**Data Dictionary: Google Street View Images**

**Background:**

Investigations into neighborhood conditions are typically conducted on small scales for only certain cities or neighborhoods [1, 2]. When conducted, neighborhood data collection is expensive and time consuming, and then only available for certain time periods. Currently, detailed neighborhood data come from neighborhood surveys, administrative data such as census data, and systematic inventories of neighborhood features. Subjective assessments of neighborhoods from community residents can help identify factors that residents believe are most important to their health and increase understanding on how individuals differentially use and interact with their environment. However, self-reported neighborhood data can be influenced by participants’ health status and cognitive function, resulting in “single source bias.”[3] The other neighborhood data we do have is mainly data on demographics (e.g., percent black). To our knowledge, our study is the largest to date using zip code level cases from 20 states to investigate associations between built environments and COVID-19 cases. Previous studies examining the distribution of COVID-19 cases are only focused on one or two states [4-6] or larger geographies like counties [7].

Google Street View (GSV) images represent a massive, publicly available data resource that has high potential but is very underutilized for health research. It can be used to extract information on physical features of the environment at point locations all over the country. Consistently constructed neighborhood quality indicators across large areas are severely lacking. While some studies have used human coders to classify environmental features seen in Google Street View images [8] this approach is not feasible on the massive scale necessary to compare thousands of U.S. neighborhoods. The development of data algorithms that can automatically analyze big data sources such as street view images will create a new national data resource for timely decision-making to mitigate the impact of COVID-19 and future outbreaks on health and health disparities. The purpose of characterizing built environments that have higher COVID-19 risk is to identify places where additional safeguards and resources are needed.

We utilize 170 million GSV images sampled at 50 meters apart and computer vision models to comprehensively characterize neighborhood conditions across the United States. From GSV images, we create indicators of urban development (non-single family home, single lane roads), walkability (crosswalks, sidewalks), and physical disorder (dilapidated building, visible utility wires). We hypothesize the built environments characterized by greater urban development and walkability will be associated with better health outcomes and physical disorder will be associated with worse outcomes.

Materials and Methods

**Street View image data collection.**We utilized Google Street View’s Application Programming Interface (API) to capture street view images of our search set. Image resolution was 640x640 pixels. We surveyed all U.S. roads and obtained 4 images from each sample location with angle views at 0, 90, 180, and 270 degrees, thus permitting fuller capture of the surrounding area of a point location. In total, 164 million images were obtained in November 2019.

**Image data processing.** Convolutional Neural Networks (ConvNets) [9-11] achieve state-of-the-art accuracy for several computer vision tasks including but not limited to object recognition, object detection, and scene labeling. For example, the state-of-the-art accuracy of ImageNet [12] with 1000 categories and over one million image samples is improved every year using ConvNet-based methods. The ImageNet dataset contains images from various categories (e.g., “moped”, “Granny Smith apple”) and corresponding category labels. Models trained on this dataset use trial and error to learn combinations of colors, shapes, and textures that are relevant to a wide variety of image interpretation tasks, and therefore can be used as a starting point for creating computer vision models for tasks where labeled training data is scarce. A ConvNet model “pre-trained” on ImageNet can be “fine-tuned” using a smaller amount of training data from the desired task, which delivers strong classification performance without requiring the vast training data and computational resources necessary to train the original ConvNet.

**Neighborhood definitions.** Zip codes and census tracts were utilized as neighborhood definitions. To arrive at the neighborhood indicators, we processed street imagery and then combined information on all street imagery within a zip code to arrive at zip code-level summaries (e.g., % of images in a zip code that contain a sidewalk).

**Built environment indicators.** To create a training dataset for our computer vision models, from December 2016-February 2017 we manually annotated 18,700 images (from Chicago, Illinois; Salt Lake City, UT; Charleston, West Virginia; and a national sample). These locations were chosen to capture heterogeneity in neighborhood environments across geographically and visually distinct places with varying population densities, urban development, and demographics. Labelers included the principal investigator and three graduate research assistants. Inter-rater agreement was above 85% for all neighborhood indicators. Each image received labels for these binary neighborhood characteristics: 1) street greenness (trees and landscaping comprised at least 30% of the image - yes/no), 2) presence of a crosswalk, 3) single lane road, 4) building type (single-family detached house vs. other), and 5) visible utility wires. Green streets were utilized to indicate lower urban development. Single lane/residential roads limit the number of cars and hence flow of people. Non-single family home was utilized as an indicator of residential and commercial mixture. Crosswalks were utilized as an indicator of walkability. Visible utility wires were utilized as indicators of physical disorder.

We randomly divided the dataset into a training set, a validation set, and a test set. The training and validation set contained 80% of total labeled images and the remaining 20% was used as a test set to evaluate the model’s performance. Once the hyper-parameters were chosen, each model architecture was trained multiple times. Note that neural network training is stochastic even when starting from the same initialization and using the same training set, therefore, multiple training runs are used to assess the mean and standard deviation of the error. The testing set remained unobserved until the best models have been selected using the training set. We assessed the final quality of the model using the test set. We first resized all the images to the size 224x224 for processing. We then trained a standard deep convolutional neural network architecture- Visual Geometry Group VGG-19 [9] in Tensorflow [13] with sigmoid cross entropy with logits as the loss function. The weights of the network are initialized from ImageNet weights. Adam optimizer was used with batch size 20. Training took 20 epochs and started with learning rate 1e-4. We considered the model saved in the last epoch as our final model. Accuracy of the recognition tasks (agreement between manually labeled images and computer vision predictions) were the following: street greenness (88.70%), presence of crosswalks (97.20%), non-single family home (82.35%), single lane roads (88.41%), and visible utility wires (83.00%). These figures were consistent with a separate, semi-supervised learning approach. Below, we describe the model building process for two additional neighborhood indicators that utilized different training datasets.

Dilapidated building indicator.Our training dataset consists of approximately 29,400 Google Street View images captured from Baltimore and Detroit based upon administrative lists from city governments on vacant buildings and buildings marked for demolition from 2014-2018. We randomly split this dataset in the ratio 80:20 for validation to obtain about 23,500 images for training and 5,900 for validation. The dataset has an equal number of normal and dilapidated buildings. We then trained a standard deep convolutional neural network architecture- ResNet-18 [11] in Pytorch [14] with NLL loss as the loss function. For the dilapidated building indicator, the ResNet-18 model produced an accuracy of 89.1% and a F1 score of 89.1.

Improved detection of dilapidated buildings. Since it is expensive and time-consuming to collect labels for GSV images, we use a multi-task neural network to improve the performance of the classification model on GSV images. Multi-task learning is a training paradigm in which a shared model is simultaneously trained on different tasks. In this process, the model learns general representations of data that are useful across different tasks. This process increases data efficiency and reduces overfitting. To create a dataset for our training paradigm, we have downloaded 59027 images from Flicker, where 37559 images are collected with the keywords "old/dilapidated building" and 21468 images are collected with the keywords "nice building." The images corresponding to "nice building" are labeled as class 0, and "old/dilapidated building" are labeled as class 1. We use a pre-trained ResNet18 network on the ImageNet dataset. We replace the last fully connected layer with two fully connected layers that separately predict labels for GSV images and Flickr images. We have randomly divided both Flickr and GSV datasets into training and validation sets with a ratio of 80:20. In our experiments, all images are resized to 512x512 and augmented using horizontal flipping and color jittering. The model is trained using both Flicker and GSV images simultaneously. We train the model end-to-end using binary cross-entropy loss function for 100 epochs with a learning rate of 1e-4 and batch size of 10 for Flicker and GSV datasets. Supplementing with Flickr data and using multitask learning led to an accuracy of 93.4% and F1 score of 94.9%.

Sidewalk indicator.Our training dataset consists of about 24,316 images captured from Google Street View from New Jersey that had been manually labeled. We randomly split this dataset in the ratio 80:20 for validation to obtain 19,452 images for training and 4,864 for validation. The minority label images were oversampled so that the dataset has an equal number of sidewalk present and absent cases. We then trained a standard deep convolutional neural network architecture- ResNet-18 [11] in Pytorch [14] with NLL loss as the loss function. For the sidewalk indicator, the ResNet-18 model produced an accuracy of 84.5% and a F1 score of 81.0.

Urban landscape indicators. We randomly sampled 18,000 images from our national collection of Google Street View images and labeled these images to provide a training and test set for new built environment indicators. We divided this data into a training (80%) and test set (20%). Quality control statistics are as follows for presence of street light (accuracy was 88% and F1-score was 60%); two or more cars (accuracy was 88% and F1-score was 79%); street signs (accuracy was 82% and F1-score was 68%); chain linked fence (accuracy was 95% and F1-score was 45%).

|  |  |
| --- | --- |
| **Data** | Google Street View derived variables\* |
| **Variable** | **Label** |
| Zip\_code | Zip code |
| census\_tract | 2010 Census tract |
| county id | county |
| state id | state |
| prop\_crosswalk | % of images for a given zip code (or census tract) with a crosswalk |
| prop\_sidewalk | % of images for a given zip code (or census tract) with a sidewalk |
| prop\_not\_single\_family\_home | % of images for a given zip code (or census tract) with any building that was not a single family home |
| prop\_green | % of images for a given zip code (or census tract) where street trees and street landscaping comprised at least 30% of the image |
| prop\_single\_lane | % of images for a given zip code (or census tract) with a single lane road |
| prop\_visible\_wire | % of images for a given zip code (or census tract) with visible utility wires overhead |
| prop\_dilapbldg | % of images for a given zip code (or census tract) with any dilapidated building |
| prop\_chainlinked\_fence | % of images for a given zip code (or census tract) with any chain linked fence |
| prop\_letters | % of images for a given zip code (or census tract) with any street signs |
| prop\_2ormorecars | % of images for a given zip code (or census tract) with 2 or more cars |
| prop\_streetlights | % of images for a given zip code (or census tract) with any streetlights |
| Tot\_num | Total number of Google Street View images |
| total\_crosswalk | Total number of images for a given zip code (or census tract) with a crosswalk |
| total\_sidewalk | Total number of images for a given zip code (or census tract) with a sidewalk |
| total\_not\_single\_family\_home | Total number of images for a given zip code (or census tract) with any building that was not a single family home |
| total\_green | Total number of images for a given zip code (or census tract) where street trees and street landscaping comprised at least 30% of the image |
| total\_single\_lane | Total number of images for a given zip code (or census tract) with a single lane road |
| total\_visible\_wire | Total number of images for a given zip code (or census tract) with visible utility wires overhead |
| total\_dilapidated\_building | Total number of images for a given zip code (or census tract) with any dilapidated building |

\* Previous publications used the Google Street View percentage variables, which were then categorized   
into tertiles.

References

1. National Archive of Criminal Justice Project on Human Development in Chicago Neighborhoods <http://www.icpsr.umich.edu/icpsrweb/PHDCN/> (September 24, 2012),

2. Baltimore Neighborhood Indicators Alliance - The Jacob France Institute Vital Signs 11 Reports. <http://bniajfi.org/vs11_report> (September 24, 2013),

3. de Jong, K.; Albin, M.; Skärbäck, E.; Grahn, P.; Wadbro, J.; Merlo, J.; Björk, J., Area-aggregated assessments of perceived environmental attributes may overcome single-source bias in studies of green environments and health: results from a cross-sectional survey in southern Sweden. *Environmental Health* **2011,** 10, (1), 4.

4. Chen, J. T.; Waterman, P. D.; Krieger, N.; Krieger, N., COVID-19 and the unequal surge in mortality rates in Massachusetts, by. *population* **2020,** 25014, (B25014\_013E), B25014\_001E.

5. Almagro, M.; Orane-Hutchinson, A., The differential impact of COVID-19 across demographic groups: Evidence from NYC. *The differential impact of COVID-19 across demographic groups: Evidence from NYC (April 10, 2020)* **2020**.

6. Almagro, M.; Orane-Hutchinson, A., The Determinants of the Differential Exposure to COVID-19 in New York City and Their Evolution Over Time. *Covid Economics: Vetted and Real-Time Papers* **2020**, (13).

7. Chin, T.; Kahn, R.; Li, R.; Chen, J. T.; Krieger, N.; Buckee, C. O.; Balsari, S.; Kiang, M. V., US county-level characteristics to inform equitable COVID-19 response. *medRxiv* **2020**.

8. Nesoff, E. D.; Milam, A. J.; Barajas, C. B.; Furr-Holden, C. D. M., Expanding tools for investigating neighborhood indicators of drug use and violence: validation of the NIfETy for virtual street observation. *Prevention science* **2020,** 21, (2), 203-210.

9. Simonyan, K.; Zisserman, A., Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.

10. Krizhevsky, A.; Sutskever, I.; Hinton, G. E. In *Imagenet classification with deep convolutional neural networks*, Advances in neural information processing systems, 2012; 2012; pp 1097-1105.

11. He, K.; Zhang, X.; Ren, S.; Sun, J. In *Deep residual learning for image recognition*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016; 2016; pp 770-778.

12. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M., Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* **2015,** 115, (3), 211-252.

13. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M. In *Tensorflow: A system for large-scale machine learning*, 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016; 2016; pp 265-283.

14. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L. In *PyTorch: An imperative style, high-performance deep learning library*, Advances in Neural Information Processing Systems, 2019; 2019; pp 8024-8035.