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

1 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 not always accurate. (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) (Figure 1) 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).

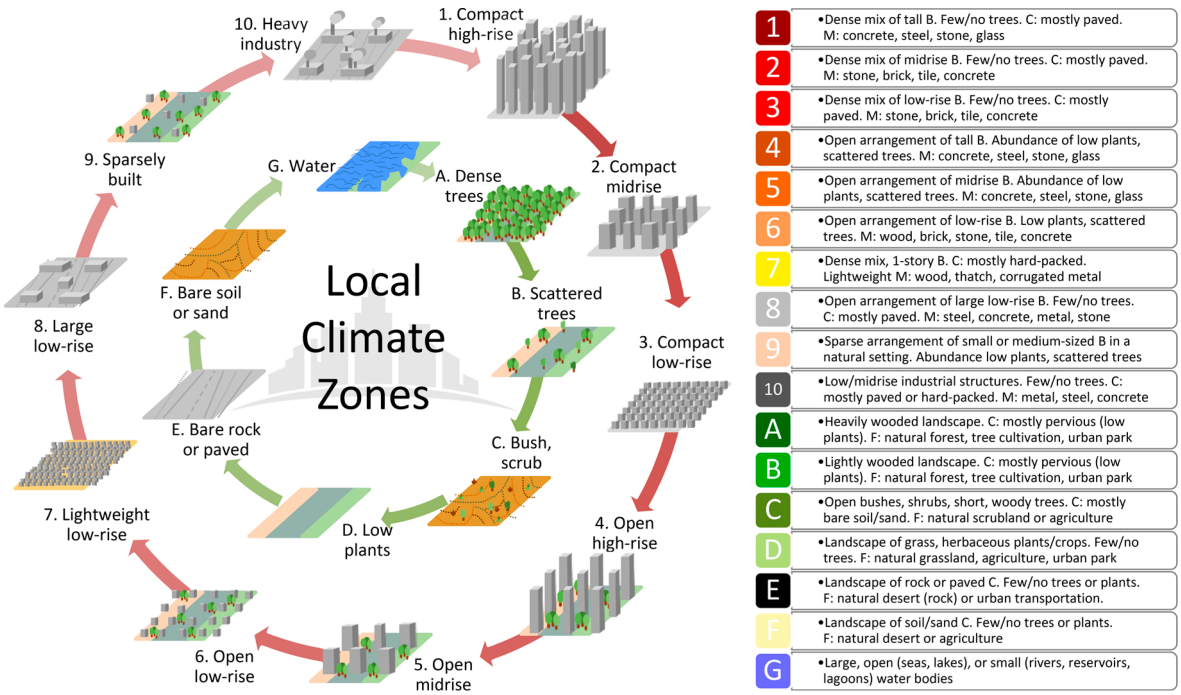

Figure 1: Local Climate Zone classes. Originally from Stewart and Oke (2012) and remade by Bechtel et al. (2017). Copyright CC-BY 4.0

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. Since they by nature address only rare classes, this complexity was deemed unnecessary for this project. Finally, 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 the focus, with comparisons between accuracy using a variety of tuning parameters.

2 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.

All 9 available bands of all 4 Landsat scenes amounted to 36 input variables. Each pixel is an observation, with 279,312 pixels total over the almost 2,800 square kilometers that are contained within the area of interest. The LCZ reference data has 179 polygons, which cover 8,846 pixels. This is approximately 88.5 square kilometers and 3% of the total area of interest. The LCZ reference data was randomly split into training and test sets by LCZ class and polygon. The resulting groups are not exactly equal in pixel numbers, but I determined that randomization of polygons takes precedence over equality in pixels due to the potential for artificially high accuracy values, as mentioned before. In order to get as close as possible a barplot was made showing the number of pixels in training and test, by class, so that equality in pixel number could be explored without having knowledge of polygon groupings. Sampling was repeated until these bars looked marginally even. (Table 1).

```
# TABLE 1
```

2.2 Random Forests

Random forests consist of many decision trees. A decision tree can be used for classification or regression. Here I will focus on classification since my goal is to predict LCZ class, which is a categorical variable. Decision trees put each observation through a series of conditional statements, or splits, in order to group it with other observations that are similar. These similar groups are expected to have similar values for the predictor variable. Since the true value of the predictor variable is known for the observations in the training dataset, it is possible to measure the accuracy of each prospective split, and therefore to build an optimal tree.

Each point in which the data could potentially be split into two groups is called a node. The first split is called the root node. Nodes are selected 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 leaf nodes are called internal nodes.

Splits are typically evaluated by Gini impurity or entropy.

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

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

Where i is a class in the predictor variable, ranging from 1 to C. C is the total number of classes represented for a particular node, t. $p(i|t)$ is the proportion of samples that belong to each i, for a particular node t.

A Gini Impurity, or entropy, of 0 indicates a completely homogeneous group, which cannot be improved upon. These metrics are used for comparison: all possible variables and thresholds within those variables are tried and their Gini impurity or entropy calculated, the variable and threshold with the best (lowest) value selected for the node. This is done recursively at each node until no split offers 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. The predictive aspect of a random forest does not only rely on one tree, it is determined by totaling up the decisions, or "votes," that all of the trees make.

Each tree is created individually and starts with bootstrapping the training dataset. Then the process is similar to the one described above for decision trees. However, 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 either done recursively until no more splits are beneficial (just as in a regular decision tree), or until minimum size of the node is reached. 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 is poor at prediction. Another important tuning parameter is the choice of how many variables to randomly select at each node. The default value for classification is typically the square root of the number of variables, and for regression it is the number of variables divided by three. Values for the minimum size of terminal nodes, which is the minimum number of observations required to make a leaf node, can also be varied. A large minimum size means smaller trees and faster completion of the random forest. This is an alternate strategy for controlling depth, which, in the randomForest package in R, cannot be directly controlled. Depth indicates the number of splits a tree has. Greater depth gives more information about the data. Another tuning parameter with a similar approach is the maximum number of terminal nodes a tree can have. Without setting this parameter, trees are grown as large as possible.

To use a random forest to make predictions for a new set of input values, the values are fed into each decision tree individually. The predictions from each tree are combined either by direct vote (for categorical variables) or an average (for quantitative variables) and this becomes the final predicted result. This process is called bagging, because it is the action of bootstrapping the data to create each tree and using the aggregate to make a decision. Bagging is useful because it reduces variance without introducing bias.

For this analysis, parameters were varied individually. This is limited in that it does not explore any potential interactions, but beneficial in that it is more clear the individual contribution of a parameter. The number of trees is the primary parameter of interest, since reducing it as much as possible can have a large impact on time and computational resources required.

2.3 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:

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

.

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

.

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

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).

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

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

$$\text{F-1 Score} = 2 * \frac{UA * PA}{UA + PA}$$

3 Results

3.1 The Number of Trees

The parameter for the number of trees, ntree, was varied between 5 and 500 at intervals of 5. The resulting overall accuracy metrics indicate a leveling off around 125 trees (Figure 2). There's also a clear distinction between accuracy in urban vs. natural classes. The plot of F-1 scores (Figure 3) explores each class individually more clearly. It's clear that there are three approximate groupings of classes. Classes 11, 12, 13, 14, and 17 have the highest F-1 scores throughout, though 12 stands out as a little lower than the others. Classes 6, 8, and 10 have much lower F-1 scores that don't seem to improve greatly as the number of trees increases. Finally, in the middle are classes 1 though 5. They are all in the middle range between the other two groupings, and do appear to respond to increased numbers of trees.

4 Conclusion

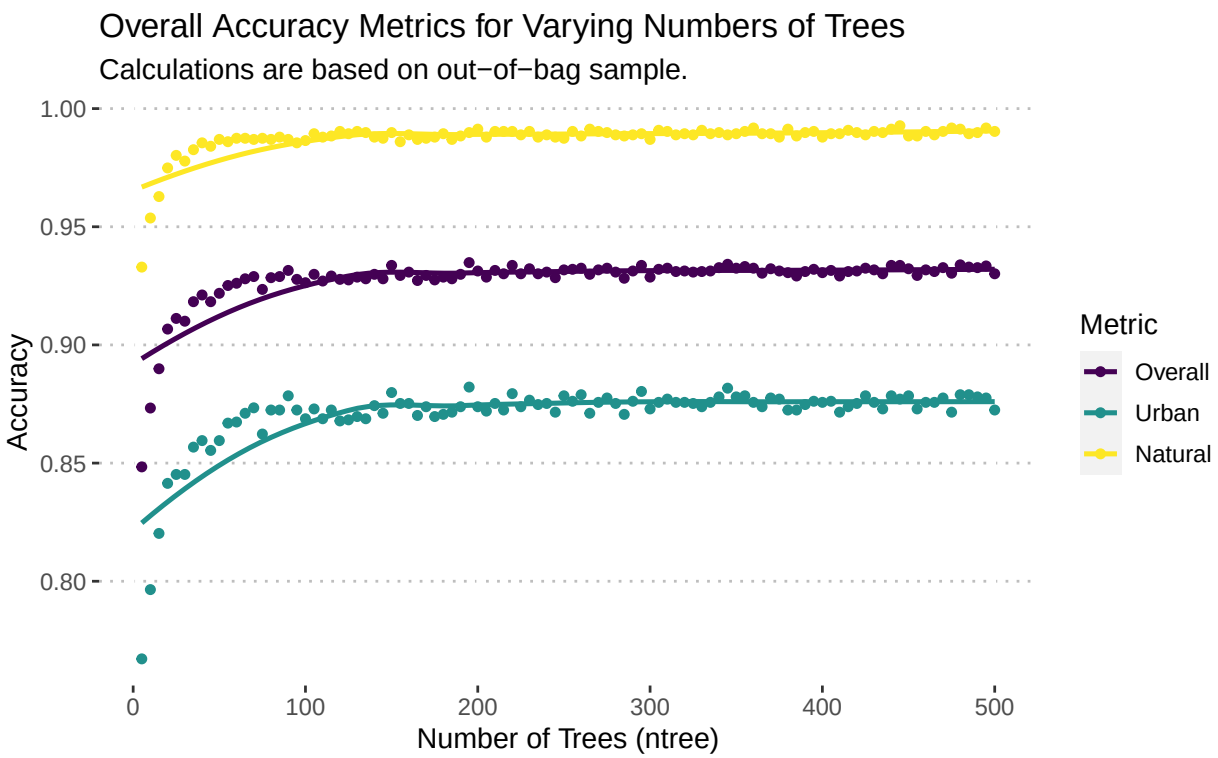

Figure 2: The increase in OA metrics levels off around 125 trees. Urban classes (1-10) have much lower accuracy than natural classes (11-17).

Figure 3: The variation between LCZ classes in F-1 score can be seen. As the number of trees in the random forest increases, F-1 score also increases. Note three approximate groupings of 1-5, 6-10, and 11-17.

5 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