Final Project Write-Up

Authors: Noah Reiner, Inthisar Kamal, Katie Patrick

Research Question

Our research question is: *Can we build a model using Cook's County Assessor's Office (CCAO) data to predict what loan grade (redling indicator) a house was given in 1930 based on the Cook's county data features today?* We also have a sub question: *Is house sale price today associated with historical redlining grade data from 1930?* These questions are crucial in determining whether or not, present-day housing data is still in any way, shape, or form, correlated to race and is based on historical data relating to surveyed housing letter grades: A indicating the "best" properties, B meaning "good" properties but not nearly as nice as A, C areas were "older, becoming obsolete" and D areas represented "poor housing conditions" or "undesirable populations." The grades were a way the federal home loan bank board segregated whites areas into homes with a better grade and blacks and immigrants to areas with lower grades. The effects of historical redlining can still be seen today as historically low graded homes are often in areas of high poverty today. We chose to look at Cook's County as they have specifically seen how racism consciously and unconsciously affected the housing appraisal process given their former assessment process systematically privileged white people. We are aiming to see if the newer Cook's county assessor machine learning algorithm is better off then the prior assessments, or if it still includes variables that trace back to historical issues of redlining, and thus, perpetuating the problems of systemic racism. Additional datasets we used for the research beyond Cook's county publicly available data are redlining maps and data about housing area grades (A,B,C, or D) which we merged into the Cook's county dataset on home longitude and latitude.

Hypothesis

We hypothesize that we can predict what grade a house was given in 1930 based on current Cook's County data. We further hypothesize the most expensive homes will be mapped over historically "green" or A areas on the map whereas the least expensive homes will be mapped over historically "red" or D areas on the map.

Related Work

Previous analysis in 2017 revealed that there was a substantial tax divide systematically disadvantaging black and latino communities, and house owners with lower and moderate priced homes. However, since new Cook County Assessor, Fritz Kaegi, took over in December of 2018, little has been done to assess how accurately his new method for calculating house prices is beyond within the assessor's office since it takes a deep understanding of algorithms and data science to investigate the new process. Although the data is publicly available, without substantial understanding of how these models are created and used, it is difficult to determine how fairly the algorithm works and how relevant all of the variables put into the algorithm are.

Specifically, no other research has looked at the association between the variables used in the house assessment model and historical redlining maps.

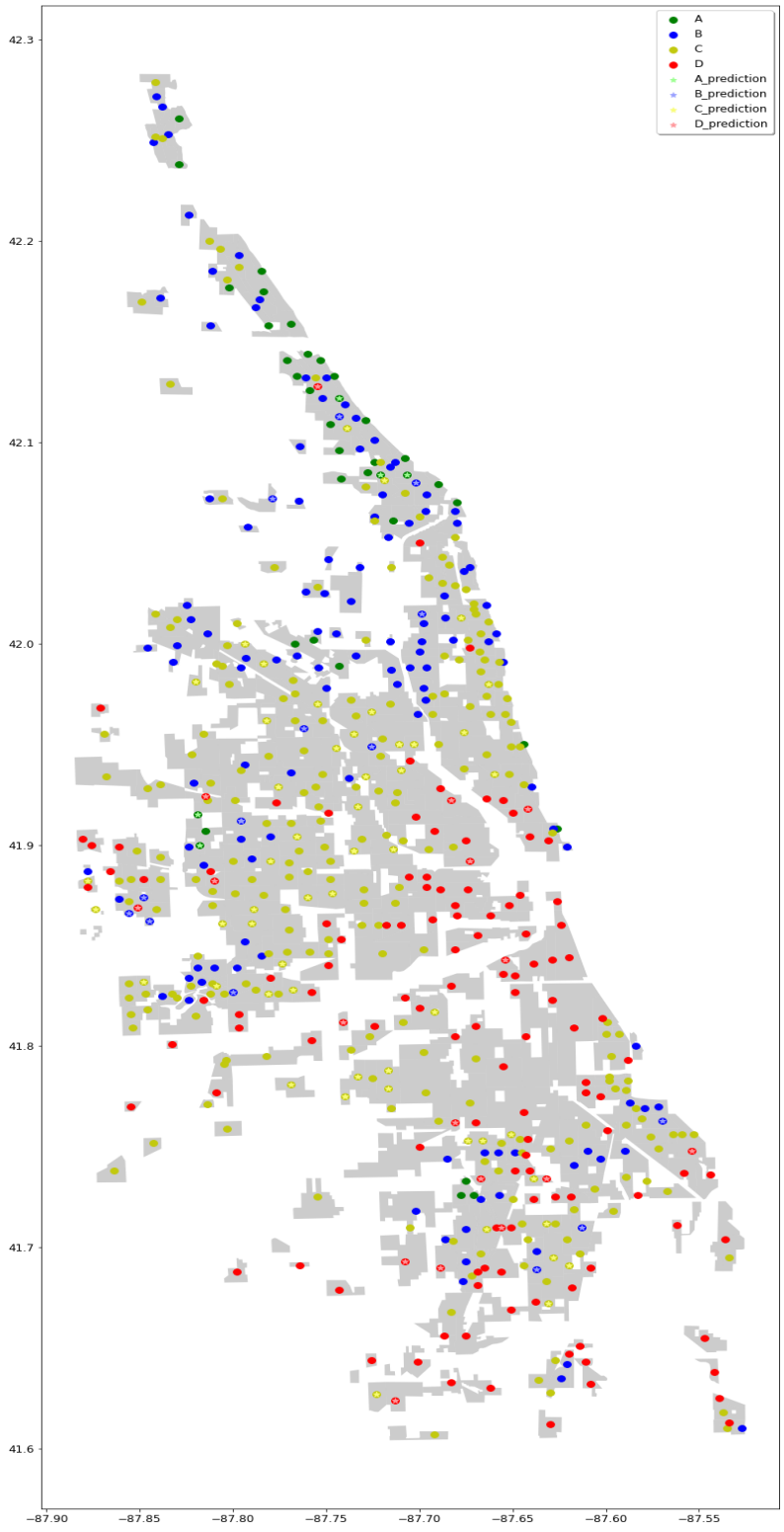
Methods

We downloaded housing data from Cook's County Assessor's Office as well as from Mapping Inequality Redlining to use for our analysis. The Mapping Inequality Redlining data was a shape file which we converted into a csv. We had to truncate decimals for latitude and longitude and drop the "names" column so longitude and latitude in the Mapping Inequality dataset would match CCAO data. We then merged Cook County historical redlining house grades with the publicly available house pricing model data on the assessor's website on house latitude and longitude data. We then looked at how accurately our model predicted a house's 1930 housing grade based on different combinations of variables in the CCAO dataset. We avoided using the variables "neighborhood code" and "town code" because they were too highly correlated with what we were trying to predict (housing grade). For example, if you know the neighborhood code you can essentially say whether or not a home was redlined. While it may have been inevitable for history to still be apparent in these data points, it is not inevitable for it to repeat itself. It is important that the CCAO analyzes their data so that they are not using historically racist data in their models. While we initially were thinking about putting "neighborhood code" and "town code" in our model, we ended up removing it because it produced too high of an accuracy and we felt it made more sense to take them out to see if other features that were not strictly neighborhood-related were correlated with historically redlined areas. Incorporating "neighborhood code" and "town code" increased our accuracy to 97% indicating that these features are closely associated with historical house grade. It should be noted, however, that having "neighborhood code" and "town code" produce such a high accuracy at predicting historical redlining grade, that this does mean there is a high association between redlined grade and neighborhood variables in the dataset. More discussion and investigation should go into how using these variables to predict housing price perpetuates the historically racist grading system in modern day. Once we removed "neighborhood code" and "town code," we added in other features and determined that the higher accuracy we had to predict historical redlining grade indicated a stronger association between the variables in the CCAO dataset and historically redlined housing grades. The features we included in the model were Property Class", "Age Decade", "Roof Material", and "Land Square Feet" which had 73% accuracy. Throughout our feature modeling process, we split the dataset into training and testing data with an 80/20 spit so that our testing data could verify our model. We also created a decision tree to assess which features were the strongest predictors of 1930's housing grade and a contour plot to visualize the impact of the features selected.

To visualize our predictive model, we plotted predicted house grades over what the grade actually was (see Figure 1). Accurate predictions are represented by the same color star inside the same colored dot. Incorrect predictions are represented by a dot with a different color star in

it. Missing data was represented by a dot with no star inside. We had mostly accurate predictions and missing data since our model had such a high accuracy. We chose to use accuracy as our model evaluation metric because accuracy depicts how well the assessment price actually evaluates the true sale price, which is how we determine fairness in this setting.

Fig. 1 Map of Predicted Housing Grade over Actual Housing Grade



In addition to creating our house grade prediction model and mapping its predictions, we also mapped the most expensive and least expensive homes on top of redlined data to see if there are still trends today from the historical data. We started by cleaning the data, dropping "1" from the "sale price" data for homes since there were so many homes with "1" as an erroneous value. We did not take out other sale price outliers because we are looking at the most extreme values of the sale price as we believe those will be the most closely mapped to historically "green" or "red" areas. We then mapped the house sale price of Cook's County with a color map and plotted housing grade overtop.

Interesting Findings

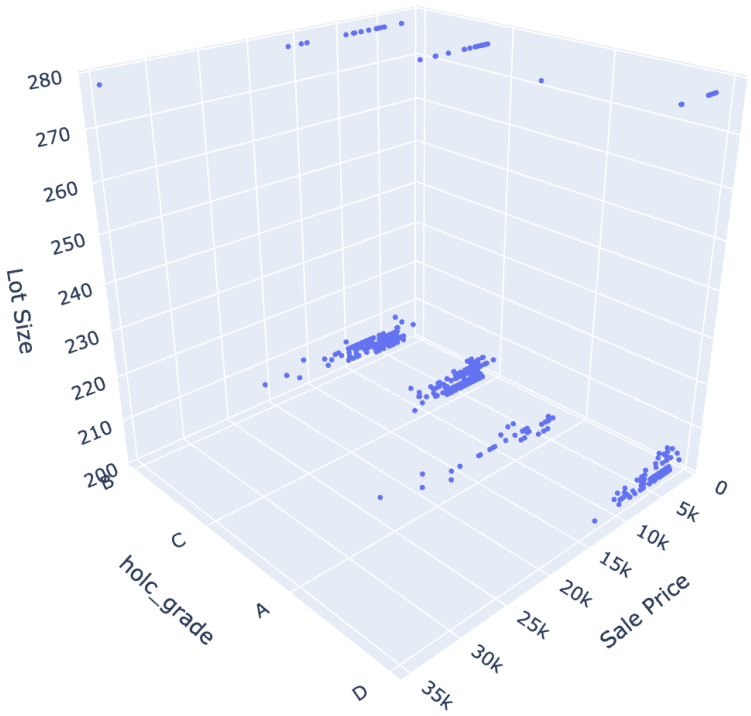
From the feature modeling, a particularly interesting finding that we noticed immediately was that the "site desirability" variable in the CCAO dataset was "2" for every home. This was surprising because we

assumed that “site desirability” would be the strongest predictor of historical housing grade. We are curious if “site desirability” was previously used as a variable to indicate how nice an area was but has been changed to make all values equal in an attempt to create a more equitable housing assessment model since how nice or desirable an area is is subjective. After further research, we found “site desirability” did indeed used to have a racial component, and therefore, they equalized it by making all of the values equal. Another interesting finding from the decision tree was that housing “square footage,” “age decade,” and “sale price” were the most predictive features for determining a house’s 1930’s grade. “Age decade” could have had such a strong correlation with historically redlined data because a greater proportion of longitude and latitude data from older homes matched the older homes in the CCAO dataset since CCAO has many more newer homes in the dataset that did not have matching 1930’s housing grades.

Analysis of Findings

The feature modeling contour plot (figure 2) shows us that while sale price and plot size are some of our strongest features, they are not strikingly different for different housing grades. We interpret this to mean historical redlined grades are not dramatically influencing sale price and lot size. We would have been much more uncomfortable had the data shown a strong correlation between sale price and lot size and historical redlined grade because that would mean the data is still rooted in these historically racist housing grades.

Fig 2. Contour Plot by Housing Grade, Lot Size, and Sale Price



From mapping the most expensive and least expensive homes (figures 3 and 4), we see most expensive homes are in historically green or grade A areas and additionally, we see the least expensive homes were all in historically red or yellow districts (grade C and D). We saw the most expensive homes are in the northern eastern part of Cook County where historically green or grade A homes used to be. However, there is also a small concentration of expensive homes in a historically redlined district in the middle eastern area of Cook County right near the port. These homes range from $2.8 million to $7 million. This could be because ports are generally urban centers where people visit often and usually expensive to buy homes in but furthermore, this location appears to be close to downtown Chicago. With the rise of urbanization especially in economically successful cities, these areas that were once redlined, have now flipped. The port area that used to be red but is now home to some of the most expensive properties in Cook's County. Overall, wealth appears to be concentrated in this county in these two pockets and compared to how spread out the least expensive homes were, this highlights a huge wealth disparity, which could be another reason for the correlation in 1930's housing grade and sale price given there is high income inequality.

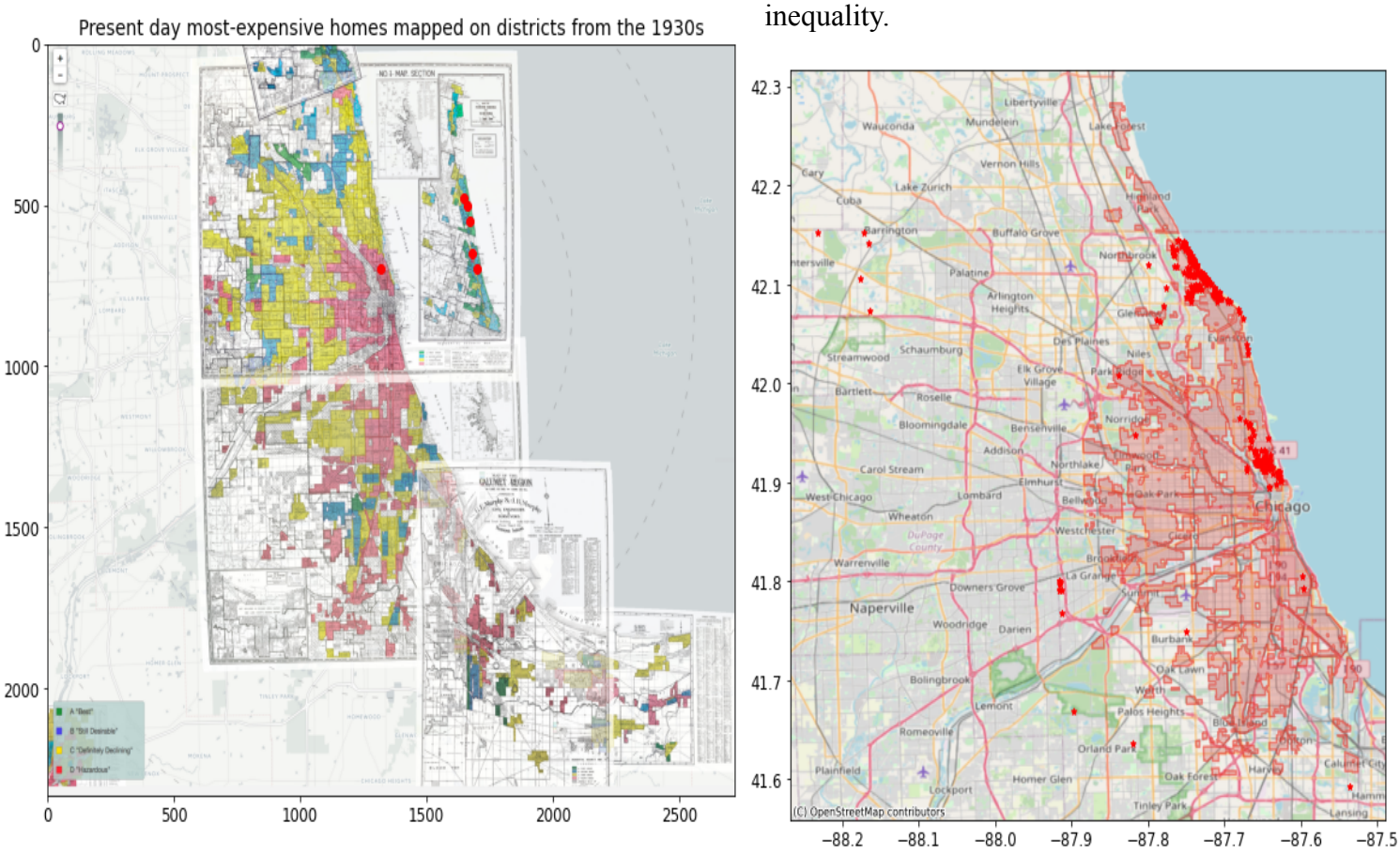


Fig 3. Present Day Most Expensive Houses Plotted Over Historical Redlining Map

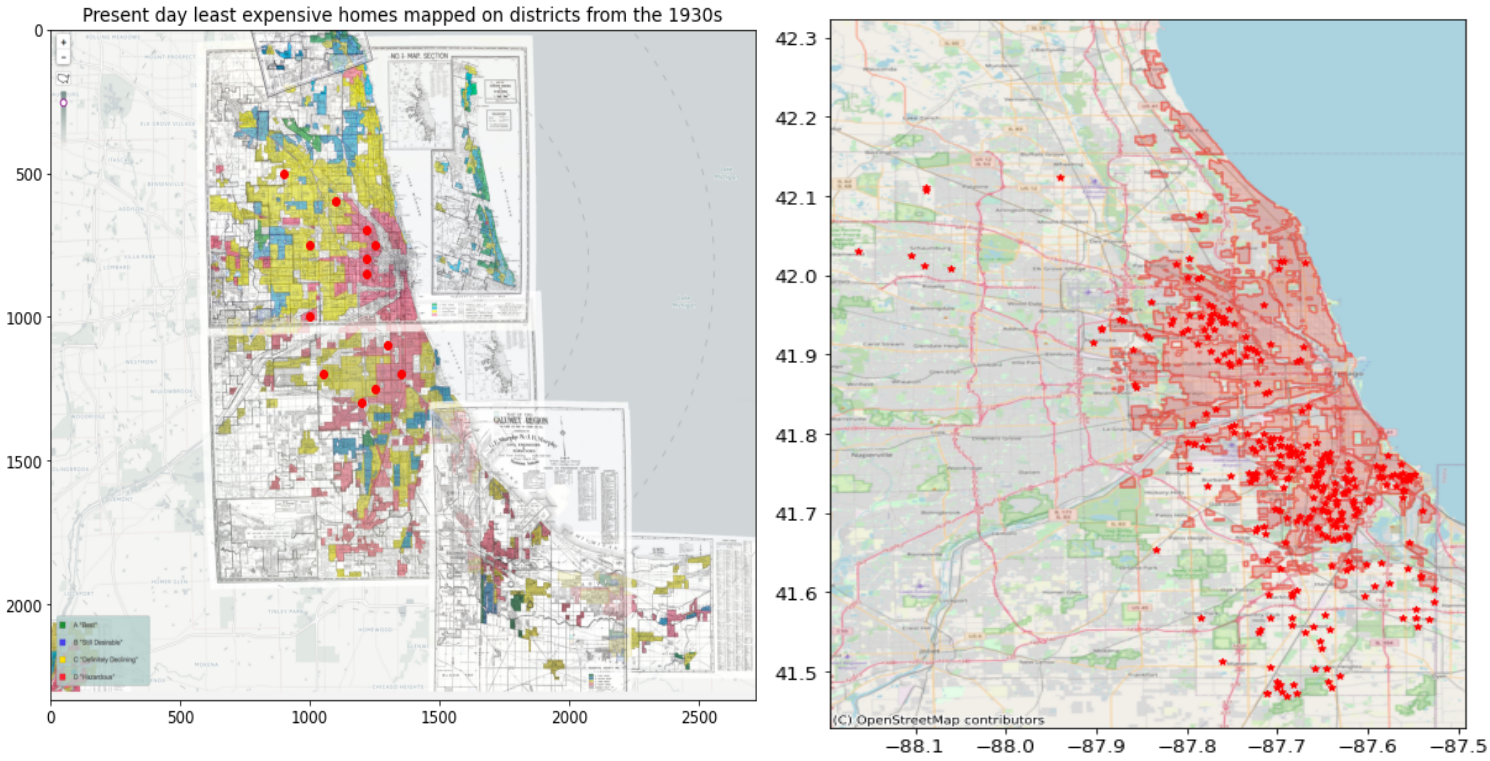


Fig 4. Present Day Least Expensive Houses Plotted Over Historical Redlining Map

Evaluation and Limitations

Although the questions we are trying to answer are different from the questions Cook's County is trying to answer, the questions we are answering are tangentially related. What the CCAO office was trying to do was make the property evaluation office more equitable so by looking at if the data they are using can predict whether a property was redlined we can say whether or not their process is still inequitable. While we expect history to still be somewhat embedded in this data, there are social, economic, and political implications of these using certain variables that are associated with historical redlining grades. Our findings that "neighborhood code" and "town code" accurately predict housing grade and that housing "sale price" and "lot size" can be used to predict house grade get at overall issues in housing policy that still need to be addressed. While we believe our model is shedding light on these issues, it will be difficult to completely separate today's data from the past. It is important that we invest in these neighborhoods as much as we invest in the data collection and analysis processes to undo the historical damage that can still be seen today.

A primary limitation of our data is that not every longitude and latitude metric from the 1930's historical redlined data mapped to the longitude and latitude metrics in the CCAO dataset since many of the homes in the CCAO dataset were not there in 1930 so we were only able to include

the houses that were included in both datasets. Additionally, it is difficult to separate the effects other than redlining associated with how and why certain variables could predict historical housing grade since housing grade was also associated with indicators other than race. We also were limited in that we had to classify *neighborhoods* instead of individual *homes* because we only have area grades and not individual home grades and this aggregated data may not be as accurate as individual data. Another limitation is we do not know which neighborhoods had rent control and which did not but we know rent control is a huge factor for desirability and without this data, we cannot say for sure if it is biased because there are other institutions at play. Furthermore, since the variable "sale price" was not in our test set so we had to use other variables as a proxy instead. In addition to this, in the training data, where "sale price" is included, there may be some errors in the data or the data may be capturing too much irrelevant information. For example, some values for sale prices are listed as $2 or $100. We initially thought this represented values in the thousands but nowhere was there such an indication and furthermore, it did not match the rest of the data where housing prices were in 6 digits. It is highly unlikely that homes sold for $2 and this could be another null value but there is even an inconsistency between the null values.

We decided to map some of these null values onto the map. What we found was that null values were pretty evenly distributed across Cook's County. There was only one pocket that didn't appear to have any null values coincided with the locations of the wealthiest homes. This may be a coincidence but also it could speak to data quality, specifically the quality of data obtained from wealthier neighborhoods versus less wealthy neighborhoods and lines of communication that are accessible to higher income neighborhoods.

Another limitation is that there is no information on CCAO's website relating to how the housing data was sampled or when it was sampled and how it is updated. There could be selection bias if housing data is only sampled when a house is sold since the most "in-demand" homes (both really cheap *and* really high valued homes) are more likely to be sold and bought more frequently. This could bias it so that only homes in the extremes are being updated regularly in the CCAO database.

APPENDIX

Fig A. Historical Map of Cook's County Redlining

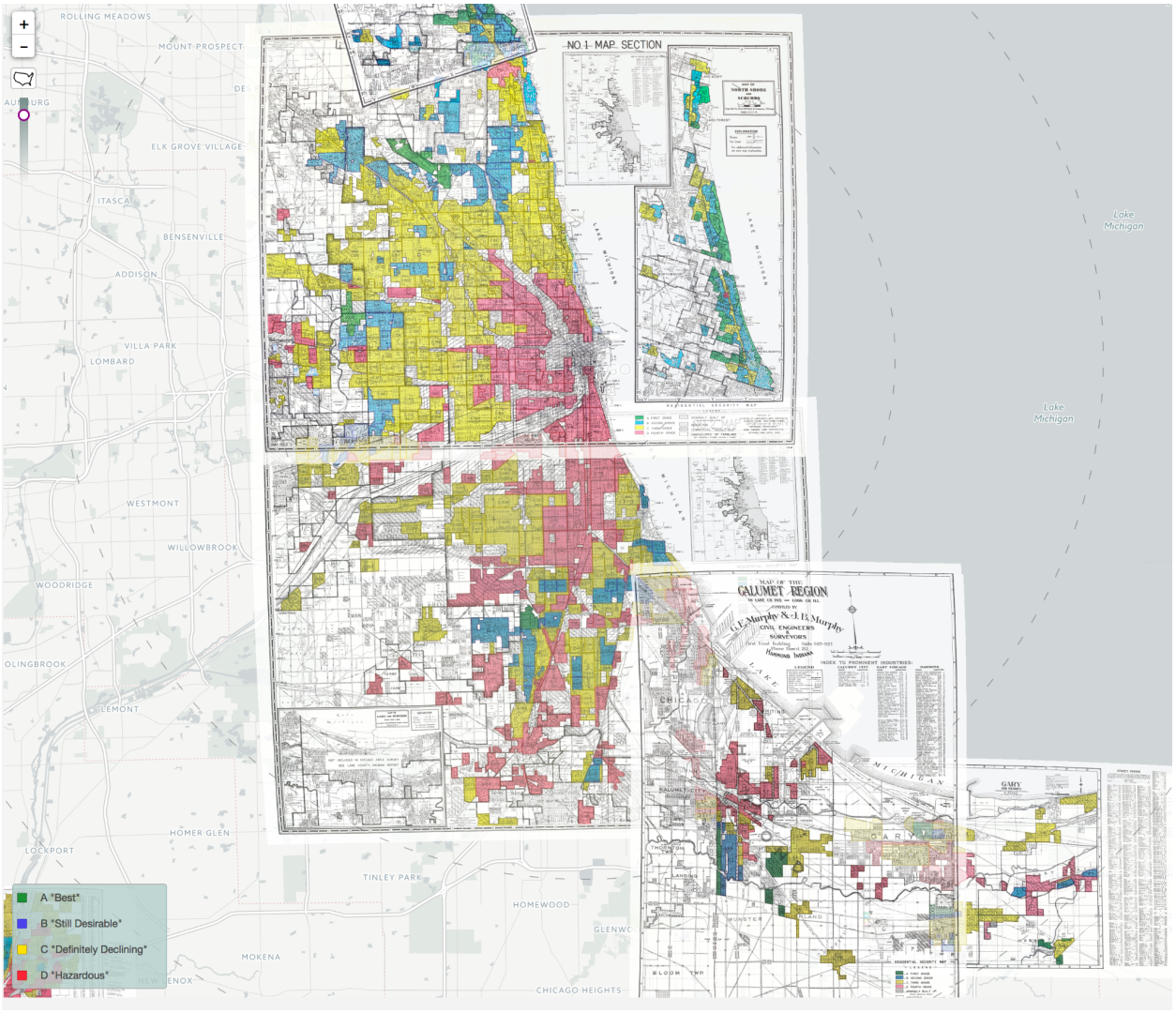


Fig B. A mapping of current day least expensive homes in Cook's County on a Base Map