Environmental Effects on Housing Prices

A study comparing housing prices with satellite imageries and air quality data

Patrick Coady, Thomas James, Jiayu Li

Introduction

We are looking into the potential effect of environmental factors on housing prices. We want to see if there is any correlation between housing prices for different city landscapes (vegetation, water, etc.) and pollution levels. This was inspired by the recent urban planning interests and EPA connecting lead pollution to economic harm in the 1980s.

Questions we want to answer:

How could we extract the landscapes / environmental features from satellite images?

Potential future usage of multispectral-imaging could utilize different wavelengths of light to allow observation of nutrient levels and plant health in the landscape's terrain. This could help officials determine the value of real-estate land before putting it on sale. Lower levels of nutrients could also be an indication that watershed or pollution is nearby and needs further investigation.

Does the landscape pattern of areas affect housing prices?

Landscape of the area will determine the amount of erosion that can possibly occur, the types of plants that can possibly grow there, and thereby will have an effect on the housing prices for that area. Water-drainage patterns and frequency of flooding can also have a potential impact on housing prices.

How do the landscapes / environmental features evolve over time?

Various environmental factors; such as volcanoes, mountains, floods, droughts, stormwater-runoff, soil-types, and erosion; can play a part in the evolution of landscape and environmental features over time. Man-made factors, such as building real-estate, building new bridges, and mining for valuable minerals also can cause environmental and landscape changes over time.

How does the change of landscapes / environmental features link to the house prices?

Increase in risk factors associated with landscape/environmental features increase the potential for damage to the real estate and thereby decrease the pricing value. Not all risk factors are readily visible when new families are searching for quality homes. More stable landscape and soil conditions can provide a longer lifespan for the real estate and increase the value of the houses. Real-estate located in areas close to convenient locations, such as populated-cities or town-centers, could also be expected to show an increase in housing prices. A correlation between housing-price levels and distance away from populated cities is expected. High levels of air pollution may also encourage people to seek housing elsewhere, depreciating the housing prices.

Goal

To better understand what makes a city popular and livable, and inform policy decisions and real estate developers to meet those ends.

Data Processing

Data sources:

Zillow House Price [1]: Monthly average house prices by zip code across most of the United States and over a large timespan.

- Format: CSV file available to pulled from Zillow research API
- Size: ~ 31MB, with over 6,882 zip codes areas each with 13 features

LANDSAT 8 [2]: Processed satellite images (collection 2 level 2). The data is available in the AWS bucket of USGS data.

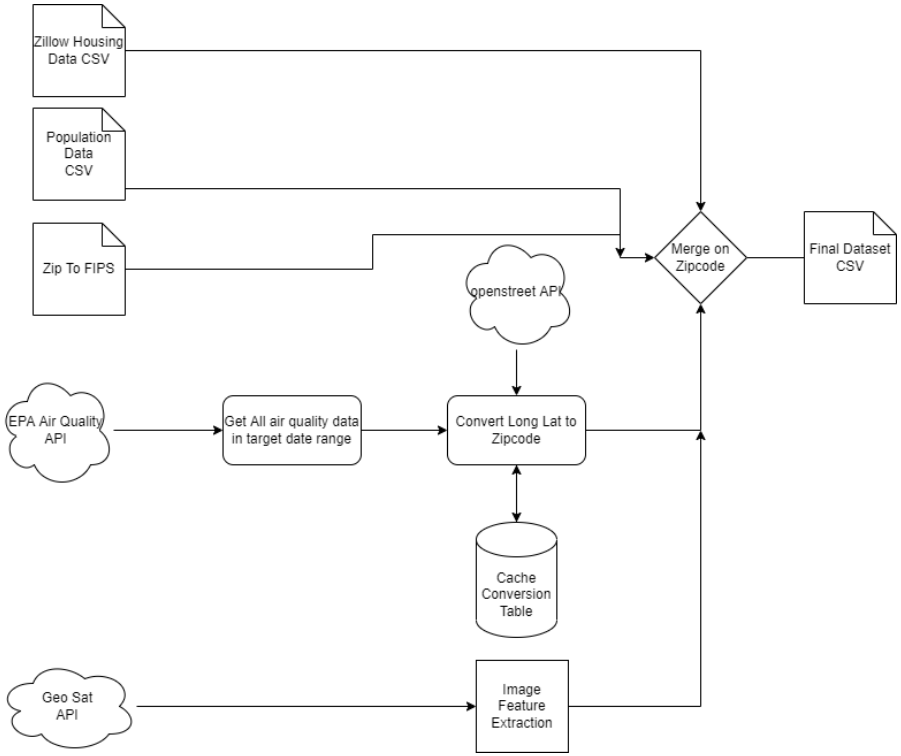
- Format: TIF image file with meta information.
- Size: ~2.7 G (30K images) for the area and time of interest. Each area is scanned every 16 days.

EPA [3]: The EPA Air Quality System (AQS) provides air quality information on layers over pollution type, date range, and geographic zone.

- Format JSON export from EPS API
- After all pollutant types are combined, we have a total of 8 features with about 5000 rows.

US zip code level data [4]: zip-code level area geographic and population data are available in Kaggle.com

- Format: CSV data available to download in the link above.
- Two key columns: the center coordinate of zip code areas and the population according to 2017 census data.



The investigation time frame is during the years 2015 to 2018. This time-frame occurs before the pandemic when house prices are less volatile, which allows better resolution. Joining all the three datasets provides the following data-of-interest:

- 1,425 zip-code areas across the US with available house prices, satellite images, and at least 1 piece of air-quality information.
- Each zip code area has an average listing house price for each month during 2015-2018.
- Each zip code area has multiple satellite images covering the 3x3 mile regions which were taken within the period.
- Not all zip codes have all kinds of air quality data available.

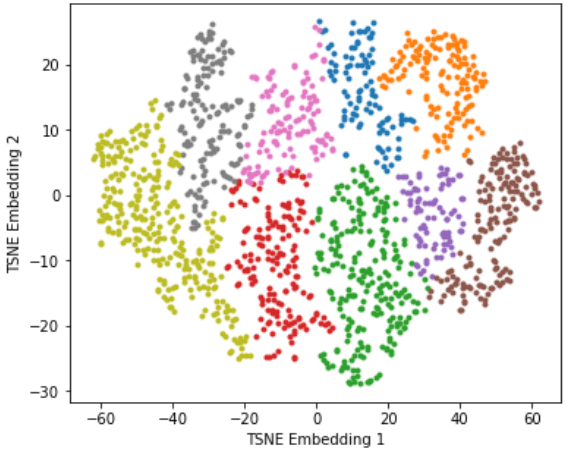
LANDSAT 8 Data Processing

- Input: Image size 128x128x6 (2.5 miles on geographical size).

- Data Cleaning
  - Images are already processed by USGS to remove most atmospheric effects.
  - **Further cleaning:** Take the median pixel value of images taken at different times to average the effect of occasional cloud cover and sunlight conditions.

- **Branch 1:** Normalized Difference Vegetation Index (NDVI) and Water Index (NDWI)
  - Recall Band: B5: Near Infrared Light, B4: Red, B3: Green
  - NDVI = (B5 - B4) / (B5 + B4)
  - NDWI = (B3 - B5) / (B3 + B5)

- **Branch 2:** Image Embedding & Clustering
  - Modified U-Net for embedding [5]
  - Dimensional reduction:
    (t-SNE [6]) +
    Hierarchical clustering
    (H-DBSCAN [7]) +
    Impute
    (KNN with k=3 [8])
  - **Resulting Cluster**: 9 different landscapes (next page)



- **Final Output Features (5 features):**

  Cluster, Mean & Std of NDVI, NDWI

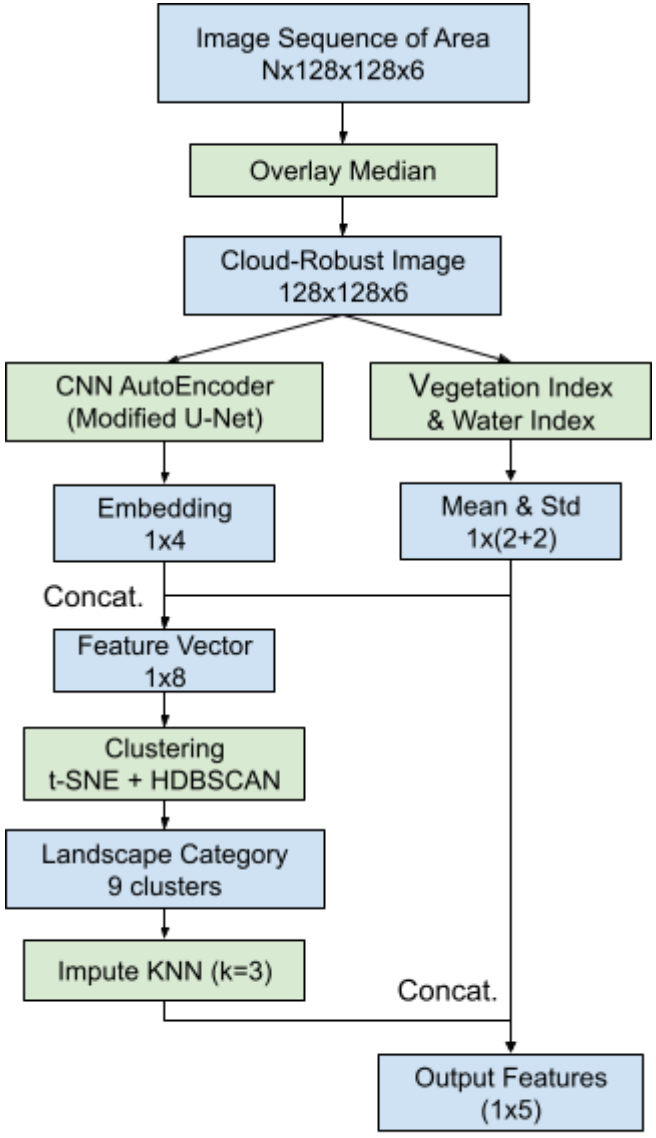


Fig. L-1 data pipeline

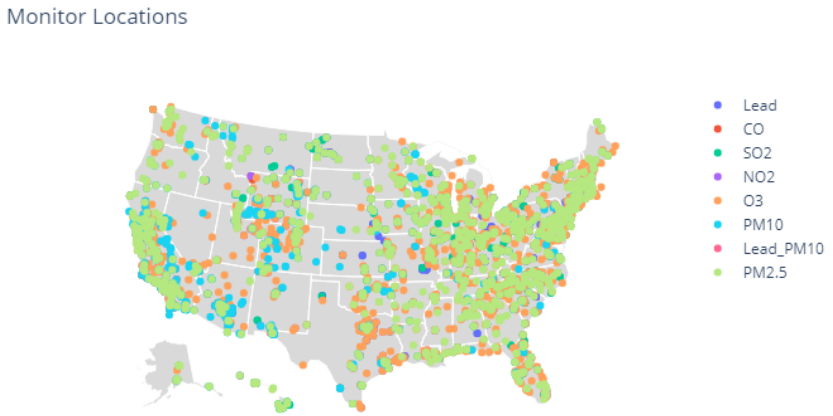
- 9 Landscape categories (cluster number) according to the clustered texture and landscape properties.

(#5) Urban: High-Density (11.2%)	(#4) Urban: Industrial (5.9%)	(#2) Urban Mid-Density (16.4%)	
(#3) Suburban: Flat (11.6%)	(#6) Suburban: Hilly (8.4%)		
(#7) Rural: Flat, Wet (10.2%)	(#8) Rural: Hilly, Wet (19.9%)	(#0) Rural: Dry (6.9%)	(#1) Desert Area (9.7%)

Visualization: (column wise with 5 samples each).

Zillow Housing Data and EPA Air Quality

EPA Air Quality:



Input:
AQS API that returns a JSON response for a single pollutant in a specified geological zone and time span.

Data Cleaning:

- Take the mean data for each pollutant over our desired time range of 2015->2018.
- Retrieve the zipcode for each monitor station using its latitude and longitude and add the feature to each pollutant dataset.
- Pivot and merge all the pollutant information together into a single dataset
- Normalize the data using the StandardScaler normalizer in sklearn
- Small amounts of missing data were dropped during analyses

Zillow Housing Data:

Input:
CSV containing monthly data from 2000->2022

Data Cleaning:

- Selected data from 2015->2018 to match the data we could retrieve in the LandSat, and pollution datasets.
- Averaged the housing prices over that time frame to remove the noise of housing price fluctuations in our correlation analysis.
- Normalized the data by taking the log of the housing prices, and then passing it through the StandardScaler.
- Small amounts of missing data were dropped during analyses

Combine:

Merged each dataset; population, pollution, landsat features, and housing prices; by zipcode to create a final dataset for analysis. This results in a rather sparse dataset for the pollution features since not every zipcode captures every pollution feature.

Analysis and Discussions

Methodology

The house prices are determined by numerous factors including the property types, amenities, school district, accessibility, and even the financial market. Environmental factors are just one of the factors. Therefore, to better address the environmental effect on the house-pricing market. We choose "Year-over-Year (YoY)" growth of housing prices as a prediction metric.

The house prices are also heavily influenced by seasons (Fig. A-1a). Therefore, we use the slope of the logarithmic trend line as our YoY growth rate. More specifically, we run a regression:

$$Price \sim \beta\, t(year) \; + \; \alpha$$

And define the YoY growth as $YoY \; = \exp(\beta) \; - \; 1$. The distribution of the target variable is shown in Fig. A-1b. The target variable is distributed in a normal distribution which means a common inference method, such as linear regression, is appropriate.

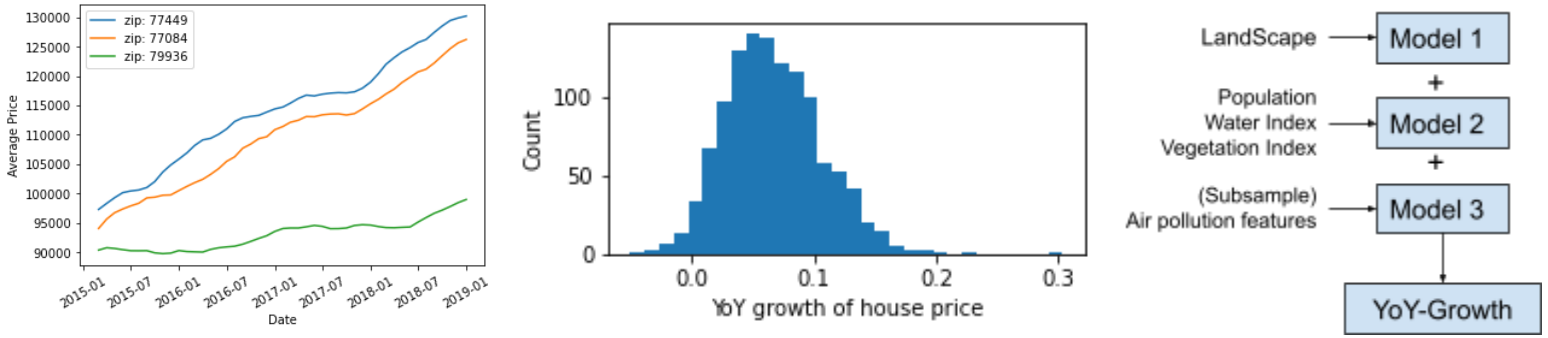


Fig. A-1. (a) Example house price trend of 3 zip codes (b) The distribution of YoY growth with zip-level average housing-prices (c) The illustration of the additive model

To achieve better inference, we employ an additive model (boosting) to link the features to YoY growth. All 3 models utilize ordinary linear regression. The first model only involves the landscape types. This model shows the difference between the mean growth rate over landscapes. The second model investigates the residual from the first model. A model was trained over population, water, and vegetation index for each landscape group. The air quality factors are also included. The residuals of the second model are fitted on the subset where the particular air quality metrics are available to address the effect of these factors.

Since confounders are not fully investigated, the above analysis only provides correlation. Possible causality will be discussed with the result to address potential future research directions.

Correlation between features

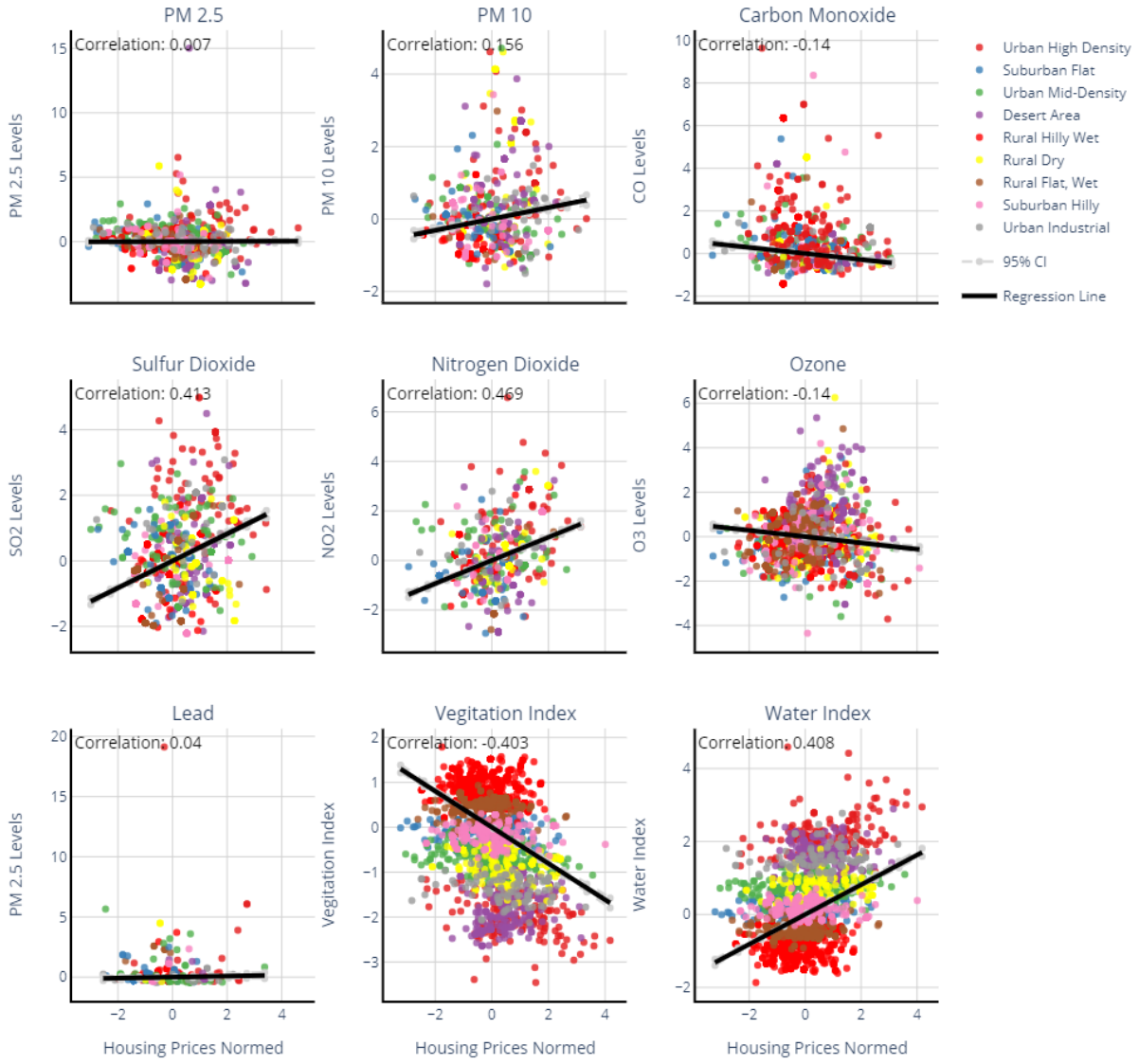
To determine if any relationship coexists between the features being analyzed and housing prices, an ordinary least squares regression was conducted between each of them. The data in graph (Fig. A-2) represents the average values over a three year period and has been further normalized to keep everything on the same scale. We found very weak to no correlation between PM (Particulate Matter) 2.5, PM10, Carbon Monoxide, Ozone, and Lead.

However, we discovered much stronger correlations with Nitrogen Dioxide(NO2), Sulfur Dioxide(SO2) and our Vegetation and Water indexes. The main sources of SO2 are power-plants, industrial facilities, and volcanoes. NO2 is released into the air from vehicle-emissions and power-plants. Both are indicators of industrialized locations and could be potential causes for the higher housing-prices. More investigation would be needed before any strong conclusions could be created.

The vegetation index tells a similar story, with more urban landscapes having less greenery and being more expensive to live in.

The water index shows that wetter locations tend to be more expensive. This may be because of the effect of waterfront property being much more expensive.

(Right) Fig. A-2 Scatter plot between features.



Model 1: Growth Rate over Landscapes

Fig. A-3 Shows the estimated mean (with 95% CI) of YoY growth rate of each landscape group. The mean YOY growth rate (6.79%) is shown in a gray dashed line.

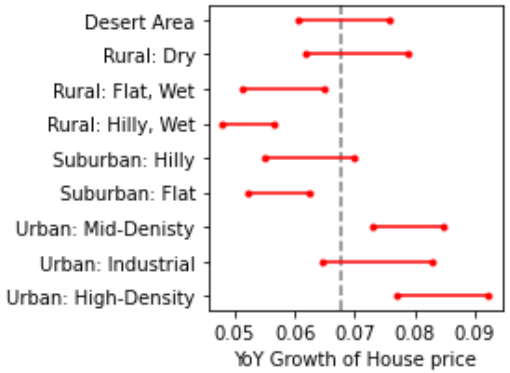


Fig. A-3. Estimated group YoY Growth (95% CI)

4 landscape groups have significant differences in growth rates: Rural areas with wet climate, suburban areas with flat terrain and two urban areas. The results suggest that during the year of 2015-2018, the growth of house prices are driven mainly by the urban areas; especially the metropolitan areas.

Model 2: Effect of Vegetation & Water Index

Table below shows only the significant coefficients in model 2 over each combination of landscape groups and features. Population coefficients (increase of growth rate per 1M people) have significant impact in both industrial urban and flat suburban areas. This reflects the pattern of urbanization: the place with higher potential for residential development (with higher population) has faster growth in house-prices.

The mean of vegetation and water index is positively correlated with house-prices in hilly, wet rural areas. This is also natural because such areas are commonly developed into high-end neighborhoods.

Interestingly, the standard deviation of vegetation and water index has a divided effect on house-prices in desert areas. The standard deviation usually represents richness and diversity of landscapes which could be connected to desert areas. The negative correlation of vegetation index variance could be due to lack of neighborhood development in areas with negative vegetation and water indexes.

Landscape	**Variable**	coef	P>|t|	[0.025	0.975]
Urban: Industrial	population	0.2717	0.005	0.085	0.458
Suburban: Flat	population	0.1384	0.022	0.021	0.256
Rural: Hilly, Wet	NDVI_mean	1.3647	0.009	0.344	2.385
Rural: Hilly, Wet	NDWI_mean	1.845	0.002	0.696	2.994
Desert Area	NDVI_std	-2.2408	0.025	-4.19	-0.292
Desert Area	NDWI_std	2.9	0.013	0.626	5.174

Model 3: Effect of Vegetation & Water Index

Similar to the analysis performed in model 2, model 3 shows the coefficient of air qualities and index values over the residuals of model 2. It is surprising that the average house price is positively correlated with pollution. This causality may be related to the density of the population area which coincides with landscape. But one thing is certain: in most cases, there is no significant correlation (and causality) between air pollution level and house prices from the analysis.

pollution	metric	coef	P>|t|	[0.025	0.975]
NO2	mean	0.0012	0.008	0.000	0.002
O3	max-value	0.4907	0	0.257	0.725

Discussion

Combining all the models, the prediction R2 score for YoY growth rate is 0.175. This helps explain the environmental features that we developed from satellite images and the 17.5% variation of the house prices. This is a remarkable result as only a few features were utilized during the analysis. The most significant environmental factor originates from the landscape features extracted from the satellite images.

However, the method has its limitations. The first limitation comes from the misclassification of landscapes. The example images observed displayed several errors in the clustering using embedded information. Such errors are due to the hierarchical clustering methods which require back and forth manual check of parameters. Switching to a supervised learning algorithm with a labeled dataset is a good solution.

Another limitation is the presence of a sampling-bias from where the pollution monitors were located. California has the most pollution monitors, and some states like Montana have hardly any. This bias could be exacerbating any correlative observations that we discovered.

A further sampling bias is present from our choice to only sample data from the United States because of the abundance and richness of the data available. The United States is a particular country with its own unique culture and climate, so what is true here may not hold true for other countries. Furthermore, other countries have far more severe air pollution problems and create larger concerns for the people living there when they purchase a house. A natural extension of our work would be to expand our analysis to other countries across the globe.

Conclusion

In this report, the average house price year-over-year by zipcode was successfully linked with the growth of two environmental datasets: satellite images and air quality data. The additive model we developed could explain 17.5% variance in the house prices' growth-rate during 2015-2018, which suggests that the environment is one of the key factors in determining an area's average house-price.

The CNN AutoEncoder unlocked the abilities to compress and embed the information in satellite images. With the extracted features, landscape types could be automatically detected by clustering the embedded features. These landscape types, interacting with the traditional features such as vegetation and water index, have discovered that the urban area has experienced a significantly higher growth-rate than suburban and rural areas. Both vegetation and water index have complicated impact conditions with the landscape. However, the causality behind the correlation is yet to be determined.

Last but not least, the air quality index did not show a significant relationship to the area housing prices due to limitations from biased samples. However, as the awareness of environment protection increases, more data will be collected and can benefit the industry.

Reference

[1] Zillow Home value. viewed 2022-05-24. URL https://www.zillow.com/research/data/,

[2] LANDSAT Mission Product: Collection 2, Level 2. viewed 2022-05-24. URL: https://www.usgs.gov/landsat-missions/landsat-collection-2-level-2-science-products.

[3] United State Environment Protection Agency. Air Quality System. viewed 2022-05-24. URL: https://aqs.epa.gov/data/api/

[4] Kaggle.com. US zip code data collection. viewed 2022-05-24. URL https://www.kaggle.com/datasets?search=Zip+code

[5] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

[6] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11).

[7] McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. J. Open Source Softw., 2(11), 205.

[8] Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003, November). KNN model-based approach in classification. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 986-996). Springer, Berlin, Heidelberg.

Statement of Work

Patrick Coady

- Consuming and cleaning the pollution API, and well as any graphs specific to that dataset.
- Combined the various datasets together.
- Code files: **pollution.ipynb**

Thomas James

- Cleaning and initial analysis of the housing dataset
- Charting bar-charts, histograms, log-histograms, and chloropleth charts using the real-estate data and pollutant particle-matter data from each state.
- Proofreading/Revision of Report
- Code files: **Zillow_Data_Analysis.ipynb**

Jiayu Li

- Develop the download, process, and feature extraction pipeline for LANDSAT 8 data. Including the U-Net auto-encoder and clustering.
- Code files: **utils.py**, **download_images.py**, **process_images.ipynb.**

Together

- All analysis of the final merged dataset, and the drafting of discoveries was done together as a team.
- Code file: **Regression.ipynb**