Improving Harris County Census Data with Density Mapping and Machine Learning

Tessa Cannon July 30th, 2021 Rice University, Google Data Science REU

Abstract

Data collection for the United States census is expensive, challenging, and prone to errors. The distinction between single family and multifamily households is important for mailing the correct number of census forms to each person, but multifamily houses often go undetected. Machine learning can be used on aerial imagery to detect precise housing information and improve census data. Automating housing verification could have significant impacts on the way that the United States census is conducted, guiding canvassers to the highest priority locations and making the overall counting process more efficient.

Background

Density mapping has typically only been used on counting crowds of people, animal populations, or cells and microorganisms. To the best of my knowledge, density mapping has not been attempted on counting housing structures. This research will therefore test the limitations of density mapping to see if it is capable of distinguishing between very similar housing structures, and if it can generate an accurate count of such houses and pinpoint their locations.

Previous Work

Computer science students at Rice University began using aerial imagery to detect multifamily houses in 2019. By testing several convolutional networks pretrained on ImageNet, they discovered over 1800 undetected multifamily houses. However, their data structure, consisting of cropped single house image tiles, limited the scope of their research to only a single zip code in Harris County (77004).

Approach

I will improve previous multi-family detection models by introducing density mapping. Through the generation of density maps that indicate the presence of different housing structures, I will be able to estimate the number and location of multi-family houses provided the aerial image of any given area. This will extend the research scope from one zip code to all of Harris County.

Methods

Procedure

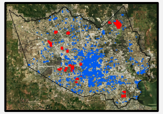
1. Find aerial images of all of Harris County
2. Crop aerial images into smaller image tiles
3. Rank areas based on data accuracy and use areas with higher label accuracy as training data
4. Label buildings in tiles based on housing type
5. Generate density map of each image tile
6. Feed image tiles and corresponding density maps as labels into convolutional neural network
7. Generate maps and predict multifamily houses

Data

- Aerial Imagery: 50x50m image tiles, 6 inch res.
- Building Labels: Multifamily vs. Non-multifamily
- Building Footprints: Vertices of households
- Bad MAF score: Block group census accuracy

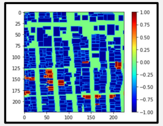
Processing

Data processed using ArcGIS, SQL, Python



- Split aerial data into 12800 (2632 x 1683 foot) image tiles
- Training Data: Tiles with >1 multifamily house and low MAF score
- Testing Data: Remaining tiles in Harris County
- Resize images into 224*224 pixels

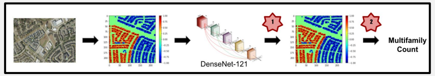
Density Mapping



Density mapping outperforms simple classification models on large area images by pinpointing the location and size of each building. Pixels corresponding to multifamily houses are assigned a value of 1 while non-multifamily houses are assigned a value of -1. Building edges are then softened while remaining pixels are set to zero.

Experimental Setup

Once the ground truth density maps are generated, they are fed into a convolutional neural network, which outputs a corresponding prediction density map. The values in the output density map are tallied to get the final multifamily count. The convolutional neural network used was Densenet-121, pretrained with ImageNet.



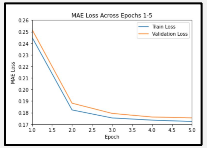
Evaluation metrics:

- Cross validation
- Loss: MAE
- Accuracy: R^2

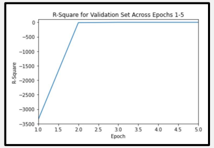
Parameters:

- Optimizer: Adam
- Learning Rate: 0.00001
- Epochs: 5
- Batch size: 7

Results

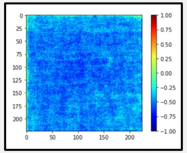


Loss decreases over time across 5 epochs for both training and validation and plateaus at 0.17.



R^2 validation test reveals that predictions are not fitting data accurately. Across epochs, the R^2 value is extremely negative, converges towards zero, and then flattens around -0.18.

Density Map Generation



Model is not generating accurate density map predictions.

- Model is not able to localize objects
- No predicted multifamily houses (red pixels)
- Random noise

Conclusion

This model is a work in progress.The neural network is currently inaccurately predicting the number of pixels that correspond to multifamily houses and unable to localize objects. This leaves us with inconclusive results about the capability of density mapping when it comes to distinguishing between housing structures. Since the model is not functioning properly, the test dataset has not been analyzed yet for potential undetected multifamily houses. Model and density methods need to be altered further before exposing test data

Future Work

- Continue to tune parameters
  - Change loss metrics
  - Explore different neural network architectures
  - Discover new methods to localize objects
- Add a feature that outputs exact address of newly identified multifamily houses.
- Once model is working properly:
  - Communicate results to census workers
  - Canvassers verify newly identified multifamily houses
- Combine improved housing data with demographic information to form a deeper understanding of what types of people live in multifamily houses.

Resources

1. Harris County Appraisal District. Retrieved from https://hcad.org/hcad-online-services/pdata/
2. Houston-Galveston Area Council. (2018). Aerial Imagery 2018. [Data set]. Rice University-Kinder Institute: UDP.
3. Kissam, E., Quezada, C., & Intili, J. A. (2018). Community-based canvassing to improve the U.S. Census Bureau's Master Address File: California's experience in LUCA 2018. *Statistical Journal of the IAOS, 34*(4), 605-619. doi:10.3233/sji-180480
4. LiDAR Imagery. Retrieved from https://www.h-gac.com/imagery/lidar

Acknowledgements

Thank you to my graduate student mentor Zhaozhuo Xu and faculty mentor Anshumali Shrivastava.
Thank you to Jean Aroom at the Fondren GIS/Data Center for help with the dataset.
Thank you to Rice University and Google for sponsoring and organizing this program.