Measuring spatial-temporal change of physical conditions in neighborhoods with street view imagery

Daniel Chen
Stanford University
danschen@stanford.edu

Tingyan Deng
Vanderbilt University
tingyan.deng@vanderbilt.edu

Evelyn Fitzgerald
Cornell University
acf67@cornell.edu

Lijing Wang
Stanford University
lijing52@stanford.edu

Jackelyn Hwang
Stanford University
jihwang@stanford.edu

1. Introduction

Neighborhood environments play a significant role in shaping the well-being of individuals and communities, consequently contributing to inequality in the U.S. There is limited empirical evidence of how physical and environmental neighborhood conditions affect well-being partly because collecting data on the conditions of places, especially across cities and over time, requires extensive time and labor. This project aims to improve understanding of the role of physical and environmental neighborhood conditions in structuring inequality in well-being in U.S. cities. By doing so, this project aims to inform solutions that can improve the well-being of residents and reduce inequities.

Specifically, the project draw on street view image data across five U.S. cities over 11 years. Drawing on the images, we used crowdsourcing and computer vision techniques to quantitatively measure physical aspects of urban neighborhood environments that have been linked to well-being (e.g., trash; levels, types, and maintenance of greenery; building materials, layouts, and maintenance; and frontage maintenance). We linked these measures with existing data on individual and community well-being to analyze how the physical and environmental conditions of neighborhoods affect the well-being of individuals and neighborhoods.

2. Related work

Previous research shows that the presence of physical disorder, poorly maintained properties, and vacant lots in neighborhoods can negatively affect neighborhood physical and mental health; it can also attract more crime and disorder. These problems lead to neighborhood disinvestment. In addition, many studies related to residential well-being have shown that greenery is associated with higher well-being.

Clearly, neighborhood environments play a significant role in shaping the well-being of individuals and communities and contributing to inequality in the U.S. However, very few studies have investigated this process because collecting data on the physical conditions of neighborhoods, especially across space and time, is labor-intensive.

3. Data collection

The census block and street names are used to sample the images. We do not download an entire city's worth of images; rather, we download a sample of them. The sampling process is as follows: we scrape four addresses from each side of each city block – so four addresses per block per street. In the image below, this would mean four addresses from each of: Street A, Block 1; Street B, Block 1; Street D, Block 1; Street E, Block 1; Street A, Block 2; Street B, Block 2; Street E, Block 2; Street F, Block 2; Street B, Block 3; Street C, Block 3; Street D, Block 3; Street E, Block 3; Street B, Block 4; Street C, Block 4; Street E, Block 4; Street F, Block 4.

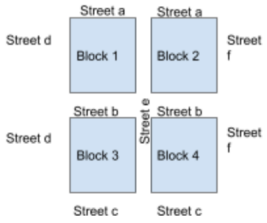

Figure 1. sampling process illustration

We scrape three different kinds of images: buildings, frontage, and trash.

The primary dataset includes 15 million scraped images from Google Street View spanning from 2007 to 2017, with the images taken every 1–3 years. Across the five U.S. cities, about 4 parcels per street segment are included in

the dataset, and photographs of each parcel are sampled at three different angles to capture the building, the exterior area, and the street and sidewalk.

Then, in order to create training data, we used the crowdsourcing platform Amazon Mechanical Turk (MTurk). We created survey and testing tools and published raw images scraped from Google Street View. We found that pairwise surveys are more reliable in producing comparisons across a set of objects during our experiment. Thus, MTurkers were asked to compare the two images and answer which one has a better upkeep after they pass a test to determine if they can make reasonable decisions given some example building images. To make our training data more comprehensive and robust, each image is compared at least 18 times to produce stable estimates. Our training set consists of 2,964 images from Boston and 3,995 images from Detroit.

3.1. Training and validation sets

The TrueSkill score algorithm was then applied after we receive the comparison of the images to rank the images and each image was given a TrueSkill score with 25 being the median for each city. We then qualitatively create cutoffs on TrueSkill score to create 4 classes (classes 0–3). An example of some output TrueSkill scores and classes figure is shown in Figure 2 below. The top row is the detroit image and bottom row is the Boston image. The first number of each image denotes the image class. The higher the class, the better the building upkeep. The second number of each image denotes the TrueSkill score associated with image. Figure 2 contains all four classes for both cities. As we go from left to right, the TrueSkill score is getting lower while the class number is getting higher and the upkeep is getting better. A smaller TrueSkill score means the building has a good upkeep and a high class number.

As a visual inspection, Boston tend to have a better upkeep compared to Detroit as you can tell from the first three images from the Figure 2. To further examine the accuracy of this assumption, the table that contains the number of images of each city in each is shown below in Table 1.

City	Boston	Detroit
Class 0	3	307
Class 1	48	690
Class 2	2,540	2,988
Class 3	373	10
Total	2,964	3,995

Table 1. The number of images from each each class for Boston and Detroit.

As can be seen in the tables detailing the class splits above, Boston has a much smaller amount of Class 0 images and a high amount of Class 3 images, which indicates Boston in general has a better upkeep than Detroit. Also, there is a massive class imbalance present amongst all labeling schemes. To mitigate this, we find it is crucial to upsample the minority classes to give the impression of balanced classes each epoch, which is discussed later in the paper. To summarize, we have 2,964 Boston images and 3,995 Detroit images, each with a TrueSkill score ranging from 0 to 50 and a class label from 0 to 3, which is determined qualitatively based on the TrueSkill scores.

3.2. Time-series data

The time-series data for Boston was also scraped in the same fashion discussed above. Table 2 contains the image information for Boston.

Year	Frequency
2007	24,094
2008	52
2009	15,281
2010	74
2011	29,293
2012	596
2013	23,564
2014	14,955
2015	673
2016	10,413
2017	9,661
Total	128,656

Table 2. The number of images from each year in the Boston time-series dataset.

From table 2, we realized that there were far fewer images for even numbered years than for odd numbered years. Thus, for our time-series analysis (discussed in section 5.3), we grouped consecutive years together (e.g. 2007–08).

4. Methodology

4.1. Classification vs regression

In past work, the challenge of predicting building upkeep was framed as a classification task. That is, past models on this project have aimed to predict building upkeep at discrete levels of upkeep, using the 4 classes described in the Data Collection section. However, this classification method had a number of limitations. Specifically, posing the problem as a classification task caused issues in class imbalance and model interpretability.

First, because the dataset (see Table 1) has extreme class imbalance (with the vast majority of images being in class 2), the model had difficulty predicting the minority classes. Methods were used to help alleviate this issue, such as upsampling of the minority classes in the training set. While such methods improved model performance, the model still performed quite poorly on these minority classes. One possible reason why this might be the case is that there is a large

Figure 2. Google street view images example for Detroit and Boston

variability in upkeep within class 2 that is not learned by the model. Because all such images are lumped into class 2, there is no way to distinguish between the best upkeep class 2 image from the worst upkeep class 2 image. This possibility is further evidenced by the fact that class 2 covers a much broader range of TrueSkill scores than the other classes. For instance, in the Detroit images from the training set, the range of TrueSkill scores for class 2 is 19.14 to 26.58, while the range for class 1 is only 26.58 to 28.55. As such, the model may struggle to learn some of the nuances of upkeep between images since class 2 covers such a large range of upkeep values.

Second, the classes used to indicate upkeep were difficult to interpret. It was apparent that changes in TrueSkill score reflected changes in building upkeep. Thus, on average, the different classes reflect different levels of upkeep. However, it was not clear on the level of individual images what differences between classes were. For instance, if two images close to each side of the boundary between class 1 and class 2 were compared, it was not apparent which image belonged to which class. Thus, differences in upkeep between images might be inconsistent given that the boundaries between classes did not reflect discrete differences in upkeep. Because there was no distinguishing feature that reflected differences among classes, it was difficult to ascertain what each class really reflected about upkeep.

To help alleviate this issue, we thought that shifting to a regression problem might help both class imbalance and model interpretability. Instead of predicting discrete classes, the goal would be to predict the continuous TrueSkill scores that were used to derive the classes. First, shifting to a regression might force the model to learn more about the nuances between images in the same class, especially within class 2. For instance, a high upkeep class 2 image would need to be differentiated from a low upkeep class 2 image. In doing so, the model would also learn more about what differentiates images originally from class 2 from images originally in the minority classes. Thus, using a regression model should help improve prediction for images previously in the minority classes. Second, shifting to a regression would make the model more interpretable. Since the output of the model would be a TrueSkill score, that score would clearly delineate the relationships between images. The model would be more interpretable since every image would be on the same continuous scale, and would reflect the incremental differences in upkeep between images. As such, it would be much clearer what the individual differences were between pairs of images, helping improve model interpretability.

Thus, our new model treats the problem as a regression, instead of a classification task. Now, the model's goal is to use Google Street View images to predict the TrueSkill score rating of each image.

4.2. TrueSkill score alignment

Since our new model aims to predict TrueSkill scores, it must be able to interpret scores from both Detroit and Boston on the same scale. However, while the discrete classes are designed to be on the same scale across cities (that is, a class 0 in Detroit should reflect the same upkeep as a class 0 in Boston), the TrueSkill scores are not on the same scale. Since TrueSkill scores are derived based on comparisons between images (performed by MTurkers), the scale of the scores depend on the images that they are compared to. However, images were only compared to other images from the same city. Thus, Detroit images were never directly compared to Boston images, which means that the two are not on the same scale. This is most clear when you consider that the TrueSkill algorithm is designed so that the average score is 25. Thus, a score of 25 in De-

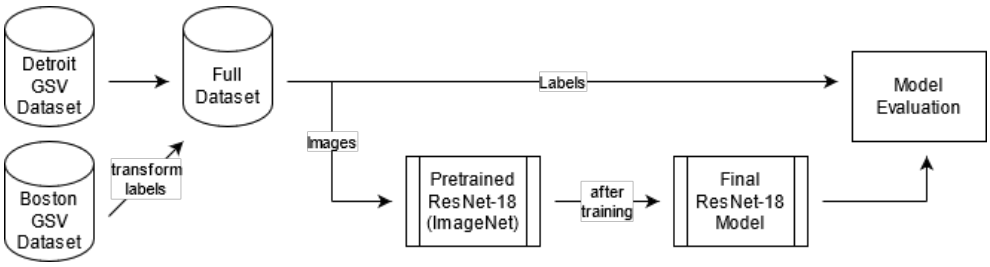

Figure 3. model architecture diagram

troit reflects the average upkeep in Detroit, and likewise for Boston. However, we know from our discrete class frequencies that Boston's distribution of upkeep is much better than Detroit's. As such, a TrueSkill score of 25 from Detroit should actually reflect poorer upkeep than a TrueSkill score of 25 from Boston.

To resolve this issue, the TrueSkill scores from Boston were aligned to the scale of TrueSkill scores from Detroit. Our current method of score alignment leverages the fact that while the TrueSkill scores aren't comparable across cities, the discrete class labels are. Specifically, we can look at the boundary TrueSkill scores for each class, and use those boundaries to align the TrueSkill scores to the same scale. Ideally, the boundary TrueSkill scores for each class in Boston should be the same as the boundary TrueSkill scores of the corresponding class in Detroit. As such, we can transform the TrueSkill scores from each class in Boston so that they line up with the corresponding class in Detroit. The exact algorithm used is described below.

For each city, we found the boundary of TrueSkill scores between each class (i.e. the average between the lower TrueSkill score of class 0 and the highest TrueSkill score of class 1, the same for class 1 and 2, 2 and 3). We also took the min and max TrueSkill score in each city as well. Thus, we have 5 scores for each city.

Then, for each score in Boston, we found which bucket it belonged to. For the sake of example, let's say an image had a class of 1, with label '29.1'. Thus, it would be in the bucket between the boundary of class 0 and 1 (let's say this value is '30.5'), and the boundary of class 1 and 2 (let's say this value is '29.0'). Note that these boundaries are the ones from BOSTON. Then, we would look for the corresponding boundaries, but this time in DETROIT (let's say the corresponding values are '31.5' and '30.5'). The transformation applied is then:

$$s_{new} = (s_b - \text{lo}_b)/(\text{hi}_b - \text{lo}_b) * (\text{hi}_d - \text{lo}_d) + \text{lo}_d$$

where s indicates the TrueSkill score, b indicates Boston and d indicates Detroit. Additionally, lo_{city} is the lower boundary of the bucket for that city, and hi_{city} is the upper boundary. For instance, in our previous example, our new score would be:

$$s_{new} = (29.1 - 29.0)/(30.5 - 29.0) * (31.5 - 30.5) + 30.5 = 30.567$$

Using this method, we create a piecewise linear transformation of the Boston scores to the Detroit scores. After running this transformation on the dataset, the TrueSkill score labels can be used as is for training and evaluating our model.

4.3. Model architecture

First, the Boston TrueSkill score labels are transformed to match the scale of the Detroit labels, as described previously. Then, the Boston training data and the Detroit training data are combined to create a final training set. Then, a ResNet-18 model that has already been pretrained on ImageNet is finetuned using the combined training (learning rate = 0.00001). This model is trained for 50 epochs, with checkpoints taken every epoch. The final model is the checkpoint that had the lowest mean-squared error based on the validation set (which contains images from both Boston and Detroit). Once this final model is created, it is then used for inference on both the validation set, as well as for inference on time-series data from various cities. For an overview of the model architecture, view Figure 3.

5. Results

5.1. Performance on validation set

The final model was used to predict the images from the validation dataset, which contains 1000 Detroit images and 741 Boston images. Using this model, there was a mean-squared error of 4.487. When the predicted TrueSkill scores were plotted against the ground truth TrueSkill scores (see Figure 4), there was an R^2 score of 0.484.

Class	Precision (naive)	Recall (naive)	Precision (current)	Recall (current)	N
0	0.05	0.05	0.42	0.27	86
1	0.09	0.09	0.25	0.22	164
2	0.81	0.81	0.86	0.90	1416
3	0.04	0.04	0.22	0.16	75

Table 3. Precision and recall of naive baseline model and current model

Class	Precision (previous)	Recall (previous)	Precision (current)	Recall (current)	N
0	0.19	0.37	0.43	0.27	84
1,2,3	0.94	0.85	0.94	0.97	916

Table 4. Precision and recall of previous model and current model using P/A split

Class	Precision (Shubhang)	Recall (Shubhang)	Precision (current)	Recall (current)	N
0,1	0.36	0.47	0.54	0.42	241
2,3	0.81	0.73	0.83	0.89	759

Table 5. Precision and recall of previous model and current model using L/H split

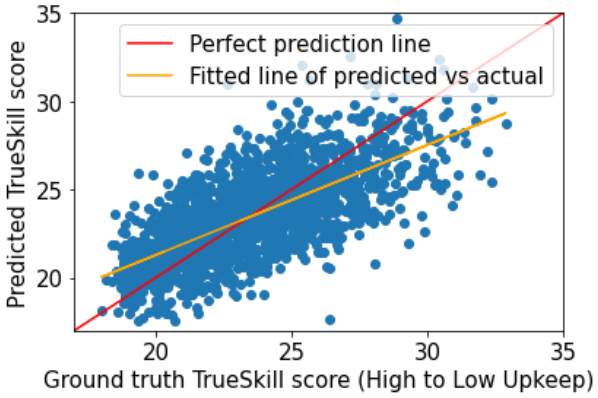

Figure 4. Predicted vs. ground truth TrueSkill score

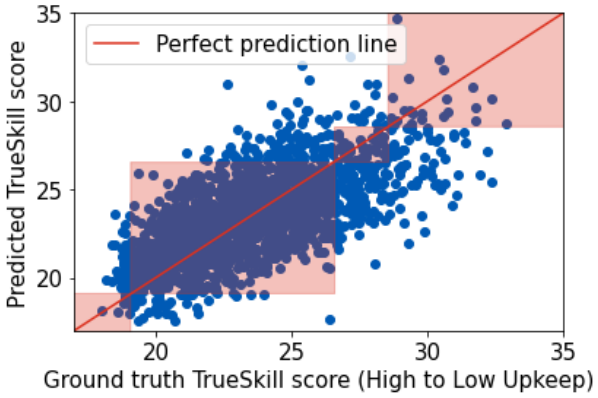

Figure 5. Correctly classified regions

5.2. Comparison to previous classification models

To determine whether the current regression model helped alleviate the issues with past classification models, the current regression results were transformed back into classes in order to compare results. To do this, we used the cutoffs that were originally applied to the Detroit training set (since all the predictions are on the scale of the Detroit TrueSkill scores) to turn TrueSkill scores into discrete classes. These cutoffs were directly applied to the predicted scores, to turn the predicted TrueSkill scores back into classes. Then, those classes were compared to the ground truth classes for those images. A plot visualizing the regions of correct classification can be found in Figure 5.

First, to compare classification accuracy across all 4 classes, the current model was compared to a naive baseline. This baseline randomly predicts each class based on the frequency of the class in the dataset. Specifically, P(Classifier chooses X) = number of images of X / total images. The precision and recall of our current model compared to this baseline can be found in Table 3. As you can see, the model has strictly better precision and recall, and this is especially the case for the minority classes (0, 1, and 3).

Additionally, we also compared our model's performance to previous classification models (created by Shubhang Desai) used for the current task. However, previous models did not use all 4 classes when classifying. Instead, there were two types of splits for the data - 'presence/absence' (P/A, class 0 vs class 1, 2, and 3), and 'low/high' (L/H, class 0 and 1 vs class 2 and 3). Additionally, previous models only trained and validated on Detroit data. Thus, to ensure that we had an equal comparison between models, we removed the Boston validation points from our model predictions, and combined our model predictions into the P/A and L/H classes.

Across both the P/A and L/H splits (see Table 4 and Table 5), our current model outperforms the previous classification model used. For the majority class for both splits, the precision and recall of the current model is higher than the previous model. For the minority classes, we can see that our current model has a tradeoff between precision and recall, and has a lower recall and a higher precision than the previous model. However, the loss in recall tends to be much smaller than the increase in precision, indicating that our current model overall outperforms the previous model for predicting minority classes as well.

5.3. Inference on time-series data

We performed a time-series analysis on the GSV Boston time-series dataset, which consisted of 128,656 images

taken between 2007 and 2017. Since there were far fewer images for even-numbered years than for odd-numbered years, we grouped consecutive years together (e.g. 2007–08, 2009–10). We took the average of predicted TrueSkill scores for each block group during each two-year period.

As shown in Figure 6, most of the block groups shown in the maps get brighter between 2008 and 2011, and then darker between 2012 and 2017, indicating that the average level of blight in Boston has increased and then decreased.

What could explain this trend? We believe that the increase in blight in Boston between 2008 and 2011, and subsequent decrease in blight after 2012, may be a result of the Great Recession, a global economic downturn that occurred between 2007 and 2009. To explore this possible link, we compared the average predicted TrueSkill scores per two-year period to data for real median household income in Boston for odd-numbered years from 2007 [1]. As shown in Figure 7, the changes in average blight in Boston are significant and are associated with opposite changes in median household income.

6. Conclusion and future work

We trained a deep neural network to measure the level of upkeep in a neighborhood using street view images of buildings. Our neural network can identify changes in average building upkeep at the block group level in different neighborhoods and cities. Our time-series analysis shows how changes in building upkeep in Boston track overall economic trends, specifically the Great Recession and the subsequent recovery.

In the future, we hope that our project will lead to a better understanding of how neighborhood upkeep in cities changes over time. Specifically, we hope to extend our time-series analysis to other cities like Austin and Detroit. We also hope to better understand the relationship between neighborhood upkeep and well-being using metrics such as crime rates, physical and mental health, income, and subjective well-being. Finally, we hope this research will lead to recommendations for policies that can reduce inequities in well-being.

Acknowledgments

The authors would like to thank Nima Dahir, Tanish Jain, Armin Thomas, and Shubhang Desai.

References

[1] B. Engebreth. Boston-Cambridge-Quincy Massachusetts household income. https://www.deptofnumbers.com/income/massachusetts/boston/. 6

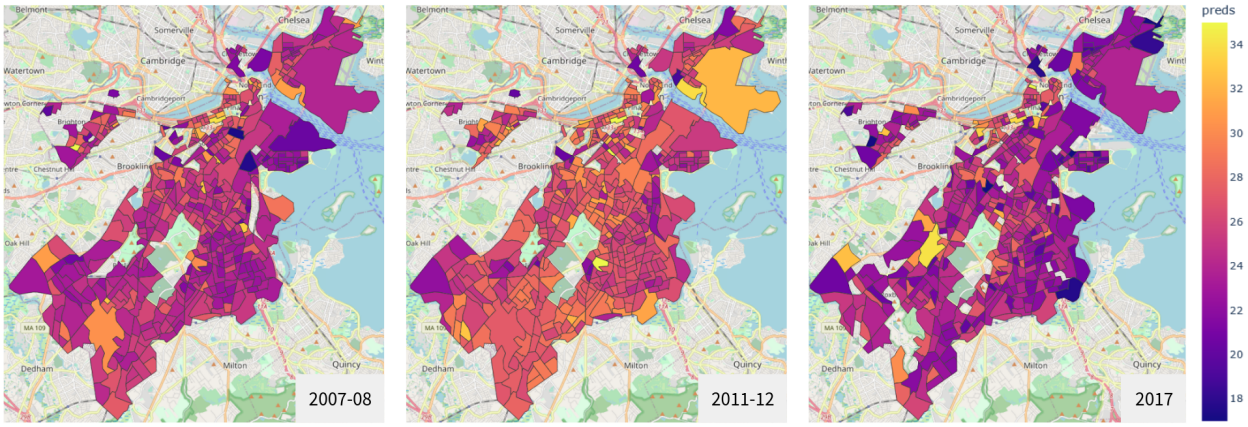

Figure 6. The average predicted TrueSkill scores by block group in Boston for the time periods 2007–08 (left), 2011–12 (center), and 2017 (right).

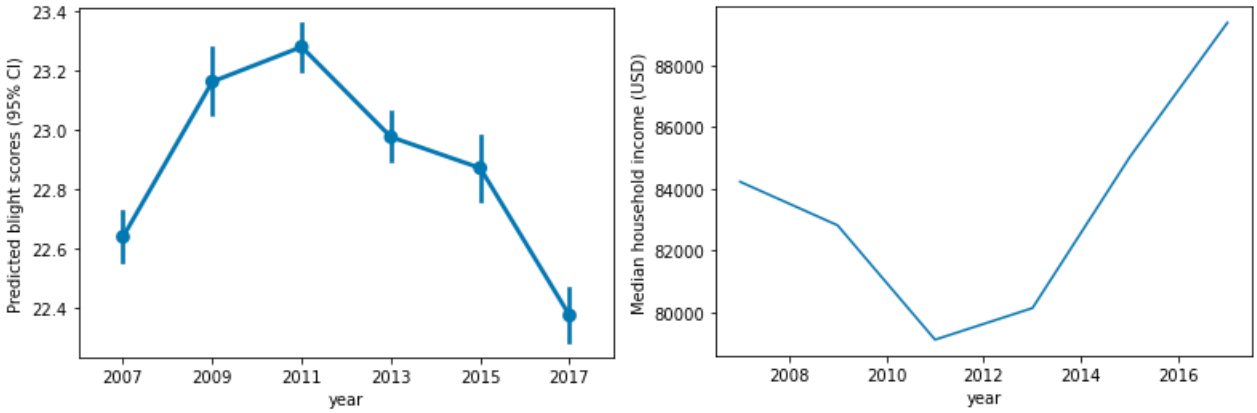

Figure 7. Average predicted TrueSkill scores for Boston (left) and real median household income (right).