CITADEL | CITADEL | Securities

TheDataOpen

Preserving the Power of Green Spaces: Using Big Data and Machine Learning to Improve Park Health in New York City

Tarun Chiruvolu[1] Junwei Huang[2] Arjun Rastogi[3] Krish Suraparaju[4]

[1]*Carnegie Mellon University, School of Computer Science*
[2]*Carnegie Mellon University, School of Computer Science*
[3]*University of Michigan, Department of Mathematics*
[4]*Carnegie Mellon University, Department of Mathematics*

February 12, 2023

1 Topic Question

From reducing stress and anxiety to improving health through increased physical activity, green spaces such as parks and gardens have been shown to exhibit positive impacts on human health. Despite this, many city legislators face challenges in maintaining and preserving these essential areas. This highlights a lack of information among lawmakers regarding the key factors that contribute to the decline of these spaces and the measures necessary to address them. In this study, we address the following questions:

1. How can publicly available and crowdsourced data be used to diagnose park health and pinpoint areas for intervention in New York City?

We seek to use the ubiquity of big data and large-scale machine learning techniques to shed further light onto the underlying factors behind urban greenspace dynamics. Our answers to these questions provide crucial insights into the importance of green spaces for promoting human health and guide efforts towards ensuring that these areas are accessible and preserved for future generations.

2 Executive Summary

Existing research has shown the value of urban greenspaces in improving both physical and mental health - however, the precise mechanisms by which they do so, and the features of such green spaces that optimally encourage healthy behavior remain unclear. Traditionally, studies have relied on manually collected surveys within a given geographical region. However, such surveys often fail to account for a number of biases present within the sampling process, such as confirmation biases in user responses or the demographics of residents. Moreover, these problems are compounded by the inherent difficulty of assigning credit to ecosystem services, as their underlying benefits are not often apparent and manifest themselves in downstream effects. As such, pinpointing the exact factors that incentivize visitation and facilitate greenspace accessibility is an open challenge for urban planners and policymakers, yet one that could potentially yield significant improvements in residents' quality of life if appropriately tackled.

To that end, the increasing availability of social media data has offered several new opportunities in population studies. As the social networks' user bases continue to grow, they paint an increasingly accurate picture of social dynamics, such as urban mobility, cultural practices, and socioeconomic trends (Ilieva and McPhearson, 2018). Researchers have begun to leverage this wealth of information through meta-analyses in a given metropolitan area - e.g., an analysis of Flickr and Twitter data in Minneapolis to disentangle the factors behind visitation frequency (Donahue et al., 2018).

To this end, we leverage existing social media data from the greater New York City metropolitan area to determine causal factors for visitation. We collected data from the social media site Flickr corresponding to the number of pictures within a given geographical area, and used this information to analyze the relationship between various infrasturctural and regional factors with park visitation and health. In addition, we use advances in modeling large-scale unstructured data - specifically, deep learning methods as predictive and analytic tools for using satellite imaging data. In particular, we collect existing high-resolution imaging data over parks in New York City and find that standard convolutional neural networks (CNNs) fail to use this imaging data to predict metrics such as commumnity eco-health index (CEHI) (Cochran et al., 2019) effectively, highlighting a paucity of standardized geospatial urban health data. We also train a convolutional autoencoder to compress satellite images into lower-dimensional latent representations, which we then project and cluster using principal components analysis (PCA) and k-means clustering to analyze if these representations concide with relevant urban ecosystem factors.

In summary, our findings highlight the following areas in which policymakers can standardize data collection procedures to encourage the adoption of large-scale data analysis techniques: (1) greater standardization of regional identification (2) collection of temporal and hierarchical socioeconomic information. While our deep learning methods achieved a baseline performance, they were limited primarily by (1), as satellite imagery was plentiful in several resolutions but could not be mapped to relevant urban ecosystem information. On the other hand, having a more diverse set of data in (2) such as time series health data would allow our neural network methods to have more informative labels, thereby improving the conditioning of the learning problem.

3 Technical Exposition

We begin by discussing our various sources of data and the statistical analysis that we performed.

3.1 Data Collection and Processing

We use the data provided by Citadel relating to the community ecohealth, public park accessibility, and the percent tree canopy coverage across U.S. counties. We then supplement this data with external crowd sourced data collected from social media sites and geosatellite images to analyze and predict CEHI indices for parks in New York City, NY.

3.1.1 Scope of Paper

The decision to limit the scope of our analysis to only New York City was based on the recognition that geographic location can play a critical role in shaping the prevalence and accessibility to greenspaces. By focusing on a single geographic location, we aimed to provide a more comprehensive and in-depth examination of correlated metrics in a well-defined context. The dense population and diverse demographic mix of New York City make it an ideal laboratory for exploring the complex interplay between various metrics and greenspace accesibility. This approach allows us to concentrate our efforts and resources on a specific area, resulting in a more precise and meaningful analysis. The data obtained from this analysis will not only provide valuable insights into the situation in New York City, but it will also serve as a reference point for future studies and policy decisions.

3.1.2 CEHI Ecohealth Index

We were provided with a dataset containing the Community EcoHealth Index (CEHI) for various communities in the United States. A version of the index, $\text{CEHI}_{\text{NIndW}}$, was calculated to allow comparisons between neighbourhoods of a single, larger community such as a metropolitan city. For this reason, we chose to use $\text{CEHI}_{\text{NIndW}}$, as the single summary metric related to a community's ecohealth since we are comparing data between the five borough of New York City (Manhattan, Bronx, Brooklyn, Queens, and Staten Island).

3.1.3 New York City Open Parks Dataset

We begin by collecting a public dataset provided by the state of New York with data on all its public parks and green spaces (found here). We obtain several features for each park described in the table 1 below.

Feature	Description
Geom	Geographical shape as defined by latitudinal and longitudinal points
Location	Park's address in NYC
Zipcode	Exact zipcode of park
Borough	Which of the five borough the park fell under
Acreage	Total acres measured for the park
Retirement States	If the park is still open to public, or if it's retired
Class	Park, Playground, or Zone
Subcategory	Neighborhood Park, Garden, etc.
Park Name	Unique Name of the specific park
Waterfront Status	If the park has a body of water or not

Table 1: Features obtained from the New York City Open Parks dataset

Then, using an online dataset, found at Kaggle Fips dataset, we obtain official census information mapping each zipcode in the US to its corresponding FIPS code(s). This information allows us to assign CEHI index to any zipcode since the dataset provided by Citadel consists of mappingss between GEOIDs and their CEHI index, and the first five digits of a GEOID are its corresponding FIPS. We notice that each zipcode has approximately 5-20 respective CEHI values, so we assume that an empirical average would be reflective of the overall CEHI in a zipcode. Thus, we assign the average CEHI of a zipcode to each park in our New York City Open Parks dataset. We use a similar process as above with the "percent_cover_tracts_with_buffer.txt" dataset provided by Citadel to quantify the accessibility to parks within a zipcode.

3.1.4 Collecting Metrics for Predicting CEHI

In order to determine if any other standard metrics can be used to predict CEHI of a given community, we collect demographic data that may serve as key metrics using information fund here. We obtain metrics such as average income, national ranking of average income, and population from each zipcode in New York City. We associate each of these demographic data to its respective park in our New York City Open Parks Dataset.

We now gather crowed sourced data from the social media app Flickr to determine if number of average posts about a park can be used as a metric to predict a community's CEHI. We collect data relating to a specific park over a period of 6 years, from January 1st 2010 to January 1st 2016. The Flickr API has the ability to make queries based on specific geolocations as well as post captions/tag, so we use relating data as the main search parameters. More specifically, utilize the bounding boxes for each geolocation query, and then pass in the name of the park as a search term. We then count the total number of posts that match the search query and compile that into a single visitation frequency metric.

	park_name	zipcode	Categorical CEHI	Population	Average Zip Income	Income National Ranking	Average Park Coverage in Zipcode	Flickr Post Counts	acres	borough	retired	category	subcategory	waterfront?
0	Seaside Wildlife Nature Park	10308	Category 2	26451	61868.000	2588	13.150	1745	20.907	R	False	PARK	Neighborhood Park	False
1	Strippoli Square	11377	Category 3	88339	37360.000	14527	7.892	0	0.061	Q	False	PARK	Triangle/Plaza	False
2	D'Emic Playground	11232	Category 4	27723	28395.000	25283	6.580	0	1.130	B	False	PARK	Playground	False
3	Harding Park	10473	Category 4	56166	27733.000	25921	11.935	1752	2.160	X	False	PARK	Neighborhood Park	False
4	Wakefield Playground	10470	Category 4	15780	38464.000	13496	10.718	85	1.104	X	False	PARK	Jointly Operated Playground	False

Figure 1: Final list of features obtained compiled into a single dataset.

3.2 Exploratory Data Analysis

3.2.1 Issues with CEHI Granularity

We start by examining how each feature interacts with the CEHI index to gain insight into the underlying relationships. However, as we mapped CEHI against other predictors, it becomes apparent that there is little to no evidence for underlying correlation between the chosen features and CEHI. Figure 2 is an example of what one of the graphs (CEHI vs. Population)

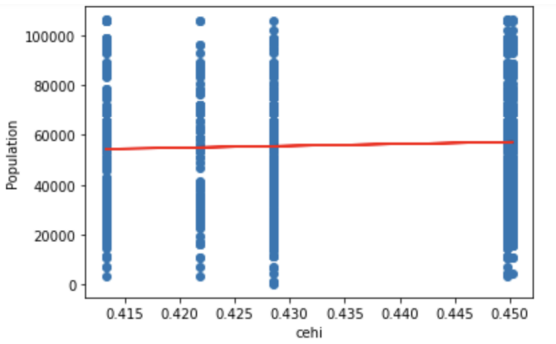

Figure 2: Sample Plot

This suggests that there may be some granularity between the CEHI indexes given to us. So, we graph the specific CEHI values on a histogram shown in figure 3 to examine the frequency of distribution.

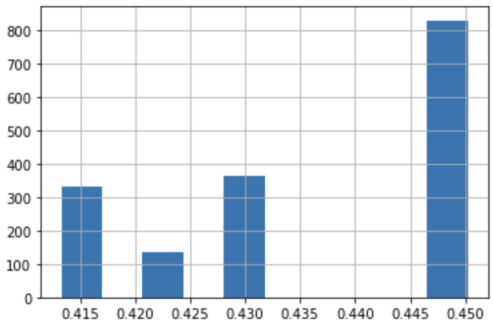

Figure 3: Frequency distribution of CEHI indexes. Note that all values fall into 4 distinct bins.

We discover that that all of our CEHI values were densely packed into 4 unique groups. We hypothesize that this may be due to the fact that our data analysis specifically concerns parks in New York City,

Categorical CEHI	borough	retired	category	subcategory	waterfront?
Category 1	B	False	PARK	Garden	False
Category 2	B	False	PARK	Neighborhood Park	False
Category 3	B	False	PARK	Garden	False
Category 4	B	False	PARK	Triangle/Plaza	False

Figure 4: Grouped Data by mode (Qualitative Variables)

Categorical CEHI	cehi	Population	Average Zip Income	Income National Ranking	Average Park Coverage in Zipcode	Flickr Post Counts	acres
Category 1	0.413	55110.596	36909.762	17838.440	9.758	460.889	19.325
Category 2	0.422	55218.331	35425.051	18763.419	9.663	574.882	15.982
Category 3	0.429	54179.086	36953.334	17782.622	9.748	537.091	14.902
Category 4	0.450	57376.286	36355.643	18080.375	9.852	500.551	12.506

Figure 5: Grouped Data by mean (Quantitative Variables) - as with the qualitative variables, the quantitative variables had little variation across clusters.

so the CEHI index for varios communities would similar to one another. Therefore, we realized that in order to run any form of predictive analysis, we need to order our CEHI bins as categorical variables ranging from category 1 (low) up to category 4 (high) and run classificaiton tasks. We used the following conditions to split the CEHI into categories:

$$\text{If } 0.42 \leq \text{CEHI} \rightarrow \text{Category 1}$$
$$\text{If } 0.42 < \text{CEHI} \leq 0.425 \rightarrow \text{Category 2}$$
$$\text{If } 0.425 < \text{CEHI} \leq 0.44 \rightarrow \text{Category 3}$$
$$\text{If } 0.44 < \text{CEHI} \rightarrow \text{Category 4}$$

3.2.2 Regression Analysis

We proceeded by grouping our dataset by our CEHI categories and got the information about our groupings seen in Figures 4 and 5.

The table made it difficult to identify any particular trend between which CEHI category a datapoint would be in and any particular predictor, except, we noticed that there was a correlation with acreage, since the average acreage seemed to decrease as CEHI rose. So, we did a basic k-nearest-neighbors classification test between the CEHI category and acreage as shown in Figure 6. We chose this model since it is known to be fruitful in low-dimension settings such as this one.

Unfortunately, it seems that the testing accuracy of this model tapers out eventually at less than 0.5 (not at all a good accuracy), thus not showing great single-variable correlation between acreage and CEHI category. However, we still want to consider acreage because in a multivariable classification it may prove useful.

Next, we decided to run a k-prototypes clustering algorithm on all our data to see what trends we could pick up. k-prototypes is an algorithm simiilar to k-means and k-modes that creates arbitrary clusters based on trends recognized by the algorithm. The reason we use k-prototypes is because we have a

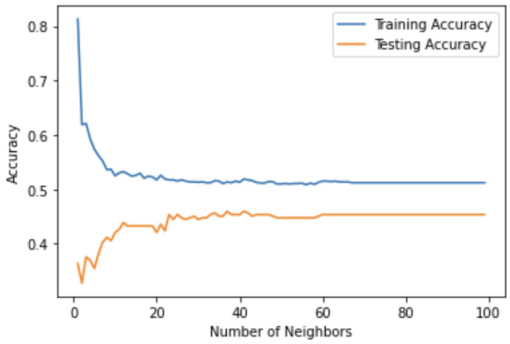

Figure 6: K-Nearest-Neighbors Accuracy Plot (Acres vs CEHI Category)

	Segment	Categorical CEHI	borough	retired	category	subcategory	waterfront?	Population	Average Zip Income	Average Tree Cover in Zipcode	Flickr Post Counts	acres
0	First	498	B	False	PARK	Garden	False	85652.187	30354.861	8.802	464.833	11.031
1	Second	424	B	False	PARK	Garden	False	61034.024	30087.290	9.816	497.448	7.974
2	Third	435	Q	False	PARK	Triangle/Plaza	False	35619.503	56077.791	10.302	606.368	29.734
3	Fourth	304	Q	False	PARK	Garden	False	29839.589	27608.299	10.667	445.424	8.447

Figure 7: K-Prototypes Clustering Information

mix of qualitative and quantitative variables, so we do need to consider both means and modes. Through this clustering, we achieved the information in Figure 7.

Through this clustering alone, we could start to see that the algorithm likely didn't identify a separation trend based on CEHI category. This is because when we compare the information of this clustering versus the information we get when grouping by the CEHI categories (as shown below), none of the clusters really seem to align with any of the groupings.

At this point, its more clear that there isn't a statistical significance in predicting CEHI category based on the predictors we gathered, however we wanted to confirm this through a series of tests.

So, we did the following tests: (NOTE before all tests: We dropped the predictor of "National Ranking of Average Income" since it is obviously correlated strongly with the predictor of "Average Zipcode Income" as justified both logically and visually by the graph in Figure 8).

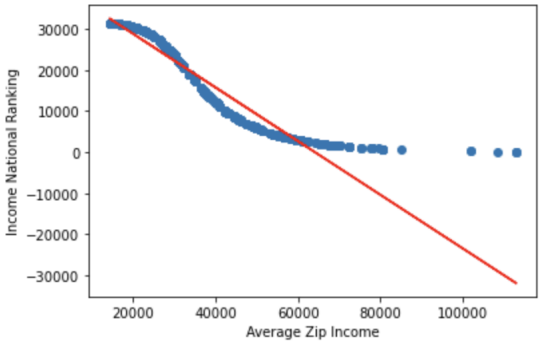

Figure 8: Correlation graph between "Income National Ranking" and "Average Zip Income"

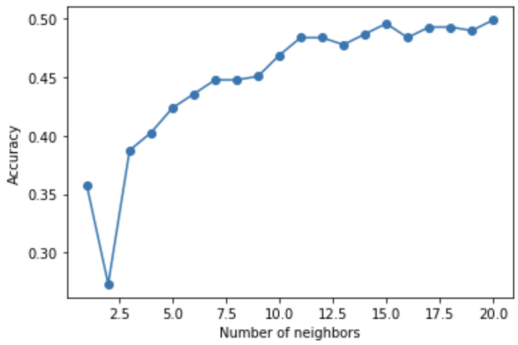

Figure 9: K-Nearest-Neighbors Classification Accuracies (CEHI Category vs all Predictors)

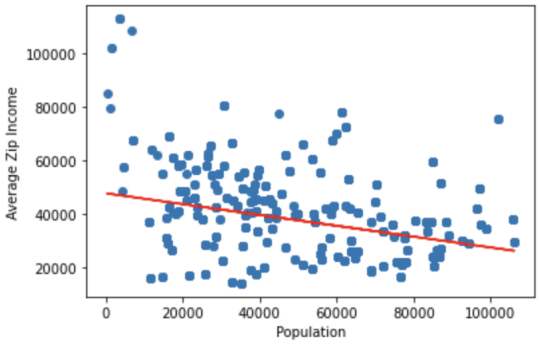

Figure 10: First Example of Uncorrelated Plot. This plot, visually, elucidates that there isn't much of a correlation between Population and Average Zip Income, somewhat implying independence between them.

3.2.3 K-Nearest-Neighbors

We did a KNN classification approach to identify the accuracy of CEHI category classification. Going into the model, even if (a big if) there is a significant correlation, we aren't expecting to justify the trend by this model since we have a significant amount of predictors and KNN loses accuracy as predictors increase due to the distancing of datapoints in higher dimensions (also known as the curse of dimensionality). We used hot-coding to quantify our qualitative variables to make this model possible. However, our model proved unsuccessful, tapering off in accuracy at less than 50 percent as shown by Figure 9.

3.2.4 Naive-Bayes Classifier

Next, we used a Naive-Bayes Classifier. We actually did have some hope for this model since it is built on the assumption that each characteristic is independent from one another, which seems to hold true based on the pair-wise graphs of our predictors. We've included a couple in Figures 10 and 11 as reference.

However, once again, our model is not very accurate, circling at an accuracy around 0.5. To confirm that the accuracy was actually that low, we did a 5 K-folds validation and generated Figure 12 that illustrates the failure of this model.

3.2.5 Decision Tree Model

We continued our thorough verification by using a decision-tree model with K-folds validation. Decision trees are good models when there is a mix of qualitative and quantitative variables, so we were optimistic

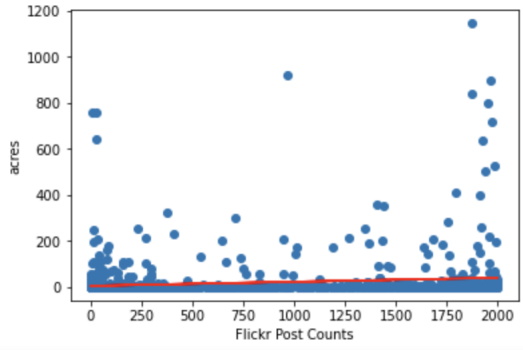

Figure 11: Second Example of Uncorrelated Plot. We also found that the size of the park was relatively uncorrelated with the number of Flickr posts. This could likely be due to the simultaneous heavy-tailedness and sparsity of the counts distribution.

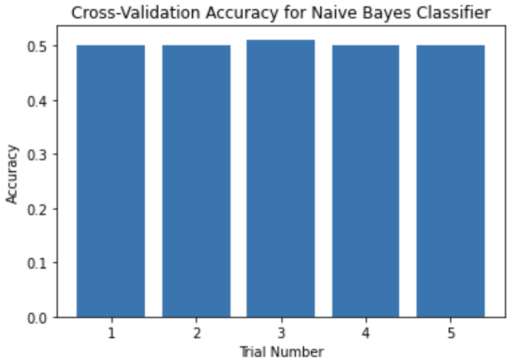

Figure 12: Naive Bayes Success over 5 Trials - We see that running across multiple initialization yields similar success rates.

for the results. We still generated an accuracy graph (Figure 13) that is equally as much of a failure as that of the Naive-Bayes classifier.

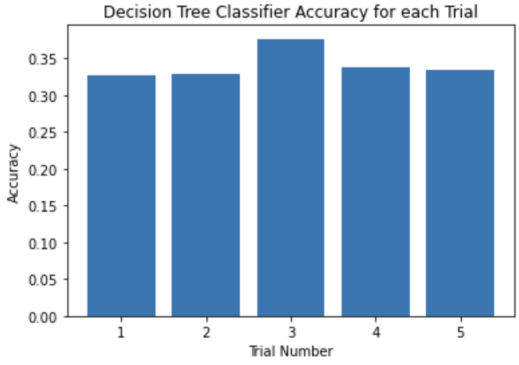

Figure 13: Decision Tree Success over 5 Trials

3.2.6 Understanding

From our data and repeated models, it became clear that predicting CEHI data using the predictors we had, including the crowd-sourced Flickr information, wasn't a success. So, it begged the question of why this didn't work. To answer this, we ran a random forests classifier model. This is because not only

does random forests reaffirm what we already know (that we aren't predicting CEHI Category well with our current predictors), but it also provides a ranking of which features are most useful for this. The reason why is because a random forests algorithm is an aggregation of running multiple decision tree algorithms, and decision trees are built on the fundamental information theory principle of weighting, evaluating, and comparing features against one another. Thus, the following weighting of features in Figure 14 holds strongly.

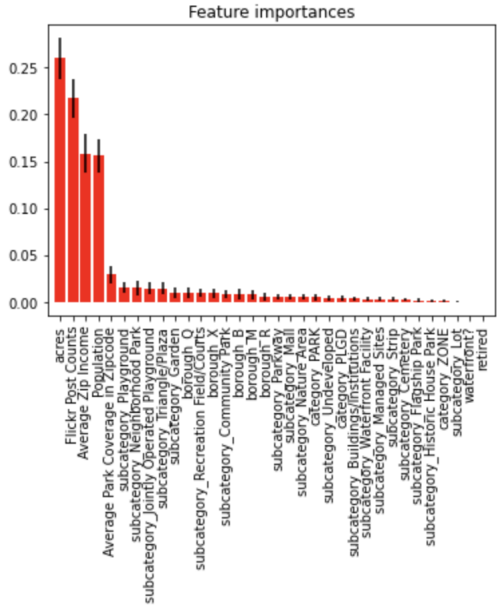

Figure 14: Ranking the Strength of Predictors by Random Forests Model

Figure 14 aligns with our previous observation that (relative to other predictors) acreage could be somewhat useful in predicting CEHI Category, however it shows that none of our predictors are very significant for this type of analysis.

3.3 Satellite Image Processing

Since analyzing Flickr data proved to not bear significance in identifying CEHI category, our next goal became to use advances in deep learning to facilitate identification of features of urban ecosystems. To do so, we used the datasets provided by the Google Earth Engine API (Gorelick et al., 2017). As described earlier, our motivation was to analyze satellite imagery, due to its prevalence and ease of analysis, even in low-resource settings where performing standard surveys or data collection procedures would be prohibitive. First, we decided to use human-readable satellite imagery, analogous to those found on Google Earth, since choosing a more specific type of imaging data could limit our selection of datasets to those that might be frequently updated or not available in different parts of the globe. Moreover, we wanted our dataset to be high-resolution, so that it could capture features of roads and trees at the level of our parks dataset.

Eventually, our search led us to the National Agricultural Imagery Program (NAIP), a nationwide dataset collecting high-resolution RGB and infrared imagery of the United States. We chose this dataset since it allowed us to query any latitude and longitude position within the continental United States

through the Earth Engine API, and because it provided data at very high resolution (0.6 meters per pixel). After initial inspection, we were able to generate low-level pictures for all 2015 parks in our dataset. Our hypothesis was that using such high-resolution imagery would give our models the flexibility to identify low-level geographic features and estimate quantities of interest in computing CEHI, such as tree cover or access to water, and require limited preprocessing compared to other image modalities.

3.4 Predicting CEHI using CNNs

As a catch-all metric for describing an urban ecosystem, CEHI offers a potentially useful diagnostic tool to flag communities in need of policy-based intervention. However, the metrics from which CEHI is calculated, such as tree cover along walkable streets or other geographical features can be difficult to calculate and vary in surveys, particularly in low-resource settings. Nevertheless, since CEHI is dependent on certain features readily visible from satellite imagery, we hypothesized that one could obtain a reasonable approximation of predicted CEHI value solely by analyzing high-resolution satellite images from various neighborhoods. Doing so would eliminate significant overhead in estimating the CEHI's constituent metrics and encourage a larger-scale, more data-driven approach to flagging possibly underserved neighborhoods.

Since our prior parks dataset contained the most specific information in terms of latitude and longitude, we focused on collecting satellite images above these parks and their surrounding neighborhood, as well as the associated CEHI index for the relevant county. We were able to isolate 1812 parks that contained both the CEHI as well as precise geographical coordinates, and used the Google Earth Engine API, as described above, to query an overhead view of the parks. To standardize inputs, we took the central 256 pixel square from each image, as visual inspection determined that this size captured both the park as well as parts of the neighborhood, but also respected our compute constraints. After this, we performed standard data whitening procedures (subtracting the pixel-wise mean mean and dividing by the standard deviation).

3.4.1 Neural Architecture

In this section, we describe the architecture used for our prediction task. The nature of our task motivated several design decisions, both in terms of the architecture as well as the training process. First, compared to even basic computer vision tasks, such as digit classification (MNIST), our dataset was very small ($\approx$1000 versus $\approx$60000 samples), so our initial experiments consisted of training on the entire dataset to first understand the properties of the learning problem before validating our architecture designs. Our compute constraints also placed an implicit restriction on the length of the training process. Ultimately, we chose to use neural networks because of their ubiquity in modern machine learning systems and improved performance over alternative methods in nearly all computer vision tasks.

As a result, our architecture was structured as follows we used two blocks of a convolutional layer, Dropout layer, MaxPool layer, and BatchNorm layer (Ioffe and Szegedy, 2015), another block without the MaxPool layer, and two fully connected layers with 175 and 25 hidden units, respectively. Each block was followed by the application of a rectified linear unit (ReLU) function. The output was a scalar corresponding to the predicted CEHI value. Both convolutional layers had a kernel size of 5×5 with

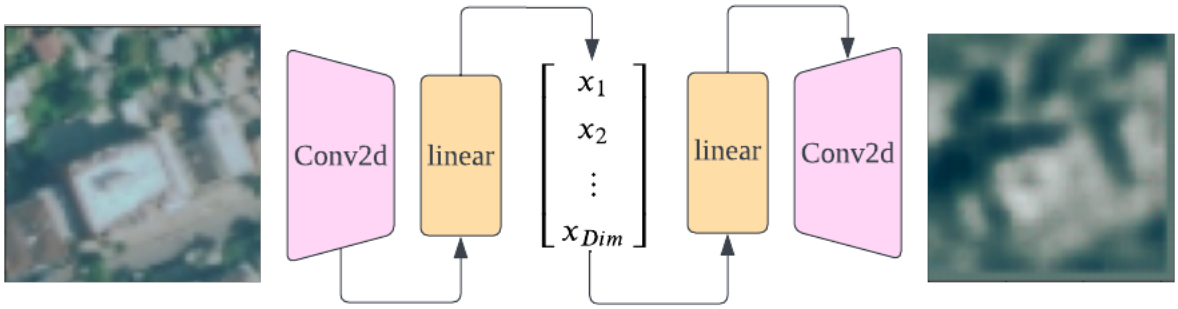

Figure 15: Our convolutional autoencoder takes in as input the raw image, uses convolutional layers to extract semantic image features from the image, and projects the features into a lower dimensional space using a series of linear mappings. The hidden layer as seen above (of dimension $\mathbb{R}^{\text{Dim}}$) corresponds to the encoding of the image. The decoder takes in as input the encoding and uses a symmetric architecture to the encoder to produce a reconstructed image.

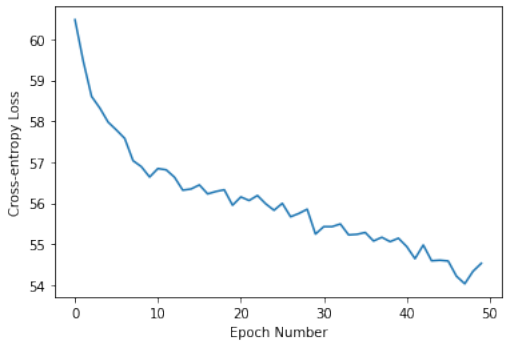

Figure 16: We trained a classifier to predict CEHI bin (equivalent to county) on the parks dataset, yet found that the model was unable to converge to a sufficiently low-loss solution.

a stride of 2 and no padding. Each MaxPool layer had a filter size of 5 as well. This was to ensure that we could aggressively downsample the size of the image so that our model could fit in memory and be trained in a time-efficient manner. Further, since our dataset was quite small, we used extensive regularization through the Dropout (Hinton et al., 2012) and BatchNorm layers.

3.4.2 Training and Results

To train the model, we used a mean squared error loss and the AdamW optimizer, an analogue of the highly-cited Adam optimizer that adds weight decay (similar to an ℓ_2 penalty in linear regression), another form of regularization. We made this choice largely due to the dataset size, which meant that we could not operate in the standard computer vision regime, where the most common optimizer is stochastic gradient descent (SGD), and models often have several blocks of convolutional layers.

We trained our first model on the entire dataset for 50 epochs and found that after lowering the learning rate to roughly $1e-5$, the training was able to converge. Higher settings caused divergence within the first several epochs, and resulted in highly suboptimal solutions. After the training converged, we plotted the training labels alongside the predicted values, and found that the labels fell into a very small number

of discrete bins, likely due to the fact that the CEHI data was provided at the granularity of county or zip code, so many parks in the same region would have the exact same label. We realized this would likely limit the applicability of our findings, since if we were not able to have low-level, accurate estimates of CEHI, we would need to modify the training problem.

Therefore, we discretized the label into one of five bins depending on the county, and modified our previous architecture to output a vector consisting of a probability distribution over the five counties. In other words, the model would predict which county the image came from, which would be correlated to the CEHI based on our labels. The only modification to the training procedure that we made was to switch the loss from mean squared error to a cross entropy loss. This model was able to converge after 50 epochs with a learning rate near $1e-5$, and achieved slightly better performance than random guessing 40 percent on the dataset. However, the fact that this accuracy was on the training set showed that our model was still heavily underfitting the data due to the small dataset size and lack of diverse labels. As such, our analysis demonstrates the need for more standardized data collection procedures at a higher resolution, so that data-hungry methods like deep learning can effectively be applied.

3.5 Autoencoder Analysis

Since our classifier was underfit, our next goal was to build a proof-of-concept that neural networks could extract meaningful data about urban greenspaces from satellite imagery. Indeed, if we could train a model to extract meaningful representations from the data in a self-supervised fashion, given sufficient scale, these representations could be used in downstream analyses, like the previous task of predicting CEHI.

3.5.1 Data Preparation and Training

Since we no longer needed to make predictions on a large context size, to increase the size of our dataset, we used the previous dataset of 256×256 images and split each image into 64×64 pixel patches. Our rationale was that doing so would allow the model to focus on extracting low-level information such as individual trees, and the aggregate features extracted for each patch within a geographical area would correspond to some overall semantic description of the region.

More formally, the architecture we used was a convolutional autoencoder, where both the encoder and decoder were symmetric. The encoder was composed of two convolutional layers with 8 and 16 layers and a kernel size of 3, respectively, followed by a BatchNorm layer and another convolutional layer with 32 filters and the same kernel size. This was then processed by two fully-connected layers to produce a compact representation of size 64 - a 192x reduction in size compared to the original input. The decoder would then map this set of latent variables to a 64×64 image. The model was trained using a standard pixel-wise ℓ_2 norm reconstruction loss, and through the course of training, we varied the optimization algorithm.

First, we trained the model using SGD with several learning rates - ranging from $1e-1$ to $1e-6$, and found that this model converged to poor solutions with highly suboptimal reconstruction fidelity. Therefore, we switched to the AdamW optimizer with a learning rate of $1e-3$, and found that this

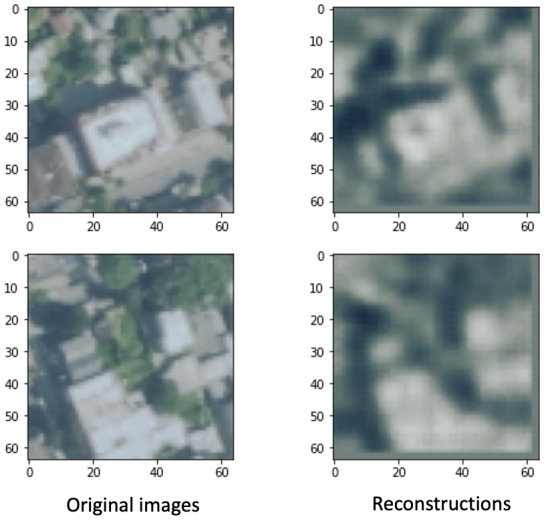

Figure 17: Our autoencoder achieves good reconstruction loss as well as reasonable reconstruction fidelity from visual inspection.

performed the best, even after further modifications such as learning rate scheduling. We plotted some sample reconstructions to validate the quality of the model, and found that while the reconstructions were relatively blurry (likely an artifact of using the ℓ_2 norm as a loss function), the reconstructions captured the basic features of buildings and roads.

Ultimately, the goal of this analysis was to further interrogate the quality of the intermediate representations learned by the model. Therefore, for the entire dataset, we computed the encodings and performed dimensionality reduction using PCA. Since the first two components had a high proportion of explained variance (≈ 50 %), we hypothesized that the visualization would be representative of the underlying structure of the data. We then overlaid the CEHI value of each park in the dataset to visually inspect whether any regions in the latent space corresponded to differences in CEHI value. However, from visual inspection, the CEHI values for each park were relatively uniform across the latent space, meaning that distance in latent space was likely a poor predictor of variation in CEHI between two parks. Training on a larger dataset to achieve better reconstructions or using more granular labels (CEHI at the level of individual blocks) may have improved the conclusions we could have drawn from this analysis as well, but we were limited by the scope of the CEHI data. Nevertheless, the fact that the autoencoder was able to successfully reconstruct the satellite imagery shows promise that larger models trained on potentially global satellite data could be used in downstream tasks in the future.

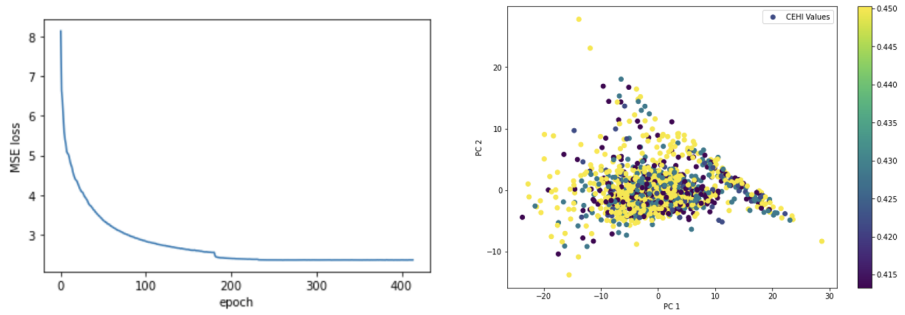

Figure 18: (a) The best model, trained with the AdamW optimizer (Loshchilov and Hutter, 2017), achieves good reconstruction accuracy and eventually converges. (b) While the PCA of the encodings of the images shows that the first dimension does indeed capture a significant proportion of the variance, it appears that distances in latent space do not correlate well with CEHI values, meaning that the underlying representation likely discriminates images on more low-level features and less on semantic information.

4 Conclusion

Our research demonstrates that while social media data contains important information about behavioral patterns of urban residents and general information about urban ecosystems, significant work remains to be done to make large-scale crowdsourced applications possible. First, a more concerted effort is needed from policy-decision makers to standardize data collection procedures via GEOID disambiguation and collaborate with academic researchers to ensure that robust analyses can be performed. For example, in our research, we found limited relationships between visitation frequency and CEHI, in part due to the difficulty of finding universal geolocated data at the resolution of the parks we were analyzing. In addition, our work highlights how satellite imagery can be a potentially powerful tool to diagnose urban ecosystem health. However, having more fine-grained information, such as urban health level at the granularity of high-resolution imaging (i.e., at the block or residential area level) can further accelerate the training and adoption of deep learning methods - especially because in our analyses, we found that a lack of data caused our model to significantly underfit to the task. Moreover, using more sophisticated self-supervised learning techniques, such as masked autoencoders (MAEs) or variational autoencoders (VAEs) could improve the quality of the representations, since the model we chose was limited in capacity due to our computational constraints. Ultimately, the goal of urban policy-makers should be to have a set of tools that allows them to quickly and accurately pinpoint relevant metrics of urban greenspaces, such as visitation frequency and general health. Integrating crowdsourced data in the future will ensure greater robustness and faithfulness to true population dynamics in this procedure.

References

Cochran, F., L. Jackson, A. Neale, J. Lovette, and L. Tran (2019, August). A community EcoHealth index from EnviroAtlas ecosystem services metrics. *International Journal of Environmental Research and Public Health 16*(15), 2760.

Donahue, M. L., B. L. Keeler, S. A. Wood, D. M. Fisher, Z. A. Hamstead, and T. McPhearson (2018).

Using social media to understand drivers of urban park visitation in the twin cities, mn. *Landscape and Urban Planning 175*, 1–10.

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*.

Hinton, G. E., N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2012). Improving neural networks by preventing co-adaptation of feature detectors. *CoRR abs/1207.0580*.

Ilieva, R. T. and T. McPhearson (2018, October). Social-media data for urban sustainability. *Nature Sustainability 1*(10), 553–565.

Ioffe, S. and C. Szegedy (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR abs/1502.03167*.

Loshchilov, I. and F. Hutter (2017). Fixing weight decay regularization in adam. *CoRR abs/1711.05101*.