its region with REC systems. Thus, the other goal of this study was to provide a method to identify and estimate the prevalence of homes in an urban retail water supplier's service area with REC systems.

The recommended method uses a random forest classifier (RF). RF's involve simple tuning and variable-selection procedures and are well suited to problems with non-linear relationships between response and predictors. RF's are also easily interpreted and provide a clear indication of the most useful predictors. The two major choices that go into tuning an RF are which predictors to provide the algorithm and *M,* the number of variables chosen at random at each split from the all the variables provided. Implementations for RF's can be found in R, Python, and other statistical and mathematical programming packages.

The predictors used in the RF depended on the data available in the given data scenario. In both data scenarios, the following variables were used in the final model: vintage; residence size; residence quality; number of bedrooms; mean daily water consumption during the summer, winter, and fall/spring; a measure of the response of water consumption to cooling-day dry-bulb temperature; and a measure of the response of water consumption to cooling-day wet-bulb temperature.

The RF model was tested in two ways: (1) a test to measure the RF model's ability to estimate URWS A regional REC prevalence using URWS A sub-sample training data and URWS A testing data; and (2) a test of the RF model's ability to identify SFR's with REC units outside of the region on which it was trained using URWS B training data and URWS A testing data. In (1), the RF predictions were compared to logistic regression (logit), another common off-the-shelf algorithm, demonstrating the reasons for recommending use of RF.

DATA SCENARIOS

Data scenarios were used to illustrate how and when to use the regression and classification methods depending on the different combinations of data resources a URWS may have at its disposal.

All data scenarios assumed that a URWS has access to its SFR customers' monthly water consumption data and tax assessor records. Each scenario then adds a different combination of the two additional data resources: whether energy consumption data is or isn't available and whether a sample of homes with REC and AC systems is or is not known.

In scenarios 1 and 2 the URWS is assumed to have a sample of SFR's in each cooling system category, AC and REC, and that the variable describing the category of cooling system is known for each SFR in the sample. The specific data used in each scenario is described in the following sections. In both scenarios the regression method is used to estimate REC water consumption and the classification method is used to estimate REC regional prevalence. The results were sensitive to sample size, thus sample-size sensitivity tests were conducted using four sample sizes: where there are 50, 100, 300, and 500 SFR's in each cooling system category.

In scenarios 3 and 4 the URWS is assumed to have no information regarding what homes in its supply region have AC or REC systems. Thus, these scenarios Illustrate the use of the classification method to impute a cooling system category variable when it is unknown, and demonstrate the efficacy of the regression method when used with the imputed variable. CWEE was unable to obtain a non-URWS A training set with energy data, and thus no results are reported for scenario 3. In scenario 4, a supplemental dataset obtained from URWS B was used to train an RF classifier and the trained classifier was then used to predict an imputed cooling system category variable for URWS A. The imputed cooling system category variable was then used in the regression method and the results based on the imputed cooling system variable were compared to results from scenario 2. More details on the specific variables used in scenario 4 are given in the following sections.

These scenarios are presented in Table 3 and are further described in the following sections.

Table 3. Data availability scenarios. The same weather, water, and property data were used in each scenario. Only energy data and address-specific AC and REC data differed between the data scenarios.

Scenario	Daily Weather Data	Monthly Water Data	Property Characteristics	Hourly Energy Data	Sample SFR's with REC & AC
1	YES	YES	YES	YES	YES
2	YES	YES	YES	NO	YES
3	YES	YES	YES	YES	NO
4	YES	YES	YES	NO	NO

Scenario 1

In this scenario, in addition to monthly water consumption data, daily weather data, and property information, the supplier is assumed to have access to account level hourly energy consumption data and have a sample of SFR's in its service region known to have REC systems and a sample of SFR's known to have AC systems.

The energy data was used to estimate how many days during a service period an SFR was occupied. This number was then used to adjust the regression method response variable (mean daily water consumption per month) by reducing the denominator in the mean calculation from total number of days in that month to total number of occupied days in that month. The estimate itself was made by counting all days with zero variation in hourly energy consumption as unoccupied. There are more advanced means of making this estimate (Kleminger et al. 2013), but they required energy data at much higher frequency than were available.

The regression analysis for determining REC water use was carried out using the adjusted measure of daily water consumption as the response variable. No other changes to the predictors or model structure were made.

The classification method for determining REC regional prevalence was trained and tested using address-specific predictors. Home vintage, home size, number of bedrooms, and home quality were taken from the URWS A Tax Assessor dataset. The mean summer, winter, and spring/fall wet- and dry-bulb temperatures and measures of the response of water consumption to cooling-day dry- and wet-bulb temperatures were derived using regional weather data and address-specific monthly water-use data. Measures of the response of energy consumption to cooling-day dry- and wet-bulb temperatures and average summer energy usage, broken down by hour of the day were derived using regional weather data and address-specific hourly energy-use data. See the Appendix for more details on how to construct predictors for the classification method.

Scenario 2

In this scenario the supplier is assumed to have access to account-level, monthly water consumption data, property characteristic information, and have a sample of SFR's in its service region known to have REC systems and a sample of SFR's known to have AC systems. The response variable (mean daily water consumption per month) is not corrected for occupancy, as energy data is assumed to be unknown.

The regression analysis was carried out using the mean daily water consumption per month as the response variable with no changes to the predictors or model structure.

The classification method was trained and tested using all predictors from scenario 1 excluding those derived from address-specific hourly energy data. See the Appendix for more details on how to construct these predictors.

Scenario 3

This scenario required training data from a suitable URWS supply region with both water and energy data. CWEE was unable to find a URWS with the required data resources that was willing to contribute their data. Thus, this method went un-tested.

Scenario 4

In this scenario, the classification method was used both to impute a home cooling system category variable and to determine REC regional prevalence. Using the imputed home cooling system category variable, the regression method was then used to estimate typical REC water usage and the result was compared to the regression result obtained in scenario 2.

The classification algorithm was trained using data from URWS B under scenario 2 and then used to predict the cooling system categories of SFR's in the URWS A dataset. The training set contained 4,277 samples. Each sample represented a single home in the URWS B region and contained all the required variables described in scenario 2. 3,428 of the URWS B training samples had AC systems and 849 had REC systems. Training sets were drawn at random in various sample sizes so that homes with REC systems and homes with AC systems were equally represented. 100 training and testing iterations were run for each sample size and the specificity, sensitivity, and detection prevalence were all recorded.

The accuracy of the regression using the imputed cooling system category variable was demonstrated using a similar random sampling procedure. 100 iterations were run and in each iteration the classification method was trained using a random sample of 500 homes in each of the cooling system categories from URWS B and used to classify the cooling system category variable in URWS A. The imputed cooling system category variable was then used to carry out the regression analysis on URWS A scenario 2 data. This process yielded 100 regression coefficient estimates for the typical REC water use in URWS A, which were then compared with the result obtained in scenario 2.

RESULTS

SCENARIO 1

REC systems for individual SFR homes were estimated to contribute 3.59 liters/day/cooling-degree with a standard error of 1.23 (Figure 2). As expected, the REC contribution to water use when temperatures were in heating range (β_2) were not statistically different from 0. The same analysis was run using smaller, random samples of homes from the overall dataset to demonstrate variance of the estimate with sample size. The mean square error from the full-dataset estimates for each sample-size were: 6.19 for 50 homes, 3.72 for 100 homes, 1.75 for 300 homes, and 1.31 for 500 homes (Figure 3).

Figure 2. Estimates of REC cooling-day (β_1) and heating-day (β_2) coefficients with 95% confidence intervals. Estimates are made using energy data as a proxy for home occupancy during variable construction.

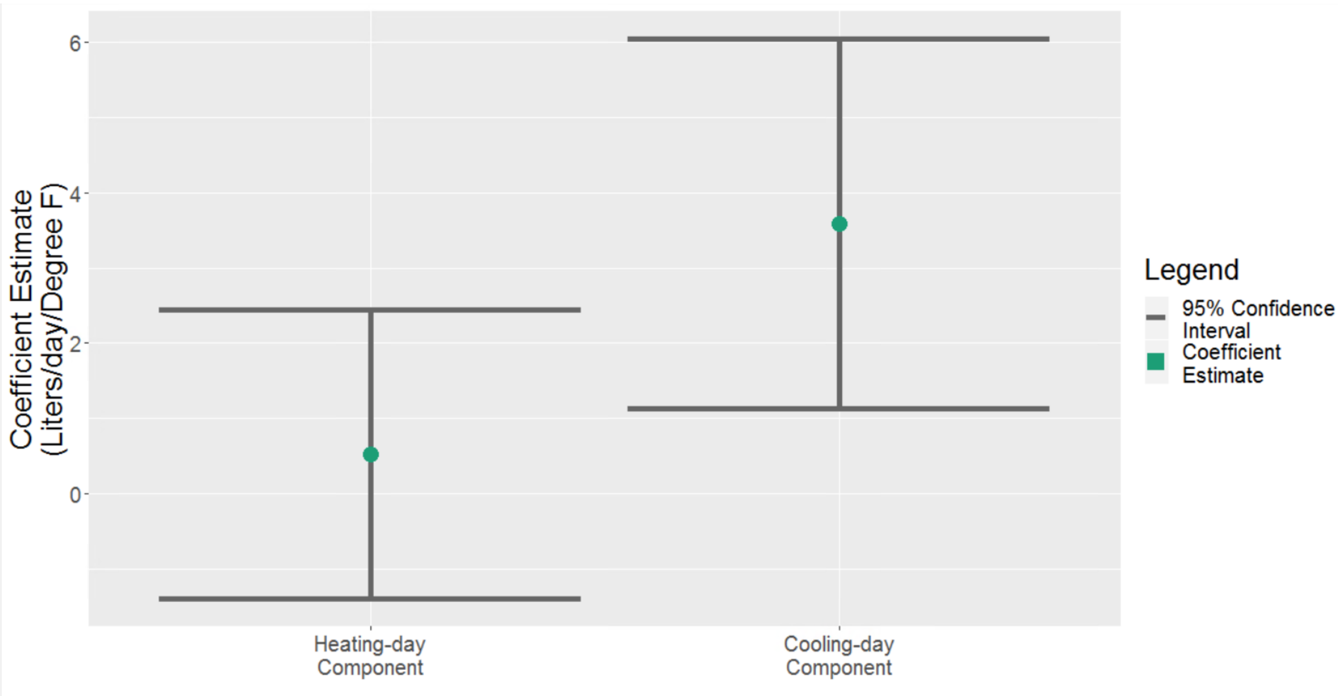


Figure 3. Distributions of errors between sub-sample estimates of β_1 and the overall sample β_1 estimate for scenario 1. A supplier with a smaller sample size of residences will have larger estimate standard errors. Bar and whisker plots show median value, interquartile range (IQR), and 1.5 times the IQR.

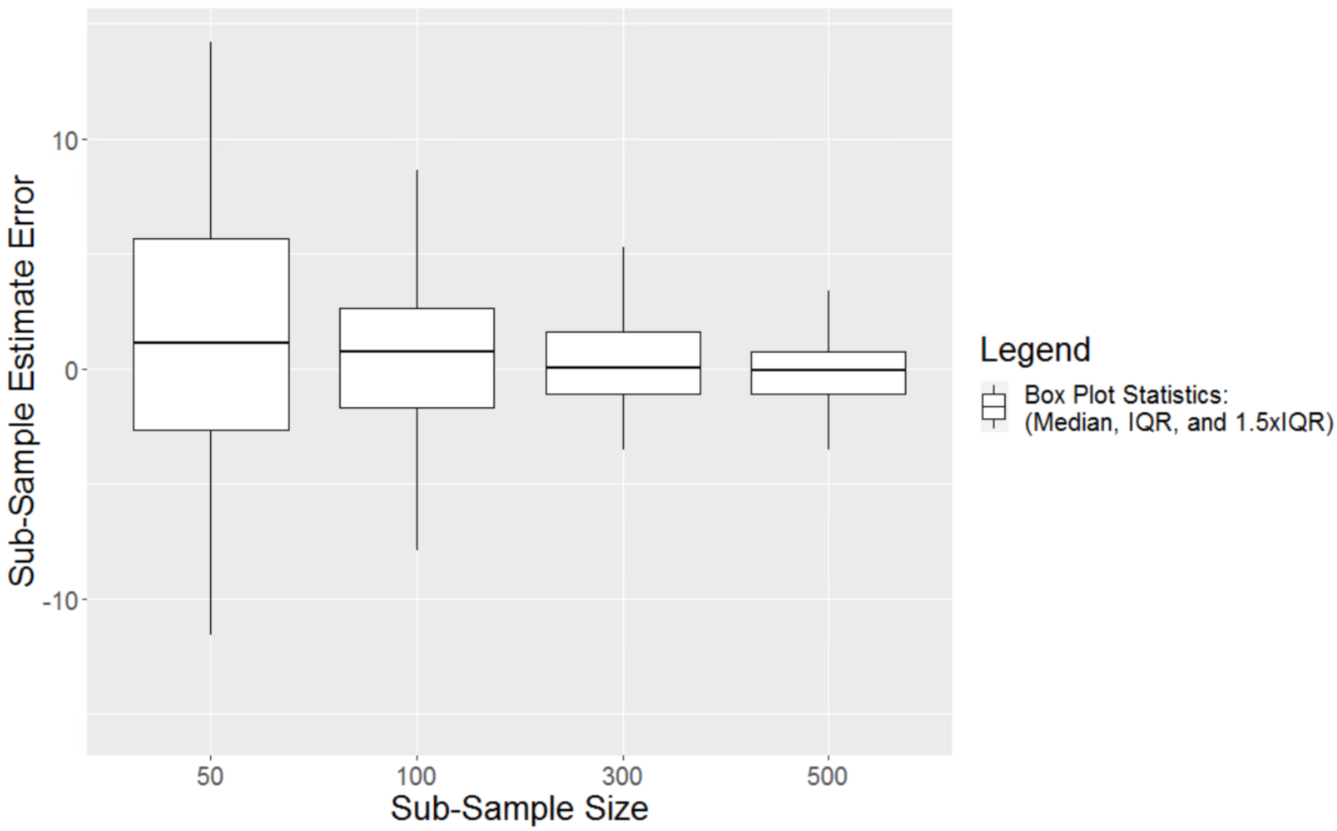


Figure 4 compares the REC prediction sensitivity of the two classification algorithms, logit and RF, trained on all scenario 1 predictors. With careful tuning, other algorithms may outperform the RF. However, RF's have simple tuning and variable selection procedures, and in this case, outperform logistic regression off-the-shelf, making them a competitive method for this use case.

Figure 5 presents the results of the RF prevalence estimate, again relative to the REC and AC labels from the tax assessor dataset. The median prevalence error with 5th and 95th quantiles for each sample size were: 50 SFR's in each category: -5.76% (-11.82%, 0.37%); 100 SFRS's in each: -4.00% (-9.96%, .13%); 300 SFR's in each: -4.84% (-7.31%, -2.23%); and 500 SFR's in each: -5.92% (-8.48%, -3 .60%). The main benefit of an increased sample size was, as expected, decreased variance in the estimate errors.

Figure 4. Distributions of classification sensitivity for different training-set sizes (50, 100, 300, and 500 in each category AC and REC) and for two classification methods, logit and random forest. Bar and whisker plots show median value, interquartile range (IQR), and 1.5 times the IQR.

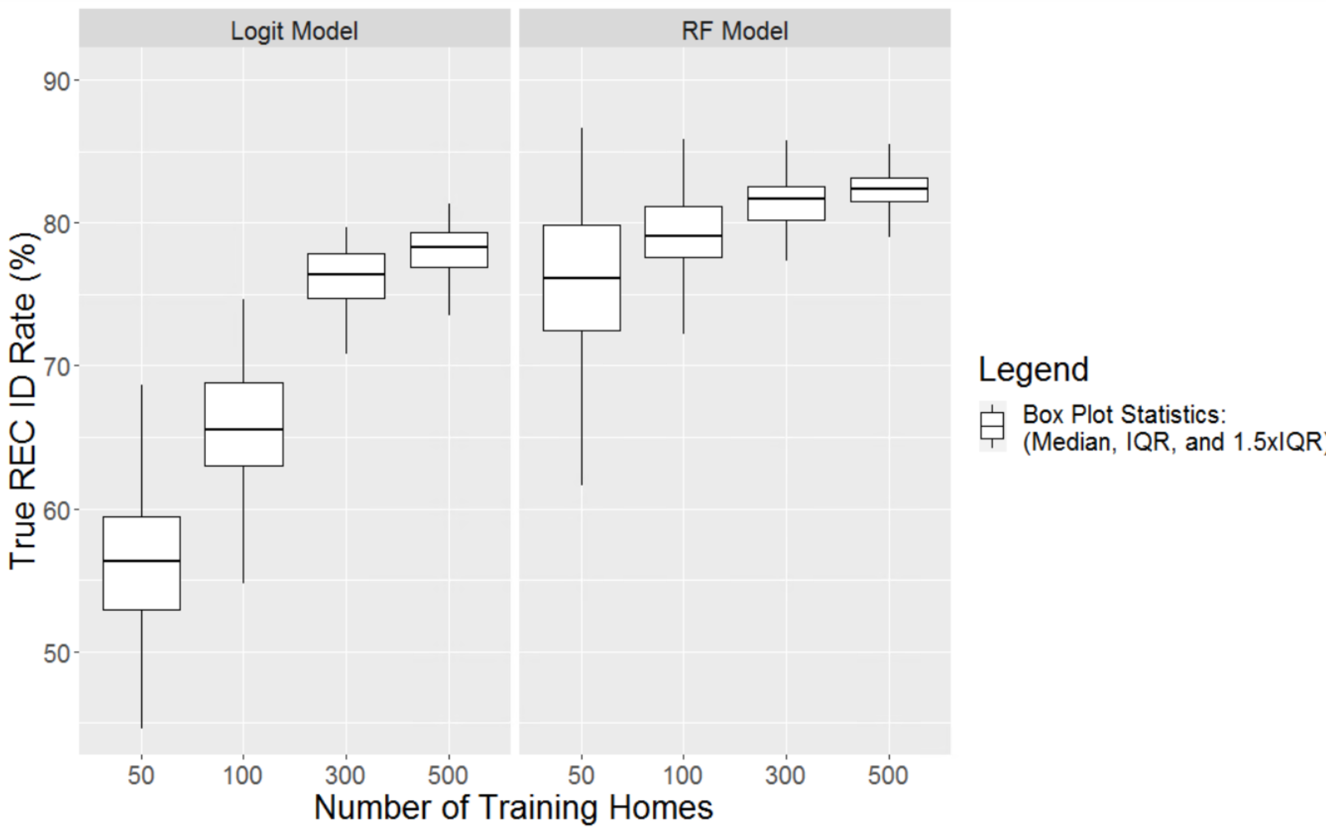


Figure 5. Distributions of classification prevalence estimates for different training-set sizes (50, 100, 300, and 500 in each category AC and REC) using the random forest model. Bar and whisker plots show median value, interquartile range (IQR), and 1.5 times the IQR.

SCENARIO 2

In this scenario, REC systems were estimated to contribute 2.65 liters/day/cooling-degree with a standard error of .92 (Figure 6). As in the previous scenario, the REC contribution when temperatures were in heating range were statistically equivalent to 0. Again, the same analysis was run using smaller, random samples of homes from the overall dataset to demonstrate variance of the estimate changes with sample size. The mean square error from the overall estimates for each sample-size were: 5.72 for 50 homes, 3.03 for 100 homes, 1.54 for 300 homes, and .91 for 500 homes (Figure 7).

Figure 6. Estimates of REC cooling-day (β_1) and heating-day (β_2) coefficients with 95% confidence intervals. Estimates are made without using energy data as a proxy for home occupancy during variable construction.

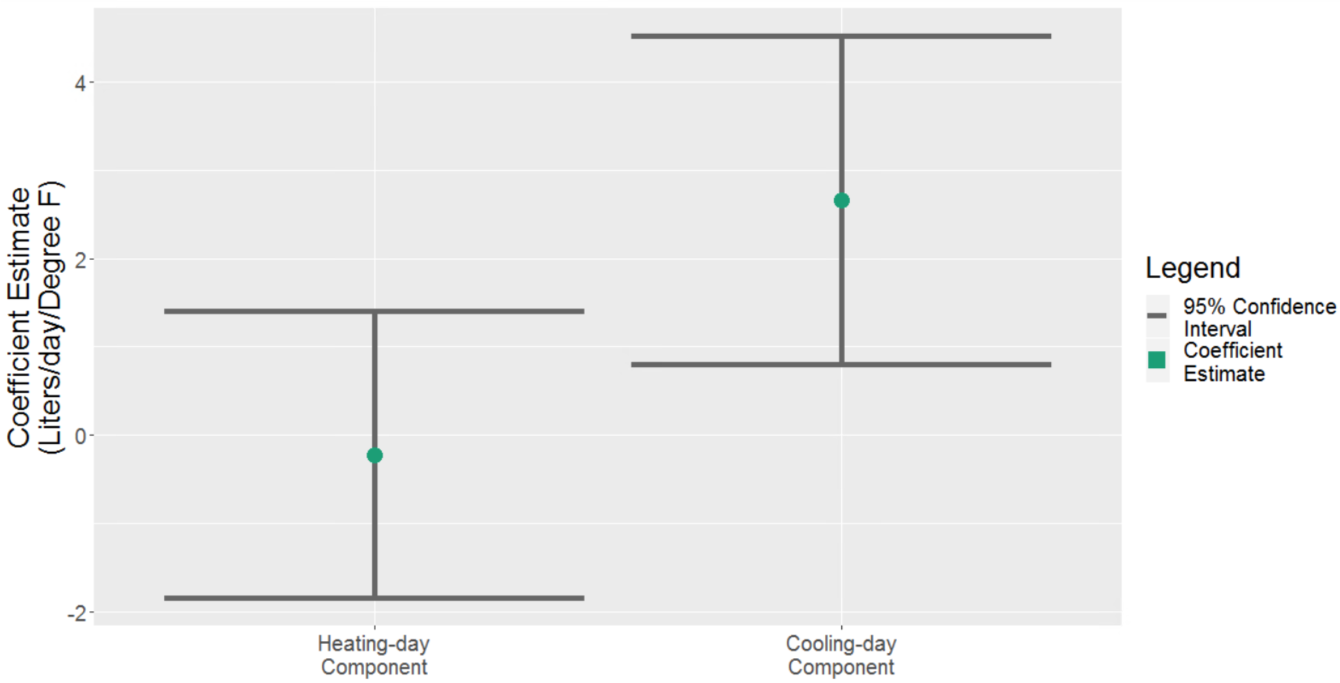


Figure 7. Distributions of errors between sub-sample estimates the overall sample estimate. A supplier with a smaller sample size of residences will have larger estimate standard errors. Bar and whisker plots show median value, interquartile range (IQR), and 1.5 times the IQR.

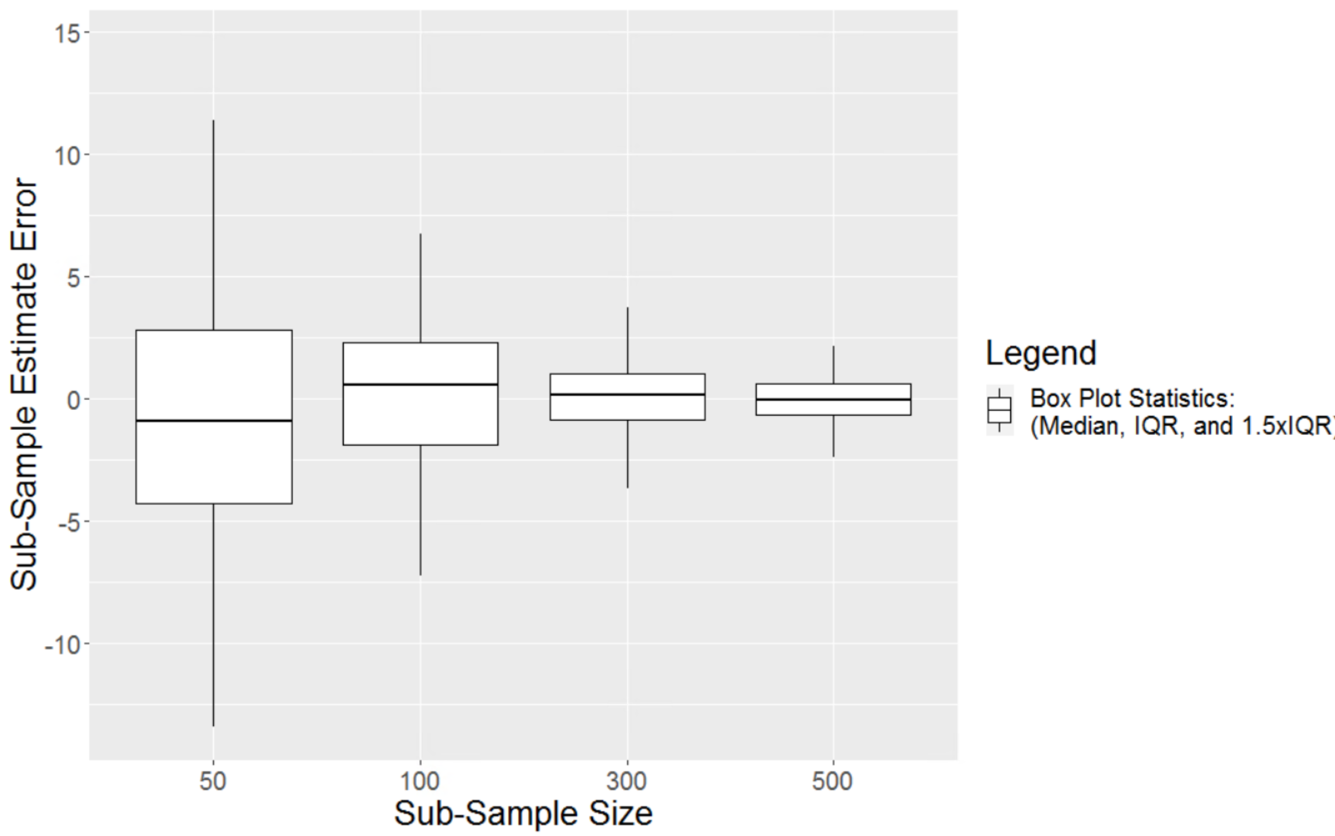


Figure 8 is equivalent to Figure 4 in scenario 1. In this case no variables based on energy data were used to train the algorithms, but the results are overall comparable.

The median prevalence error (Figure 9) with 5th and 95th quantiles for each sample size were: 50 SFR's in each category: -3.32% (-9.97%, 2.46%); 100 SFRS's in each: -3.10% (-8.55%, 1.37%); 300 SFR's in each: -3.44% (-6.61%, -.68%); and 500 SFR's in each: -4.05% (-6.30%, -1.45%). Again, the main benefit of an increased sample size was increased consistency in the estimated error.

Figure 8. Distributions of classification sensitivity for different training-set sizes (50, 100, 300, and 500 in each category AC and REC) and for two classification methods, logit and random forest, trained using all non-energy-based predictors. Bar and whisker plots show median value, interquartile range (IQR), and 1.5 times the IQR.

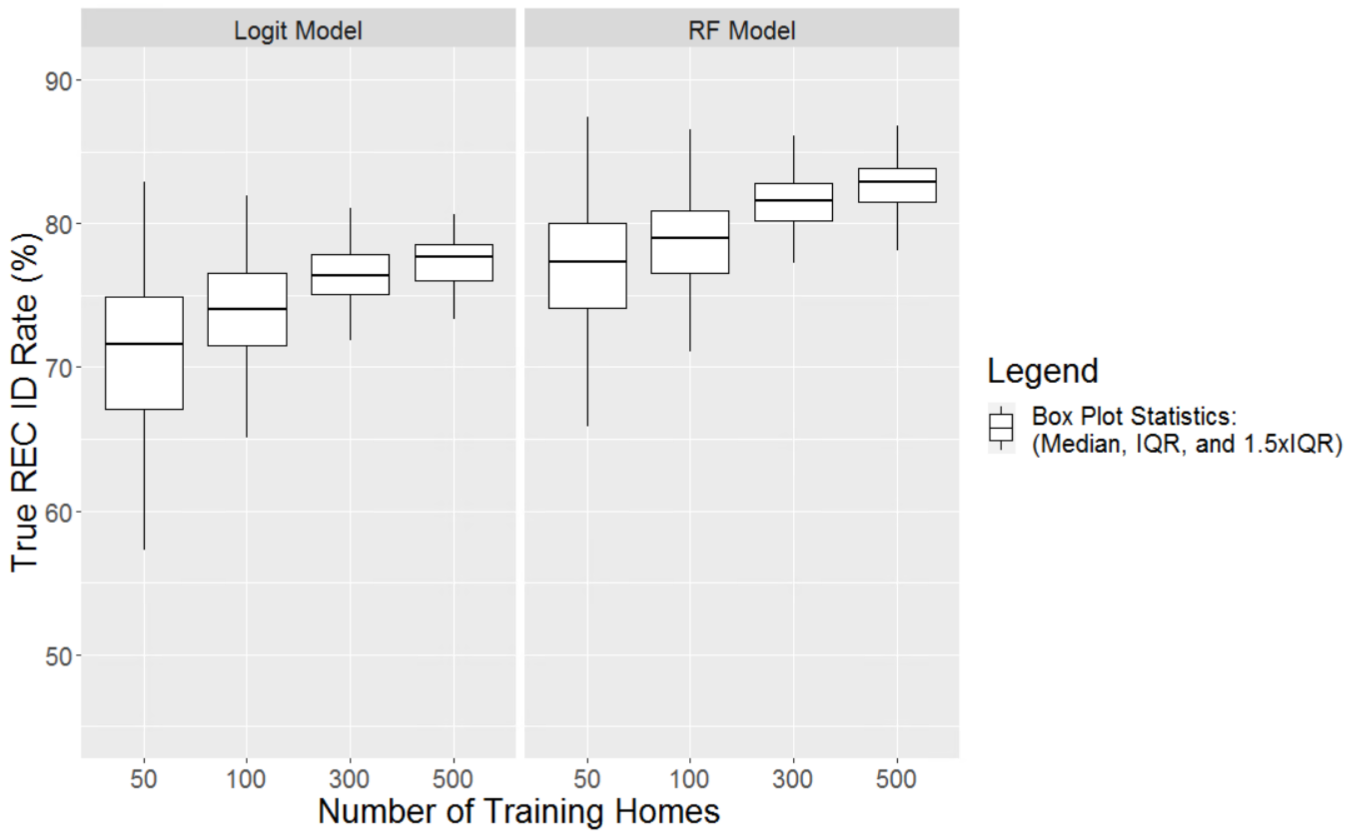
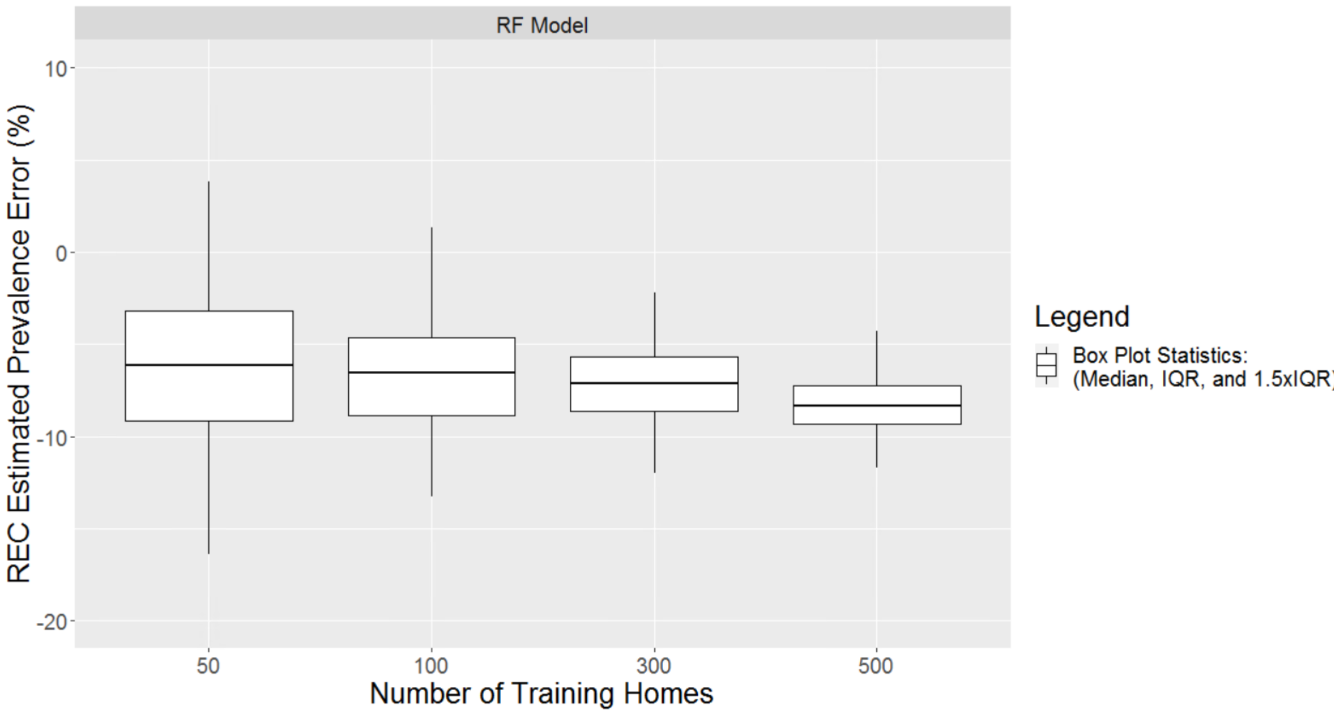


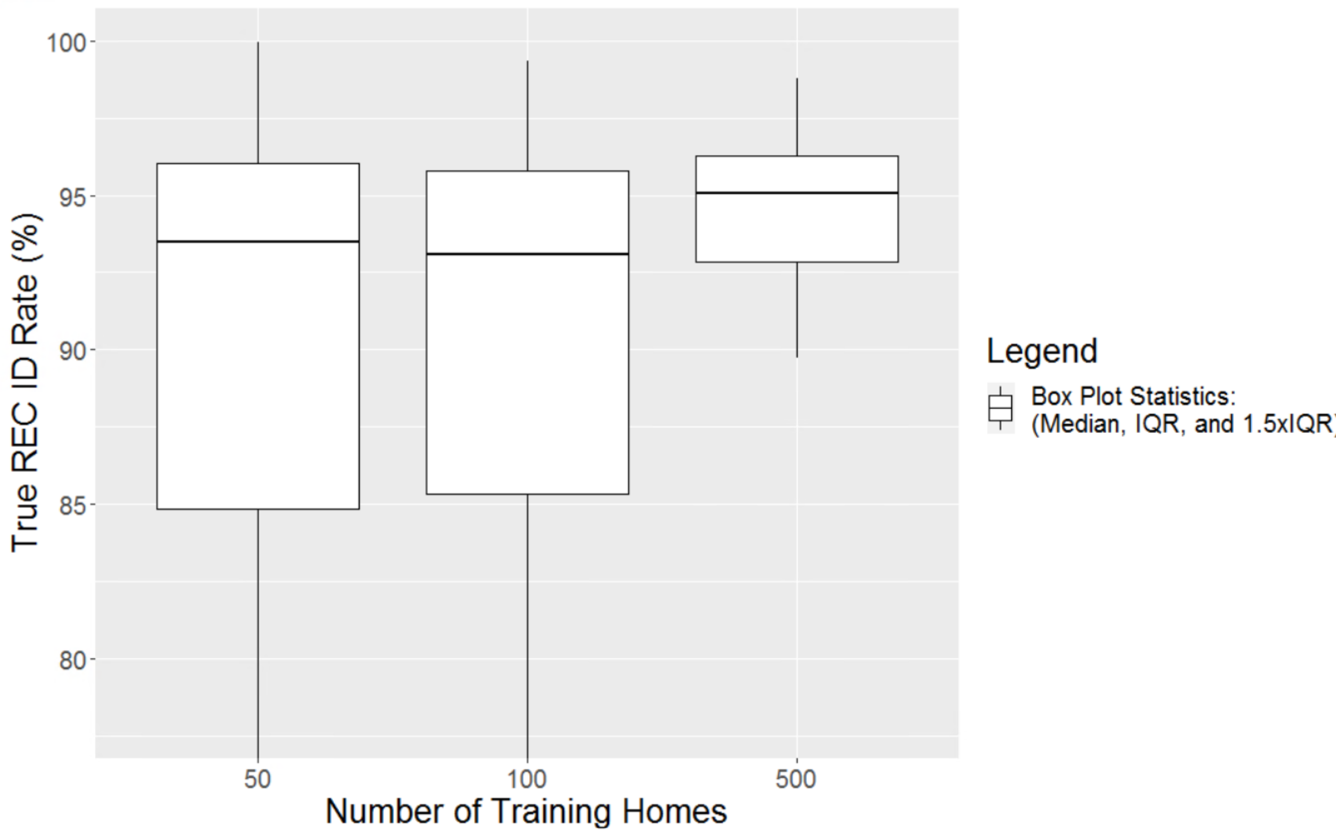
Figure 9. Distributions of classification prevalence estimates for different training-set sizes (50, 100, 300, and 500 in each category AC and REC) using the RF model. Bar and whisker plots show median value, interquartile range (IQR), and 1.5 times the IQR.



SCENARIO 4

The URWS B dataset was comprised of SFRs, mobile homes, and manufactured homes and most homes with REC systems were mobile or manufactured homes. In contrast, the URWS A dataset was comprised entirely of SFRs, sixty percent of which had REC systems according to tax assessor data. This provided an immediate cause for concern, as prediction accuracy was likely to be poor given the categorical differences between the training and testing datasets. Nevertheless, the resulting distributions of sensitivity and prevalence error for each training set size are shown in Figures 10 and 11. The distribution of coefficient estimates from the imputed regressions are shown compared with the scenario 2 result in Figure 12.

Figure 10. Distributions of URWS B test classification sensitivity for different URWS A training set sizes (50, 100, 300, and 500 in each category AC and REC) used to train the RF. Bar and whisker plots show median value, interquartile range (IQR), and 1.5 times the IQR.



The attempt to classify SFR cooling system category was unsuccessful. The RF test sensitivity rate was over 90% when 500 sample homes in each category were used. However, this is due to rampant misclassification of AC units, hence the roughly 20% prevalence error. In addition, regression results from the imputed cooling system variable failed to approximate the scenario 2 coefficient estimate. The imputed regression results suggest SRF's with REC units use no more water on a hot day than SFR's with AC units, which is consistent with the number of SFR's with AC units misclassified as having REC units.

Given the strength of the RF model in scenarios 1 and 2, it is likely that URWS B was simply not an appropriate training set in this situation, despite its proximity to URWS A.

Figure 11. Distributions of URWS B test classification prevalence for different URWS A training set sizes (50, 100, 300, and 500 in each category AC and REC) used to train the random forest model. Bar and whisker plots show median value, interquartile range (IQR), and 1.5 times the IQR.

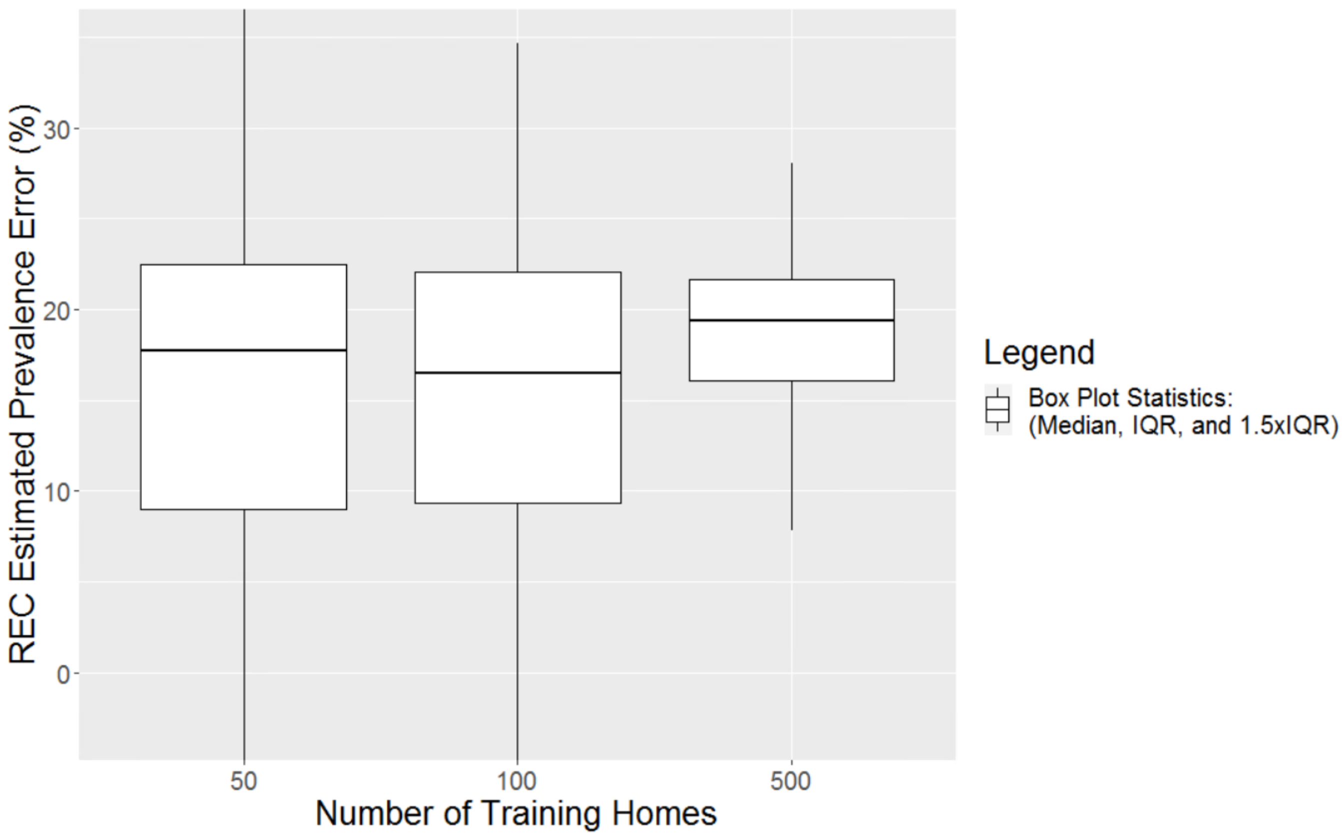
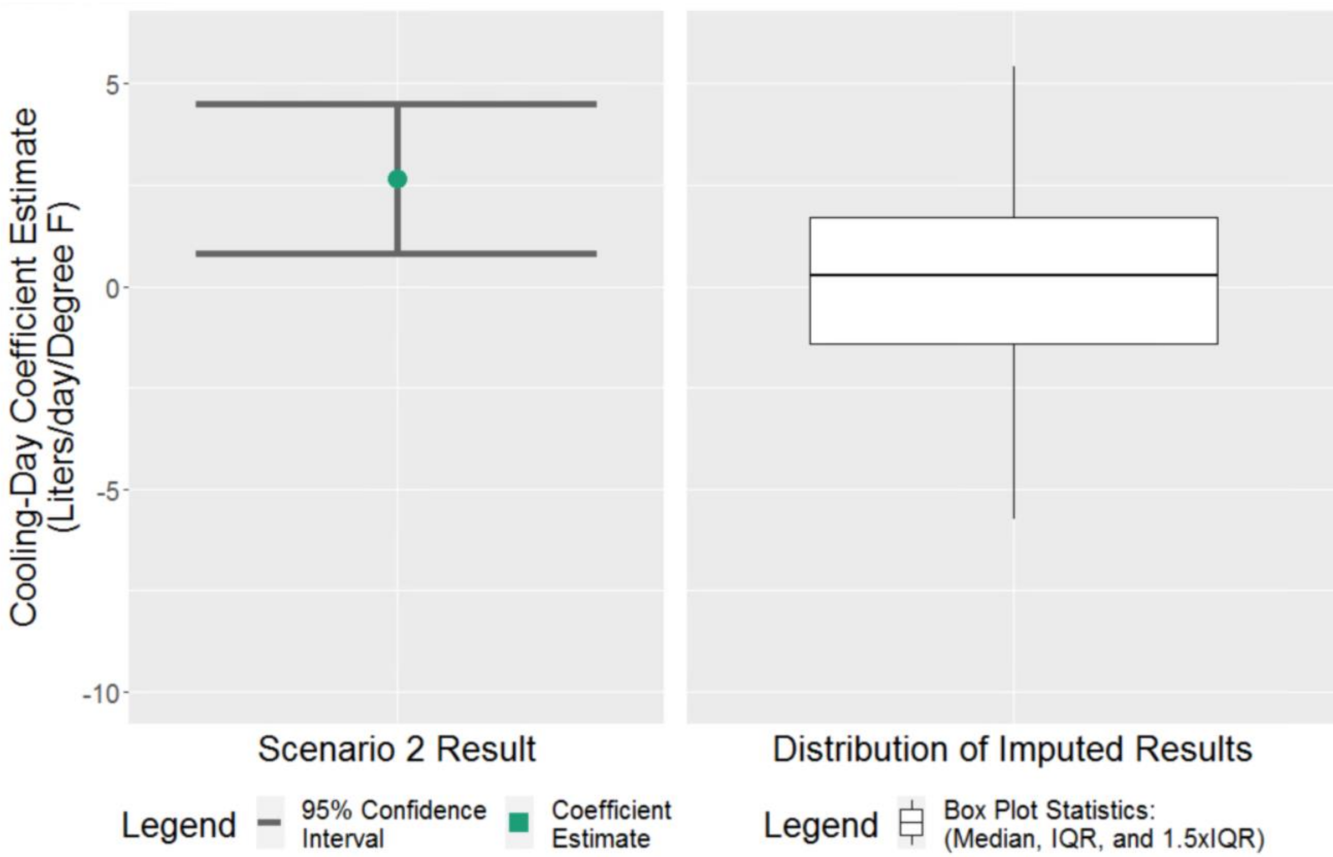


Figure 12. Scenario 2 coefficient estimate with 95% confidence interval compared with the distribution of coefficient estimates made using the imputed cooling system variable from the RF Model trained on URWS B data. Bar and whisker plots show median value, interquartile range (IQR), and 1.5 times the IQR.



DISCUSSION

There is a significant discrepancy between the expected daily REC water usage found in literature and the results shown above. Lab and field measurements found cooling-season usage for RECs on the order of 200-300 liters per unit per day. In addition, after carrying out the above analysis, CWEE was granted access to study of REC water consumption made using metered REC units in three URWS regions in Building Climate Zone 14. That study found an average of 197 liters per day per unit REC. In contrast this study estimated usage on the order of 40-60 liters per unit per day.

A hypothesis for this difference, supported observations by researchers working in the URWS A region, is this: the tax assessor dataset of property characteristics used to determine homes with REC and AC units is incomplete. Affordable, small to medium sized window-mounted REC units, installed by residents are almost certainly missing from the dataset. In fact, it is likely that there is near-ubiquitous use of RECs in the URWS A region. Given the significant savings on cooling season energy bills (on the order of hundreds of dollars per month), it makes financial sense for a resident to purchase and use an REC even if their home already has an AC system of some kind.

Thus, instead of comparing two groups of homes, one with REC systems the other with AC systems, the CWEE model may be comparing homes exclusively cooled by REC systems with those being cooled by both REC and AC systems. If this is the case, the results have a different interpretation: the coefficient estimates in scenarios 1 and 2 quantify the additional cooling-based water consumption of a home cooled exclusively by an REC over a home cooled by both an AC and REC.

These findings suggest that the identified methods are not nonconclusive for intended applications without additional research and data availability. Determining the amount of water used by RECs is complicated by the lack of data availability, privacy concerns, and uncertainty in the estimates with available data. It is recommended that individual customer survey and applying aggregated engineering calculations as presented in this technical document will provide suppliers the most robust method of calculating REC water use.

Given the above findings, an alternative for calculating water use volume for the REC variance may be based on a combination of home REC surveys and “engineering calculations” based on weather conditions, REC unit specifications, and physical properties associated with evaporative cooling (see Appendix for the description of the engineering calculations) . Home surveys can be used to estimate the frequency and size of REC’s and an expected range of REC water usage can be determined from the engineering calculations.

REFERENCES

American Society of Heating Refrigeration and Air-conditioning Engineers (ASHRAE). 2012. *HVAC Systems and Equipment* (SI ed.). 2012 ASHRAE Handbook Ch. 41.1.

Karpiscak M, Babcock T, France G, Zauderer J, Hopf S, Foster K. 1998. Evaporative Cooler Water Use in Phoenix. Journal AWWA 90(4):121-130.

Karpiscak M, Marion M H. 2016. Evaporative Cooler Water Use. The University of Arizona College of Agriculture Cooperative Extension.

Kleminger W, Staake T, Santini S. 2013. Occupancy Detection from Electricity Consumption Data. BuildSys 2013(14-15).

Pistochini T, Modera M. 2011. Water-use efficiency for alternative cooling systems in arid climates. Energy and Buildings 43(2011):631-638.

Pistochini T, Drwencke A, Tucker C, Narayanan V, Stevens M, Farbman D. 2018. Improving Water and Energy Efficiency in California's Dairy Industry (Contract No. EPC-16-010). UC Davis Western Cooling Efficiency Center.

Spartz P, Jenkins N, Surles T, Therkelsen R. 2004. Advanced Evaporative Cooling White Paper (Contract No. 600-98-022). Davis Energy Group.

Sahai R, Shah N, Phadke A. 2012. Addressing Water Consumption of Evaporative Coolers with Greywater. Lawrence Berkeley National Laboratory Environmental Energy Systems Division International Energy Studies Group.

UC Davis Western Cooling Efficiency Center 2020. Performance Evaluation of a Residential Indirect-Direct Evaporative Cooler (IDEC). ET19SDG1031, UC Davis Western Cooling Efficiency Center.

APPENDIX

The following sections provide supplemental information supporting the various methods and analyses undertaken above. Details are given for the regression and classification methods and the engineering calculations. In addition, the supplemental dataset showing metered REC unit water usage in the URSW A region is presented.

APPLYING THE METHODS

To apply the methods, a supplier must first calculate a series of variables using the weather, property, and water/energy consumption data from their service area. The variables are then passed through two generic statistical tools to generate the results. The statistical tools (an ordinary least squares regression and a random forest classifier) are available in most statistical software, such as R or Python. The technical details for constructing the variables for each of the two methods are given in the following sections.

Constructing the Variables for Regression

Average Cooling and Heating Degrees

These values are computed from the daily average wet- and dry-bulb temperatures (F) recorded at a given service location. The regression method uses only the daily mean dry bulb temperatures to calculate average cooling and heating degrees. They are calculated as:

$$CD_j = mean\{max(0, T_{db,k} - 70) : \text{day k in billing period j}\}$$

$$HD_j = mean\{max(0, 70 - T_{db,k}) : \text{day k in billing period j}\}$$

Where $T_{db,k}$ is the mean dry-bulb temperature on day k. The classification method uses cooling and heating degrees based both on wet- and dry-bulb temperatures. Using wet-bulb temperature, the cooling and heating degrees are calculated as:

$$CD_j = mean\{max(0, T_{wb,k} - 50) : \text{day k in billing period j}\}$$

$$HD_j = mean\{max(0, 50 - T_{wb,k}) : \text{day k in billing period j}\}$$

Where $T_{wb,k}$ is the mean wet-bulb temperature on day k.

Binned Number of Bedrooms

This value requires knowledge of the number of bedrooms in each residence. It is calculated as:

$$R_i = min(5, B_i)$$

Where B_i is the number of bedrooms in residence i.

Binned Home Quality

This value requires knowledge of a home quality metric for each residence (on a scale of 1-100). Let Q_0 be the lowest quality measure in the sample and let residence i have quality Q_i. The quality bin for residence i is calculated as:

$$HQ_i = 1 + Floor((Q_i - Q_0)/10)$$

where *Floor* is the floor function.

Binned Vintage

This value requires knowledge of the vintage of each residence. Let Y_0 be the earliest vintage in the sample and let residence i have vintage Y_i. The vintage bin for residence i is calculated as:

$$V_i = 1 + Floor((Y_i - Y_0)/10)$$

where *Floor* is the floor function.

Binned Home Size (Square Feet)

This value requires knowledge of the size of each residence. Let S_0 be the smallest homes size in the sample and let residence i have size S_i. The size bin for residence i is calculated as:

$$A_i = 1 + Floor((S_i - S_0)/100)$$

where *Floor* is the floor function.

Average Daily Water Consumption per Billing Period

For residence i, average daily water consumption (liters) for billing period j is calculated as:

$$Cons_{ij} = (\text{Total Water Use in Residence i During Period j})/(\text{Total Number of Days in Period j})$$

If hourly energy data is available, the total number of days in a given period is limited only to days with non-zero variance in hourly energy-use.

Address Fixed Effect

For each address, create a dummy variable defined to be 1 if residence i has that address and 0 otherwise. In R, this can just be a factor or character variable containing each residence's address. Assuming the use of R, call this factor *Addr*.

REC indicator Variable

Create a dummy variable that is 1 if residence i has and REC system (central or non-central) and 0 otherwise. Call this variable *Cool*. In R, this can be a factor or character variable containing the string representation of the residence's cooling system, either 'AC' or 'REC'.

Period Variable

The period variable indicates in which month and year a billing period took place (ideally, each billing period is a single month in a given year). The period is not used in regression, but as a clustering variable for calculating robust standard errors. Construct it as the month and year of the billing period concatenated together. For example, a record from billing period September 2018 could have period "9-2018".

Carrying out the Regression

Using statistical programming software such as R, Python, or Stata, carry out the following regression (shown here in R):

```
mod = lm(Cons ~ Addr + CD + HD + Cool:CD + Cool:HD +
                V:CD + V:HD +
                A:CD + A:HD +
                HQ:CD + HQ:HD +
                R:CD + R:HD,
                data = <supplier_data>)
```

using the dry-bulb *CD* value. After fitting the model, the next step is to make the standard errors robust to heteroskedasticity and two-way cluster by premise and period. This can be done in R using the cluster.vcov() function from the multiwayvcov package with the following code:

```
V = cluster.vcov(mod, <study_data>[,c('Addr','Period')])
```

And then extracting the resulting standard errors from:

```
coeftest(mod, V)
```

Using coeftest() from the R lmtest package.

Constructing the Predictors for Classification

Seasonal Averages

Each of these predictors is the average of a specific variable in a given season and for a given residence. First, define the seasons as summer, winter, and spring/fall. Then define summer as May through September, winter as November through February, and spring/fall as the remaining months. Then for each residence *i*, season *S,* and variable *V* calculate:

$$V_{Si} = mean\{V_{ij} : \text{service month j in S}\}$$

The list of variables from which to select *V* includes average daily water consumption, average daily dry- and wet-bulb temperature, and the average energy consumption during each hour of the day 0-23 (when hourly energy data are available). For example, the average summer noon energy use for residence *i* is calculated as the average of all noon-hour energy-use measurements during the summer months over all years.

Water Use per Square Foot per Degree

This value represents the relationship between water use per square foot in an individual residence and the outdoor dry- and wet-bulb temperature. It is calculated for each residence separately using only that residence's monthly water use data and the associated monthly CD and *HD* values as calculated above in the regression section (dry- and wet-bulb, separately). First, for residence *i* and for each billing period *j*, calculate:

$$ConsSqft_{ji} = Cons_j / Sqft_i$$

Then carry out the two linear regressions:

```
mod_db_i = lm(ConsSqft ~ CD + HD, data = <residence_i_db_data>)
mod_wb_i = lm(ConsSqft ~ CD + HD, data = <residence_i_wb_data>)
```

Separately save the coefficient estimates of CD_{db} and CD_{wb} for each residence and concatenate them over all residences to form predictor variables $H2O_{db}$ and $H2O_{wb}$, respectively.

Energy Use per Square Foot per Degree

This predictor is constructed in the same way as $H2O_{db}$ and $H2O_{wb}$, but using hourly energy data and hourly temperature data aggregated to the daily scale. For each residence and each day, $ConsSqft$ is that residence's total daily energy use divided by its size in square feet. And the *CD* and *HD* (both dry-and wet-bulb) values are calculated from daily average temperatures and not aggregated up to the service period. Then the same regression is carried out and the coefficients are saved and concatenated to form predictors KWH_{db} and KWH_{wb}.

Property Characteristic Variables

Raw values from the property characteristics dataset are used as predictors in the classifier. These include home square-footage, gross property acreage, home vintage, number of bedrooms, and home quality.

Training the RF Classifier

Letting X be the matrix of predictors, Y the response vector, 'REC' the name of the response variable, and 'Vars' be a string variable containing the names of the predictors concatenated by '+', use the R package randomForest and functions tuneRF() and randomForest() to run:

```
mtry = tuneRF(X, Y, ntree = 1000, stepFactor = 1.5, improve = .01,
                    trace = F, plot = F)

rf = randomForest(as.formula(paste0('REC ~', Vars)),
                    cbind(X,Y, ntree = 1000,
                    mtry = (mtry[mtry[,2] == min(mtry[,2]),1]))
```

or run the equivalent in Python. Then *rf* is the fitted model and can be used to make predictions and assess variable importance.

ENGINEERING CALCULATIONS

The following section provides details on the engineering calculations used to predict REC water use rates. The calculations only require local weather data (dry- and wet-bulb temperature and humidity), and values for atmospheric pressure, home-size, and air flow.

Expected REC Water Use

1. Given the outdoor dry-bulb temperature ($T_{db,out}$), outdoor wet-bulb temperature ($T_{wb,out}$) and evaporative efficiency chosen from between .8-.95 (*eff*) calculate the supply dry-bulb temperature:

$$T_{db,sup} = T_{db,out} - eff \cdot (T_{db,out} - T_{wb,out})$$

2. Use the supply/outdoor dry- and wet-bulb temperatures and atmospheric pressure to calculate the supply/outdoor humidity ratios. These are in units of Liter of H2O per kg of Air.
3. The difference in supply humidity ratio and outdoor humidity ratio describes the amount of water evaporated in the process of cooling. Call this value Δ_{evap}. It is also in units of Liter of H2O per kg of Air.
4. Calculate the specific volume of the supply air given the dry-bulb temperature of the supply air and the atmospheric pressure. Call this s_{air}. This is in units of cubic feet of air per kg of air.
5. Choose a standard *CFM*. Typical values are in the range 500-4000.
6. Then the liters of H2O consumed per hour due to Evaporative cooling is calculated as:

$$cons = 60 \cdot CFM \cdot \Delta_{evap}/s_{air}$$

SUPPLEMENTAL DATA

The Evaporative Cooler study was collaboration between the two Urban Retail Water Suppliers, URWS A and URWS C, but included homes from the URWS B region. The purpose was to quantify water used to cool homes that use evaporative coolers. Many homes in both URWS service regions use evaporative coolers due to the low purchase and operation costs.

The method used to collect data was installation of a water meter installed on the intake side of the evaporative cooler. The water meter used was an IPERL I2S1FLXX ¾''. The meter was reduced to ¼'' to accommodate the evaporative cooler connection and capable of reading water flows used in evaporative cooling.

Meters were installed with homeowner permissions and a final reading was taken. Homeowners were asked to keep track of any days the cooler was not in regular use. Down draft and window mounted coolers were used in this study. 4 customers were disqualified due to meters being removed.

The follow data was collected during the cooling season of 2020.

Customer	Region	Days in Use	Final Reading ft3	Gallons used	Daily Average
01	URWS C	101	812.223	6075	60
02	URWS A	71	435.444	3257	45
03	URWS C	101	882.526	6601	65
04	URWS A	71	432.994	3228	45
05	URWS A	71	458.569	3430	48
06	URWS C	42	225.614	1687	40
07	URWS C	92	654.705	4897	53
08	URWS A	71	733.482	5486	77
09	URWS A	71	409.586	3063	43
10	URWS A	71	679.257	5080	71

Customer	Region	Days in Use	Final Reading ft3	Gallons used	Daily Average
11	URWS A	71	425.095	3179	44
12	URWS A	71	580.372	4341	61
13	URWS C	101	784.316	5866	58
14	URWS A	71	472.085	3531	49
15	URWS C	92	405.382	3031	32
16	URWS A	71	445.367	3331	47
17	URWS C	92	822.300	6150	67
18	URWS C	92	440.866	3297	35
19	URWS A	71	676.328	5058	71
20	URWS A	71	620.751	4643	65
21	URWS C	101	570.842	4269	42
22	URWS A	85	610.744	4568	53
23	URWS A	85	656.244	4908	58
24	URWS B	89	702.361	5253	59
25	URWS B	89	691.235	5170	58
26	URWS B	74	587.373	4393	59
			Total Daily Average 52 GPD		