ORNL/TM-2018/780

Evaluation of Factors that Influence Residential Solar Panel Installations

April M. Morton
Olufemi A. Omitaomu
Susan M. Kotikot
Elizabeth L. Held
Budhendra L. Bhaduri

March 2018

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE US DEPARTMENT OF ENERGY

ORNL/TM-2018/780

Computational Sciences and Engineering Division
Geographic Information Science and Technology Group

Evaluation of Factors that Influence Residential Solar Panel Installations

April M. Morton
Olufemi A. Omitaomu
Susan M. Kotikot
Elizabeth L. Held
Budhendra L. Bhaduri

Date Published: March 2018

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831-6283
managed by
UT-BATTELLE, LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The authors would like to thank Courtney Grosvenor and Nathaniel Horner with the US Department of Energy's Office of Energy Policy and Systems Analysis for their constructive comments and suggestions on the initial draft of this report. They also thank Bandana Kar and Melissa Allen for their comments within the Oak Ridge National Laboratory's internal review system.

ABSTRACT

Though rooftop photovoltaic (PV) systems are the fastest growing source of distributed generation, detailed information about where they are located and who their owners are is often known only to installers and utility companies. This lack of detailed information is a barrier to policy and financial assessment of solar energy generation and use. To bridge the described data gap, Oak Ridge National Laboratory (ORNL) was sponsored by the Department of Energy (DOE) Office of Energy Policy and Systems Analysis (EPSA) to create an automated approach for detecting and characterizing buildings with installed solar panels using high-resolution overhead imagery. Additionally, ORNL was tasked with using machine learning techniques to classify parcels on which solar panels were automatically detected in the Washington, DC, and Boston areas as commercial or residential, and then providing a list of recommended variables and modeling techniques that could be combined with these results to identify attributes that motivate the installation of residential solar panels. This technical report describes the methodology, results, and recommendations in greater detail, including lessons learned and future work.

1. INTRODUCTION

Rooftop photovoltaic (PV) systems are the fastest growing source of distributed generation; small-scale PV installed capacity and generation have roughly doubled from 2014 to 2016 [Solar Energy Industries Association (SEIA), 2017; US Energy Information Administration (EIA), 2017]. Hence, there has been a significant increase in the number of installed solar panels in recent years in the United States. However, detailed information about where these solar panels are located and who their owners are is known only to installers and utility companies. This lack of information about installed solar panels is a barrier to policy and financial assessment of solar energy generation and use.

To bridge the described data gap, Oak Ridge National Laboratory (ORNL) was previously sponsored by the Department of Energy (DOE) Office of Energy Policy and Systems Analysis (EPSA) to create an automated approach for detecting and characterizing buildings with installed solar panels using high-resolution overhead imagery. ORNL applied the developed approach to 10 study areas provided by EPSA and delivered a candidate set of automatically detected solar panels in each of the study areas. ORNL also provided socioeconomic attributes of buildings detected with potential solar panels using US census data.

To enhance the utility of the developed datasets, ORNL was further tasked to provide recommended data and methods for using the candidate set of automatically detected solar panels to evaluate the socioeconomic, policy-related, and environmental factors that motivate residential homeowners to install solar panels in the Washington, DC, and Boston areas. To achieve this goal, this current project is divided into two tasks:

i. **Task One:** ORNL's automatic solar panel detection algorithm produces a map of detected solar panels and the parcels they belong to. Because we are currently only focused on the factors which motivate residential homeowners to install solar panels, we must first determine which solar panels belong to residential parcels only. Consequently, our first task is to classify the candidate set of automatically detected parcels with solar panels as commercial or residential using advanced machine learning techniques. Homeowners of units in multifamily homes and condominiums often do not have the option to install solar panels, so we define residential parcels as single-family homes and commercial parcels as non-single-family homes and buildings that generate income from businesses, such as offices, retail space, and apartment buildings.

ii. **Task Two:** In order to guide future research based upon ORNL's dataset of parcels with detected residential panels, we summarize previous work that has used PV-related datasets to identify attributes that motivate the installation of residential solar panels. Based on this work and the set of parcels classified as residential with detected solar panels, we provide a list of recommended variables and modeling techniques for identifying attributes that motivate the installation of residential solar panels in the Boston and DC areas.

This technical report describes the automated approach for detecting and characterizing buildings with installed solar panels, as well as the methodology and results of the residential/commercial parcel classification task for the Washington, DC, and Boston study areas. It concludes with a summary of previous work that has used PV-related datasets to identify attributes that motivate the installation of residential solar panels. Based upon this work, we provide a list of recommended variables and modeling techniques that can be combined with ORNL's PV-related dataset to identify attributes that motivate the installation of residential solar panels in Boston and DC. In addition, we discuss lessons learned and future work.

2. METHODOLOGY

This section is divided into two sub-sections. Section 2.1 reviews the methodology and results for the previous effort to identify solar panels from imagery. Section 2.2 presents the methodology for the current tasks.

2.1 DETECTION AND CHARACTERIZATION OF SOLAR PANELS

Automatic detection of solar panels using aerial images is a challenging task. Solar panels are considerably smaller than objects that are often targeted in aerial image analysis, such as roads and buildings. A large number of solar panels occupy less than 100 pixels in standard satellite imagery and provide very few image features; hence, they can be easily confused with other objects. Moreover, the appearances of solar panels in images vary vastly. In addition to image variations caused by differences in acquisition conditions, solar panels have a variety of types, sizes, and shapes. To understand how different images will perform, we experimented with aerial images (obtained from the National Geospatial-Intelligence Agency) and satellite images with RGB bands (obtained from WorldView-3) of approximately 0.3 meter resolution. This sub-section summarizes the developed methodologies for automatic detection of solar panels and their subsequent characterization.

2.1.1 Automatic Detection of Buildings with Solar Panels

A supervised automatic detection algorithm using images with a spatial resolution of 0.3 meters was developed for the detection of buildings with solar panels. The algorithm implemented deep convolutional neural networks to achieve detection. As a supervised approach, the performance of the algorithm heavily depends on the number of training samples and how representative those training samples are. Hence, we collected a set of training samples (ground truth data) and a set of validation (testing) samples. The samples were collected by manual delineation and digitization of solar panels in some images. We then trained the algorithm to learn solar panels features in the training samples. We did a preliminary study using satellite and aerial images of four cities—**Washington, DC, Chattanooga, North Boston, and San Francisco**. The images were collected between 2012 and 2014. We trained the networks using a training set containing 2,040 training samples (500-by-500 image tiles). We applied the trained network to two images for testing, each of which was 40,000-by-30,000 pixels (about 108 square kilometers). One image covered the entire San Francisco area, and the other an area in north Boston. In total, 4,500 solar panels were extracted in San Francisco and 1,300 in Boston. For quantitative validation, we selected a new image tile of 5,000-by-5,000 pixels that was not within the training and test sets in each city. The metric

of performance used for the evaluation of these validation samples was high detection accuracy. However, measuring accuracy at the pixel level is not practically meaningful for this application due to the small size of solar panels. We needed to be able to detect a whole solar panel, not some portion of it. Therefore, we computed two performance scores—completeness and correctness, which are often used to compare road vectors. Completeness is defined as the number of manual labels containing center points divided by the total number of manual labels. Correctness is the number of center points inside manual labels divided by the total number of center points. The procedure used to quantify these scores was to compute centers of detected solar panels and dilate the manual label by 1 meter. Using aerial images, the completeness score for San Francisco was 87.3% and for Boston was 84.0%. The correctness score for San Francisco was 85.5% and for Boston was 81.2%. Using satellite images, the completeness score and correctness score for San Francisco were 67.5% and 61.8%, respectively. After we identified a good model based on the results of the validation set, we deployed the model for large-scale implementation, which is the stage at which we performed the city-scale automatic solar panels detection study. Furthermore, we decided to use aerial images for the city-scale implementation. The city-scale implementation was completed for six additional cities—**Fresno, Modesto, Stockton, Sacramento, South Boston, and Phoenix**. For detailed information about the developed algorithm and the detection results, please see Yuan et al. (2016).

2.1.2 Characterization of the Detected Buildings Using Parcels and Census Data

To understand the characteristics of the owners of detected buildings with solar panels, we needed to match those buildings to social, economic, and demographic datasets. Hence, we used the 2008–2012 American Community Survey (ACS) Summary Tables published by the US Census Bureau (US Census Bureau, 2012). A geographic information system approach was then used to achieve the matching. The overall steps used can be summarized as follows:

i. Output of the PV Detection Algorithm is converted from raster to vector format data (i.e., polygon shapefiles). Geometry of the solar panels is calculated.
ii. Parcel data is obtained from city and county GIS portals respective to each city. A unique identifier, "PARCID" (i.e., Parcel ID), is generated for each parcel that contains a solar panel.
iii. The output and parcel data are spatially joined in two configurations:
    a. First, the attributes of the tax parcels are joined to the panel polygons—this tag is a PARCID for each individual solar panel so that cross-referencing is possible in the future.
    b. Second, the resulting solar panel data layer is then joined to the parcels layer.
    c. The output at this point is a parcel layer with attributes of the spatially respective detected solar panels. The join process results in a count of how many panels are joined to each parcel, as well as the geometry of the total panels by sum, mean, max, and min in each parcel.
iv. The data is transformed into a point layer based on the centroid of each parcel.
v. Census county, tract, block group, and block GEOIDs are spatially joined to the centroids to facilitate appending demographic data, and addresses are geocoded for each point. A minor portion of city parcel layers contains address data. In these cases, geocoding addresses are unnecessary and the original parcel addresses are kept.
vi. The final output is a point shapefile as a database of addresses where solar panels were detected by the algorithm, along with panel counts, metrics, and Census GEOIDs.

2.2 CLASSIFICATION OF THE DETECTED PARCELS INTO RESIDENTIAL AND COMMERCIAL PARCELS

This section presents the methodology for Tasks 1 and 2 defined above. Data classification is a machine learning technique used to guess the class or category of a data item. A data item could be a loan

application in the banking sector, a patient's medical report in the healthcare sector, or parcels with rooftop solar panels, such as those discussed in this report. To accurately guess the category of a data item, we need past and present examples of items in the same category as the item of interest. In the loan application example, the loan officer would like to decide whether the loan application should be approved, approved with conditions, or denied. To make such a critical decision, the loan officer uses a data classification algorithm (lines of computer code), with known information about the applicant's past credit records and known information about other applicants with backgrounds similar to the applicant's. This information is then used to categorize the applicant as "credit worthy," "credit worthy with conditions," or "not credit worthy." In this task, we used a data classification algorithm to determine whether a parcel with installed solar panels is a residential parcel or a commercial parcel.

The data classification algorithm used in this task follows this five-step process:

i. **Choose distinctive features:** When humans group items, they typically consider only those qualities that best differentiate one item from another. The data classification algorithms also depend on selecting distinctive characteristics, or "features," that will allow the algorithm to distinguish between the different items. Therefore, the first step for this task is to identify the features that distinguish a residential parcel from a commercial parcel. Some of these features are the area of the parcel and the number of installed solar panels. Since some of these features may not provide complete distinction, we may need to include information that characterizes the surrounding environment, such as the number of commercial businesses within a certain distance to the parcel.

ii. **Choose algorithms to test:** Just as humans have different techniques for guessing the class of an item, data classification algorithms also have different techniques for guessing item classes. Thus, the second step is to choose and compare appropriate algorithms for the classification problem at hand. This depends on the available data, the data type, number of available examples, and the power of the computer being used. To classify parcels containing solar panels as residential versus commercial, we applied three data classification algorithms.

iii. **Manually classify a subset of the dataset:** To obtain examples of items (parcels) with known classes (residential or commercial parcels), we manually identify and group some parcels into residential and commercial parcel groups. To make this determination, we compute the latitude/longitude coordinates of each parcel and then manually look up the building addresses using Google Maps (Google Maps, 2017). We then search for each address using the Google Search Engine to determine if it corresponds to a single-family home. If it shows up as a single-family home on a real estate company's website, we classify it as residential, and if it does not, we search for the business or other website it corresponds to and classify it as commercial. As the number of examples in each class increases, the accuracy of the data classification algorithm increases.

iv. **Split the labeled data into training and testing datasets:** Once we have enough examples, we divide the examples into two groups—training and testing groups. The training group is what is given to the data classification algorithm as known examples. The algorithm uses these examples to develop a computer classification model. To test how well the developed computer model performs, we validate the model using several testing and training groups. We then calculate the number of correct guesses and the number of wrong guesses in each testing/training group. In our case we randomly choose 625 different testing and training groups. In each testing/training group, we select 75% of the examples for the training group and 25% of the examples for the testing group. If the average number of correct guesses is

high enough for our problem, we say that the model is useful. By selecting several different training and testing groups, with a 75%/25% split, we increase the chances that the model has enough data to be adequately trained and tested. Otherwise, we will either provide more known examples to the algorithm to generate a new model or will adjust which variables are included in the model. We will continue to do this until we have a useful model.

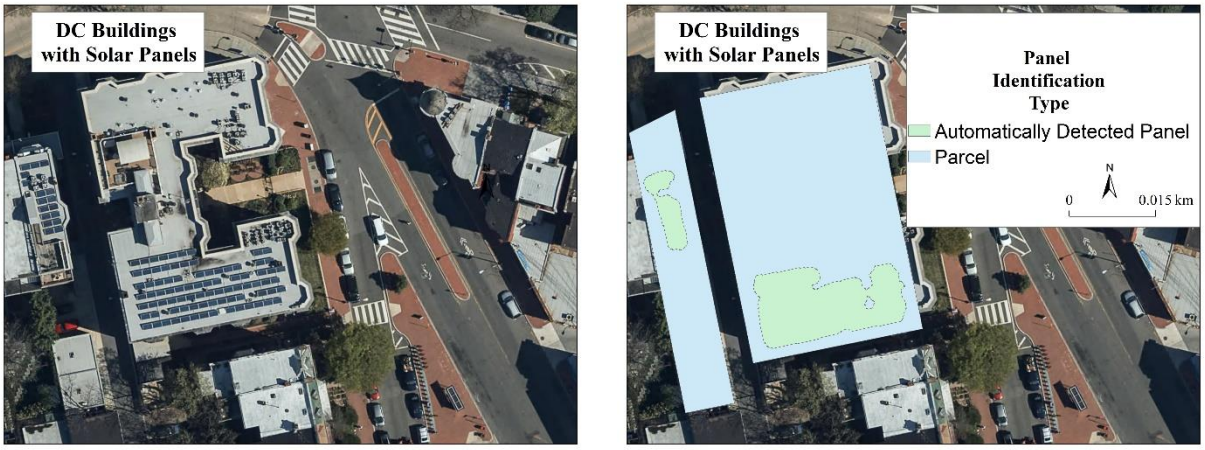
v. **Apply the useful model to additional items:** Once we have a useful model, we can then apply the model to guess the class of new (additional) items—i.e., the parcels in the dataset that were not manually classified. While we are not able to manually validate the error when applying the model to the full dataset including the new items, we assume the error is equal to the error achieved in the testing group. We may use this model for a long time if the features that distinguish the items are still valid. In the case of residential versus commercial parcel classification, the model may become invalid if the features describing each parcel change significantly. For example, the model may need to be retrained if a large number of newly designed homes are built with features that are much different from the previously built homes. Once the model is no longer valid, we will go back to the appropriate steps above to repeat the process.

3. EXPERIMENT

Our study areas for this task are Washington, DC, and Boston. In this section, we describe the dataset used, feature selection process, the data classification algorithms, and experimental setup. The data, code, and a copy of this report can be downloaded from the following link: https://github.com/GIST-ORNL/ornlsolar.

3.1 THE DATASET

The parcel dataset used in this project was obtained by merging the candidate set of automatically detected solar panels with parcel data from the Boston and DC areas. Note that although the Boston and DC solar panels were detected in the same manner as described in Section 2.1.1, the derived parcel datasets were processed a bit differently. More specifically, the Boston parcels were obtained by merging the parcel data from the BostonGIS's "Parcels 2016 Data Full" dataset (BostonGIS, 2016) with the "L3_TAXPAR_POLY_ASSESS.shp" file that is part of the MassGIS's "Standardized Parcel Data" dataset (MassGIS, 2013). The DC parcels were all obtained from the DCGIS Open Data's "Common Ownership Lots" dataset (DCGIS Open Data, 2017). Only parcels that contained an automatically detected solar panel candidate were retained in the parcel dataset. In addition, if a solar panel detection intersected two parcels, we only retained the parcel that it intersected with most (i.e., whose intersected area was largest). Furthermore, we deleted any duplicated parcel polygons so that every parcel was only represented once. Figure 1(b) shows examples of two automatically detected solar panels and their parcel boundaries, for the same buildings shown in Figure 1(a). There are some 10,000 and 40,000 detected solar panels for DC and Boston, respectively. Because it is more expensive to generate ground truth data to validate these results, we randomly selected 60 of the parcels with detected solar panels for the DC and Boston areas, respectively, and manually checked detection accuracy (i.e., percentage of detected parcels that actually had solar panels) using the same imagery used for the automatic detection implementation. We found that in Boston, 18.3% of the checked parcels had solar panels (11 out of 60), and in DC, 25% of the checked parcels had solar panels (15/60). In Fresno, 57 parcels out of 60 had solar panels (95% accuracy); in Modesto, 52 parcels out of 60 had solar panels (86.7% accuracy); in Stockton, 32 parcels out of 60 had solar panels (53% accuracy); and in Sacramento, 46 parcels out of 60 had solar panels (76.7% accuracy).



(a) **(b)**

Figure 1. (a) An image of two parcels with solar panels in the DC area; (b) the automatically detected solar panels and their parcel boundaries for the same buildings shown in (a).

3.2 FEATURE SELECTION

To aid the classification algorithm in distinguishing between commercial and residential parcels with solar panels, we investigated several variables related to each parcel, such as characteristics of the parcel, installed solar panels, buildings contained within the parcel, and the surrounding area. From this superset of features, we chose four distinctive features. Table 1 contains the selected features chosen for parcel classification. Table B1 in Appendix B contains a list of all features considered, the data sources they came from, whether they were selected or not, and the reason they were or were not selected. To reach the final selection of features, we considered several factors, such as the distinctiveness of features (we did not want to choose features that were correlated or closely related), the feasibility of processing the data within the project timeline, the cost of the data, and how much value each feature would likely add to each algorithms' performance. Future efforts may want to reconsider those features that were not selected, as they did show signs of improving the algorithm but were excluded mainly because of the time required to process and validate the data. In addition, one might want to reconsider the features related to business points, as they also showed promise of improving the algorithm but were excluded mainly because of data licensing issues.

Table 1. Selected features for classifying parcels containing solar panels as commercial and residential.

Feature Abbreviation	Feature Description	Source
pn_sqm_sum	Total area of all solar panels on roof (sum of individual solar panel areas)	ORNL Detected Solar Panels
num_panels	Number of solar panels on roof	ORNL Detected Solar Panels
parc_sqm	Area of parcel	Parcel Data (BostonGIS, 2016; MassGIS, 2013; DCGIS Open Data, 2017)
num_emps	Number of employees in block group containing parcel	LEHD Origin-Destination Employment Statistics (LODES) Data (US Census Bureau, 2013)

3.3 CLASSIFICATION ALGORITHMS

To classify parcels containing solar panels as residential or commercial parcels, we applied three data classification algorithms to the set of parcels containing the candidate set of automatically detected solar panels in the Washington, DC, and Boston areas. More specifically, we implemented, compared, and assessed the performance of the random forest, neural network, and logistic regression algorithms using all 15 combinations* of the features (every possible set of two or more features) described in Table 1 and summarized in Tables B2–B6. All algorithms and code were implemented using R, an open source software environment for statistical computing and graphics (R Core Team, 2013). The following subsections describe the algorithms in greater detail.

i. **Random Forest (RF):** The RF method is used for classification and regression; it provides predictions by aggregating results from many decision trees (Boulesteix et al., 2012). In our experiments, we implemented the RF method using R's randomForest function within the randomForest package (Breiman et al., 2011). We used the default parameters for the algorithm except for the number of trees, which we set to 1000, and importance, which we set to "TRUE". All default parameter settings can be found in the package documentation (Breiman et al., 2011).

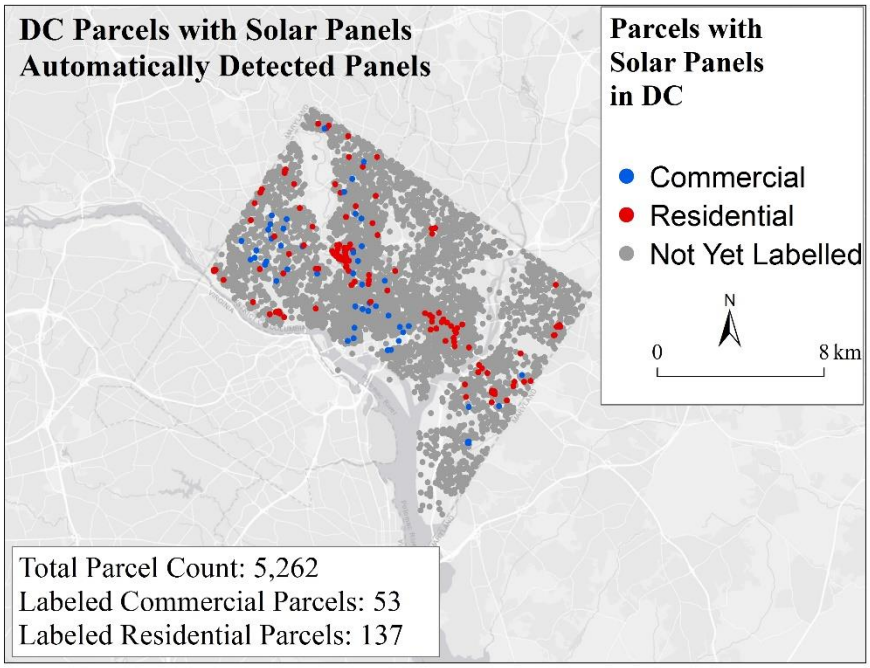
ii. **Neural Network (NN):** The NN method is a highly flexible function approximator first used in the fields of cognitive science and engineering (Kaastra & Boyd, 1996). In our experiments, we implemented the NN method using R's neuralnetwork function within the neuralnetwork package (Fritsch et al., 2016). We used the default parameters except the number of hidden neurons, which we set to 10, the threshold of the partial derivatives of the error, which was set to 0.3, and the maximum steps of the training of the network, which was set to 10 million. All default parameter settings can be found in the package documentation (Fritsch et al., 2016).

iii. **Logistic Regression (LR):** The LR method is a variant of ordinary least-squares regression where the dependent variable (whether a parcel is residential or commercial) is categorical (Peng et al., 2002). In our experiments, we implemented the LR method using R's glm function and step function within the stats package (R Core Team, 2013). We used the default parameters in the glm function except for the family type, which we set to "logit". We used the default parameters in the step function except for the mode of stepwise search, which we set to "both". All default parameter settings can be found in the package documentation (R Core Team, 2013).

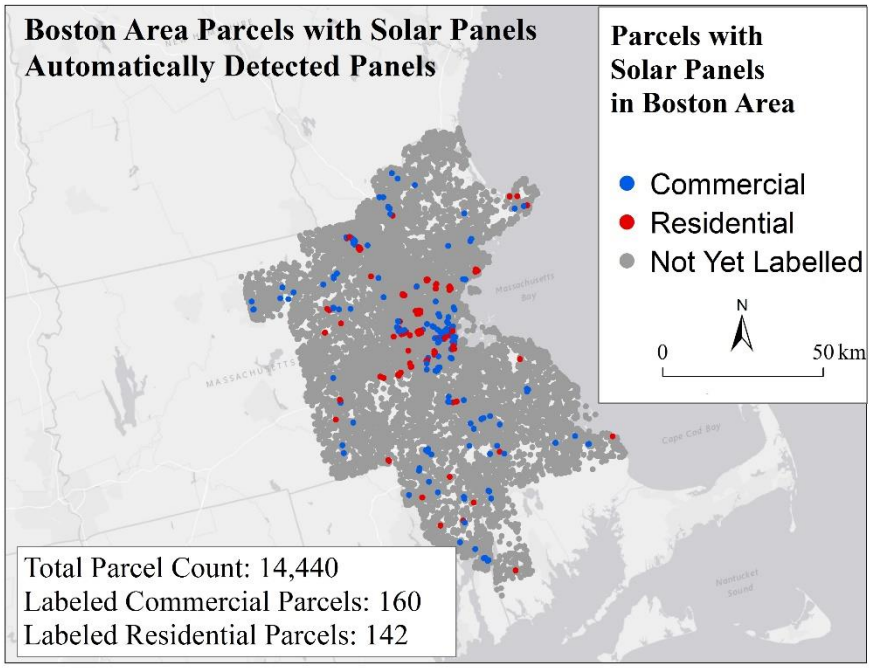
3.4 EXPERIMENTAL SETUP

To train, test, and compare each algorithm, we first manually labeled a subset of commercial and residential parcels in the DC and Boston areas. To make this determination, we computed the latitude/longitude coordinates of each parcel and then manually looked up the building addresses using Google Maps (Google Maps, 2017). We then searched for each address using the Google Search Engine to determine if it corresponded to a single-family home. If it showed up as single-family home on a real estate company's website, we classified it as residential; if it did not, we searched for the business or other website it corresponded to and classified it as commercial. Figures 2(a) and 2(b) show the labeled subsets in greater detail. Throughout the labeling process, we attempted to maintain as much geographic and characteristic variability as possible, as shown in Figures 3(a) and 3(b).

* Every possible set of two or more features from Table 1.

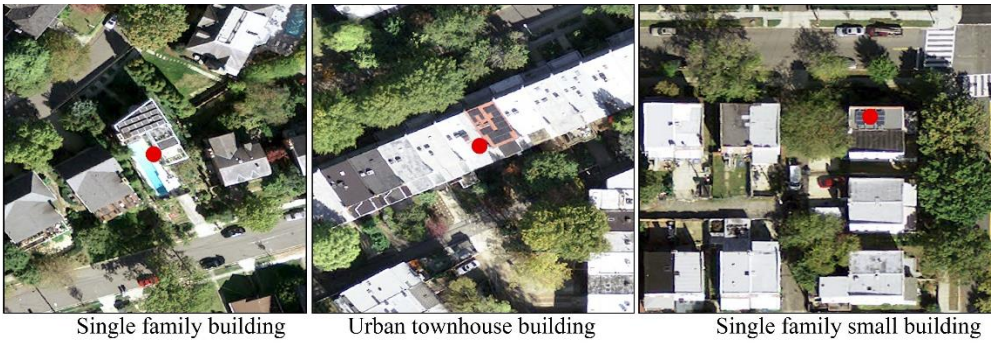


(a)

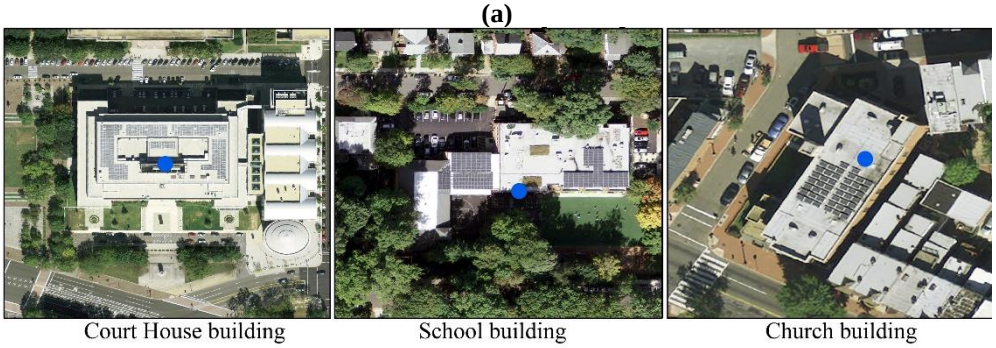


(b)

Figure 2. (a) The set of parcels containing the candidate set of automatically detected solar panels in the DC area, with a subset of parcels from the candidate set of automatically detected solar panels that were labeled (manually classified) as commercial and residential; (b) the set of parcels containing the candidate set of automatically detected solar panels in the Boston area, with a subset of parcels from the candidate set of automatically detected solar panels that were labeled (manually classified) as commercial and residential.



(a)



(b)

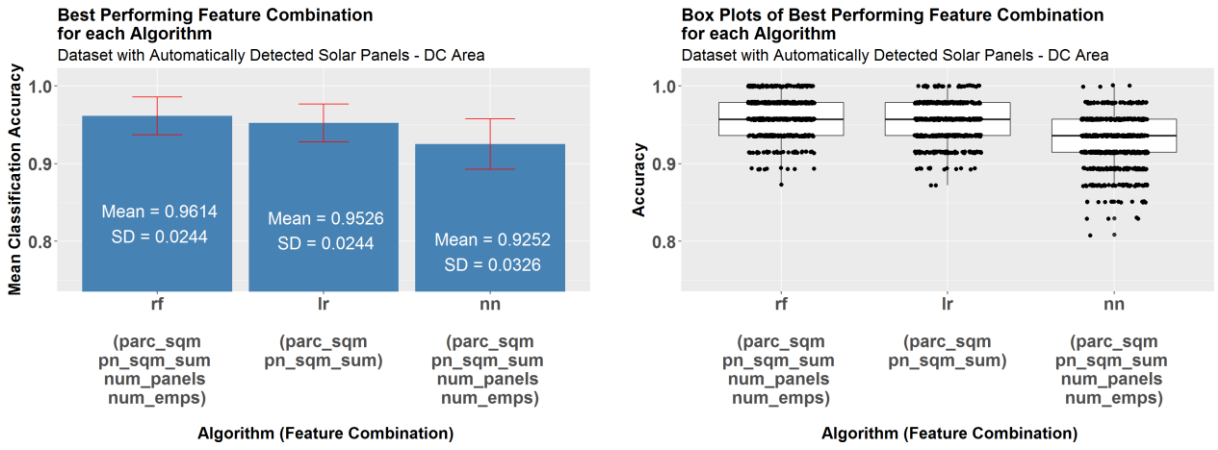
Figure 3. (a) Closeup of three labeled (manually classified) residential parcels in the DC area; (b) closeup of three labeled (manually classified) commercial parcels in the DC area.

To train and measure the performance of each residential/commercial classification algorithm, we first scaled the data by subtracting the mean of each feature from each feature value and then dividing each feature value by the feature's standard deviation. We then divided our labeled data for both DC and Boston into testing and training datasets. However, because different training/testing subsets often result in slightly different "trained" algorithms and accuracies, researchers typically divide their labeled data into several different training and testing subsets. Following this best practice, we randomly divided our labeled dataset into 625 different testing/training sets for each city, where 75% of the labeled data in each study area went to the training set and 25% of the labeled data went to the testing set. We then ran each algorithm/combination of features on each of the 625 different testing and training sets, measuring the resulting accuracy, or percentage of total predicted classes that were correct, each time. Because this resulted in 625 slightly different accuracies for each algorithm and feature combination, we summarized the performance of each by calculating the mean and standard deviation accuracy over all 625 accuracies. In addition, we computed the mean and standard deviation of commercial and residential accuracies for the algorithm/feature combination with highest mean accuracy in each study area. Commercial accuracy refers to the percentage of commercial predictions that were correct, and residential accuracy refers to the percentage of residential predictions that were correct. In addition, in order to examine how well an algorithm trained on one area might perform on another area, we repeated the procedure described above by training the algorithm using the DC data and testing it on the Boston data. The following section summarizes the results in greater detail.

4. RESULTS

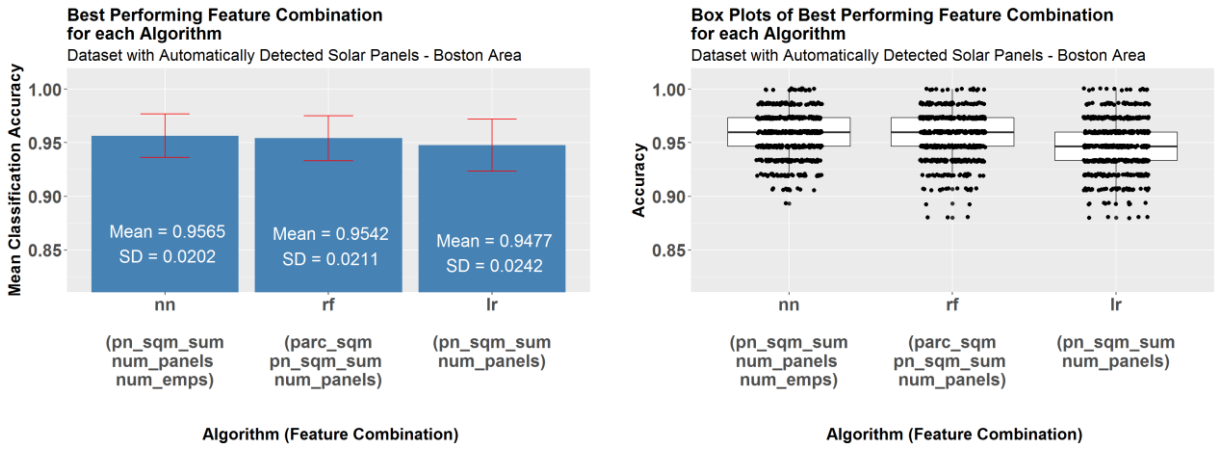
Figures 4(a) and 5(a) show the feature combinations with the highest mean accuracies for each residential/commercial classification algorithm for the DC and Boston areas, respectively. The RF algorithm trained on the features related to parcel size, total panel area, number of panels, and number of employees marginally out-performed the best LR and NN algorithms on the DC dataset, with an average accuracy of 96.14% and a standard deviation (SD) of 2.44%. The NN algorithm trained on the total panel area, number of panels, and number of employees marginally out-performed the best RF and LR algorithms on the Boston dataset, with an average accuracy of 95.65% and a SD of 2.02%. In addition, although we cannot quantitatively assess the accuracy of the classification algorithm on the unlabeled points, the visual results seem to intuitively make sense. Figure 6(a) shows that most of the downtown DC area consists of commercial buildings, 6(b) shows a closeup of DC residential buildings that visually look residential, and 6(c) shows a closeup of DC commercial buildings that visually look commercial. Figure 7 shows very similar results for the Boston area, correctly identifying the residential homes shown in 7(b) and accurately picking out a group of commercial buildings in 7(c). It is important to note that although the residential/commercial parcel classification accuracies reported in this task are not directly related to the solar panel detection accuracies, they are indirectly related to the them when characteristics of the detected solar panels are used as features. More specifically, if the algorithm/feature combinations reported depend on characteristics of the detected solar panels, such as the total area of the detected solar panels, inaccurate solar panel detections could influence whether a parcel is correctly classified as commercial or residential. For example, a very small false solar panel detection on a commercial building could lead the algorithm to misclassify a parcel as residential.

There are several other observations we can make from these results. First, the top-performing algorithm for DC depends on a mix of features related to solar panels, parcels, and surrounding areas. This makes sense, as the solar panel characteristics, parcel features, and surrounding areas tend to have distinct patterns in each class, which the algorithm can exploit to make more informed decisions about the classes that different parcels belong to. For example, the algorithm can learn that commercial buildings in the DC area generally have larger solar panels designed with more cells (Lets Go Solar, 2017), larger parcels, and are surrounded by more businesses, while residential buildings tend to have smaller solar panels, smaller parcels, and are not typically surrounded by many businesses. Figure 8(a) shows an example of a residential building in the DC area that was correctly classified as residential, while Figure 8(b) shows an example of a commercial building correctly classified as commercial. The algorithm likely did well with these buildings because they clearly resemble the “average” parcel of their type. More specifically, the correctly classified residential parcel in 8(a) is a small parcel surrounded by few businesses, while the correctly classified commercial parcel in 8(b) is a large parcel surrounded by many businesses. Figures 8(c) and 8(d) show incorrectly classified commercial and residential parcels, respectively, that do not seem to follow the average patterns expected of their respective classes. For example, the incorrectly classified commercial parcel in 8(c) is surrounded by more houses than businesses, while the incorrectly classified residential parcel in 8(d) is larger than the typical residential parcel.



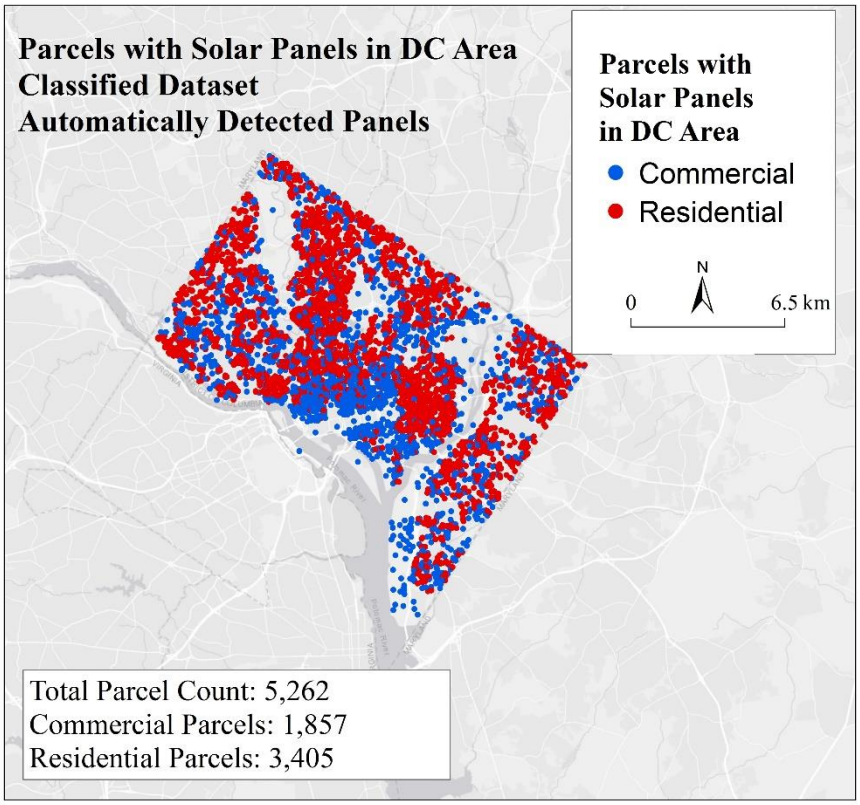
(a) **(b)**

Figure 4. (a) Bar charts with SD error bars for the best performing set of features for each of the three tested algorithms on the testing dataset in the DC area; (b) box plots with jittered accuracies for the best performing set of features for each of the three tested algorithms on the testing dataset in the DC area.



(a) **(b)**

Figure 5. (a) Bar charts with SD error bars for the best performing set of features for each of the three tested algorithms on the testing dataset in the Boston area; (b) box plots with jittered accuracies for the best performing set of features for each of the three tested algorithms on the testing dataset in the Boston area.

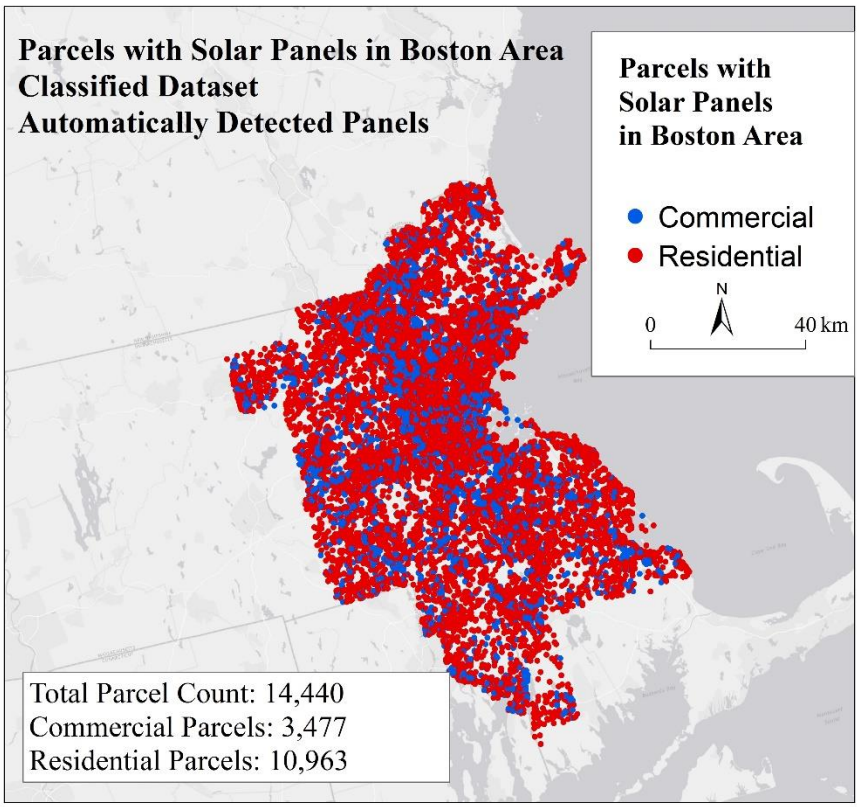


(a)

(b)

(c)

Figure 6. (a) Classification results for all parcels containing automatically detected solar panels in the DC area; (b) closeup of residential area with accurate classifications in DC; (c) closeup of commercial area with accurate classifications in DC.



(a)

(b)

(c)

Figure 7. (a) Classification results for all parcels containing automatically detected solar panels in the Boston area; (b) closeup of residential area with accurate classifications in Boston; (c) closeup of commercial area with accurate classifications in Boston.

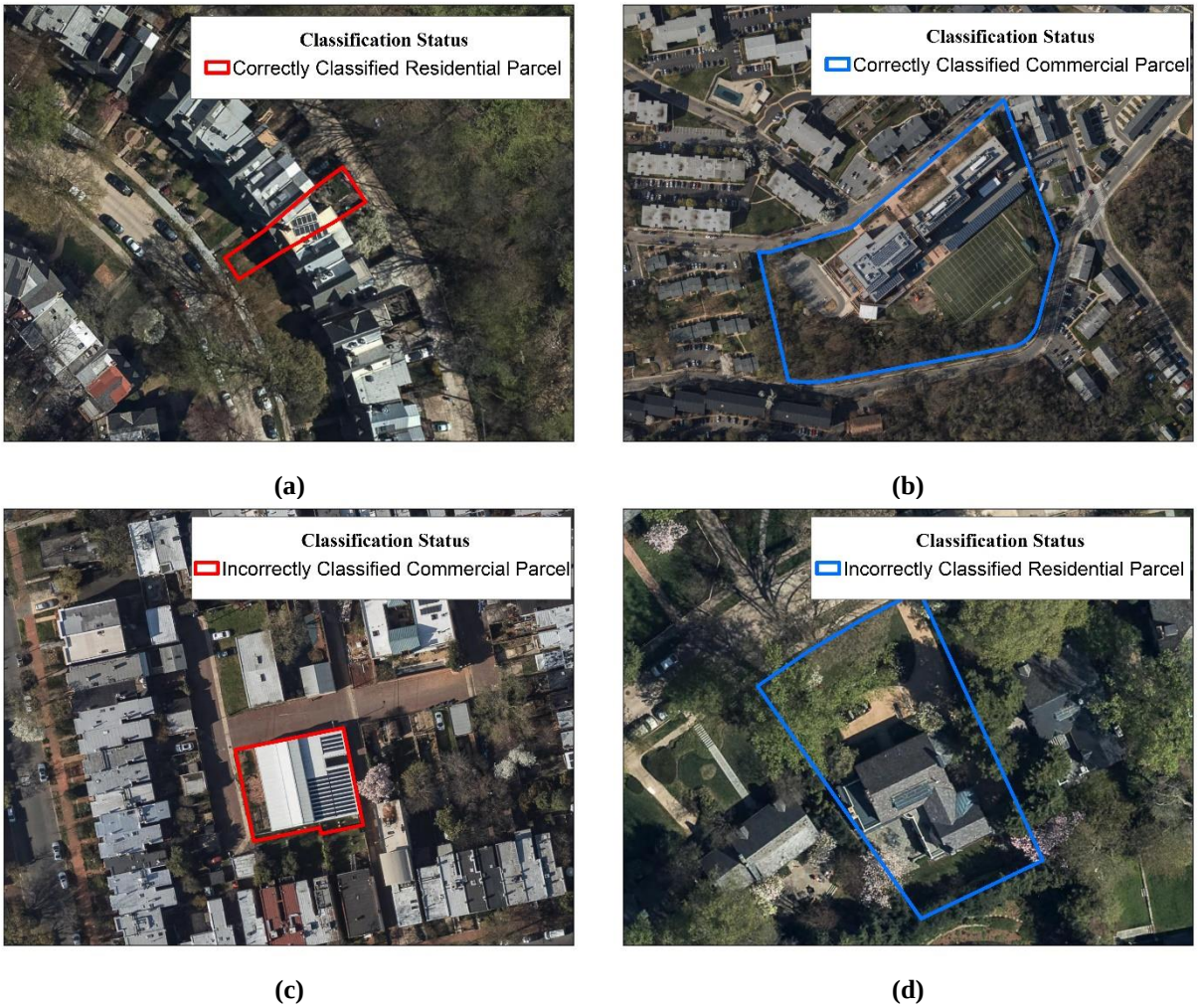


Figure 8. (a) Closeup of an image of a residential parcel in DC correctly classified as residential; (b) closeup of an image of a commercial parcel in DC correctly classified as commercial; (c) closeup of an image of a commercial parcel in DC classified incorrectly as residential; (d) closeup of an image of a residential parcel in DC incorrectly classified as commercial.

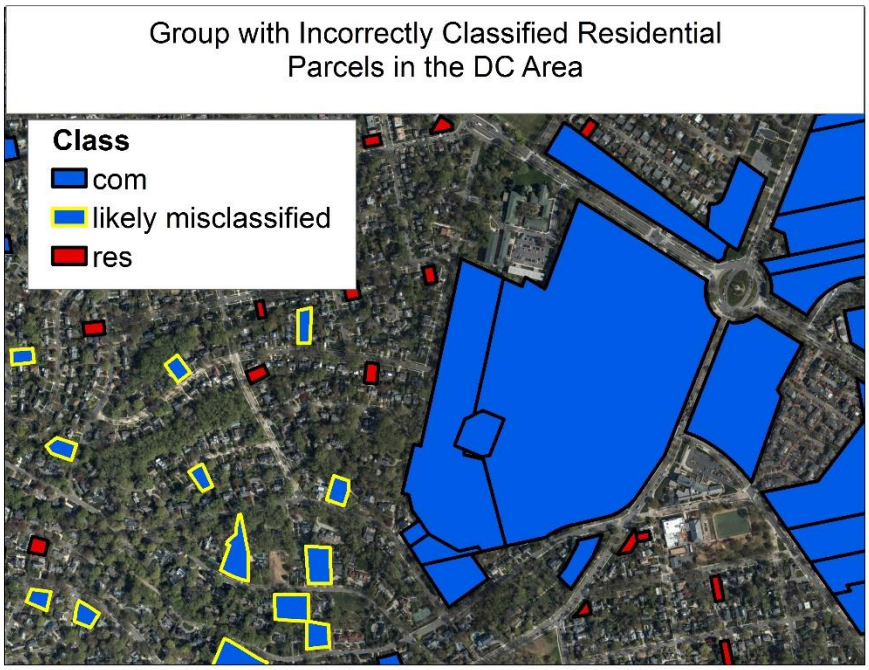
We also observe that the best performing algorithms in DC and Boston have relatively high residential and commercial accuracies. Table 2 shows that the best performing algorithm in the DC area has a mean residential accuracy of 98.31% and a mean commercial accuracy of 91.40%. Boston also has relatively high accuracies per class for its top algorithm, with a mean residential accuracy of 92.56% and a mean commercial accuracy of 98.91%. Interestingly, we notice that the best algorithm trained and tested on the DC area has a much better residential accuracy than commercial accuracy, while the best algorithm trained and tested on the Boston area has a much better commercial accuracy than residential accuracy. DC may have a lower commercial classification accuracy because there are several residential buildings in the DC testing/training set that have large parcel areas and are misclassified as commercial buildings. This trend can be seen in Figure 9(a), which displays a group of parcels in a suburban neighborhood with large residential parcels that have been misclassified as commercial parcels. Figure 9(b), on the other hand, shows a group of smaller parcels in a wealthy suburban neighborhood that have been classified correctly. To improve the algorithm, it may be worth adding more training samples of large residential parcels. However, even if this is done there is still a chance that these types of residential buildings are simply too difficult to distinguish when using parcel area as a feature. Boston, on the other hand, may have a lower residential classification accuracy because there may be several commercial buildings that

have small solar panel areas that closely resemble solar panel areas of residential parcels. To improve the algorithm, it may be worth adding more training samples of commercial buildings with small solar panels. However, even if this is done there is still a chance that commercial parcels in the Boston area simply have smaller solar panels, which may still make these types of commercial buildings difficult to distinguish when using total solar panel area as a feature. In addition, many of the false positive detections in the Boston area are very small panels, even on commercial buildings, which likely leads to more misclassifications of commercial buildings. Because the second best performing algorithm and feature combination for the Boston area is only 0.23% less than the best performing algorithm, and depends on the parcel area, it may be a better choice when classifying the entire region in order to avoid misclassification based on small solar panels coming from false positive detections.

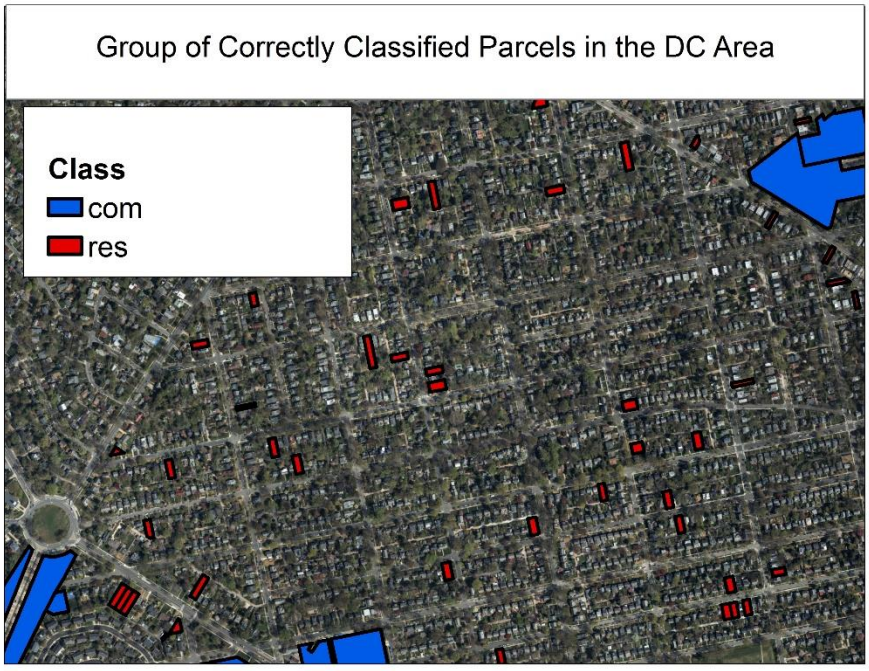
Table 2. Overall mean accuracies and standard deviations along with mean accuracies and standard deviations of commercial and residential classes for the algorithms and feature combinations with highest overall mean accuracies in each study area. Mean accuracies are the first numbers reported in the table, followed by standard deviations in parentheses.

Study Area	Overall Accuracy	Residential Accuracy	Commercial Accuracy
DC	0.9614 (0.0244)	0.9831 (0.0215)	0.9140 (0.0613)
Boston	0.9565 (0.0202)	0.9256 (0.0359)	0.9891 (0.0153)
Boston (Trained with DC)	0.9158 (0.0277)	0.8535 (0.0425)	0.9929 (0.0153)

Additionally, in order to examine how well an algorithm trained on one area might perform on another area, we applied the algorithms trained with the DC training data to the Boston testing data. In addition, we classified the entire Boston parcel dataset using the DC-trained algorithm. Figure 10(a) shows the feature combinations with the highest mean accuracies for each residential/commercial classification algorithm for the Boston area trained with the DC data. The LR algorithm trained on the total panel area, number of panels, and number of employees marginally out-performed the best RF and NN algorithms, with an average accuracy of 91.58% and an SD of 2.77%. Though the best performing algorithm is different (changing from NN to LR), it depends on the same features. In addition, the overall mean accuracy noticeably drops from 95.65% to 91.58%, the residential accuracy drops from 92.56% to 85.35%, and the commercial accuracy very slightly increases from 98.91% to 99.29%. The residential accuracy likely drops even lower when Boston is trained with DC data because the commercial parcels in the DC training set probably have larger solar panels than those in Boston, causing even more commercial parcels in Boston to be classified as residential. In addition, many of the false positive solar panel detection shapes in the Boston area are very small, even on commercial buildings, which likely adds to the misclassifications of commercial buildings when using the DC-trained algorithm on Boston. This trend can be seen in Figure 11. Figure 11(a) shows a group of unclassified commercial Boston parcels, Figure 11(b) shows the same group of Boston parcels correctly classified as commercial when the Boston algorithm was trained with the Boston Data, and Figure 11(c) shows the same group of parcels with three parcels that were misclassified as residential when the Boston algorithm was trained with the DC data. Though we do lose accuracy when the DC-trained algorithm is applied to Boston, depending on the resources available to train a model for a new city, these new accuracies might be "good enough." For example, if a researcher had a low budget project and had already developed a model for a city that was as similar to the new city under consideration as the DC study area is to the Boston study area, he or she might decide that a 5% drop in overall accuracy is better than spending more resources or time building a new model. Thus, overall, these new results show that models can likely be applied to other cities but must be done with a consideration of the similarities of the cities as they relate to the features used in the algorithm and the overall accuracy goals of the project.

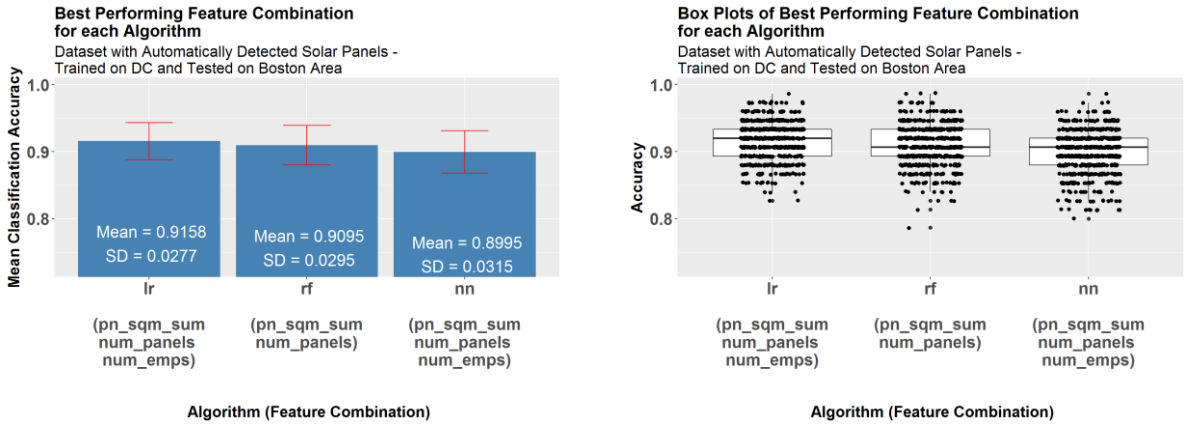


(a)



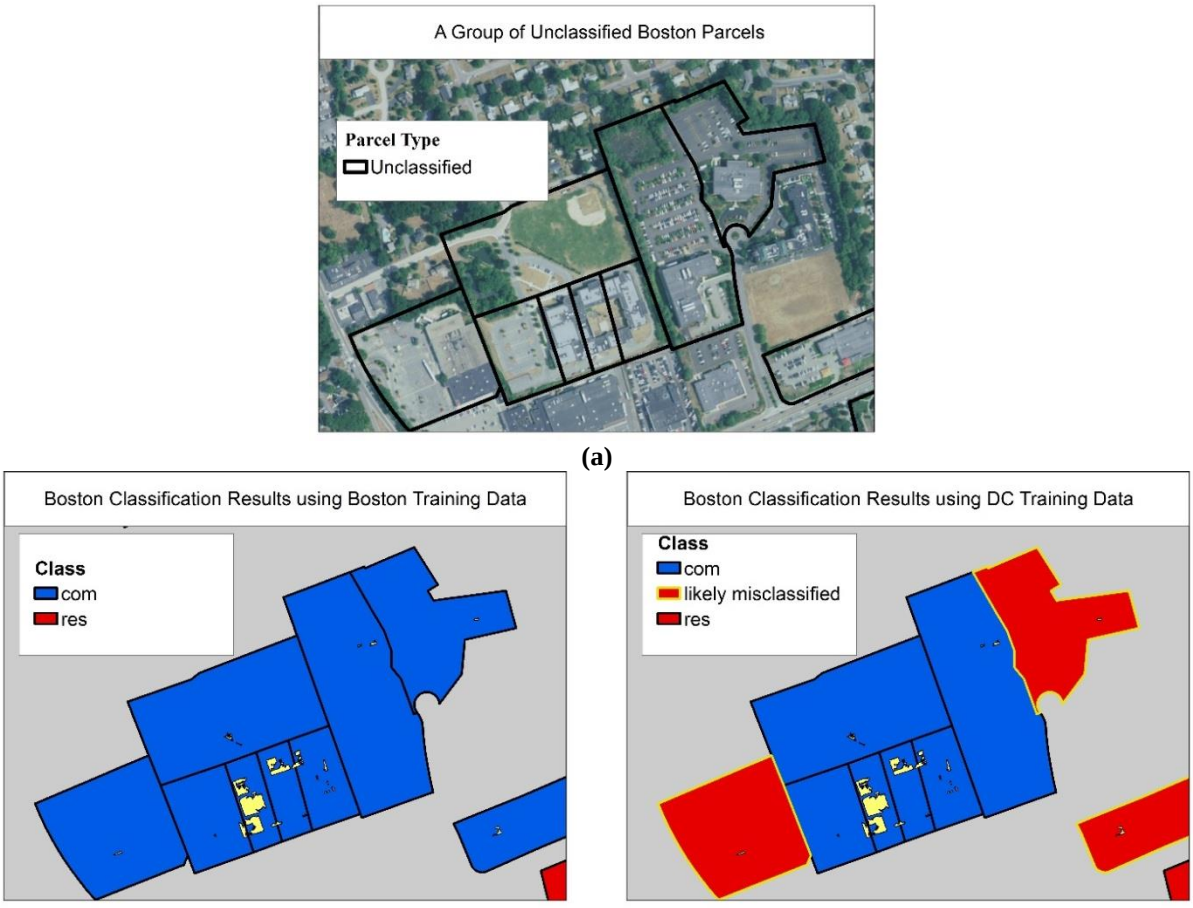
(b)

Figure 9. (a) Group of correctly classified residential and commercial parcels, along with some incorrectly classified residential parcels, in the DC area; (b) group of correctly classified residential and commercial parcels in the DC area.



(a) **(b)**

Figure 10. (a) Bar charts with SD error bars for the best performing set of features for each of the three algorithms that were trained using the DC labeled data and tested using the Boston testing data; (b) box plots with jittered accuracies for the best performing set of features for each of the three algorithms that were trained using the DC labeled data and tested using the Boston testing data.



(a)

(b) **(c)**

Figure 11. (a) A group of unclassified Boston parcels in a commercial area; (b) the same group of Boston parcels classified using the Boston training data; (c) the same group of Boston parcels classified using the DC training data.

Thus far we've only discussed the best performing algorithms and feature combinations. However, one may have noticed that the second and third best algorithms and feature combinations have mean accuracies that are only slightly lower (usually not more than a 1% difference) than the best performing models. Although most of the mean accuracies are likely significantly different due to the large number of simulations we conducted, they are not very different in a practical sense and many of the individual simulated accuracies for each algorithm were actually the same. More specifically, Figures 4(b), 5(b), and 10(b) all show box plots of the best three models that were trained/tested on DC, trained/tested on Boston, and trained on DC and tested on Boston, respectively, along with accuracies that were randomly adjusted (or jittered) by up to 0.2 on the x-axis and 0.001 on the y-axis to more easily show the spread of the accuracies. Each of these plots shows that many simulations resulted in the same accuracy for all of the models and had a similar range of values. Similarly, Tables B2, B3, and B4 display the overall mean accuracies and standard deviations for all models trained/tested on DC, trained/tested on Boston, and trained on DC and tested on Boston, respectively, where the rows are sorted by the best performing algorithm for each testing/training combination. For each of the testing/training combinations and each of these ranked models, we again see very small differences between consecutive algorithm/feature combinations. Therefore, researchers might choose a model not necessarily based on its accuracy but on whether or not the algorithm is "good enough" for the specific area or level of accuracy they are comfortable with. For example, one who prefers simple models that do not require as much data preparation might prefer to go with the RF model trained only on the DC parcel area when developing the DC model, since this model (Table B2) still has an accuracy of 91.14%, which is only about 5% lower than the best performing model. Thus, choosing a model is both an art and a science, as sometime sacrifices in accuracy are preferred to researchers with different comfort levels and preferences.

Because the detected panels were often fuzzy outlines of the actual panels, we also tested the algorithm on a subset of manually detected solar panels, or solar panels outlined by hand, in the DC and Boston areas. This was done to better understand the impact of the fuzzy detections on the algorithms' performances. Figure A1(b) shows examples of two manually detected solar panels and their parcel boundaries, while Figures A2(a) and A2(b) show the labeled subsets in greater detail. Figures A3(a) and A4(a) show the best performing sets of features for each of the three tested algorithms applied to the DC and Boston area datasets with manually detected panels. In addition, Figures A3(b) and A4(b) show box plots for each of the three tested algorithms applied to the DC and Boston area datasets with manually detected panels, along with accuracies that were randomly adjusted (or jittered) by up to 0.2 on the x-axis and 0.001 on the y-axis to more easily show the spread of the accuracies. Furthermore, Tables B4 and B5 display the overall mean accuracies and standard deviations for all models trained/tested on the DC and Boston area datasets with manually detected panels, respectively, where the rows are sorted by the best performing algorithm for each testing/training combinations. Following the same protocol outlined in this section, we found that the RF algorithm marginally out-performed the LR and NN algorithms for both the DC and Boston areas, with an average accuracy of 95.12% for the DC areas and 92.04% for the Boston area. These accuracies are very similar to those obtained from our candidate set of automatic detections, so we can conclude that the fuzzy outlines do not drastically affect our accuracy results.

5. PREVIOUS WORK IN SOLAR PANEL MAPPING

Now that we have presented our methodology and results for determining which parcels containing solar panels are residential, we provide a list of recommended variables and modeling techniques for using the classified parcel data to identify attributes that motivate the installation of residential solar panels in the Boston and DC areas.

Residential homeowners that install solar panels may be motivated by several factors. Overall, we can classify these factors into five groups: the environment, financial savings, peer influence, personal

comfort, and personal energy security groups. In this section, we present a summary of previous studies to understand these factors and others that might have motivated homeowners to adopt energy-saving technology and/or solar panels.

Schelly (2014) references three models that have been developed to understand the human decision-making process as it relates to energy technology adoption. The first and arguably the most popular understanding of technology adoption is based on the body of work within the field of social psychology that focuses on **environmental motivations** for purchasing behavior (Stern, 1992). Solar technology is frequently referred to as a "green"* source of energy, and scholars and policy-makers assume that adopting green technology is shaped by environmental values that can be encouraged through green power marketing initiatives (Schelly, 2014). Several authors have studied the effects of environmental concern on market response for green products by analyzing detailed survey data. For example, Clark, Kotchen, and Moore (2003) used a logit model† to analyze data from a mail survey of consumers who had the opportunity to participate in a green electricity program that required individuals to pay a fee to lease at least one 100 W block of solar electricity service from a centralized facility in Michigan. They found that willing participants were primarily motivated by biocentric intentions, or intentions focused on altruism toward the environment. Whitmarsh and O'Neill (2010) conducted a similar study by using a postal survey with 551 participants to determine, through a linear regression analysis,† the influence of pro-environmental self-identity across a range of behaviors and found that self-identity was a significant behavioral determinant for carbon-offsetting behavior. Several other researchers conducted similar surveys and analyses to understand the connection between pro-environmental attitudes and purchasing decisions (Pickett-Baker and Ozaki, 2008; Schlegelmilch, Bohlen, and Diamantopoulos, 1996; Young, Hwang, McDonald, and Oates, 2010). Although all of these studies report correlations between biocentric intentions and attitudes, the strength or existence of these correlations varies, suggesting that environmental concerns and attitudes are not sufficient to explain the adoption of residential photovoltaic systems (Schelly, 2014).

A second theory aims to understand the decision to adopt a new technology in terms of **economic rationality** (Schelly, 2014). Microeconomic theory suggests that a decision to invest results from rationally calculated decision-making (Shwom & Lorenzen, 2012). Given the up-front cost, scholars in this field believe homeowners arrive at the decision to adopt rooftop PV after conducting rational cost-benefit analyses, usually related to estimated returns on investment. Many current policies, including tax credits, incentives, and rebates, are based on this economic rationality model and assume that lowering up-front investment costs will increase consumer purchases. Many scholars have studied the effects of tax credits and other policy-driven economic incentives on green energy and solar purchases. For instance, Carpenter and Chester (1984) used log-linear contingency tables† based on a survey of 8,369 mail questionnaires to determine the extent to which conservation decisions were contingent on the availability of tax credits and found that particular groups, such as owners of older, conventional homes, were most likely to take advantage of tax credit conservation programs. In addition, Durham, Colby, and Longstreth (1988) applied a *probit* model† to survey data on 2,751 solar installers and non-installers in the western states and found that state tax credits, along with other variables, were significantly related to the adoption of solar water heating devices. Hassett and Metcalf (1995) used a *logit* model† based on panel data from the Michigan Tax Research Database to understand the impact of government tax policies on residential energy conservation investment. In their study, they define the tax price of one dollar's worth of

* The term "green" often has slightly different meanings to different authors reviewed in this study. Green electricity technology is typically used to describe electricity that is generated from solar, wind, or other renewable energy sources. Green products, on the other hand, are often thought to more generally contribute to sustainable patterns of consumption. Because of the subjectivity of the word, readers are encouraged to use caution when interpreting its meaning in the reviewed literature.

† Model, algorithm, or other method for analyzing or predicting data. The *logit* and *probit* models predict dichotomous categorical outcomes, such as the decision of a homeowner to install a solar panel system or not, whereas the other methods typically predict counts or other quantitative outcomes, such as the number of solar panel systems installed in a neighborhood.

investment as one minus the marginal tax rate. They found that a 10% decline in the tax price for energy investment led to a 24% increase in the probability of making the investment.

A third model for understanding technology adoption is by viewing new technology as innovations whose adoptions can be understood through the "**Diffusion of Innovations**" theory. The Diffusion of Innovations theory proposes that the adoption of innovations is a result of "communication through certain channels, over time, and among members of a social system" (Faiers & Neame, 2006). Models based on this theory commonly consider the demographic characteristics of adopters, their perceptions of risk, and the observability of adoption as a system for diffusion (Schelly, 2014). Richter (2013) found that since the observability of PV systems increases as their density increases, policies based on the Diffusion of Innovations theory might encourage a high number of new installations in neighborhoods that are inclined, based on their demographics, to participate in communication channels. For example, Richter (2013) used an econometric model* to analyze neighborhood-level demographic and PV system installation data and found that higher educated neighborhoods installed more PV systems than neighborhoods with, on average, lower educated populations (Richter, 2013). Additionally, she found a significant relationship between the number of solar PV systems previously installed in an area and the number of PV systems installed 3 months later. Taken together, these results may indicate that higher educated neighborhoods are more inclined to promote the spread of technology within their neighborhoods. Similarly, Bollinger and Gillingham (2010) used daily solar panel adoption data from three investor-owned utility regions in California to perform a zip-code-level analysis of solar panel adoption using an ordinary least-squares regression* containing time-based variables. They found significant evidence that one household's choice to adopt PV systems may be influenced by other nearby households' previous decisions to adopt PV systems (Bollinger and Gillingham, 2010). Graziano and Gillingham (2014) developed a slightly different custom model that used socioeconomic, demographic, political affiliation, built environment, and other variables to predict the demand for residential PV systems within block groups. They found that demographic and socioeconomic variables, such as median household income and median age, all significantly impacted PV adoption. They also found that higher numbers of previously installed systems significantly increased the number of subsequent adoptions nearby. Overall, the body of work related to the Diffusion of Innovations theory suggests that early adopters influence each other through networks of communication and are typically "younger, more highly educated, have a higher income and occupational status, and are earlier in the family life cycle than non-adopters" (Schelly, 2014).

5.1 RECOMMENDED DATA AND METHODS

Although studies related to all three green technology adoption models have proven valuable in specific contexts, most have been based on unique surveys unavailable for many areas of the world. This has limited residential PV adoption studies to only those regions with available data. Despite these limitations, however, we believe it is still possible to use ORNL's growing capabilities in automated solar panel mapping to gain valuable insights related to PV adoption. More specifically, because the Diffusion of Innovations theory has been linked to demographic and socioeconomic variables reported by the census for small neighborhoods, it is possible to explore solar adoption patterns by modeling the relationship between demographic variables and PV system installation counts at the neighborhood level. Furthermore, since our current study area includes multiple cities and states, with varying economic- and environmental-related policies, it is possible to add variables indicating whether neighborhoods have had an opportunity to participate in rebate or green-marketing programs.

* Model, algorithm, or other method for analyzing or predicting data. The *logit* and *probit* models predict dichotomous categorical outcomes, such as the decision of a homeowner to install a solar panel system or not, whereas the other methods typically predict counts or other quantitative outcomes, such as the number of solar panel systems installed in a neighborhood.

The literature summarized in the previous section suggests that there are two general modeling approaches taken to understand solar adoption motivations. The first approach combines survey data with binary regression models to predict outcomes that fall into two categories, such as the decision of a homeowner to install or not install a solar panel system. Most of the studies based on neighborhood-level data, on the other hand, take a different, second approach by using either existing or custom modeling approaches based on time-dependent data to predict the number of solar panels installed in neighborhoods. Interestingly, in the scope of this literature review, we have not come across papers that predict the percentage, rather than count, of PV systems in a particular neighborhood. This may be the case because there are several regressions available to choose from based on count data that often meet the assumptions of solar panel data more easily than those available for percentage-based dependent variables. We do not have detailed survey data that easily lend themselves to using a model to predict a binary outcome, so it makes more sense to take the second approach by predicting neighborhood-level solar panel system counts using a regression model. Furthermore, because we do not have data over multiple time periods, it is appropriate to use a more standard regression model rather than a custom regression model that includes time-based variables. In light of these considerations, we recommend using a spatial zero-inflated negative binomial regression model using census block groups as the unit of analysis. This type of model is likely most appropriate because negative binomial regression models work well for predictor variables that are count data. Furthermore, adding a zero-inflated component to the model will likely lead to a better fit because initial results indicate that several block groups will likely have zero solar panel detections. Additionally, including a spatial component is important because block groups that are close to one another are likely more related than block groups that are further away. Using census block groups as the unit of analysis is recommended because it is the smallest geographic unit with a wide range of available socioeconomic and census data. The following subsections summarize our recommended independent variables in greater detail.

5.1.1 Non-Policy Variables Based upon the Diffusion of Innovations Theory

This section provides a list and description of recommended census variables based on Schelly's claim that early adopters are typically "younger, more highly educated, have a higher income, and are earlier in the family life cycle than non-adopters" (2014). It also includes recommended control variables, based on work done by Bollinger and Gillingham (2010) and Graziano and Gillingham (2014), related to the number of occupied housing units and percentage of owner-occupied housing units in a block group. All variables are available at the block-group level from the 2008–2012 American Community Survey (US Census Bureau, 2012) and are summarized in Table 3.

i. **Median household income:** Median household income in the past 12 months (in 2012 inflation-adjusted dollars). Corresponds to variable ID b19013001 from the 2008–2012 American Community Survey (ACS) summary tables.

ii. **% of homeowners (25–44):** Percentage of householders of owned homes who are between 25 and 44 years old. Computed by summing over variable IDs b25007004 and b25007005 from the 2008–2012 ACS summary tables and then normalizing by variable b25003002.

iii. **% of Pop 25+ with at least college degree:** Percentage of population 25 years and older who have obtained at least a bachelor's degree. Computed by summing over variable IDs b15003022, b15003023, b15003024, and b15003025 from the 2008–2012 ACS summary tables and then normalizing by variable b15003001.

iv. **% of households with children 6+:** Percentage of households with children over 5 years old. Computed by summing over variable IDs b11003005, b11003006, b11003012, b11003013, b11003018, and b11003019 from the 2008–2012 ACS summary tables and then normalizing by variable b25003001.

v. **Percentage of owner-occupied housing units (control variable):** Percentage of occupied housing units that are owner occupied. Calculated by normalizing the 2008–2012 ACS variable ID b25003002 by variable ID b25003001.

vi. **Number of occupied housing units (control variable):** Number of occupied housing units. Corresponds to variable IDs b25003001 in the 2008–2012 ACS summary tables.

Currently, it is difficult to include variables directly related to the part of the theory that suggests neighbors influence each other's decisions to install solar panels because the ORNL parcel dataset does not yet have a time component. However, one might be able to include these variables in the future if the imagery becomes available for multiple years. Furthermore, even though we do not know the order of PV system installations, one might still consider including a variable related to PV system density as a proxy to measuring peer influence. For example, one might consider adding a variable that describes the average density of potential solar-panel-containing parcels, per block group, within a 1 mile radius of each potential solar-panel-containing parcel in that block group. Although we do not know exactly when the surrounding parcels installed solar panels, and therefore cannot quantify factors such as the rate of diffusion, we can assume that if they are densely concentrated, they probably did influence each other.

Table 3. Non-Policy Variables Based upon the Diffusion of Innovations Theory.

Variables	ACS Variable IDs
Independent Variables of Interest	
Median Household Income	b19013001
% of Homeowners (25–44)	b25007004 b25007005 b25003002
% of Pop 25+ with at Least College Degree	b15003022 b15003023 b15003024 b15003025 b15003001
% of Households with Children 6+	b11003005 b11003006 b11003012 b11003013 b11003018 b11003019 b25003001
Independent Controls	
% of Owner-Occupied Housing Units	b25003002 b25003001
Number of Occupied Housing Units	b25003001

It is important to note that real-world behavior is not likely to conform solely to one of the three models in isolation but rather may be driven by a combination of factors. Demographic variables such as those above would likely also be useful regressors in models designed to detect environmental and

economically rational behavior. In addition, non-policy variables related to the economic-rationality theory, such as retail electricity costs and PV system and installation costs, would likely add value to the model as well. Furthermore, one would probably want to add, if available, environmental variables, such as the average percentage of shade (or sunshine) in each block group.

5.1.2 Policy Variables

As noted above, policies can amplify or dampen PV installation behaviors, and policy analysis often employs regression models that encode policies as well as other descriptors of the environment and economy to determine the magnitude of this effect.

Policies that improve access to or knowledge of solar options may boost drivers to install rooftop solar suggested by both the environmental installation theory and the behavioral theory. Such policies include education programs and information campaigns that disseminate knowledge about the environmental performance of various power sources and the availability of distributed options; mandates that utilities offer renewable power options, including connection of distributed solar systems; and generation disclosure ("green labeling"), in which utilities are required to provide customers with information about power sources and emissions. Variation in the existence of these policies geographically or over time could be exploited to determine their effects.

Many policy incentives effectively reduce the cost of solar systems, so the following policies may increase installations for customers responding on an economic basis: tax deductions or credits, rebates, net metering compensation programs, grants, accelerated depreciation, and assisted financing. Geographic or temporal variation in these mechanisms could be modeled to determine response to policy drivers.

6. LESSONS LEARNED

In this section, we summarize some of the lessons learned during the project execution, which provides directions for future studies.

6.1 AUTOMATIC DETECTION OF BUILDINGS WITH SOLAR PANELS

Although we are confident in the deployed model, the assessment is based on the validation data and the convergence of the training process, which is determined by the training data. At large-scale image inferencing, there is an important factor that ultimately will have impact on the results—the consistency of image quality and the radiometric characteristics of the images covering the extended geographical area. If the image to be processed is significantly different from the training data/validation data, in terms of the color tone and image quality, from a supervised learning point of view, we can expect unsatisfactory performance for such images, as the trained model has not learned from such data before. Sources of inconsistency in radiometric characteristics include multi-temporal data collection and various image preprocessing procedures. In addition, some images (especially for Boston) exhibit relatively low color contrast and lower spatial resolution. Even with our bare eyes, we cannot identify the solar panels in the images easily. In this case, the quality of the images does limit the performance of the algorithm.

In the past, we have successfully exploited the use of auxiliary information to improve the solar panel detection results as part of the post-detection data processing, such as the size of the solar panels and road networks. In these two cases, they are useful in reducing the number of false positive detections. In future studies, we would like to explore the use of additional auxiliary datasets such as building layers, height information from the LiDAR data, and others for post-detection processing. From a supervised learning perspective, the more training datasets that we have, the better the generalization quality of the developed algorithm will be. Therefore, we also would like to generate additional representative datasets from

various parts of the country so that we can capture the variability in building roof material and roof design types.

6.2 RESIDENTIAL/COMMERCIAL CLASSIFICATION OF PARCELS CONTAINING SOLAR PANELS

Though the commercial/residential classification results have shown promise for automatically classifying buildings as residential or commercial, we have also observed a few notable opportunities for improvement. First, since the DC algorithm seemed to misclassify many of the large residential parcels, one might improve the algorithm by adding more training samples of large residential parcels. Boston, on the other hand, may have a lower residential classification accuracy because there may be several commercial buildings that have small solar panel areas that closely resemble solar panel areas of residential parcels. To improve the algorithm, it may be worth adding more training samples of commercial buildings with small solar panels. In addition, since many of the automatic detections that were false positives were small in nature, they likely caused misclassifications for commercial buildings that had small false positive detections. Thus, it might be better to use the second best performing NN algorithm that depends on parcel area, total panel area, number of panels, and number of employees for the Boston area, as this algorithm still has a high accuracy of 95.21%. As the detection algorithm improves with time, and the shape and size of the solar panels becomes more reflective of the true solar panels, one could again rely on the better performing algorithm that does not consider parcel area.

We also learned a few lessons related to the use of open source parcel data. Though open source parcel data is certainly valuable, it became clear through the QA/QC portion of this project that parcel data often varies city to city and must be thoroughly examined before use. Though we did carefully review the documentation before utilizing the data, there were a few tricky exceptions that were difficult to find. For example, in our final QA/QC process, we discovered that a small percentage of parcel polygons were duplicated in the MassGIS Standardized Parcel Data (MassGIS, 2013). Fortunately, once we knew this was a potential issue, we were able to automatically identify and delete these duplicate polygons. In addition, when deciding whether to include New Hampshire parcel data (New Hampshire's Statewide Geographic Information System (GIS) Clearinghouse, 2014) in our study area, we found that a few counties reported parcel data using line features rather than polygon features. Since automatically and accurately converting these features to polygons proved difficult and time consuming, we opted to leave out the New Hampshire area during this project period. We recommend that future work anticipate these types of issue and budget their time and money accordingly.

7. REFERENCES

Bollinger, B., and Gillingham, K. (2010). "Environmental preferences and peer effects in the diffusion of solar photovoltaic panels," Stanford Working Paper.

BostonGIS. (2016). Parcels 2016 Data Full. Retrieved from http://bostonopendataboston.opendata.arcgis.com/datasets/f3d274161b4a47aa9acf48d0d04cd5d4_0.

Boulesteix, A. L., Janitza, S., Kruppa, J., and König, I. R. (2012). "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6), 493–507.

Breiman, L., Cutler, A., Liaw, A., and Wiener, M. (2011). Package 'randomForest'. Software available at URL: http://stat-www. berkeley. edu/users/breiman/RandomForests.

Carpenter, E. H., and Chester, S. T. (1984). "Are federal energy tax credits effective? A Western United States survey," *The Energy Journal*, 5(2), 139–149.

Clark, C. F., Kotchen, M. J., and Moore, M. R. (2003). “Internal and external influences on pro-environmental behavior: participation in a green electricity program,” *Journal of Environmental Psychology*, 23(3), 237–246.

DCGIS Open Data. (2015). Building Footprints. Retrieved from http://opendata.dc.gov/datasets/building-footprints.

DCGIS Open Data. (2017). Common Ownership Lots. Retrieved from http://opendata.dc.gov/datasets/1f6708b1f3774306bef2fa81e612a725_40?geometry=-77.185%2C38.934%2C-77.022%2C38.996&mapSize=map-maximize&uiTab=Table.

District of Columbia Office of the Chief Technology Officer. (2015). District of Columbia – Classified Point Cloud LiDAR. Retrieved from https://aws.amazon.com/public-datasets/dc-lidar/.

Durham, C. A., Colby, B. G., and Longstreth, M. (1988). “The impact of state tax credits and energy prices on adoption of solar energy systems,” *Land Economics*, 64(4), 347–355.

Faiers, A., and Neame, C. (2006). “Consumer attitudes towards domestic solar power systems,” *Energy Policy*, 34(14), 1797–1806.

Fritsch, S., Guenther, F., and Guenther, M. F. (2016). Package ‘neuralnet’.

Google Maps. (2017). Retrieved from https://www.google.com/maps.

Graziano, M., and Gillingham, K. (2014). “Spatial patterns of solar photovoltaic system adoption: the influence of neighbors and the built environment,” *Journal of Economic Geography*, 15 (4), 815–839.

Hassett, K. A., and Metcalf, G. E. (1995). “Energy tax credits and residential conservation investment: evidence from panel data,” *Journal of Public Economics*, 57(2), 201–217.

Kaastra, I., and Boyd, M. (1996). “Designing a neural network for forecasting financial and economic time series,” *Neurocomputing*, 10(3), 215–236.

Let’s Go Solar. (2017). Residential vs. Commercial Solar Systems. Retrieved from https://www.letsgosolar.com/solar-panels/commercial-solar-panels/.

MassGIS. (2017). Building Structures. Retrieved from https://docs.digital.mass.gov/dataset/massgis-data-building-structures-2-d.

MassGIS. (2013). Standardized Parcel Data. Retrieved from http://massgis.maps.arcgis.com/apps/View/index.html?appid=4d99822d17b9457bb32d7f953ca08416.

New Hampshire’s Statewide Geographic Information System (GIS) Clearinghouse. (2014). New Hampshire Parcel Mosaic. Retrieved from http://www.granit.unh.edu/data/downloadfreedata/category/databycategory.html

Peng, C. Y. J., Lee, K. L., and Ingersoll, G. M. (2002). “An introduction to logistic regression analysis and reporting,” *The Journal of Educational Research*, 96(1), 3–14.

Pickett-Baker, J., and Ozaki, R. (2008). “Pro-environmental products: marketing influence on consumer purchase decision,” *Journal of Consumer Marketing*, 25(5), 281–293.

Pitney Bowes. (2010). U.S. Business Points Data. Available for purchase at https://www.pitneybowes.com/us/data/poi-database/business-location-data.html.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/.

Richter, L. L. (2013). “Social effects in the diffusion of solar photovoltaic technology in the UK,” Cambridge Working Paper.

Schelly, C. (2014). “Residential solar electricity adoption: what motivates, and what matters? A case study of early adopters,” *Energy Research & Social Science*, 2, 183–191.

Schlegelmilch, B. B., Bohlen, G. M., and Diamantopoulos, A. (1996). “The link between green purchasing decisions and measures of environmental consciousness,” *European Journal of Marketing*, 30(5), 35–55.

Shwom, R., and Lorenzen, J. A. (2012). “Changing household consumption to address climate change: social scientific insights and challenges,” *Wiley Interdisciplinary Reviews: Climate Change*, 3(5), 379–395.

Solar Energy Industries Association (SEIA). (2017). Solar Industry Data. Retrieved from https://www.seia.org/solar-industry-data.

Stern, P. C. (1992). “What psychology knows about energy conservation,” *American Psychologist*, 47(10), 1224.

US Census Bureau (2012). 2008–2012 American Community Survey (ACS). Retrieved from https://www.census.gov/programs-surveys/acs/.

US Census Bureau. (2013). 2013 LEHD Origin-Destination Employment Statistics (LODES). Retrieved from https://lehd.ces.census.gov/data/.

US Energy Information Administration (EIA). (2017). More than half of small-scale photovoltaic generation comes from residential rooftops. Retrieved from https://www.eia.gov/todayinenergy/detail.php?id=31452.

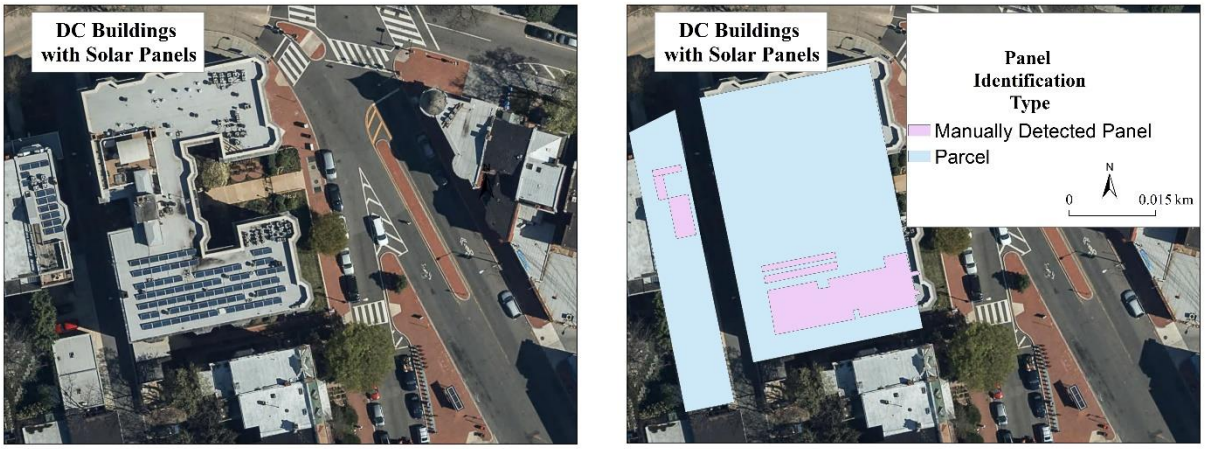
Whitmarsh, L., and O’Neill, S. (2010). “Green identity, green living? The role of pro-environmental self-identity in determining consistency across diverse pro-environmental behaviors,” *Journal of Environmental Psychology*, 30(3), 305–314.

Young, W., Hwang, K., McDonald, S., and Oates, C. J. (2010). “Sustainable consumption: green consumer behavior when purchasing products,” *Sustainable Development*, 18(1), 20–31.

Yuan, J., Yang, H.-H. L., Omitaomu, O. A., and Bhaduri, B. L. (2016). “Large-scale solar panel mapping from aerial images using deep convolutional networks,” 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, 2016, pp. 2703–2708. doi: 10.1109/BigData.2016.7840915.

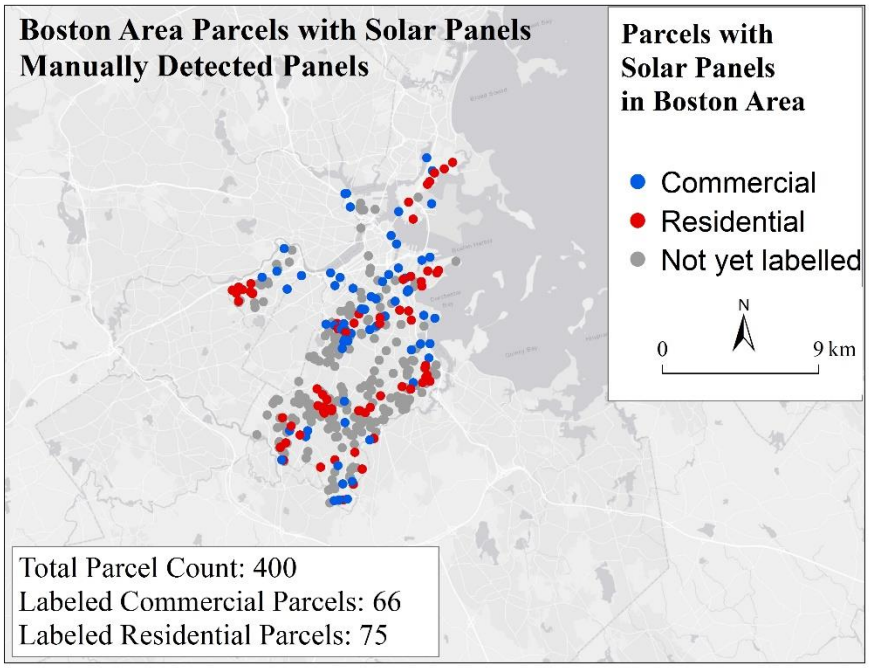
APPENDIX A. SUPPLEMENTARY FIGURES

APPENDIX A. SUPPLEMENTARY FIGURES

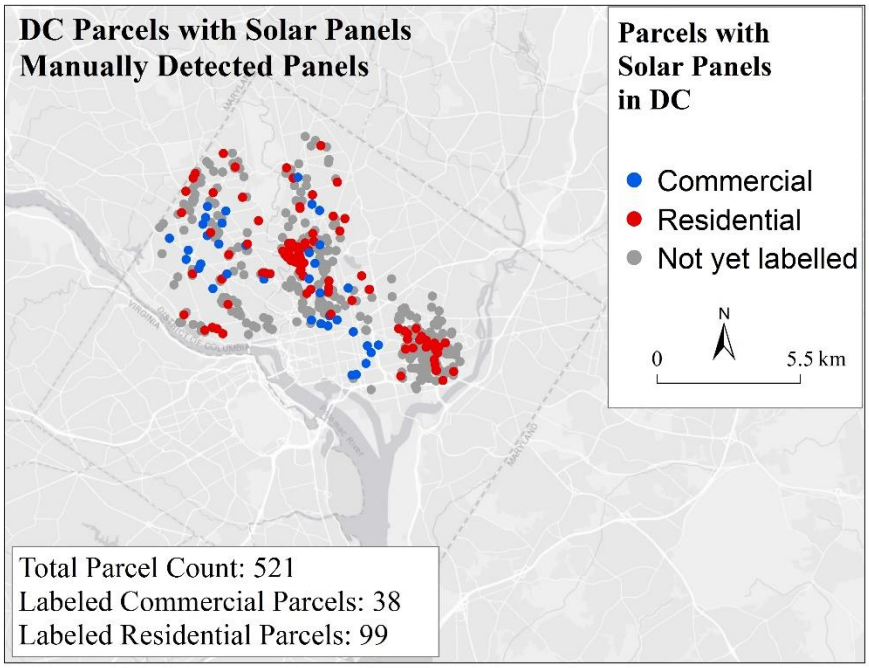


(a) **(b)**

Figure A1. (a) An image of two parcels with solar panels in the DC area; (b) the manually detected solar panels and their parcel boundaries, for the same buildings shown in (a).

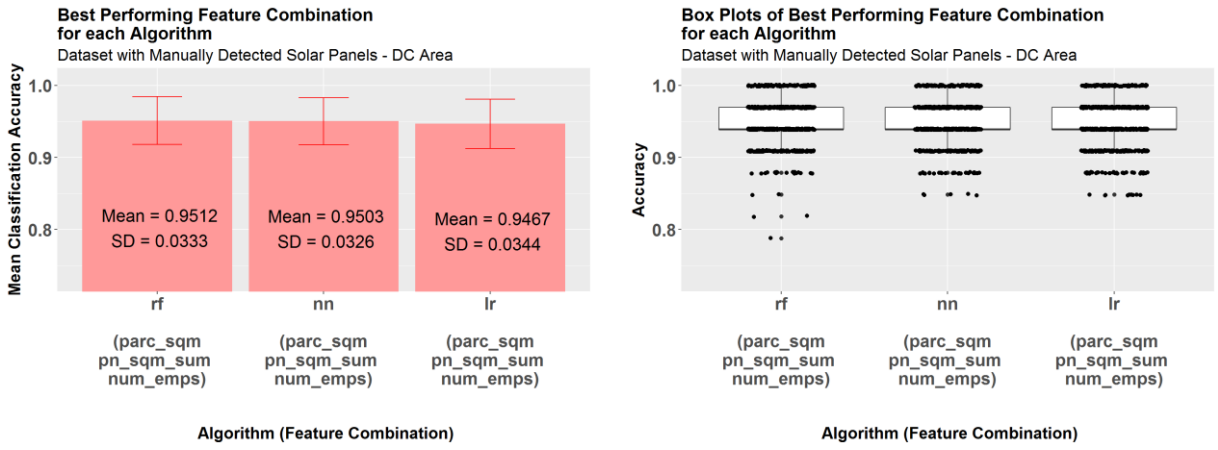


(a)



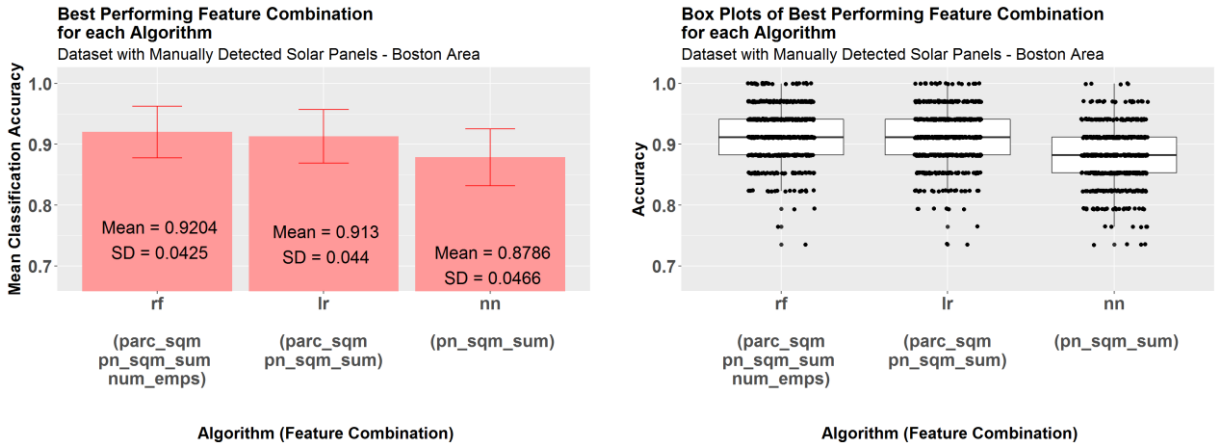
(b)

Figure A2. (a) The set of parcels containing manually detected solar panels in the DC area, with a subset of data manually classified as commercial and residential; (b) the set of parcels containing manually detected solar panels in the Boston area, with a subset of data manually classified as commercial and residential.



(a) **(b)**

Figure A3. (a) Bar charts with SD error bars for the best performing set of features for each of the three tested algorithms on the testing dataset with manually detected solar panels in the DC area; (b) box plots with jittered accuracies for the best performing set of features for each of the three tested algorithms on the testing dataset with manually detected solar panels in the DC area.



(a) **(b)**

Figure A4. (a) Bar charts with SD error bars for the best performing set of features for each of the three tested algorithms on the testing dataset with manually detected solar panels in the Boston area; (b) box plots with jittered accuracies for the best performing set of features for each of the three tested algorithms on the testing dataset with manually detected solar panels in the Boston area.

APPENDIX B. SUPPLEMENTARY TABLES

APPENDIX B. SUPPLEMENTARY TABLES

Table B1. Superset of features for classifying parcels containing solar panels as commercial and residential.

Feature Abbreviation	Feature Description	Source	Selected	Reason Selected or Rejected
parc_sqm	Area of parcel	Parcel Data (BostonGIS, 2016; MassGIS, 2013; DCGIS Open Data, 2017)	Yes	Consistently ranked high in mean accuracies
sum_bldare	Sum of building areas within parcel	Building Footprint Data (MassGIS, 2017; DCGIS Open Data, 2015)	No	Only sometimes ranked high in top mean accuracies; time-consuming to process data
max_hgtbld	Max height of building within parcel	DC Lidar Data (The District of Columbia Office of the Chief Technology Officer, 2015)	No	Only sometimes ranked high in top mean accuracies; time-consuming to process data
numbld	Number of buildings within parcel	Building Footprint Data (MassGIS, 2017; DCGIS Open Data, 2015)	No	Only sometimes ranked high in top mean accuracies; time-consuming to process data
pn_sqm_sum	Total area of all solar panels on roof (sum of individual solar panel areas)	ORNL Detected Solar Panels	Yes	Consistently ranked high in mean accuracies
num_panels	Number of solar panels on roof	ORNL Detected Solar Panels	Yes	Consistently ranked high in mean accuracies
pn_per_sum	Total perimeter of all solar panels on roof (sum of individual perimeters)	ORNL Detected Solar Panels	No	Highly related to pn_sqm_sum
pn_sqm_avg	Average solar panel area on the roof	ORNL Detected Solar Panels	No	Highly related to pn_sqm_sum
pn_per_avg	Average solar panel perimeter on the roof	ORNL Detected Solar Panels	No	Highly related to pn_sqm_sum
pn_sqm_max	Max solar panel area on roof	ORNL Detected Solar Panels	No	Highly related to pn_sqm_sum
pn_per_max	Max solar panel perimeter on roof	ORNL Detected Solar Panels	No	Highly related to pn_sqm_sum
pn_sqm_min	Min solar panel area on roof	ORNL Detected Solar Panels	No	Highly related to pn_sqm_sum
pn_per_min	Min solar panel perimeter on roof	ORNL Detected Solar Panels	No	Highly related to pn_sqm_sum

Table B1. Superset of features for classifying parcels containing solar panels as commercial and residential (continued).

Feature Abbreviation	Feature Description	Source	Selected	Reason Selected or Rejected
num_houses_bg	Number of houses in block group containing parcel	US Census Summary Table Data (US Census Bureau, 2012)	No	Consistently ranked low in mean accuracies
num_emps	Number of employees in block group containing parcel	LEHD Origin-Destination Employment Statistics (LODES) Data (US Census Bureau, 2013)	Yes	Consistently ranked high in mean accuracies
num_bus_bg	Number of businesses in block group containing parcel	Pitney Bowes Business Points Data (Pitney Bowes, 2010) (Suggested Source)	No	Data price not within project budget
dist_bus_m	Distance of closest business point	Pitney Bowes Business Points Data (Pitney Bowes, 2010) (Suggested Source)	No	Data price not within project budget
bdist_20m	Number of businesses within 20 meter radius of parcel	Pitney Bowes Business Points Data (Pitney Bowes, 2010) (Suggested Source)	No	Data price not within project budget
bdist_40m	Number of businesses within 40 meter radius of parcel	Pitney Bowes Business Points Data (Pitney Bowes, 2010) (Suggested Source)	No	Data price not within project budget
bdist_60m	Number of businesses within 60 meter radius of parcel	Pitney Bowes Business Points Data (Pitney Bowes, 2010) (Suggested Source)	No	Data price not within project budget
bdist_80m	Number of businesses within 80 meter radius of parcel	Pitney Bowes Business Points Data (Pitney Bowes, 2010) (Suggested Source)	No	Data price not within project budget

Table B2. Overall mean accuracy and standard deviations for all algorithms and feature combinations for the automatically detected solar panels in the DC area.

Feature Combination	RF*	LR	NN
parc_sqm, pn_sqm_sum, num_panels, num_emps	0.9614 (0.0244)	0.951 (0.0249)	0.9252 (0.0326)
parc_sqm, pn_sqm_sum, num_emps	0.9581 (0.0237)	0.9516 (0.0246)	0.9145 (0.0339)
parc_sqm, pn_sqm_sum, num_panels	0.9568 (0.0243)	0.9525 (0.0247)	0.9142 (0.0338)
parc_sqm, pn_sqm_sum	0.9525 (0.0248)	0.9526 (0.0244)	0.9038 (0.0349)
parc_sqm, num_panels, num_emps	0.9398 (0.0284)	0.923 (0.0331)	0.8768 (0.0364)
parc_sqm, num_emps	0.9362 (0.0287)	0.9325 (0.032)	0.8555 (0.0445)
parc_sqm, num_panels	0.9345 (0.0309)	0.9241 (0.0331)	0.8618 (0.0356)
pn_sqm_sum, num_panels, num_emps	0.9148 (0.0333)	0.9037 (0.0344)	0.9096 (0.0333)
parc_sqm	0.9114 (0.0374)	0.9299 (0.0304)	0.8495 (0.0509)
pn_sqm_sum, num_panels	0.9005 (0.0345)	0.8945 (0.0364)	0.8955 (0.0332)
pn_sqm_sum, num_emps	0.8995 (0.0375)	0.9033 (0.033)	0.8992 (0.033)
num_panels, num_emps	0.8546 (0.0366)	0.8666 (0.0362)	0.8632 (0.0362)
num_panels	0.8478 (0.0353)	0.8478 (0.0353)	0.8475 (0.0353)
pn_sqm_sum	0.7984 (0.048)	0.8898 (0.0337)	0.887 (0.0332)
num_emps	0.7295 (0.047)	0.7513 (0.0348)	0.7614 (0.0286)

*Algorithm with highest mean accuracy and feature combination. Rows are sorted by feature combination with highest to lowest mean accuracy based off of this column.

Table B3. Overall mean accuracy and standard deviations for all algorithms and feature combinations for the automatically detected solar panels in the Boston area.

Feature Combination	RF	LR	NN*
pn_sqm_sum, num_panels, num_emps	0.9537 (0.0212)	0.9476 (0.0221)	0.9565 (0.0202)
parc_sqm, pn_sqm_sum, num_panels, num_emps	0.95 (0.0216)	0.9439 (0.0227)	0.9521 (0.0219)
pn_sqm_sum, num_panels	0.9406 (0.0229)	0.9477 (0.0242)	0.9512 (0.0218)
parc_sqm, pn_sqm_sum, num_panels	0.9542 (0.0211)	0.9446 (0.0242)	0.9495 (0.0221)
pn_sqm_sum, num_emps	0.9353 (0.0232)	0.9466 (0.022)	0.9478 (0.0222)
parc_sqm, pn_sqm_sum, num_emps	0.9415 (0.0225)	0.9431 (0.0219)	0.9446 (0.0228)
pn_sqm_sum	0.899 (0.0289)	0.9411 (0.0231)	0.9411 (0.0233)
parc_sqm, pn_sqm_sum	0.9391 (0.0233)	0.9408 (0.0234)	0.9398 (0.0237)
parc_sqm, num_panels, num_emps	0.9177 (0.0282)	0.8871 (0.0326)	0.8769 (0.0353)
num_panels, num_emps	0.8627 (0.0341)	0.8751 (0.0337)	0.8751 (0.0318)
parc_sqm, num_emps	0.8831 (0.0317)	0.8543 (0.0333)	0.8719 (0.0331)
num_panels	0.8668 (0.0347)	0.8668 (0.0347)	0.8668 (0.0347)
parc_sqm, num_panels	0.906 (0.0303)	0.8731 (0.0378)	0.8623 (0.0353)
parc_sqm	0.8579 (0.0352)	0.8253 (0.0361)	0.836 (0.0466)
num_emps	0.7619 (0.0419)	0.7276 (0.0415)	0.7358 (0.0429)

*Algorithm with highest mean accuracy and feature combination. Rows are sorted by feature combination with highest to lowest mean accuracy based off of this column.

Table B4. Overall mean accuracy and standard deviations for all algorithms and feature combinations for the automatically detected panels trained on the DC area and tested on the Boston area.

Feature Combination	RF	LR*	NN
pn_sqm_sum, num_panels, num_emps	0.9083 (0.0294)	0.9158 (0.0277)	0.8995 (0.0315)
pn_sqm_sum, num_panels	0.9095 (0.0295)	0.9044 (0.0289)	0.8917 (0.0336)
pn_sqm_sum, num_emps	0.9037 (0.0293)	0.8883 (0.0304)	0.8849 (0.0302)
pn_sqm_sum	0.817 (0.0451)	0.8781 (0.0309)	0.8733 (0.032)
parc_sqm, pn_sqm_sum, num_panels, num_emps	0.8638 (0.0373)	0.8375 (0.0371)	0.8872 (0.0344)
parc_sqm, pn_sqm_sum, num_panels	0.8714 (0.0353)	0.8369 (0.0363)	0.8783 (0.0355)
parc_sqm, pn_sqm_sum, num_emps	0.8604 (0.0383)	0.8328 (0.0357)	0.8786 (0.0313)
parc_sqm, pn_sqm_sum	0.8637 (0.0382)	0.8327 (0.0353)	0.8619 (0.0344)
num_panels, num_emps	0.8371 (0.0354)	0.8323 (0.0345)	0.8195 (0.0354)
num_panels	0.8067 (0.035)	0.8067 (0.035)	0.8056 (0.0358)
parc_sqm, num_panels, num_emps	0.8067 (0.0384)	0.8018 (0.0384)	0.8163 (0.0401)
parc_sqm, num_panels	0.7925 (0.0435)	0.7887 (0.0383)	0.7981 (0.042)
parc_sqm, num_emps	0.7899 (0.0375)	0.7828 (0.0355)	0.7117 (0.0486)
parc_sqm	0.7631 (0.0368)	0.7668 (0.0366)	0.6762 (0.0692)
num_emps	0.6218 (0.0375)	0.6114 (0.0368)	0.5847 (0.0334)

*Algorithm with highest mean accuracy and feature combination. Rows are sorted by feature combination with highest to lowest mean accuracy based off of this column.

Table B5. Overall mean accuracy and standard deviations for all algorithms and feature combinations for the manually detected panels in the DC area.

Feature Combination	RF*	LR	NN
parc_sqm, pn_sqm_sum, num_emps	0.9512 (0.0333)	0.9467 (0.0344)	0.9503 (0.0326)
parc_sqm, pn_sqm_sum, num_panels, num_emps	0.9475 (0.0354)	0.9389 (0.0351)	0.9488 (0.0343)
parc_sqm, pn_sqm_sum	0.9447 (0.0358)	0.9418 (0.0328)	0.9387 (0.034)
parc_sqm, pn_sqm_sum, num_panels	0.9436 (0.0322)	0.9374 (0.0327)	0.9382 (0.0335)
pn_sqm_sum, num_panels, num_emps	0.9432 (0.0355)	0.9396 (0.0401)	0.9448 (0.0357)
parc_sqm, num_emps	0.9369 (0.0357)	0.908 (0.044)	0.889 (0.0432)
parc_sqm, num_panels, num_emps	0.9334 (0.0352)	0.9134 (0.0415)	0.8958 (0.0428)
pn_sqm_sum, num_emps	0.93 (0.0378)	0.931 (0.0393)	0.9287 (0.0377)
parc_sqm, num_panels	0.9282 (0.0383)	0.9162 (0.0371)	0.897 (0.0422)
pn_sqm_sum, num_panels	0.9268 (0.0356)	0.9357 (0.0367)	0.9337 (0.0378)
parc_sqm	0.9039 (0.0417)	0.9139 (0.0386)	0.8897 (0.0417)
pn_sqm_sum	0.894 (0.0447)	0.9232 (0.0404)	0.9208 (0.0394)
num_panels, num_emps	0.8223 (0.0515)	0.8106 (0.052)	0.8081 (0.0523)
num_panels	0.7661 (0.0528)	0.7634 (0.0518)	0.7579 (0.0497)
num_emps	0.7179 (0.0624)	0.784 (0.0419)	0.7818 (0.0361)

*Algorithm with highest mean accuracy and feature combination. Rows are sorted by feature combination with highest to lowest mean accuracy based off of this column.

Table B6. Overall mean accuracy and standard deviations for all algorithms and feature combinations for the manually detected panels in the Boston area.

Feature Combination	RF*	LR	NN
parc_sqm, pn_sqm_sum, num_emps	0.9204 (0.0425)	0.9082 (0.0423)	0.8737 (0.0469)
parc_sqm, pn_sqm_sum, num_panels, num_emps	0.9036 (0.0468)	0.902 (0.0459)	0.8721 (0.051)
pn_sqm_sum, num_emps	0.8988 (0.0442)	0.8956 (0.0439)	0.8758 (0.0466)
parc_sqm, num_emps	0.8932 (0.048)	0.9081 (0.0452)	0.7609 (0.0567)
pn_sqm_sum, num_panels, num_emps	0.8852 (0.0457)	0.8877 (0.0467)	0.8736 (0.0499)
parc_sqm, pn_sqm_sum, num_panels	0.8812 (0.0489)	0.9019 (0.0462)	0.8719 (0.0496)
parc_sqm, pn_sqm_sum	0.8804 (0.0479)	0.913 (0.044)	0.8771 (0.0477)
parc_sqm, num_panels, num_emps	0.8758 (0.0492)	0.8998 (0.0459)	0.7926 (0.059)
pn_sqm_sum, num_panels	0.8725 (0.0463)	0.8898 (0.0449)	0.8729 (0.0496)
parc_sqm, num_panels	0.862 (0.0499)	0.8779 (0.0494)	0.7169 (0.0631)
pn_sqm_sum	0.8331 (0.0547)	0.8992 (0.0419)	0.8786 (0.0466)
parc_sqm	0.8227 (0.054)	0.8821 (0.0477)	0.6339 (0.0506)
num_panels, num_emps	0.7811 (0.0601)	0.7847 (0.0564)	0.7768 (0.0592)
num_panels	0.7066 (0.0656)	0.7023 (0.0587)	0.7005 (0.0593)
num_emps	0.6866 (0.0669)	0.7225 (0.0585)	0.7294 (0.0554)

*Algorithm with highest mean accuracy and feature combination. Rows are sorted by feature combination with highest to lowest mean accuracy based off of this column.