Kai McNamee
May 6, 2022
Gov 94DN
Map link

Mapping LA's Oil Wells

Introduction

Interactive map link

Los Angeles is home to almost 8,500 active or idle oil wells. They dot the hills of parks, line the beaches, and can even be found on school grounds. Several studies have found that living near these wells can pose adverse health effects — proximity to polluting wells has been linked to increased rates of asthma, health complications at birth, and cardiovascular problems. According to an investigation by the LA Times and Center for Public Integrity, about 2 million California residents live within half a mile of an unplugged well, and low-income, Black and Hispanic people live closer at a slightly higher rate. Even after a well is no longer being actively developed, it can sit for years unplugged, exposing residents to harmful pollutants — these idle wells, writes the Times, are California's toxic "multibillion dollar problem."

As more and more drilling companies collapse in the decline of California's oil industry, wells are abandoned, and the costly process of plugging a well is left to state and city officials. Plugging the 19 oil wells on the campus of Beverly Hills High School, for example, cost the city $40 million. The wells have been the subject of public controversy since the early 2000s, when Erin Brockavich-Ellis' law firm sued the school district on behalf of former students, claiming the wells on campus had caused them to develop Hodgkins disease and cancer. The case was ultimately dismissed, the defendants claiming the plaintiffs lacked evidence linking their illness to the well. The wells had been temporarily closed multiple times before the case, once after city officials found the wells leaking benzene (a known carcinogen), and once after the operating company Veneco Inc. was found illegaly venting natural gas.

Veneco Inc. went bankrupt in 2016, and the wells sat idle until the city launched a remediation project that plugged the wells in 2020. This four year delay between the day the well went inactive to the day it was plugged is comparably short. According to the Times/Public Integrity analysis, half of California's idle wells have gone unplugged for more than a decade.

So how long would it have taken to plug the Beverly Hills High School wells if they were not the subject of intense public scrutiny? If a famous lawyer had not taken up the case? Or if the idle well was not in such an affluent, predominantly White school district, but in a working-class, predominantly Hispanic community? Previous research has repeatedly confirmed these kinds of race and class-based environmental and public health disparities exist. The 2014 water crisis in Flint, Michigan was found to have the greatest impact on socioeconomically disadvantaged communities. Toxic waste sites across the US are disproportionately found in lower-income, minority neighborhoods.

This analysis aims to further explore the question of who is most impacted by LA's oil wells and to what extent oil drilling in LA perpetuates environmental injustice. Using data from the LA County Open Data hub and American Community Survey, I applied spatial and statistical analysis techniques to model the effects of demographic and economic indicators on LA residents' proximity to plugged, active, or idle wells. I tested two hypotheses: (1) non-White and lower income communities are disproportionately impacted by active and idle wells; (2) plugged wells are more concentrated in whiter, wealthier areas, where political and economic influences can encourage or expidite the cleanup process. I found that neither of these hypotheses were clearly supported by the data, leading me to further follow-up research questions that could yield interesting results if this analysis were continued.

Data and Methods

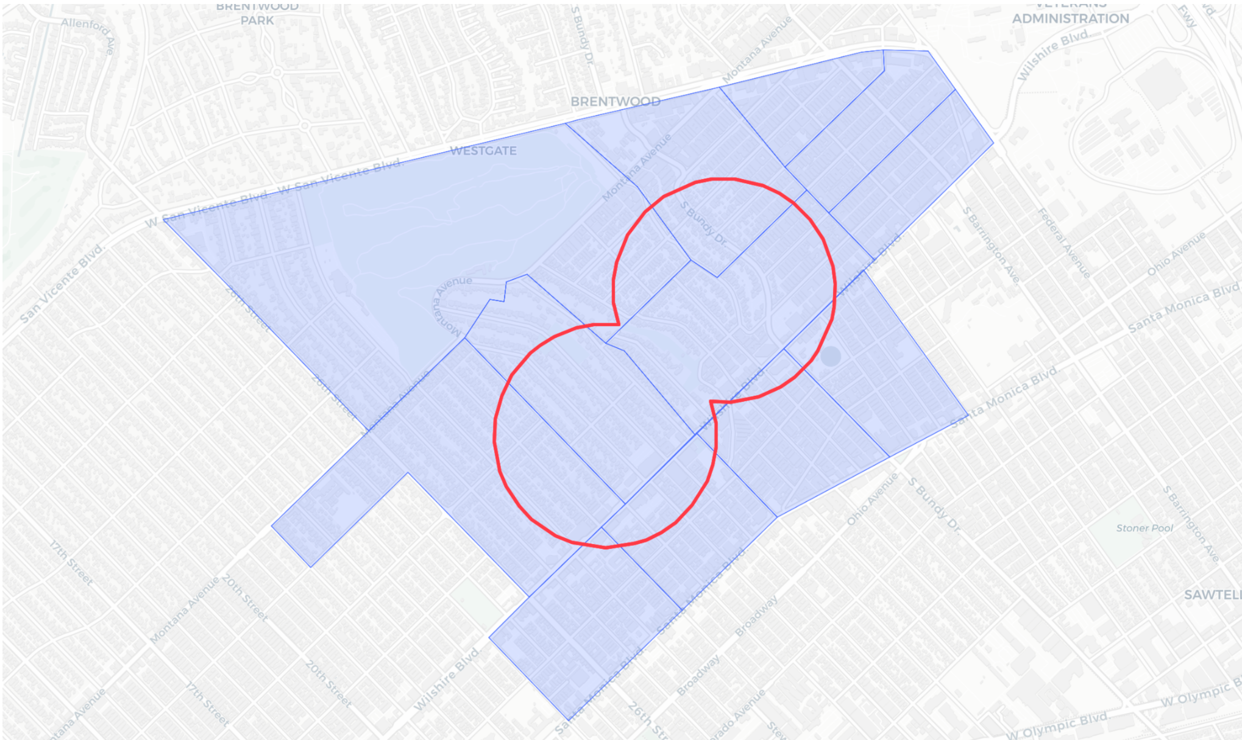
This study utilizes two methods of spatial analysis, differing primarily by the unit of observation. The first half uses buffer zones as the unit of analysis, while the second uses census block groups subdivided into hexagons. Point locations for all wells in LA county were obtained from the county's public data hub. The dataset includes coordinates for 16,883 plugged wells, 5,056 active wells, and 3,435 idle wells (total = 25,374), last updated December 2019. Demographic data was retrieved from the 2019 ACS, summarized at the block group level.

A. *Buffer Zone Analysis*

For the buffer zone analysis, I plotted each group of wells and generated buffer zones of differing sizes. I used 0.25 mi, 0.5 mi, 0.75 mi, and 1 mi radius buffers and joined all overlapping buffer zones to generate target polygons for areal interpolation. I then interpolated demographic and income data from ACS block groups into the target polygons, weighting each block group by its percentage overlap with the target polygons. The resulting dataset includes joined buffer zone polygons of different radiuses for each type of well, and demographic and income statistics contained within the polygons. I then conducted a descriptive analysis to find any relationships between well presence and demographic breakdowns and income statistics of the interpolated data within the four defined buffer zones.

Figure 1 illustrates an example of a 0.25 mile buffer drawn around two idle wells. To calculate the total population of Black residents within the buffer zone, for example, count data contained within any block groups touching the buffer is summed and weighted by the percentage overlap with the target geometry.

Figure 1: Example 0.25 mi buffer zone around 2 idle wells



This method of interpolation is based on the strong assumption of even population distribution throughout all census block groups — while this assumption likely does not hold for more rural parts of the county, most of the wells are found near urban and suburban areas, where there is likely more even population distribution. In Lancaster, for example, block groups can span several miles of rural farmland. Even in the example shown in Figure 1, the block group in the northwest corner contains a golf course, meaning the population count interpolated into the target geometry is likely lower than the actual number of residents living within the buffer zone. The percentage overlap between the buffer zone and the block group geometry is rather small, but the percentage of the block group's residential buildings contained in the buffer zone is much higher.

B. Hexagonal Subdivision Analysis

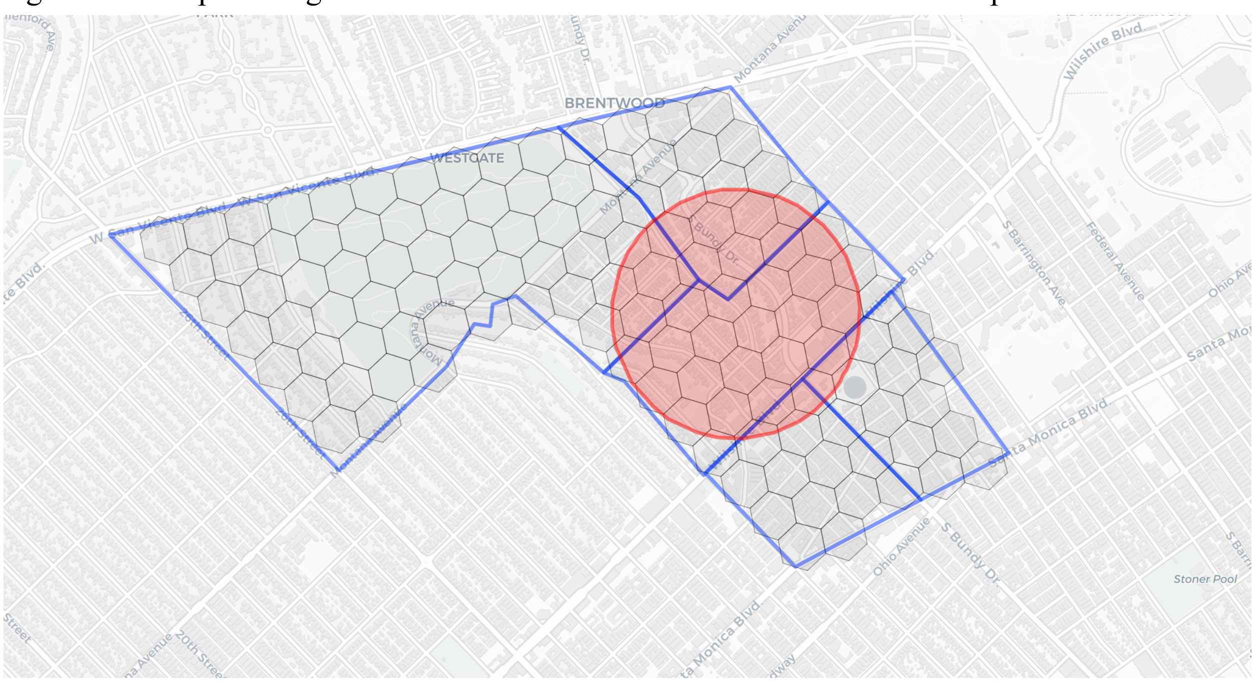
Since the buffer zone analysis only examines trends within predefined buffer zones, I created subdivisions of census block groups to model how demographic and income statistics of a given subdivision are correlated with its distance to the nearest well.

For this analysis, I used [H3](), a hierarchical geospatial indexing system, to interpolate block group data into smaller hexagonal units, and calculated each hexagonal unit's minimum distance to each type of well. Subdividing block groups increases the resolution by which distance to the nearest well can be observed, thus increasing variation in the observed minimum distances to oil wells. This also allows for more refined control over which geographic areas are included in the model. If I wanted to filter the model input data to include only residents living within a 0.25 radius of active wells, for example, filtering by the distance between wells and census tract centroids would be imprecise — the irregular shape of block group polygons makes it possible to include block groups where significant portions of the geographic area lie outside of the 0.25 radius. Subdividing block groups into hexagons helps solve this problem in a manner similar to that of the buffer zone analysis, but it still relies on the assumption of even population density.

Since generating block groups' constituent hexagons is computationally expensive and slow, I used a random sample of 2,000 wells for my models. I created a 0.25-mile buffer around all sampled wells, then generated fill hexagons for all block group geometries that intersected with the 0.25-mile buffer.

Figure 2 illustrates the process I used to subdivide block groups; the figure shows the same geographic area as Figure 1, but only one of the two wells made it into the random sample of 2,000. Filtering the model data to include only hexagons within a 0.25-mile radius results in a dataset including only the hexagons in the red circle.

Figure 2: Example hexagonal subdivisions within a 0.25-mile radius of a sampled well



Using only hexagonal subdivisions within 0.25 miles of the nearest well, I used the following model specification:

$$Distance_{si} = \beta_0 + \beta_1 White_i + \beta_2 Black_i + \beta_3 Asian_i + \beta_4 Hispanic_i + \beta_6 Population_i + \beta_7 Predominant_i + \beta_8 Per\ capita\ income_i + \varepsilon_i$$

Where Distance is the minimum distance between well type s and the centroid of hexagonal subdivision i; White, Black, Asian, Hispanic are the population proportions of subdivision i; Population is the total population of subdivision i; Predominant is categorical variable indicating the group with the highest population proportion in subdivision i; Per capita income is the aggregate income divided by the total population of subdivision i; and ε represents random error in the model.

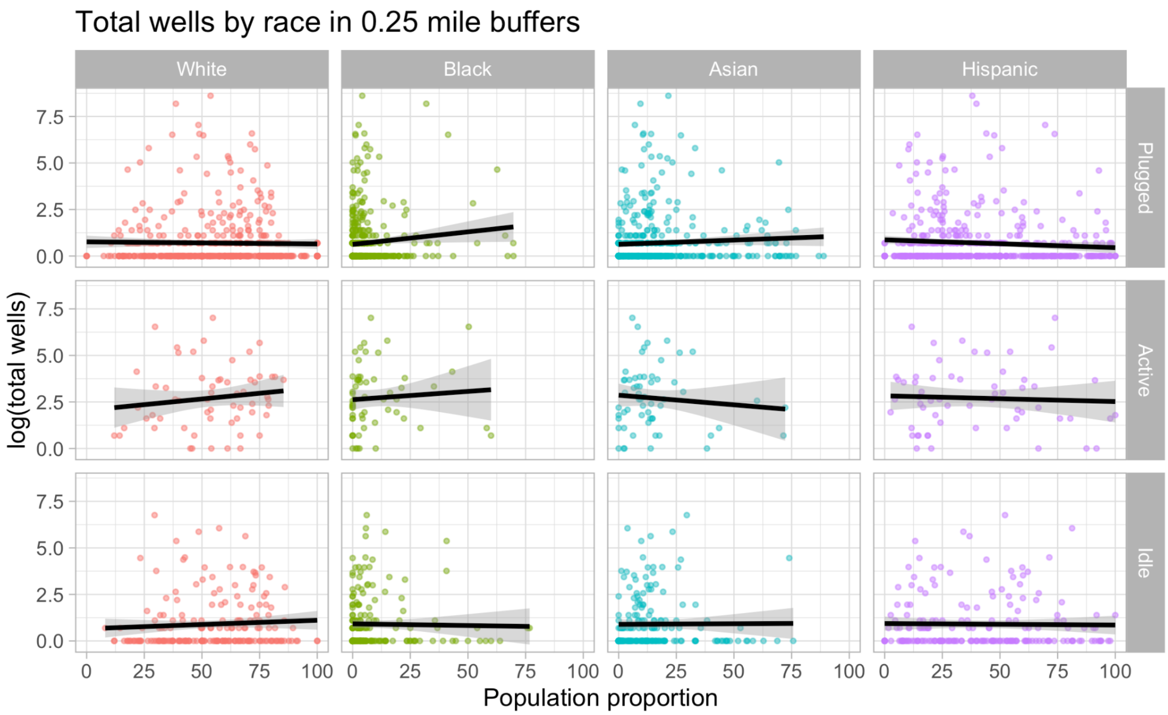
Results

A. Buffer Zone Analysis

Descriptive analysis of the data suggests there are no strong relationships between well exposure and racial demographics or income levels. An initial look at the data finds that the majority of residents within all buffers are white and/or hispanic — demographic breakdowns do not change much between buffer zones, and increasing the buffer size increases the amount of noise in the analysis (ex: most of the city of LA exists within 1 mile of a well of any type).

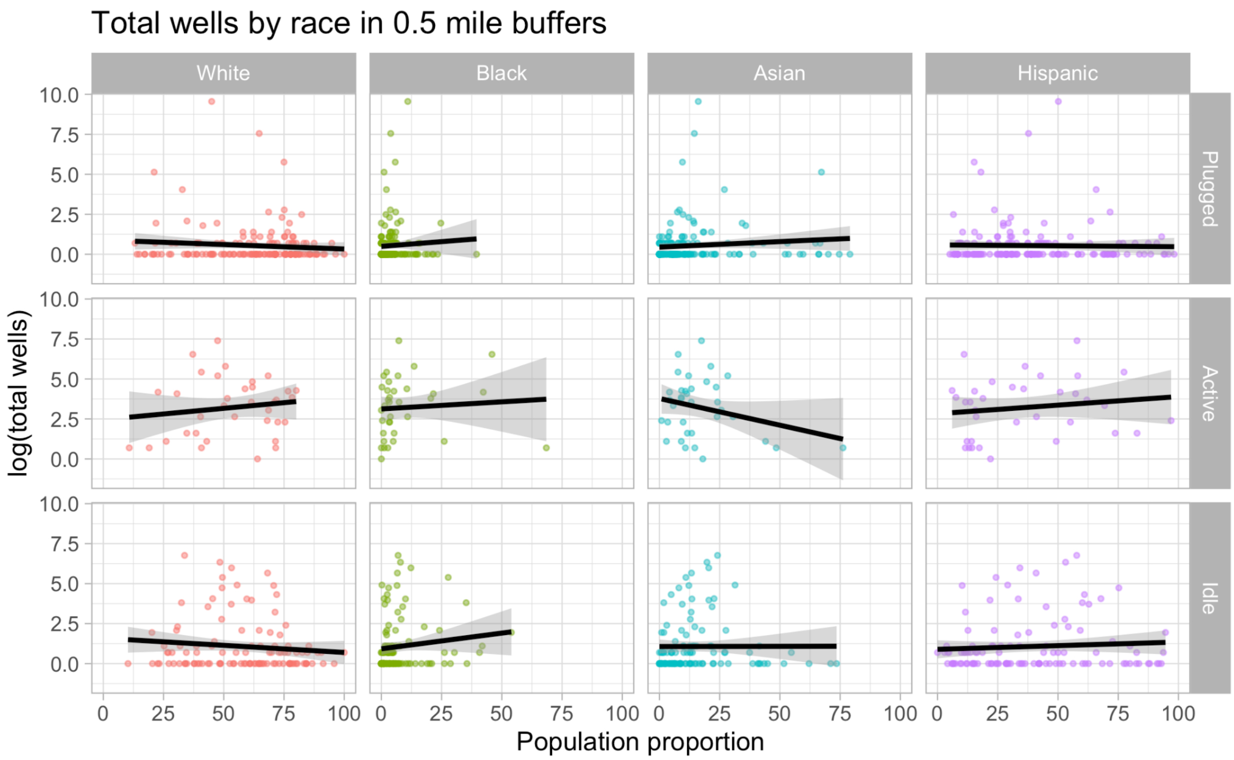
Figure 3 shows the log-transformed total wells in a given 0.25-mile buffer over the percentage population of White, Black, Asian, and Hispanic residents. There are slight positive correlations between the percentage White population and the percentage Black population with the number of active wells in a given buffer, suggesting there might be more active wells in areas with higher proportions of White or Black residents than in buffers with lower densities.

Figure 3



Changing the buffer size to 0.5 miles (Figure 4) reverses some of the correlations observed in Figure 3, suggesting there is a cutoff buffer distance after which the effects of racial demographics on the number of wells in a buffer cannot be observed. For example, the percentage of White residents in a given 0.25-mile buffer is positively correlated with the number of wells in the buffer, but this trend is reversed when examing 0.5-mile buffers.

Figure 4



Per capita income in a buffer zone is slightly correlated with the number of wells in a 0.25-mile buffer (Figure 5); the correlations are again weakened when the analysis is expanded to 0.5-mile buffers (Figure 6).

Figure 5

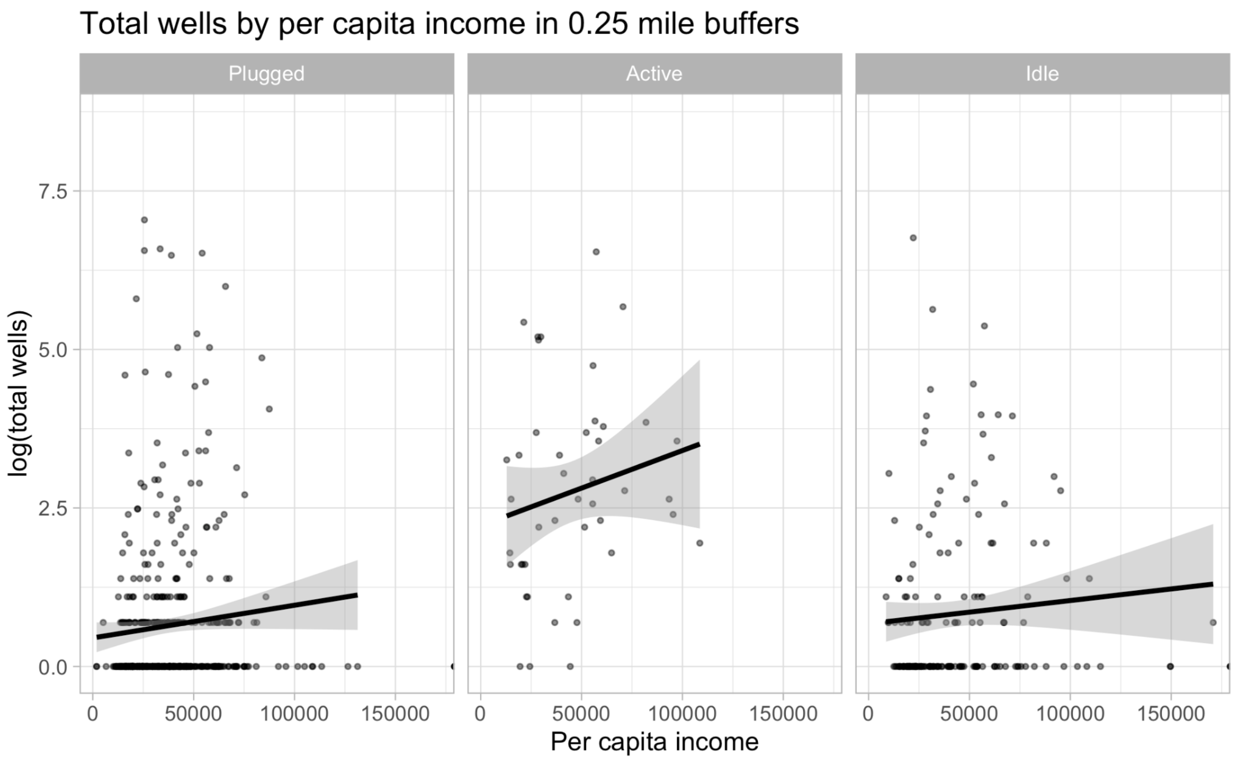
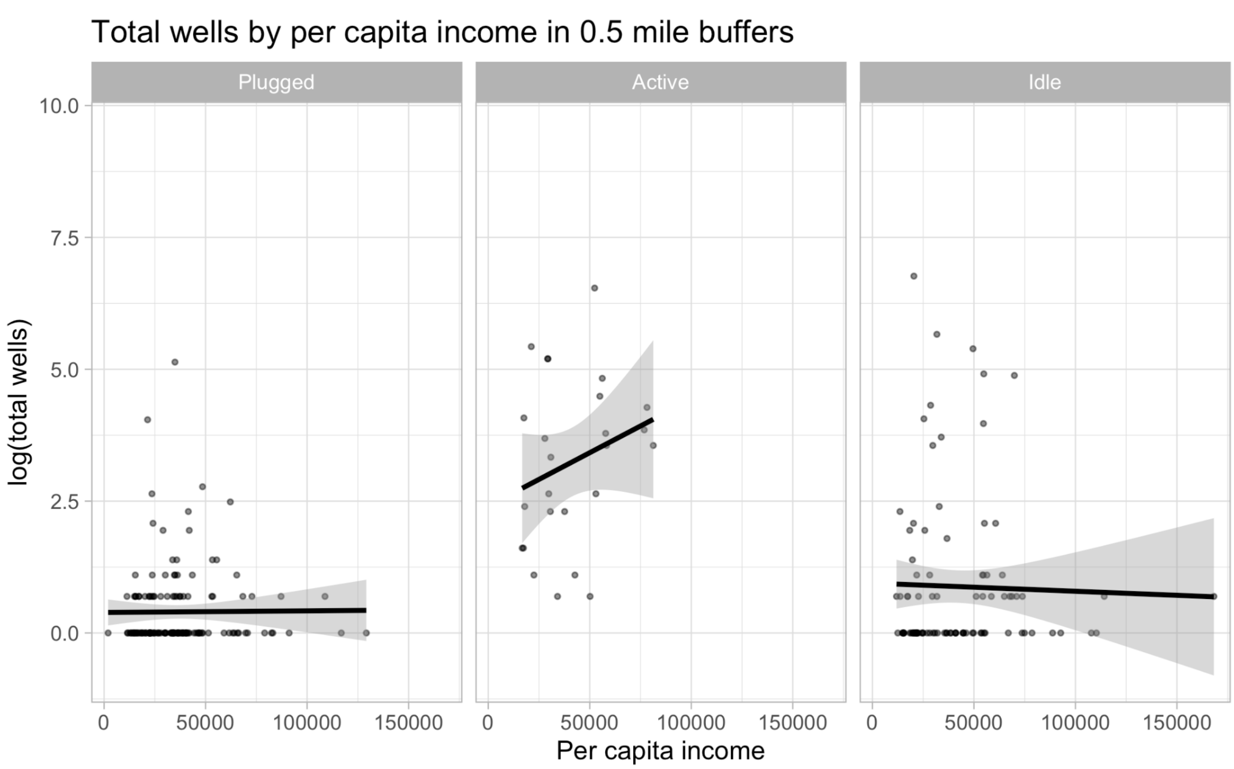
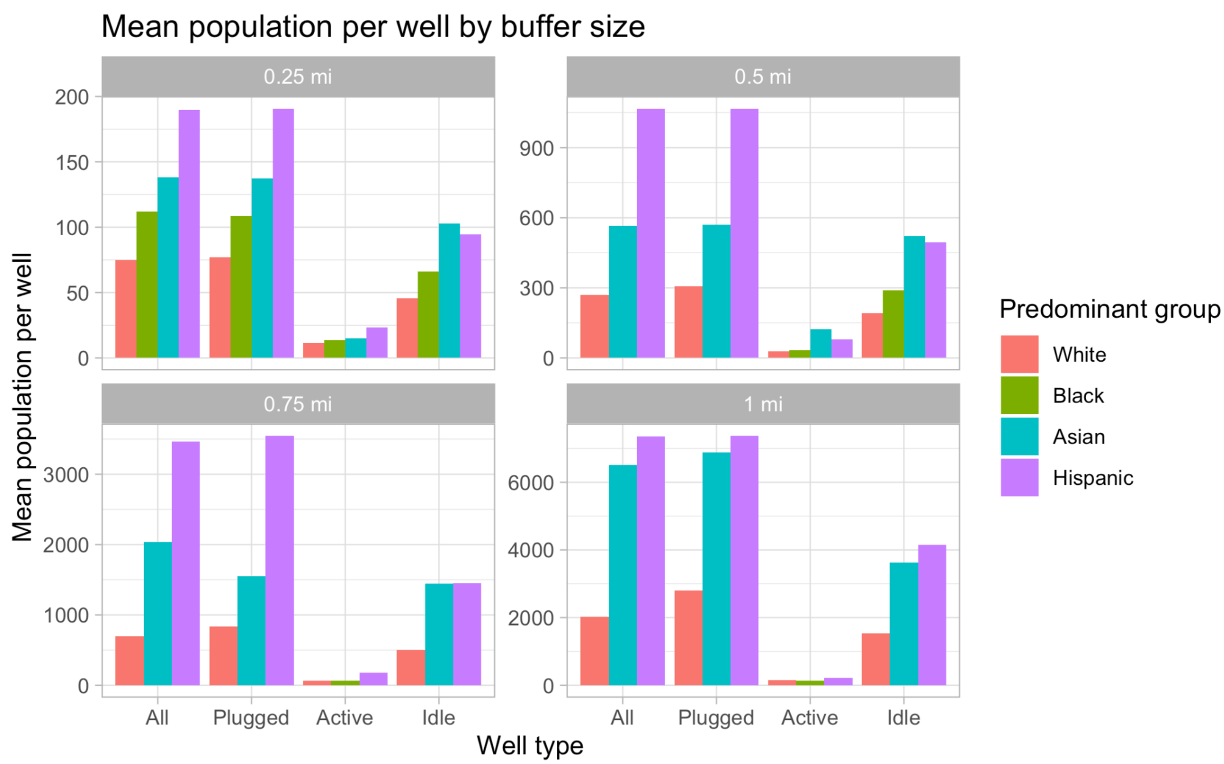


Figure 6



The buffer zone analysis also revealed disparities in the total population per well in a given buffer based on the predominant racial group. Figure 7 illustrates the average population per well for predominantly White, Black Asian, and Hispanic buffers; the plot suggests that wells impact more people in predominantly non-White buffers.

Figure 7



B. Hexagonal subdivision analysis

Modeling the impact of demographics and income on a hexagonal subdivision's distance to the nearest well similarly yielded ambiguous results. The linear regression models produced statistically significant results, but the models have very low explanatory power, as indicated by the low R squared values. Figure 7 indicates the regression coefficients resulting from 4 models trained on a random sample of 2,000 wells:

Figure 8

	Dependent variable:			
	Distance to active (1)	Distance to idle (2)	Distance to plugged (3)	Distance to any (4)
prop_white	0.527	1.867***	-0.243**	0.152
	(0.357)	(0.286)	(0.122)	(0.107)
prop_black	-1.171**	1.207***	-1.002***	-0.724***
	(0.463)	(0.428)	(0.184)	(0.157)
prop_asian	-0.479	1.854***	-0.989***	-0.601***
	(0.467)	(0.384)	(0.154)	(0.136)
prop_hispanic	0.560*	0.639**	-0.332***	-0.260***
	(0.297)	(0.248)	(0.096)	(0.084)
population	0.603***	-0.019	-0.036*	-0.017
	(0.098)	(0.026)	(0.021)	(0.015)
predominantblack	26.497	50.879**	14.953	18.932**
	(31.063)	(23.203)	(10.066)	(8.778)
predominanthispanic	-55.603***	57.949***	-16.549**	-6.790
	(19.680)	(14.452)	(6.631)	(5.742)
predominantwhite	-46.275**	31.595**	-14.232**	-12.636**
	(18.489)	(14.175)	(5.694)	(5.071)
income_per	0.001**	-0.00001	-0.0001	-0.0002**
	(0.0002)	(0.0002)	(0.0001)	(0.0001)
Constant	192.167***	30.818	300.292***	264.158***
	(44.697)	(37.237)	(14.153)	(12.613)
Observations	1,321	1,839	10,965	14,125
R2	0.057	0.037	0.008	0.008
Adjusted R2	0.050	0.032	0.007	0.007
Residual Std. Error	105.406 (df = 1311)	105.092 (df = 1829)	104.232 (df = 10955)	104.911 (df = 14115)
F Statistic	8.784*** (df = 9; 1311)	7.726*** (df = 9; 1829)	9.625*** (df = 9; 10955)	11.855*** (df = 9; 14115)
Note:				*p<0.1; **p<0.05; ***p<0.01

Some notable findings from these models:

Model 1: Predicting a subdivision’s distance to the nearest active well.

The proportion of Black residents is negatively correlated with distance to the nearest active well. A 1% increase in a hexagonal subdivision’s Black population decreases the distance to an active well by 1.17 meters. Both predominantly Hispanic and White subdivisions are more likely to be closer to an active well than predominantly Black or Asian subdivisions.

Model 2: Predicting a subdivision’s distance to the nearest idle well.

All of the demographic variables are positively correlated with distance to the nearest idle well. The regression coefficient for the proportion of White resident in a subdivision is the highest among the race variables, but the predominantly Hispanic indicator variable has the biggest impact on distance to an idle well.

Model 3: Predicting a subdivision's distance to the nearest plugged well.

The third model shows a negative correlation between all demographic variables and the distance to the nearest idle well. While the proportion of White residents decreases the distance to idle wells, the proportion of Black, Asian, and Hispanic residents has a greater impact.

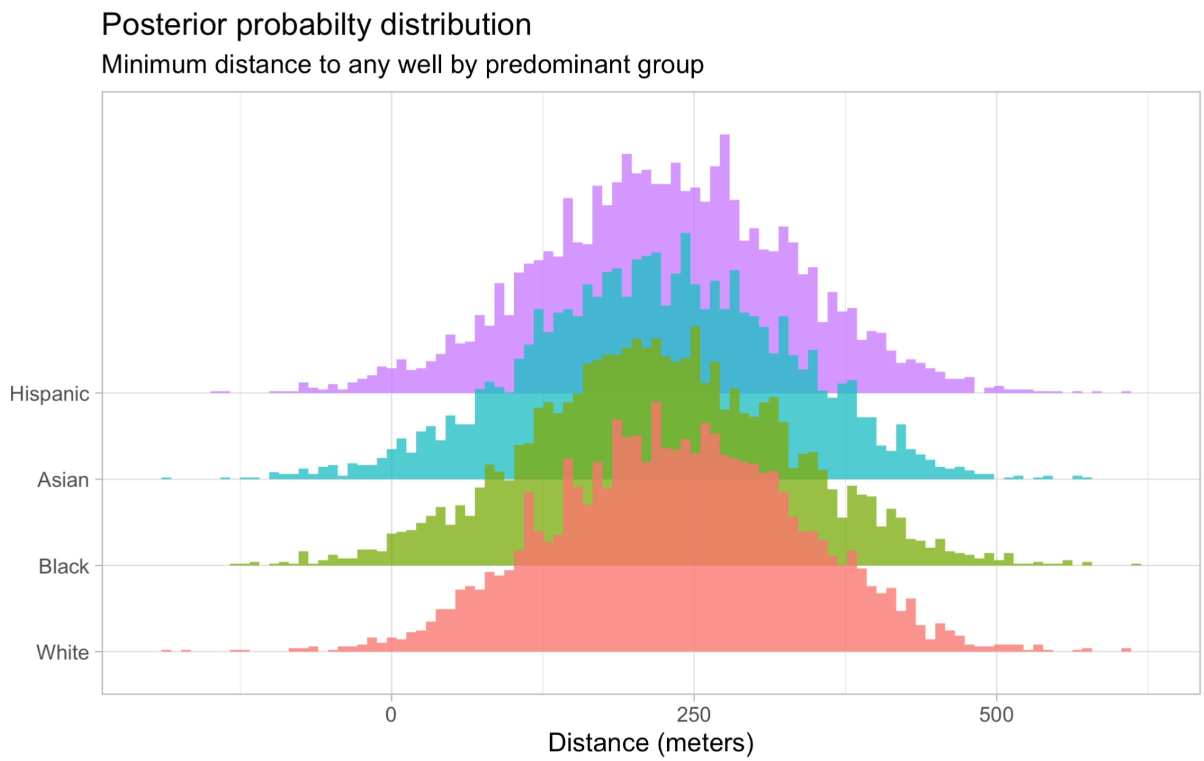
Model 4: Predicting a subdivision's distance to the nearest well, regardless of status.

Lastly, when combining the data to model the effects of demographics and income on the nearest distance to any well, the proportions of Black, Asian, and Hispanic residents have a negative effect. The proportion of white residents does not have a statistically significant effect on the dependent variable, but the predominantly White indicator variable shows a negative correlation with distance.

While many of the regression coefficients generated from the linear models indicate statistically significant results, the overall findings appear contradictory and unclear. Visualizing the model's estimates helps communicate how likely certain outcomes are, and whether the findings can support or contradict my initial hypotheses.

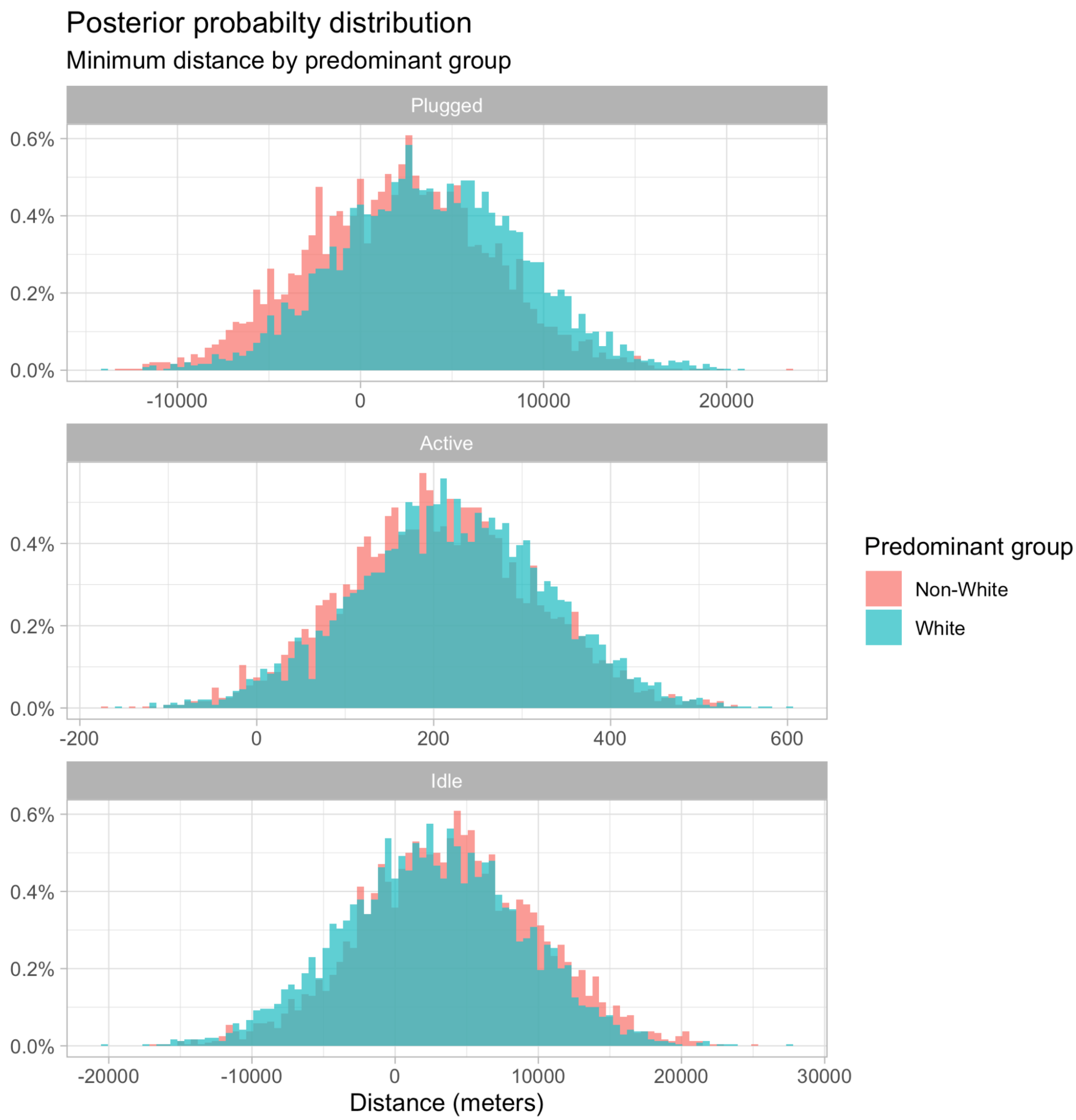
Figures 9 and 10 plot probability distributions of well proximity for the average hexagonal subdivision based on the predominant group. The histograms are generated by plugging in the average demographic proportions and per capita income for majority White, Black, Asian, and Hispanic subdivisions into the models shown in Figure 8. The horizontal axis indicates a given subdivision's predicted distance to a well, and the vertical axis indicates the probability of that outcome. While the predominant group was a statistically significant predictor of distance in some of the models, the probability distributions for all groups overlap substantially — the distributions for subdivisions of predominantly White, Black, Asian, and Hispanic residents all approximately center around 250 meters, suggesting that the predominant group does not likely predict a change in average distance to the nearest well.

Figure 9



Grouping the data into predominantly White and non-White subdivisions similarly produces unclear results. The distribution of possible values for a predominantly White subdivision with average population demographics and income almost entirely overlaps with the average predominantly non-White subdivisions.

Figure 10



Conclusion

The results of this analysis do not clearly support the proposed research hypotheses. As both methods of analysis showed, the relationship between demographic or economic factors and the impact of LA's oil wells is unclear. The buffer zone analysis found only slight relationships between the number of wells in a buffer and the racial breakdown of that area; there was also a slight positive relationship between per capita income and the number of wells. These findings are further complicated by the models using block group subdivisions — while the model outputs statistically significant results, the model specification explains little of the observed variance in well proximity. The probability distributions in Figure 9 and 10 illustrate how uncertain the effect the predominant racial group has on well proximity in average subdivisions.

Some possible limitations of the analysis include:

1. The precision of the data. Since areal interpolation requires the assumption of equal population distribution over a given geography, subdividing block groups in the second half of the analysis could be misleading. More precise data might lead to better predictions, especially while analyzing the smaller buffer zones. Similarly, I would like to redo the subdivision analysis using fill hexagons of different sizes — the resolution at which I subdivided the block groups might have been to high, which could have introduced noise without actually creating relevant data via areal interpolation.

2. The method of analysis. Due to computational and time constraints, I had to drop one of my methods of analysis, which involved comparing data intepolated into buffer zones with data interpolated into geographies outside of buffer zones. I would have conducted difference in means tests to examine the differing demographic breakdowns inside and outside of buffer zones, but my computer could not reliably handle the spatial operations needed to generate the data