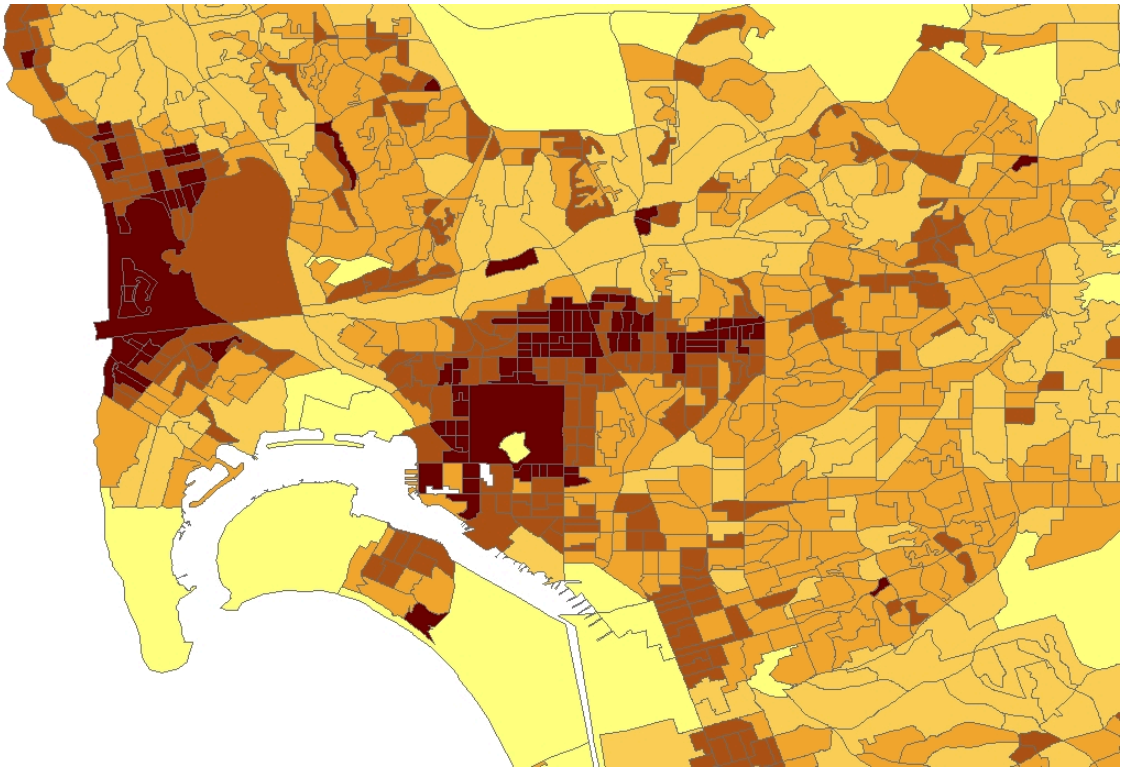
DUI, Demographics, and Urban Density



Eric Braga, Farah Farah, Peter Lenz, Esteban Lopez

04.23.2024

STAT 596 Spatiotemporal Analysis

INTRODUCTION

Crashes due to intoxication are still prevalent in the Southern California cities of San Diego and Los Angeles. Is there a pattern or contributing factors to these crash locations? If there are, it can lead to a better understanding to help prevent DUIs with road designs, urban planning and traffic stops. This paper will look at potential factors such as income, educational attainment and race as well as walkability score for census tracts, and urban density.

METHODOLOGY

In California, the Statewide Integrated Traffic Records Systems (SWITRS) is a database that collects and processes data gathered from a collision scene. The data includes date, time, and location based on roads or road intersections. The Transportation Injury Mapping System (TIMS) developed by SafeTREC offers geocoded data based on the SWITRS database. It provides x and y coordinates based on the 1984 World Geodetic System (WGS84). The data goes as far back as 2015 and is constantly updated, although 2022-2023 data is provisional as it is subject to change. We have focused our research on San Diego and Los Angeles County for comparison. With the help of Python and the Pandas library we mapped out the collision sites for each corresponding county.

Walkability is a term that arose in the 1960s in connection with Jane Jacobs and her work on urban design. In the last 20 years it has been increasingly popularized as it has a strong correlation with resident health as well as safety and economic opportunity. A walkable city is one where someone can go about a large amount of their life without the need for a car. There are grocery stores, recreation and employment all within walking distance and streets and sidewalks are safe and encourage pedestrian use.

Since walkability encourages walking and public transportation as a means of getting around, they discourage the use of single driver vehicles for the majority of the population which is currently the mode of transportation for the majority of Americans. It would stand to reason that this would have a negative

correlation with DUI deaths since less total drivers on the road would result in less drunk drivers in general.

It is clear that walkability is an asset in city planning and that it is something worth studying but how can it be measured? It is defined in an inherently subjective manner because it is based on physical factors that impact human decision making. There have been multiple different ways that walkability has been quantified and in this paper we will use the National Walkability Index released by the EPA. The index is calculated using a weighted average of 4 variables: intersection density, proximity to transit stops, the mix of different types of employment and the mix of residential and commercial land use. All four of these categories are correlated with more walk trips by residents. In figures 1-4 transit distance and walkability index are plotted for LA and SD.

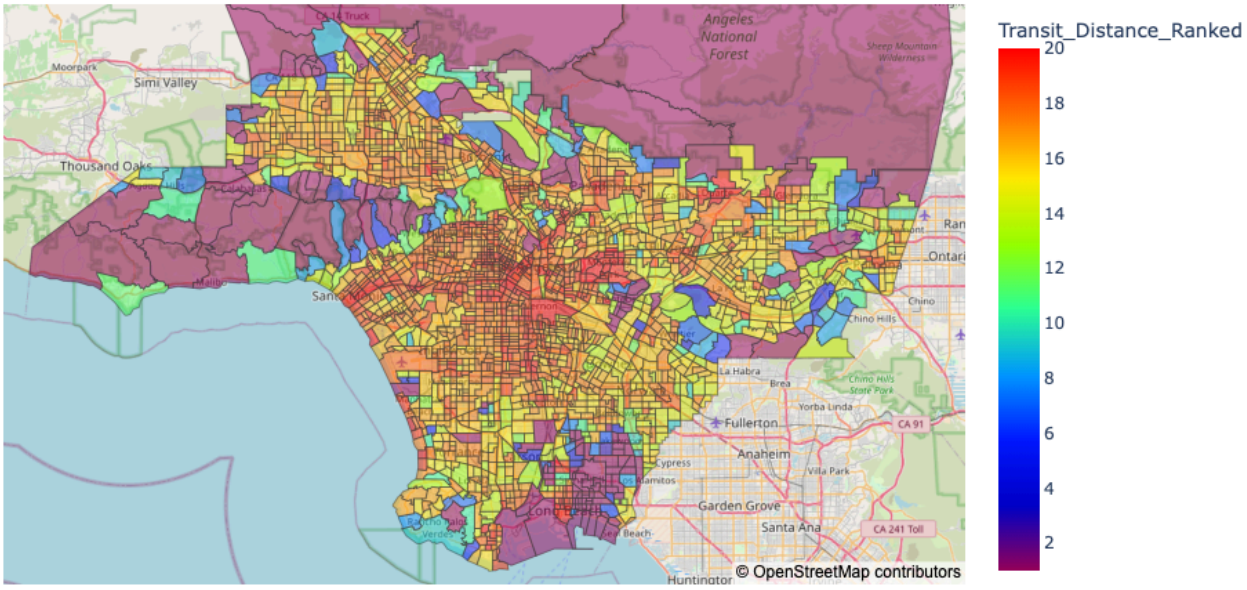


Fig. 1 Transit Distance (ranked) in LA county

LA County

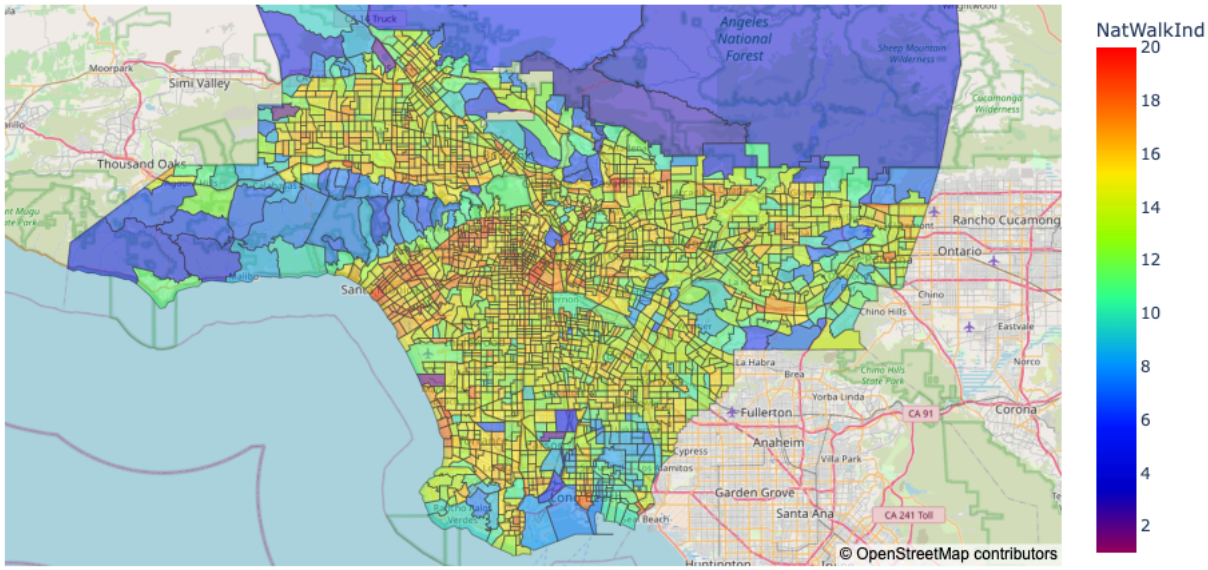


Fig. 2 National Walkability Index in LA County

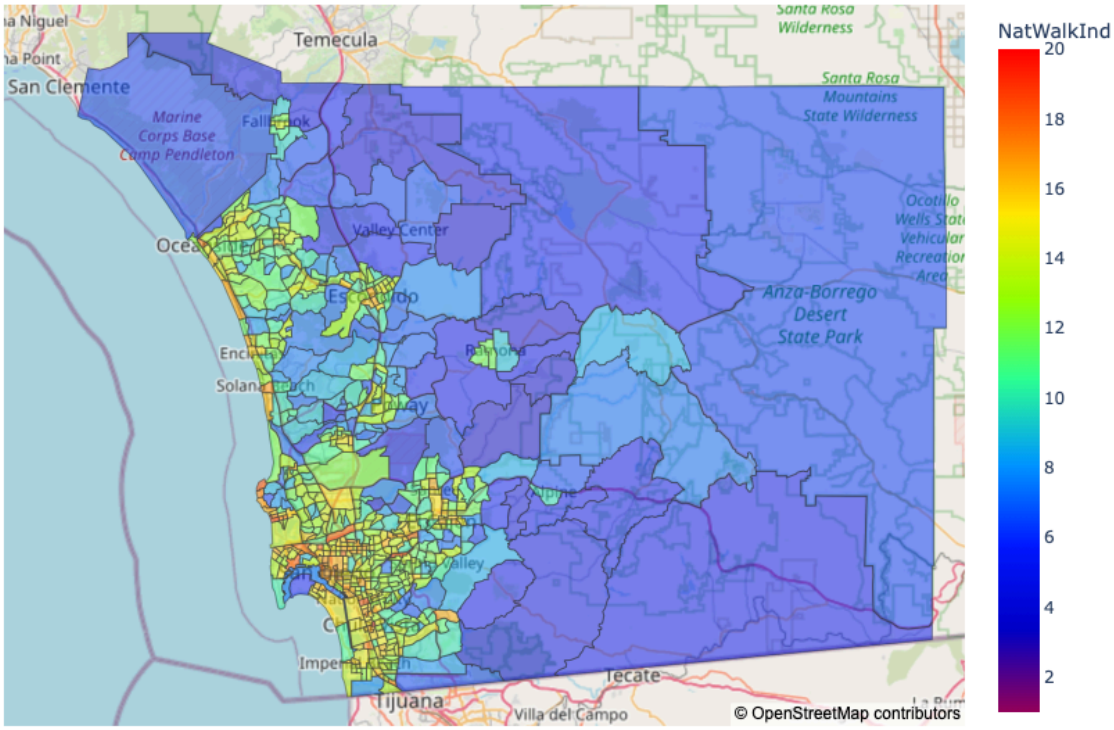


Fig. 3 National Walkability Index in SD County

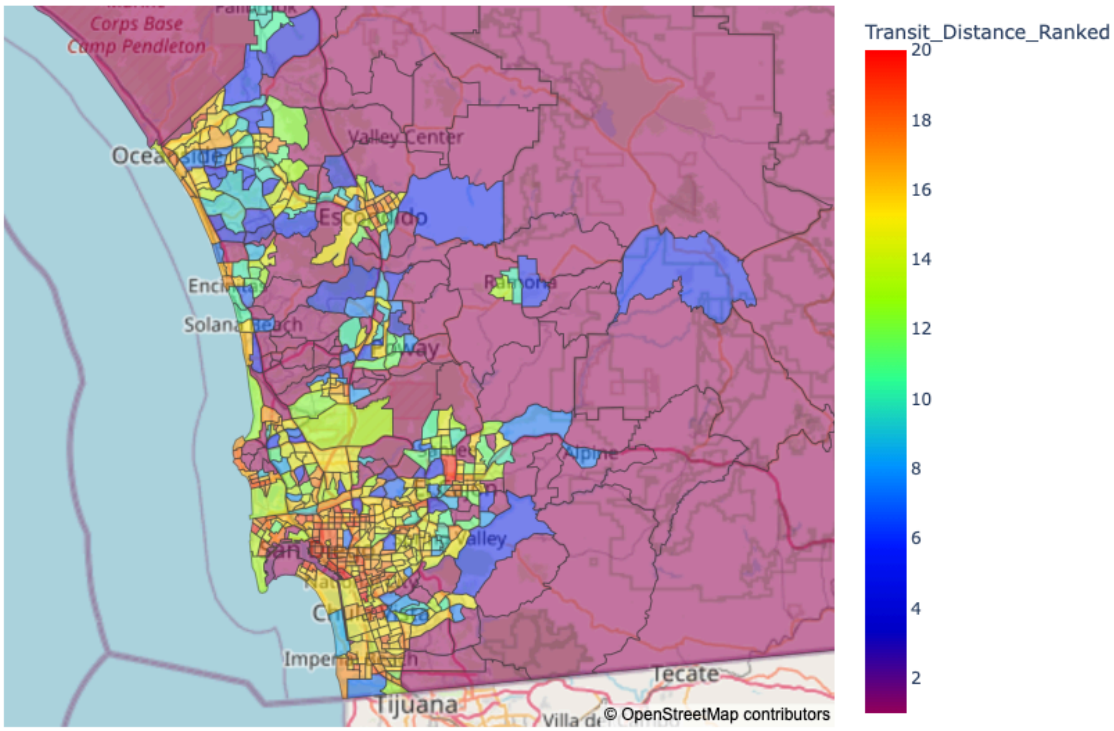


Fig. 4 Transit Distance (ranked) in SD County

Using census data can provide valuable information about demographics relating to DUI accidents. It can help us better understand the patterns and trends of DUI's by combining the maps that contain demographic data such as educational attainment, income levels, and race. Census data is a critical tool for the government and it informs their decision regarding enforcement efforts and other activities by the local government. In order to map out variables that we thought might be important for our research, we decided on using census api focusing on census tracts for the American Community Survey 5-Year Data in 2022. This was done for both Los Angeles County and San Diego County census tracts.

The urban density information comes as a raster data file from GHSL: Global settlement characteristics on Google Earth Engine. This data is all from the year 2018 with bands that describe the function and height of buildings at 10 m resolution worldwide.

EXPLORATORY DATA ANALYSIS

DUI Data

Using the Seaborne library we plotted the kernel density estimate (KDE) of both counties. It is apparent that the majority of crashes occurred closer to downtown city centers. In Figure 1 hotspots include the cities of San Diego, Chula Vista, Escondido, Oceanside, and El Cajon. Figure 2 shows a large hotspot near downtown LA with a large radius of about 5 km. Other areas with high DUI crashes include Long Beach and near Inglewood.

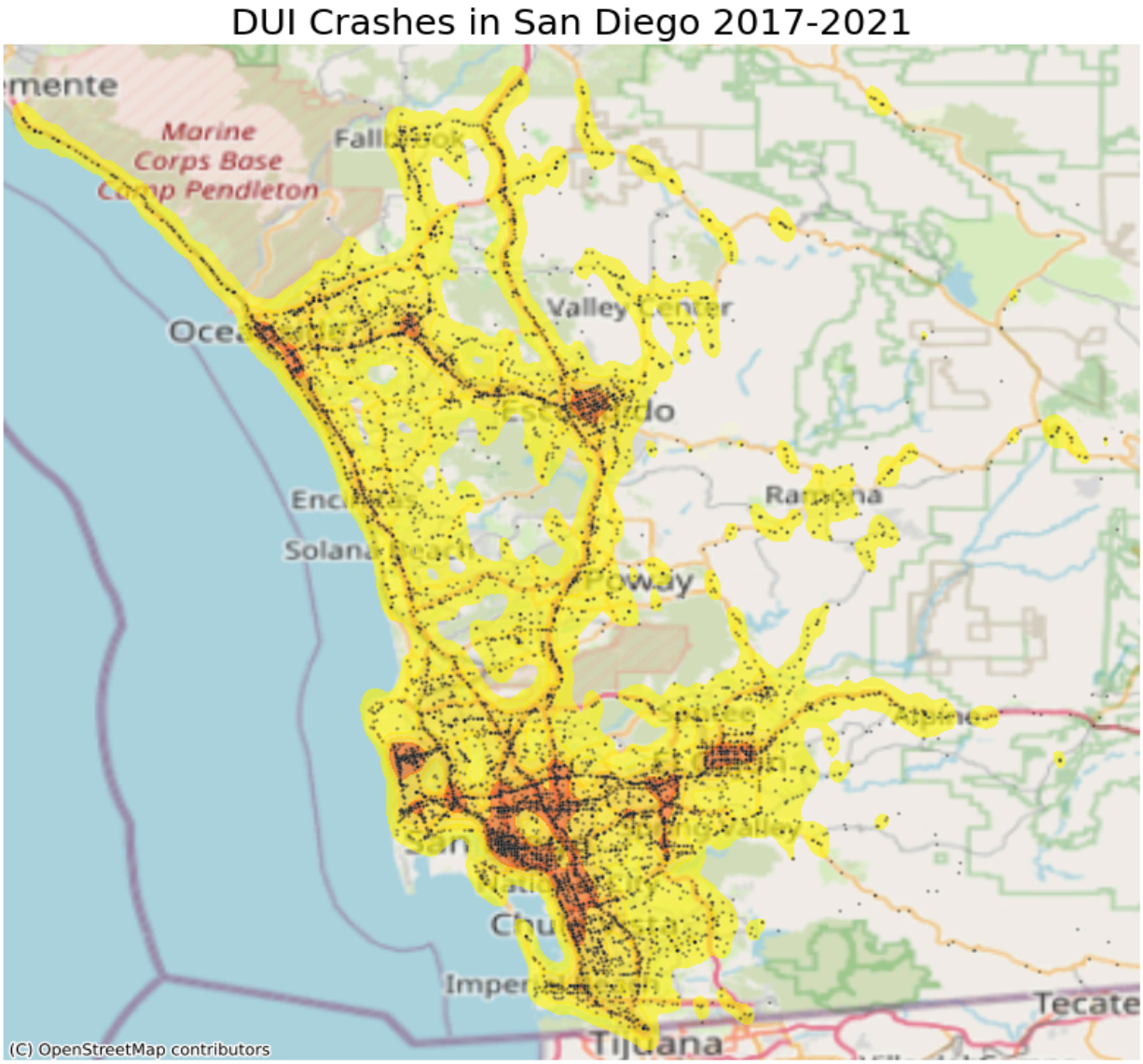


Fig. 1 San Diego Kernel Density Estimate

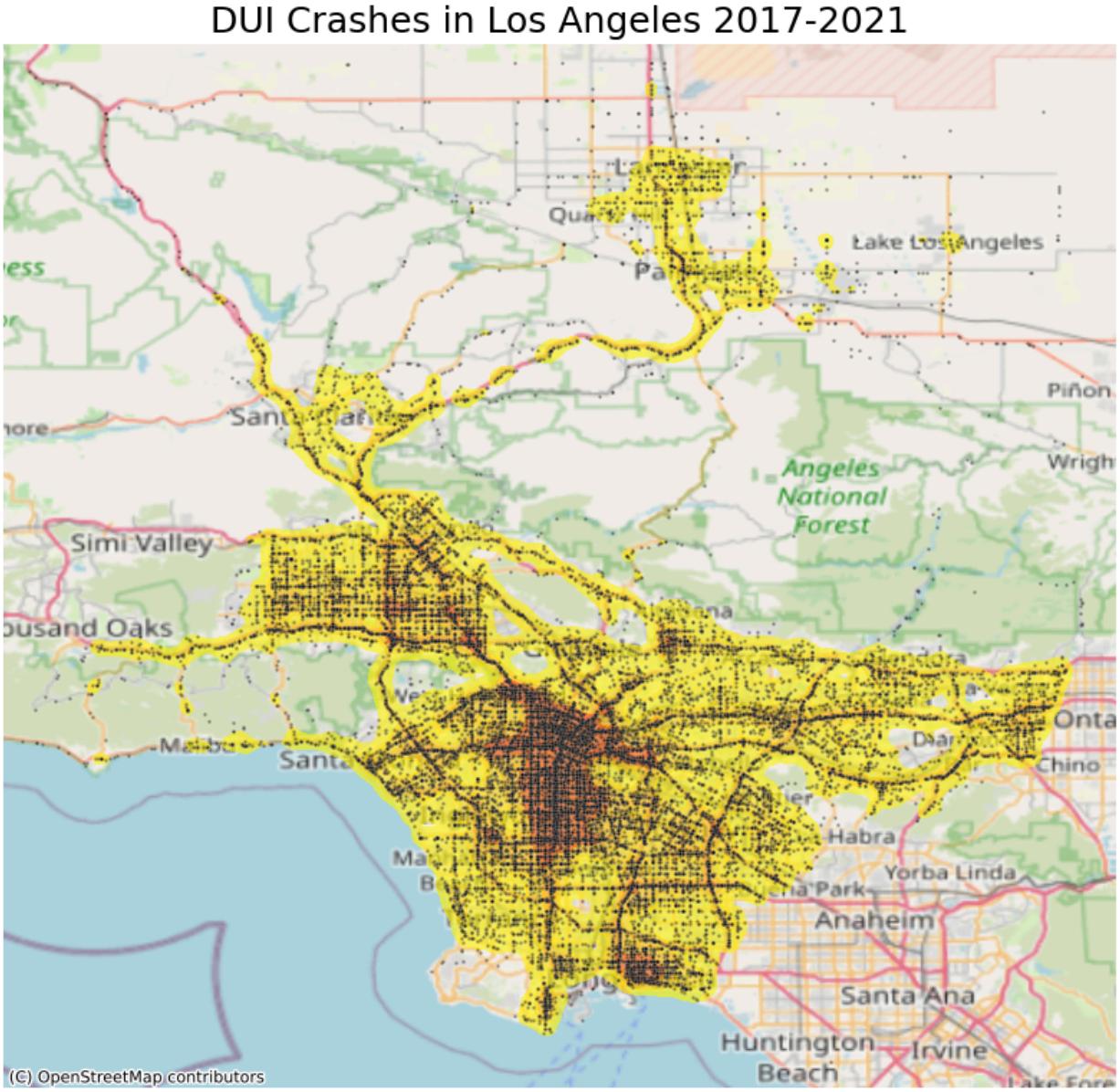


Fig. 2 Los Angeles Kernel Density Estimate

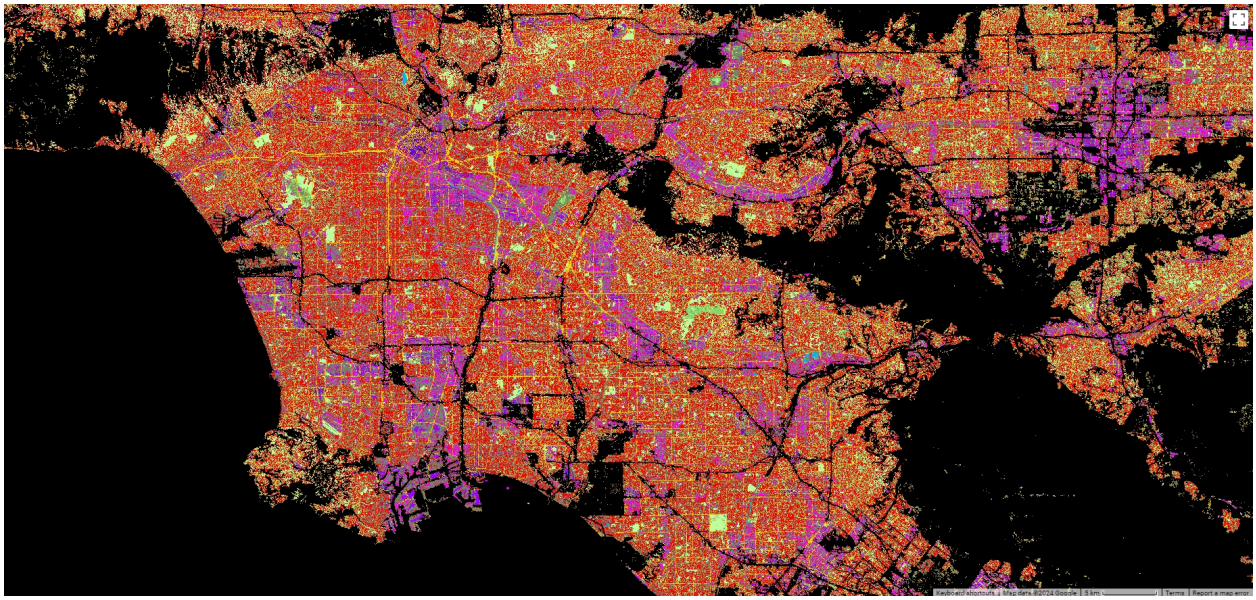


Fig 3. Los Angeles Density

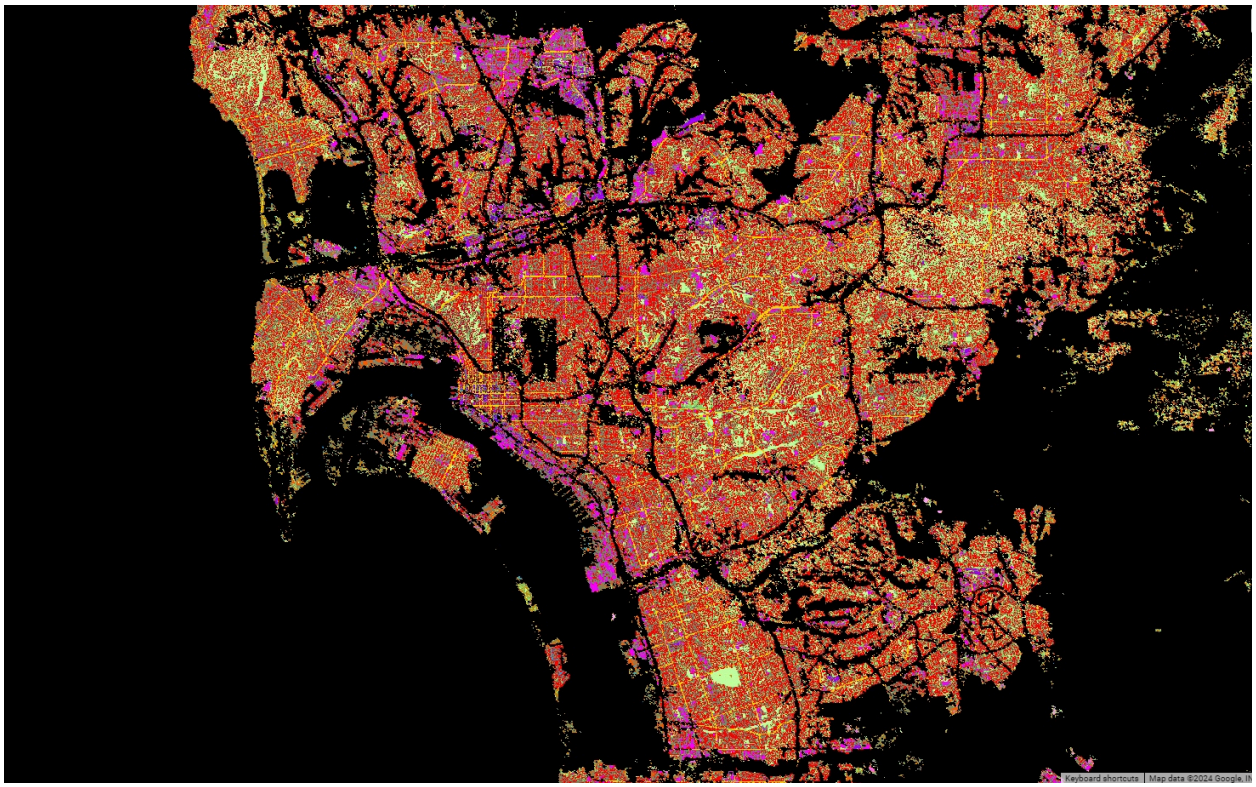


Fig. 4 San Diego Density

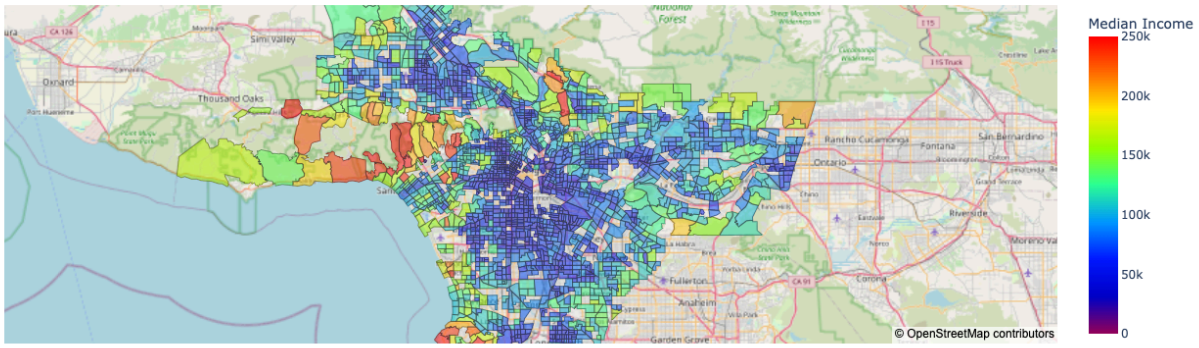
Fig. 5 Density Image Legend

#ffd501	open spaces, road surfaces
#d28200	built spaces, residential, building height <= 3m
#fe5900	built spaces, residential, 3m < building height <= 6m
#ff0101	built spaces, residential, 6m < building height <= 15m
#ce001b	built spaces, residential, 15m < building height <= 30m
#7a000a	built spaces, residential, building height > 30m
#ff9ff4	built spaces, non-residential, building height <= 3m
#ff67e4	built spaces, non-residential, 3m < building height <= 6m
#f701ff	built spaces, non-residential, 6m < building height <= 15m
#a601ff	built spaces, non-residential, 15m < building height <= 30m
#6e00fe	built spaces, non-residential, building height > 30m

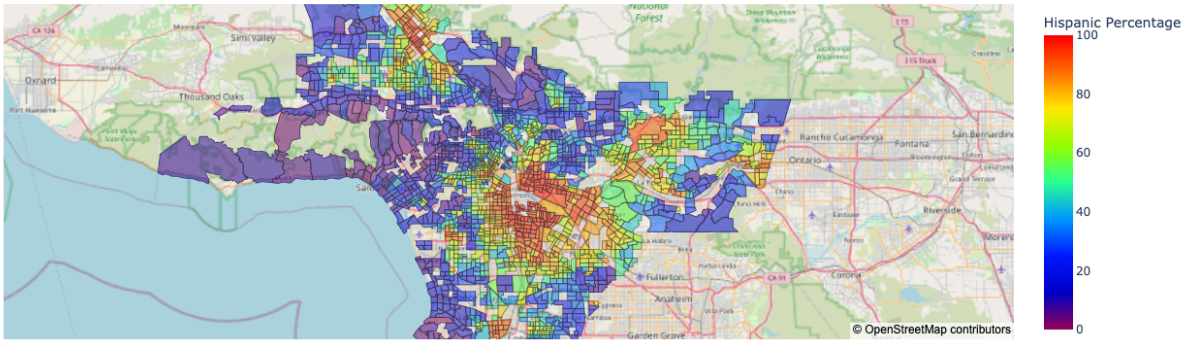
As seen in Figure 3 Los Angeles seems more dense as shown with darker colors and more dense commercial areas shown in the purple. We hypothesize there may be differences in results between the two especially considering the black areas which in LA are hills/mountains that still get passed through while driving and have more dangerous roads.

Census Data

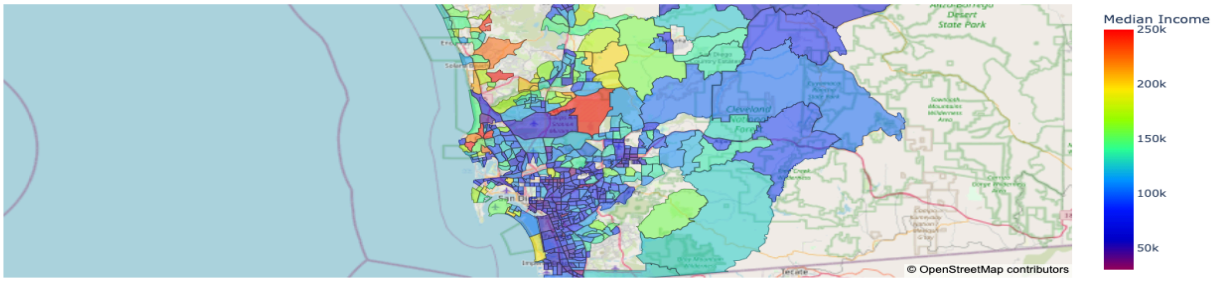
LA County



LA County



San Diego County



MODELING

Figure 9 is an image of the interactive map that compiled all layers (Urban density, census tract and dui crash points) in San Diego. The html file to this and the Los Angeles can be found on our github which is linked at the end of this report.

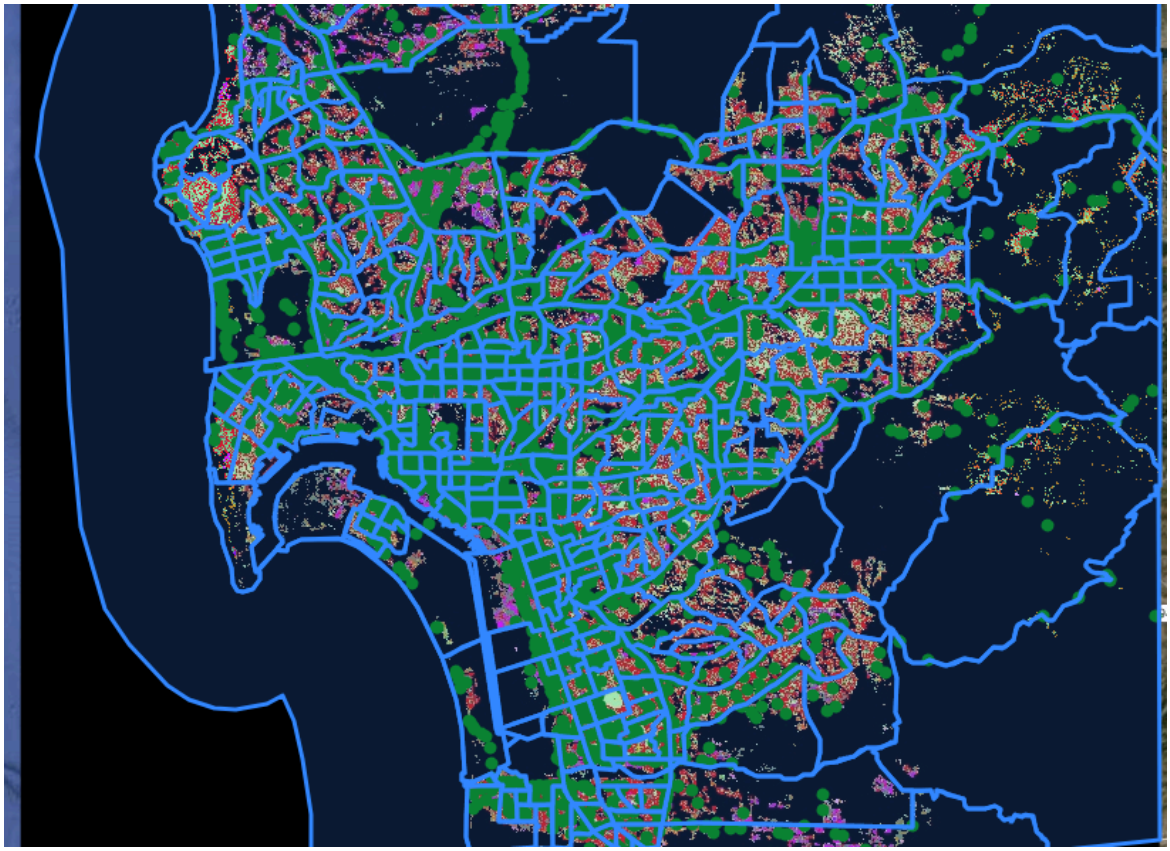


Figure 9 San Diego Interactive Map

ANALYSIS

To try to quantify if there is correlation between the crash sites and our selected data we created some decision trees. For this all data was grouped together for more data points for the model to learn from. To be able to compare the crash point data and census tract information a new column was made on the census data called points_count. This was created by using a function that was looped through all the crash

sites and determined if they intersected any of the tract polygons. This new points_count column was used as the target variable with census data and walkability being the features. The pipeline architecture for this can be found in figure 10. This resulted in an R squared value near 0 which is much worse than expected. To follow up a RandomizedSearchCV was used in case the hyperparameters of max depth and min sample split would create a big improvement. No significant improvement was found. However, an interesting find was that the feature that held the most importance was intersection density from the walkability data.

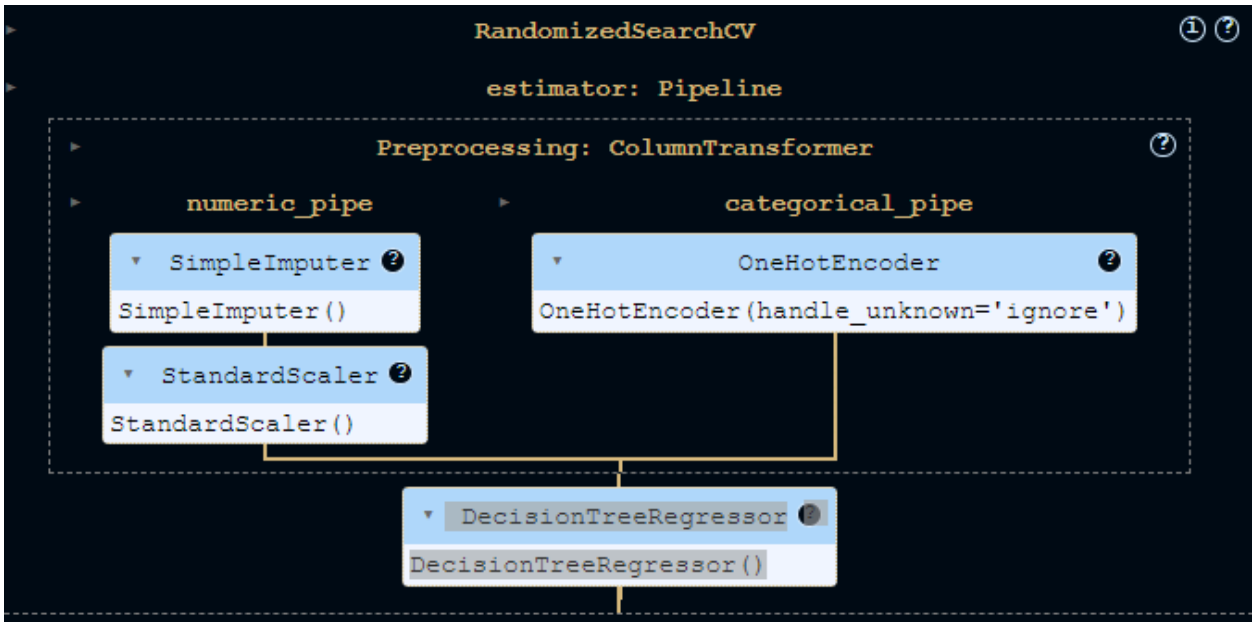


Figure 10 Decision Tree Structure

Unfortunately, our findings did not align with our expected outcome. Crash sites did not correlate with census tracts or walkability index, but we did find other interesting outcomes. Crash sites were centered around downtown city centers as well as major freeways that connected the city. This makes sense that DUIs tend to be closer to nightlife centers and businesses that sell alcohol. San Diego County is large and spreads wide, being able to drive as long as an hour from edge to edge. Most drivers will drive to and from nightlife centers toward the more populated residential areas. In fact, residential areas had much less crash sites, revealing a negative correlation between sites and residential areas. Important to note is the factor of police activity, as hit and runs would not be recorded.

Although our expected outcome was not met, our model did highlight the

importance of intersection density. The more traffic flow that passes through an intersection the higher the chances of a crash site. This makes sense that an increased volume of traffic flow leads to more collisions regardless of location.

There could be some value in developing these models further from the patterns we saw. The points_count was not able to include all the crash site data and further data could paint a better picture for these relationships. Furthermore, the urban density image data was not able to be quantified and incorporated into our model for this report which is suspected to help better locate the crash sites.

LIMITATIONS

DUIs, although accurate, are dependent on actual traffic stops where officers record data. Even if authorities are called, hit and runs cannot determine whether or not the driver was under the influence, when it happened, or other important information. This may be prevalent in less frequented roads or areas with lower police activity. Another limitation to the data provided by TIMS is that the more recent years of 2022-2023 do not have finalized data. It is provisional and subject to change. We thus shifted our focus to the years of 2017-2021.

For our census data we used American Community Surveys which relies on samples rather than the entire population. Due to this each census variable has a margin of error associated with it. There may also be different response rates for some demographic groups who may be less willing to participate in the ACS. In the past five years the response rate across the country has decreased so this may lead to less accurate data with a larger margin of error. While there is a rate limit on the Census API, it is only limited to 500 per ip address per day, so this should not be a problem for our research.

The raster data for urban density is only for 2018 but our assumption is that this kind of data would not change in the short term in our small range of 5 years.Although this is a very high spec satellite image at 10m it is still a satellite image where some of the information can be lost when zoomed in.

APPENDIX

DUI Data

https://tims.berkeley.edu/tools/dui/

https://www.chp.ca.gov/programs-services/services-information/switrs-internet-statewide-integrated-traffic-records-system

Census Data API

GHSL: Global settlement characteristics on Google Earth Engine (https://developers.google.com/earth-engine/datasets/catalog/JRC_GHSL_P2023A_GHS_BUILT_C#description)

GitHub Link: https://github.com/ecbraga/Stat-596-DUI-Census-and-Density-Project

About the Authors

Eric Braga is a master's student in the Masters of Science Big Data Analytics program at San Diego State University set to graduate in spring of 2025. He previously graduated with a bachelor's degree in Economics from California State University Northridge. He has worked for various businesses as a business analyst including a chocolate company in the Los Angeles area. He has an interest in the fields of urban planning, transportation, machine learning engineering, computer vision, and deep learning integration. He is working with The Smart Transportation Analytics Research (STAR) Lab. His email is provided below.

Email: ebraga890@gmail.com

Project Tasks: Raster and Modeling

Farah Farah is a masters student in the Masters of Science Statistics program at San Diego State University and is set to graduate in 2025. He graduated from San Diego State University in spring 2023 with a Bachelors in History. He is interested in using Machine

Learning in sports research, statistical analysis, and statistical modeling. He hopes to pursue a career in Data Science after graduation.

Email: ffarah9640@sdsu.edu

Project Tasks: Census Data

Peter Lenz is a masters student in the Masters of Science Applied Mathematics program and San Diego State University and is set to graduate in spring of 2025. He graduated with a bachelors of science in Physics from Cal Poly San Luis Obipso. He has done research on epidemic modeling and molecular dynamics simulations. He has interests in urban planning, epidemiology, machine learning and mathematical modeling.

Email: plenz0730@sdsu.edu

Project Tasks: Walkability Data and Merging

Esteban Lopez is an undergraduate student in Applied Mathematics with an emphasis in Computational Science at San Diego State University graduating in the fall of 2024. He takes an interest in the fields of climatology, sustainability, and mathematical modeling. With love for the outdoors and passion for the environment, his goal is to transition into an environmental science career after graduation.

Email: elopez3974@sdsu.edu

Project tasks: DUI Data