Using Machine Learning Methods to Classify Urban Heat Islands

Jennifer Xu and Susan Chen
Stanford University
jennxu23@stanford.edu; peeka5@stanford.edu

1 Introduction

1.1 Background

In past decades, global warming has had a significant effect on society. More extreme weather events, such as the intense forest fires in California or the unusually cold weather in Texas, has illustrated the urgency of stopping climate change. Increasing urbanization, which involves both the movement of people into urban areas as well as construction of urban infrastructure, has contributed to climate change. The term "urban heat island" refers to a metropolitan area that is significantly warmer than its surrounding rural counterparts.

Specific characteristics of urban areas that contribute to the urban heat island effect. For example, urban areas are more dense than rural areas. People, buildings, and vehicles are packed together and constantly burning energy, generating heat. Man-made structures, such as buildings and parking garages, reflect light more than natural structures eventually warming up the surrounding air. As a result, urban areas also tend to be significantly warmer at night compared their rural neighbors (National Geographic 2012). As cities develop and more people move into urban areas, the urban heat island effect could have dire consequences. Additionally, a higher population density as well as warmer temperatures lead to resource strain within urban heat islands.

For our project, we decided to study the influence of population, land surface temperature, land usage, and reflectivity as possible contributors to the urban heat island effect. Using relevant data sets and applicable machine learning algorithms, our goal for this project is to build a model that can classify a given city as an urban heat island.

1.2 Related Work

Existing research has been done on specific urban regions. In a recent study, researchers used multi-temporal Landsat satellite images to simulate land cover and temperature changes overtime in Bangladesh (Kafy et. al 2021). From the LandSat data, Kafy used a Support Vector Machine (SVM) model to simulate urban development and changes in land surface temperature. Another study used machine learning methods to determine which factors contribute most to the development of urban heat islands (Yoo, 2018). Using data from the cities of Marion County, Indiana, Yoo used Principal Component Analysis (PCA) to determine that imperviousness, vegetation, and building footprint were significant contributors to the Urban Heat Island Effect. A similar study was done on a larger scale with urban areas in China, using a linear regression model to illustrate the trend between urban development and increases in surface temperature (Zhou et. al, 2015)

We decided to take an approach similar to the above studies and build an urban heat island classifier based on environmental data. Similar to Yoo's study, we also analyze the relative influences of particular environmental factors.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

2 Methodology

2.1 Data set and Feature Extraction

The data for this project is sourced from three different data sets: The Global Urban Heat Island (UHI) data set v1 (2013) from the Socioeconomic Data and Applications Center (SEDAC), the LandSat 8 data set from the United States Geological Survey (USGS), and the Moderate Resolution Imag-154ing Spectroradiometer (MODIS) Land Cover Type (MCD12Q1) Version 6 data set from the USGS.

Each entry in the Global UHI data set represented a specific urban area. In addition to identifying information, the Global UHI dataset included population data from 1990, 1995, and 2000; area of urban extent in square kilometers, daytime average temperature data, and nighttime average temperature data. To develop the training labels, we used the absolute difference between the daytime urban average temperature and the daytime buffer (average temperature of an urban area and a 10km buffer) average temperature. If a given area had greater than a 1°difference in daytime average temperature as compared to its buffer zone, it was classified as an urban heat island (NASA, 2013).

Using Google Earth Engine, we located 2013 LandSat images and composed an extraction script to obtain bands 4, 5, and 6 data corresponding to each city. Bands 4, 5 and 6 represent the reflectance from short-wave infrared (1.55 - 1.75 µm), near-infrared (0.77 - 0.90 µm), and visible (0.63 - 0.69 µm) light respectively. For the land usage data, we used the National Land Cover Database (NLCD). A city normally has multiple land cover types. Each land cover type is represented as a number. Dominant land cover types were selected for analysis. To analyze the impact of each dominant land cover type, we used hot-encoding. For a given city and a given land cover type, a 1 indicated that the land cover type was present in the city and a 0 indicated otherwise.

The extracted data from the LandSat and NLCD data sets were appended onto the the Global UHI data set. Due to the geographical and temporal restrictions of the three original data sets, we limited the scope of our final data set to be urban areas in the United States as measured in 2013. Our final data set included the original features of the Global UHI data set (excluding identifying information and the daytime average difference); band 4,5, and 6 reflectance data; and presence of land cover type 82, 90, 41, 52, 81, 71, 42, 11, 95, 43, 21, 23, 22, 24, and 31. The final data set has a total of 28 features.

Finally, we randomly split our data set into a training set, a testing set, and a validation set based on a 70:15:15 ratio. The training set contains 3500 samples, the testing set contains 750 samples, and the validation set contains 749 samples. Each sample represents one metropolitan area.

2.2 Models

Because our ultimate goal is to build a classifier for urban heat islands, and because we have categorical features, we decided to use three basic classification algorithms: logistic regression, Gaussian discriminant analysis, and a support vector machine (SVM) model.

Logistic Regression Our logistic regression model directly predicts whether or not a city is an urban heat island based on a probabilistic analysis. The basis of a logistic regression model is the Sigmoid function, which maps our real valued predictions to a value between 0 and 1. We can write the probability of a given data point as

$$p(y|x;\theta) = h_\theta(x)^y(1-h_\theta(x))^{1-y} \tag{1}$$

where x is a given training example, y is the associated label, and $h_\theta(x)$ is the sigmoid function.

From the above equation, we can derive he weights of each feature by gradient descent, where the log-likelihood function is

$$\ell(\theta) = \sum_{i=1}^{n} y^{(i)} \log h(x^{(i)}) + (1-y^{(i)}) \log(1-h(x^{(i)})) \tag{2}$$

In the above equation, $x^{(i)}$ is a training example, $y^{(i)}$ is its associated label, and $h(x^{(i)})$ is the sigmoid function. We used SciKit-Learn's Logistic Regression class to generate our logistic regression model.

Gaussian Discriminant Analysis We also developed a Gaussian discriminant analysis model (GDA). Since the features we collected are environmental data, we can possibly fit a Gaussian Distribution to our model. GDA is predicated on the assumption that the probability of the training set given the labels ($p(x|y)$) can be modeled as a multivariate Gaussian distribution. From the law of total probability and the underlying Gaussian assumption, we can derive and maximize the following likelihood function:

$$\ell(\phi, \mu_0, \mu_1, \Sigma) = \log \Pi_{i=1}^{n} p(x^{(i)}|y^{(i)}; \phi, \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi, \mu_0, \mu_1, \Sigma) \tag{3}$$

where ϕ is the probability that a training example is an urban heat island ($p(y = 1)$), μ_0 and μ_1 are the expected values of the negative (non-urban heat islands) and positive (urban heat islands) classes respectively, and Σ is the covariance matrix for both classes. We used Scikit-Learn's Linear Discriminant class to generate our Gaussian Discriminant model. We also used Scikit-Learn's Gaussian Discriminant class to explore the possibility of a quadratic decision boundary.

Support Vector Machine We also developed an SVM model based on the data set. The incentive to use SVM is to find a separating hyperplane that maximizes the total distance from the data points. Our project can be treated as a linear classifier for a binary classification problem.Therefore, we use weights w and intercept b, and write our classifier as

$$h_{w,b}(x) = g(w^T x + b) \tag{4}$$

The functional margin of (w, b) with respect to the training example can be written as:

$$\gamma^{(i)} = y^{(i)}(w^T x^{(i)} + b) \tag{5}$$

The SVM model minimized the sum of functional margins for the training set. We tried implementing both a polynomial kernel with degree 3 and radial basis function kernel to explore different model complexities.

Feature Analysis As a part of our analysis, we wanted to see if feature selection would lead to an increase in model performance. We used a Random Forest Classifier to select features. Random decision forests are an ensemble learning method for classification. The model operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) (Ho, 1995). Since we have multiple features belong to different data types, random forests serve as a an effective tool to rank the importance of features and narrow down to the features that matters. Our Random Forest model ranks features based on Mean Decrease in Impurity (MDI), which measures the average decrease in impurity based on the Gini impurity across all the trees (Louppe et. al, 2013).

3 Results

For our preliminary experiments, we built a Logistic Regression Model, a GDA model, and a SVM model on the entire data set. To explore hyper parameters and evaluate model fit, we utilized SciKit-Learn's's GridSearch CV algorithm to exhaustively test and determine optimal hyper parameters for each model. We evaluated the effectiveness of our model based on the accuracy and the ROC-AUC score (generated from SciKit-Learn's metrics package). Error was caluclated by taking the norm of the differences between the predicted and true labels.

Next, we explored the possibility of feature selection to both improve our models and to determine which features have the greatest impacts on the models. We used a Random Forest model to select the most impactful features. Then, we re-built each model and calculated optimal hyper parameters based on the reduced-dimensionality data set. Finally, we used SHAP to anlayze the impact of individual features on individual models.

Model Results Before hyperparameter tuning, the Logistic Regression model had an accuracy of 0.536 with an error of 18.65. Using GridSearchCV, combinations of different degrees of the inverse of regularization ($\frac{1}{\lambda}$ and the penalty term $l1$ and $l2$) were tested over 5 folds. The optimal hyper parameter combination was determined to be the $l2$ penalty term with an inverse of regularization of 10^{-2}. Optimizing for the best accuracy, the optimal model had an increased accuracy of 0.644 and a reduced error of 15.71.

Similarly, the Linear Discriminant Analysis model had an accuracy of 0.524 with an error of 18.894. In order to improve the accuracy of the Linear Discriminant analysis, we tried two different solvers: Singular value Decomposition (SVD) and Least Squares (LSQR). Using GridSearchCV over 5 folds, the optimal solver was determined to be SVD. With SVD, the accuracy of the Linear Discriminant Analysis was determined to be 0.620 with a reduced error of 16.248.

Finally, the SVM model has an accuracy of 0.536 and an error of 18.65. For our initial SVM model, we chose a radial basis function kernel. After using GridSearchCV and exhausting all combinations of gamma and regularization strength vaues ($\frac{1}{\lambda}$), the accuracy of SVM model increased to 0.647 and error reduced to 16.03. The optimum hyper parameters for accuracy were an inverse regularization value of 10^{-4} and a gamma value of 10^{-3}.

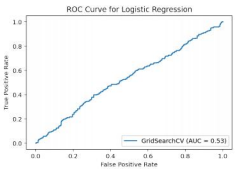

Figure 1: ROC Curve for Logistic Regression

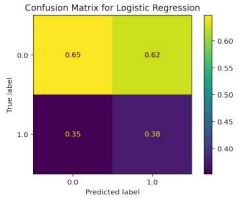

Figure 2: Normalized Confusion Matrix, Logisitic Regression

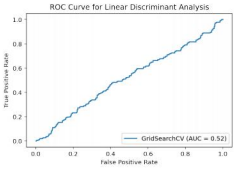

Figure 3: ROC Curve for LDA

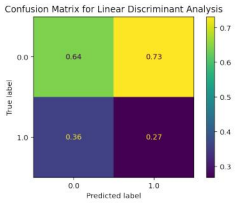

Figure 4: Normalized Confusion Matrix, LDA

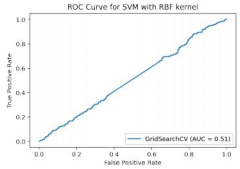

Figure 5: ROC Curve for SVM

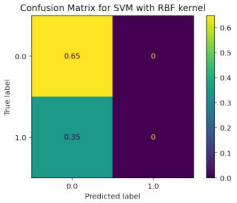

Figure 6: Normalized Confusion Matrix, SVM

Based on the ROC-AUC cures, all three models perform about the same as a random classifier. Closer analysis via the confusion matrices show that all three models correctly classify negative examples (non urban heat islands) a majority of the time, but poorly classifies positive examples (urban heat islands).

Feature Analysis We explored feature reduction as a way to possibly improve our models and hyper parameters. Using a Random Forest model, we found that land cover data doesn't significantly impact the model. Then, we retrained our models on a reduced data set including only population, temperature, and reflectivity data. The reduced data set did not improve the fit or accuracy of the models. However, analyzing the features that were most influential on the models reveal interesting trends.

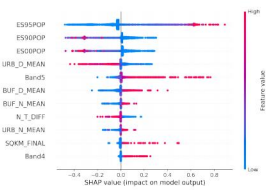

Figure 7: SHAP, LDA

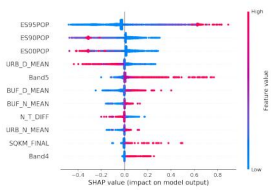

Figure 8: SHAP, SVM

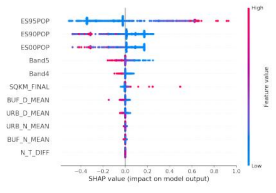

Figure 9: SHAP, Logistic Regression

We used the SHAP python library for model explainability. SHAP (SHapley Additive exPlanation) leverages the idea of Shapley values for model feature influence scoring. A Shapley value represents the average marginal contribution of a given feature across of all features. This exhaustive approach would guarantee SHAP's consistency and local accuracy.

4 Discussion

Our initial models performed poorly, with an accuracy within the .50 - .60 range. With hyperparameter tuning, our models' accuracies improved into the .60-.70 range with lower error. There is not a significant difference in accuracy between the models. As shown by the ROC-AUC curves, all three models perform about as accurately as a random classifier. Trying more complex models, feature reduction, and hyperparameter tuning did not improve our models' performances. The confusion matrices indicate that all three models perform decently at classifying negative examples (non UHIs), but do not accurately classify positive examples (UHIs).

However, SHAP analyses reveals useful information about the impact of particular features. We see from the random forest analyses that land cover data is not particularly impactful. The SHAP analyses of all three models show population data as highly impactful. Specifically, a larger population in 1995 seems to be correlated with a greater chance of being classified as an urban heat island [ES95POP]. The opposite relationship is true for the population in 1990 and 2000 [ES90POP and ES00POP]. Population data seems to be particularly impactful in the Linear Regression model, as it accounts for most of the variance in the model. The impact and positive correlation is expected, as a higher population within an urban area would mean more movement. As a result, more energy and heat is generated.

Regarding temperature, we see that, across all models, daytime temperature metrics [URB_D_MEAN and BUF_D_MEAN] seem to have a greater impact on models than nighttime daytime temperature metrics [URB_N_MEAN, BUF_N_MEAN and N_T_DIFF]. This may be due to the fact that there's a greater difference in activity, and therefore generated energy and heat, between urban areas and their surroundings during the day as compared to the night.

Band 5 data, which corresponds with the intensity of reflectivity of Short-wave Infrared (1.55 - 1.75 μm) light, was the most impactful reflective feature wavelength as compared to Near-Infrared (Band 4) or visible light (Band 6). Interestingly, high Band 5 values were positively correlated with a positive urban heat island classification in the SVM and LDA models, while it was negatively correlated in the Logistic Regression model. Previous studies indicated that reflectivity is positively associated with the urban heat islands. Urban heat islands often have artificial materials with high reflectance such as buildings and pavements, that can lead to an increase in temperature.

5 Conclusion

Main takeaway In summary, we attempted to build an urban heat island classifier using three models: Logistic Regression, Gaussian Discriminant Analysis, and Support Vector Machine. We build our models based on remote sensing data. While none of our models were particularly good classifiers of urban heat islands, feature analysis revealed interesting and novel conclusions.

Future Directions As shown by our previous work, we tested for different models with and without feature selection. Based on our results, more features should be integrated into this model. Examples of such parameters include a different population dataset (i.e population density, a more recent population data set, etc.), reflectivity of light from other wavelength in the spectrum, economic parameters, vegetation indices, or calculated measures of urbanization. This would increase model complexity and potentially lead to more accurate models and a more informative feature reduction.

Additionally, more research should be done to integrate existing features in a way that reduces colinearity and better reflects their impact on given models.For example, different reflectivity bands can be mathematically combined and used to indicate different land cover types.

Furthermore, our model indicates that hot-encoded dominant land cover do not greatly impact classification. Our models could possibly be improved by using a larger data set in which cities are classified as urban and rural, rather than being separated by specific land cover types. This could generate clearer data clustering.

Lastly, remote sensing data is often generated and transformed by complicated algorithms, whose complexities could be difficult to capture with a linear model. This could've contributed to the high bias in our models. We could try different, possibly more complex, methods of analysis such as deep learning or unsupervised training.

6 Code and Contributions

Both Susan and Jennifer worked equally on the project. Susan extracted and cleaned the reflectance and land use data. Jennifer cleaned the Global UHI dataset. Both members worked together to determine and fit the best classifiers, as well as do the feature analysis. Jennifer focused on the logistic regression and GDA model. Susan worked on the SVM and Random Forest model. Both members did hyperparameter tuning and feature explanation together.

The code can be found at the following link: `https://colab.research.google.com/drive/1u7sGnJNoVKsp_oAr3edPJsUn8eprpSMU?usp=sharing`

References

[1] Ho, Tin Kam (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

[2] Kafy, A. A., Shuvo, R. M., Naim, M. N. H., Sikdar, M. S., Chowdhury, R. R., Islam, M. A., ... Kona, M. A. (2021). *Remote sensing approach to simulate the land use/land cover and seasonal land surface temperature change using machine learning algorithms in a fastest-growing megacity of Bangladesh*. Remote Sensing Applications: Society and Environment, 21, 100463. [3] Louppe, G., Wehenkel L., Sutera A., Geurts P. (2013). *Understanding variable importances in forests of randomized trees*. NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1. December 2013. pp. 431-439

[4] National Aeronautics and Space Association. (2000). Population Density [Data File]. Retrieved from `https://neo.sci.gsfc.nasa.gov/view.php?datasetId=SEDAC_POP`.

[5] National Aeronautics and Space Association. (2013). Global Urban Heat Island (UHI) Data Set, v1 [Data File]. Retrieved from `https://sedac.ciesin.columbia.edu/data/set/sdei-global-uhi-2013/data-download`.

[6] National Geographic Society. (2012, October 09). *Urban heat island*. Retrieved from `https://www.nationalgeographic.org/encyclopedia/urban-heat-island/`.

[7] United States Geological Survey. (2013). Landsat 8 Surface Reflectance Tier 1 [Data File]. Retrieved from `https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C01_T1_SR#bands`.

[8] United States Geological Survey. (2019). Terra and Aqua combined Moderate Resolution Imaging Spectroradiometer (MODIS) Land Cover Type (MCD12Q1) Version 6 [Data File]. Retrieved from `https://lpdaac.usgs.gov/products/mcd12q1v006/`.

[9] Yoo, S. (2018). *Investigating important urban characteristics in the formation of urban heat islands: a machine learning approach*. Journal of Big Data, 5(1), 1-24.

[10] Zhou, Decheng; Zhang, Liangxia; Hao, Lu; Sun, Ge; Liu, Yongqiang; Zhu, Chao. (2016). *Spatiotemporal trends of urban heat island effect along the urban development intensity gradient in China*. Science of The Total Environment, Vol. 544: 10 pages.: 617-626. doi:10.1016/j.scitotenv.2015.11.168