Local Climate Zone classification in mega urban areas using Landsat images and random forests: A case study of Hong Kong

Ericka B. Smith

Winter 2021

Introduction

1.1 Background

The population worldwide is increasing and urbanizing (United Nations et al., 2019). In addition, issues surrounding environmental change and pollution are exacerbated in cities (Bechtel et al., 2015) (Santamouris, 2020). These changes compound the effects of each other and, as a result, a greater understanding of these processes is more important now than ever before (Bechtel et al., 2015).

Urban Heat Islands, where urban areas are warmer than the neighboring rural areas, are of particular concern. This is for four primary reasons: increased energy consumption, elevated emissions of air pollutants and greenhouse gasses, compromised human health and comfort, and impaired water quality (US EPA, 2014). A multitude of processes combine to cause Urban Heat Islands. Broadly speaking, urban built structures hold more heat than the structures and vegetation in surrounding areas (Hibbard et al., 2017). In particular, cities have an overabundance of impervious surfaces, along with a lack of vegetation (Hibbard et al., 2017). Knowledge about microclimates within cities, especially in the context of prospective sites for climate risk adaptation efforts, has the potential to ease some of these disturbances (Lempert et al., 2018) .

Unfortunately, information on these sites can be difficult to obtain. Though there is a plethora of satellite imagery, the current methods to classify this imagery are somewhat lacking (Yokoya et al., 2018). Historically the focus was on the broad categories of urban vs. rural, which does not give enough information about the nuanced climate within a large urban area (Stewart & Oke, 2012). Local Climate Zone (LCZ) classification was created by Stewart and Oke (2012) to alleviate this problem, but it often requires a significant investment from individuals with specialist knowledge to successfully classify a city (Bechtel et al., 2015). Even the best models have low accuracy outside of their training city (Verdonck et al., 2017) (Yokoya et al., 2018).

Therefore, this issue with generalizing a model to other cities, transferability, is a topic of great interest in current Urban Heat Island research (Tuia et al., 2017) (Yokoya et al., 2018).

1.2 Objective

The goal of this project is to recreate aspects of the article Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using Landsat images (Yoo et al., 2019), where methods for predicting LCZ classes for four large cities throughout the world were compared. To do so, a small training dataset from the 2017 Institute of Electrical and Electronics Engineers (IEEE) Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest (Tuia et al., 2017) was used as ground truth for LCZ classes. This was combined with satellite input data to create a series of models, which were then compared to a larger, full LCZ layer for each city and assessed for accuracy. The primary types of models considered in the Yoo et al. (2019) work were random forests and convolutional neural networks. However, for this project the focus will be only on random forests. In addition, rather than four cities, this investigation will focus on just Hong Kong. This city was chosen due to its lack of "red-star classes" (LCZ classes with 3 or fewer polygons) in the study area. The red-star classes add complexity to the splitting up of the data into training and test sets. In addition, since they by nature address only rare classes, this complexity was deemed unnecessary for this project. Here, a classification scheme like the one used by the World Urban Database and Access Portal Tools (WUDAPT) project, denoted as Scheme 1 (S1) in Yoo et al. (2019) will be compared with the classification scheme from Danylo et al. (2016) , denoted as Scheme 2 (S2) in Yoo et al. (2019) which includes more spectral information as input variables.

Methods

2.1 Data

Data for this analysis was accessed from the 2017 IEEE GRSS Data Fusion Contest. Only the Hong Kong LCZ reference and Landsat 8 datasets were used. The LCZ reference data from the contest set was taken from the WUDAPT database, checked for correctness, and then provided as a 100m resolution raster layer. In this provided layer the LCZ classes are numbered 1-17, rather than 1-10 and A-G, and initial inspection of the data verified that numbers 11-17 directly match classes A-G. In Yoo et al. (2019) the first step they took with the reference data was randomly dividing the polygons of each class into training and testing groups. This required some preprocessing since I only had access to the data in raster form and without any sort of polygon identification. Pixels within the same polygon cannot be put into both training and validation groups because it will artificially increase accuracy metrics (Zhen et al., 2013).

In Yoo et al. (2019) they downloaded and preprocessed their own Landsat 8 data to facilitate some of their classification schemes, but the contest data was sufficient for the narrower scope of this project. The Landsat 8 data provided by the contest includes four different dates, or "scenes," each of which were downloaded from the USGS EarthExplorer portal. Then the atmospheric band (band 9) and the panchromatic band (band 8) were left out, and the data were resampled using the area weighted average to 100 m x 100 m grids (Yokoya et al., 2018). To get these data into a usable format they were loaded into R by band (each band is a separate raster layer) and stacked by scene. All the scenes were then stacked with the LCZ reference data and converted to a dataframe, ready to be used as input data.

_____(MORE) v _____ All bands of all four Landsat 8 scenes amounted to 36 input variables. #land area, #pixels, # of polygons Train vs test. Pixels = observations, etc _____(MORE) ^ _____

2.2 Random Forests

Random forests consist of many decision trees. A decision tree can be used for categorization or regression. Here I'll focus on categorization since my goal is to predict LCZ class, which is a categorical variable. Decision trees work by putting each observation in the original ("training") dataset through a series of conditional statements, or tests, in order to group it with other observations that are similar. These similar groups are expected to have similar values for the variable you're interested in predicting. Since the true value of that variable is known for the observations in the training dataset it's possible to measure the accuracy of each prospective structuring of these conditional statements, and therefore to pick the best tree.

Each point in which a conditional statement is evaluated and the data could potentially be split into two groups is called a node. The first split is called the root node. Nodes are created such that each one uses the conditional statement which subsets the data into the best possible split. When any more subsetting does not increase accuracy, the path ends, and this is called a leaf node. Any nodes between the root node and any leaf nodes are called internal nodes. The resulting structure looks resemblant of a tree.

Splits are typically evaluated by Gini impurity or entropy.

$$Gini\ Impurity\ =\ I_G(t)\ = 1 - \sum_{i=1}^{C} p(i|t)^2$$

$$Entropy\ =\ I_H(t)\ = -\sum_{i=1}^{C} p(i|t) log_2 p(i|t)$$

Where $p(i|t)$ is the proportion of samples that belong to class i for a particular node t.

A Gini Impurity of 0 or an entropy value of 0 indicates a completely homogeneous group, which cannot be improved upon. A split doesn't need to have a value of 0 to create the best node though, these metrics are used for comparison. All possible different variables and different cutoffs within those variables are all tried and their Gini impurity or entropy calculated, then, the one with the best (lowest) value is the split used at that node. This is done recursively at each node until all possible splits no longer offer a decrease in the metric of choice. The resulting collection of nodes is a decision tree.

Decision trees perform quite poorly with new samples. Including a threshold value for the accuracy metric can help, as it keeps trees more simple, but they are still prone to overfitting. Collecting them into a random forest addresses this issue. True to its name, a random forest is a collection of decision trees that have a component of randomness in their creation. The predictive aspect of a random forest doesn't only rely on one tree, it is determined by totaling up the decisions, or "votes," that all of the trees make.

To create each tree the first step is bootstrapping the training dataset. The decision tree is then created with that bootstrapped data, but with slightly different rules. For each node, only a subset of the variables are randomly selected and used as candidates to split up the bootstrapped data. Not all the variables are used at each node because doing so reduces correlation and introduces randomness into the model. Whichever variable best splits the data will be the variable that is kept at that node in that specific tree. This is done recursively until no more splits are beneficial, just as in a regular decision tree. To create the next tree the process is the same, but starts completely over with a new bootstrap sample of the training data. This is repeated for a chosen number of trees.

This choice of the number of trees to create is a tuning parameter. Too many trees can be computationally expensive, but too few can create a model that isn't useful. Another important tuning parameter is the choice of how many variables to randomly select at each node in the trees must be made. The default value for this is typically the number of variables squared for classification or the number of variables divided by 3 for regression. A less common, but useful, way to tweak the model is by trying different values for the minimum size of terminal nodes. A large minimum size means smaller trees and faster completion of the random forest. In addition, the maximum number of terminal nodes can be set to a specific number. Otherwise trees are grown as large as possible. Depth, which indicates the number of splits a tree has, can also be controlled. Greater depth gives more information about the data, but can cause overfitting.

Once the model has been tuned to optimality, the random forest is used to decide about a new set of input values. This is called bagging, because you're bootstrapping the data and using the aggregate to make a decision . Bagging is useful because it reduces variance without introducing bias. The decision itself is made by putting the new set of input values into each decision tree individually and seeing which outcome each tree classifies the variable of interest into. These are all combined either by direct vote (for categorical variables) or an average (for quantitative variables) and this becomes the final predicted result.

2.4 Accuracy Assessment

In line with the methods used in our reference paper and the remote sensing field, accuracy metrics will be based on predictions for the test dataset and will include the following:

$$Overall\ Accuracy\ =\ OA\ = \frac{number\ of\ correctly\ classified\ reference\ sites}{total\ number\ of\ reference\ sites}$$

.

$$Overall\ Accuracy\ in\ Urban\ Areas\ =\ OA_{urb}\ = \frac{number\ of\ correctly\ classified\ urban\ reference\ sites}{total\ number\ of\ urban\ reference\ sites}$$

.

$$Overall\ Accuracy\ in\ Natural\ Areas\ =\ OA_{nat}\ = \frac{number\ of\ correctly\ classified\ natural\ reference\ sites}{total\ number\ of\ natural\ reference\ sites}$$

Additionally, the F_1 score will be used. This is the harmonic mean of precision (user's accuracy, UA) and recall (producer's accuracy, PA).

$$UA(z)\ = \frac{number\ of\ correctly\ identified\ pixels\ in\ class\ z}{total\ number\ of\ pixels\ identified\ as\ class\ z}$$

$$PA(z)\ = \frac{number\ of\ correctly\ identified\ pixels\ in\ class\ z}{number\ of\ pixels\ truly\ in\ class\ z}$$

$$F_1\ Score\ = 2 * \frac{UA * PA}{UA + PA}$$

Results

Conclusion

References

- Bechtel, B., Alexander, P., Böhner, J., Ching, J., Conrad, O., Feddema, J., Mills, G., See, L., & Stewart, I. (2015). Mapping Local Climate Zones for a Worldwide Database of the Form and Function of Cities. ISPRS International Journal of Geo-Information, 4(1), 199–219. https://doi.org/10.3390/ijgi4010199
- Danylo, O., See, L., Bechtel, B., Schepaschenko, D., & Fritz, S. (2016). Contributing to WUDAPT: A Local Climate Zone Classification of Two Cities in Ukraine. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9(5), 1841–1853. https://doi.org/10.1109/JSTARS.2016.2539977
- Hibbard, K. A., Hoffman, F. M., Huntzinger, D., West, T. O., Wuebbles, D. J., Fahey, D. W., Hibbard, K. A., Dokken, D. J., Stewart, B. C., & Maycock, T. K. (2017). Ch. 10: Changes in Land Cover and Terrestrial Biogeochemistry. Climate Science Special Report: Fourth National Climate Assessment, Volume I. U.S. Global Change Research Program. https://doi.org/10.7930/J0416V6X
- Lempert, R. J., Arnold, J. R., Pulwarty, R. S., Gordon, K., Greig, K., Hawkins-Hoffman, C., Sands, D., & Werrell, C. (2018). Chapter 28: Adaptation Response. Impacts, Risks, and Adaptation in the United States: The Fourth National Climate Assessment, Volume II. U.S. Global Change Research Program. https://doi.org/10.7930/NCA4.2018.CH28
- Santamouris, M. (2020). Recent progress on urban overheating and heat island research. Integrated assessment of the energy, environmental, vulnerability and health impact. Synergies with the global climate change. Energy and Buildings, 207, 109482. https://doi.org/10.1016/j.enbuild.2019.109482
- Stewart, I. D., & Oke, T. R. (2012). Local Climate Zones for Urban Temperature Studies. Bulletin of the American Meteorological Society, 93(12), 1879–1900. https://doi.org/10.1175/BAMS-D-11-00019.1
- Tuia, D., Moser, G., Le Saux, B., Bechtel, B., & See, L. (2017). 2017 IEEE GRSS Data Fusion Contest: Open Data for Global Multimodal Land Use Classification [Technical Committees]. IEEE Geoscience and Remote Sensing Magazine, 5(1), 70–73. https://doi.org/10.1109/MGRS.2016.2645380 United Nations, Department of Economic and Social Affairs, & Population Division. (2019). World urbanization prospects: The 2018 revision. US EPA, O. (2014, June 17). Heat Island Impacts [Overviews and Factsheets]. US EPA. https://www.epa.gov/heatislands/heat-island-impacts
- Verdonck, M.-L., Okujeni, A., van der Linden, S., Demuzere, M., De Wulf, R., & Van Coillie, F. (2017). Influence of neighbourhood information on 'Local Climate Zone' mapping in heterogeneous cities. International Journal of Applied Earth Observation and Geoinformation, 62, 102–113. https://doi.org/10.1016/j.jag.2017.05.017
- Yokoya, N., Ghamisi, P., Xia, J., Sukhanov, S., Heremans, R., Tankoyeu, I., Bechtel, B., Saux, B. L., & Moser, G. (2018). Open Data for Global Multimodal Land Use Classification: Outcome of the 2017 IEEE GRSS Data Fusion Contest. IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH

OBSERVATIONS AND REMOTE SENSING, 11(5), 15.

- Yoo, C., Han, D., Im, J., & Bechtel, B. (2019). Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using Landsat images. ISPRS Journal of Photogrammetry and Remote Sensing, 157, 155–170. https://doi.org/10.1016/j.isprsjprs.2019.09.009
- Zhen, Z., Quackenbush, L. J., Stehman, S. V., & Zhang, L. (2013). Impact of training and validation sample selection on classification accuracy and accuracy assessment when using reference polygons in object-based classification. International Journal of Remote Sensing, 34(19), 6914–6930. https://doi.org/10.1080/01431161.2013.810822