Analysis of the Impact of Urban Development in Los Angeles

Kijan Daniel[1] [0000-0001-6695-7911] and Avery Stubbings[2] [0000-0003-0927-5775]

[1] University of Tennessee, Knoxville TN 37920, USA
[2] Illinois Institute of Technology, Chicago IL 60616, USA
kdanie32@vols.utk.edu
astubbings@hawk.iit.edu

Abstract. The use of building and weather data is essential to the understanding of socioeconomics and urban morphology. By processing and analyzing the ORNL AutoBEM data for the Greater Los Angeles Area [1] one can see a clear correlation between different factors of urbanization, their effects on the environment, and their impacts on socioeconomics. We took a four-pronged approach to dissect the given data. First, we gathered location data and area on each individual building and tile by combining different sources of data such as a building area file. Second, we analyzed the combined data to create correlations between different aspects of urban development. Next, we used NASA's climate data and US Census data to investigate potential correlations between the density of buildings, land temperature, and county's demographics. Finally, we visualized the data to illustrate any correlations between the different characteristics of the buildings, weather, and demographics. The discovered relationships can be implemented in future urban development projects to improve the lives of residents.

Keywords: Socioeconomics, Urban Morphology, and Meteorology

1 Introduction:

Most urban areas grow much larger than their designers had originally imagined. This leads to issues arising throughout communities such as overpopulation, poor traffic patterns, and too much energy consumption. These issues can all be combated by first knowing where and why the issues are arising. One solution to identify these problematic areas is through big data to analyze urban development and morphology. This paper's purpose is to show how one could analyze urban data through the analysis of building, energy, and weather data using the Los Angeles area as an example.

1.1 A Look at the Data

The Greater Los Angeles area is in the southwest United States and is part of a warm climate region. They have a high population density relative to other cities in the USA because in recent years it has become a melting pot of cultures. The Los Angeles area has five different counties with a large variance in housing prices, demographics, and

temperature patterns. Even more telling is that there are over forty cities in the greater Los Angeles area that also exhibit these variances. The data sets mentioned below provide information about these areas.

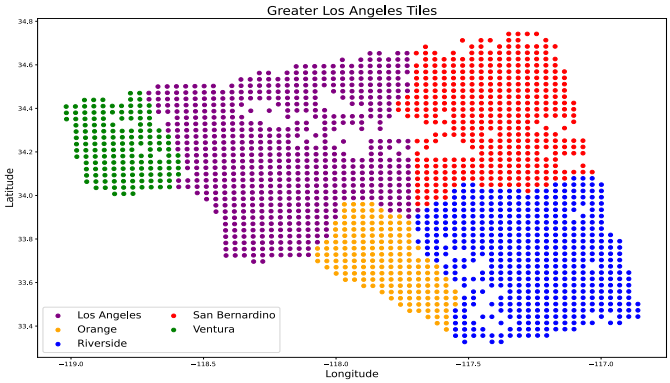

Fig. 1. The data used for this paper.

Tiles [1]. One of the data files that we used was the Los Angeles tile data. This data set contained 1,761 files each representing a tile of 3.2 by 3.2-kilometers. Each tile file contains building ID, density, height, and aerodynamic properties for each building.

Buildings [1]. This data was one large Comma Separated Values (CSV) file with over 4,200,290 buildings. Each building was identified with the same unique building ID that was also present in the tile data. The building data contains the square footage of each building.

Locations [1]. This data was a CSV file containing Latitude and Longitude for each tile in Los Angeles.

Building Archetypes [1]. This data was a CSV file with different building archetypes for each county in Los Angeles. Each building archetype was defined with a building type, average height, average area, and building standard.

NASA Weather [3]. This data was a CSV file that contained weather reports for the areas in and around Los Angeles for the first three months of 2021. This data was used to identify warmer and cooler areas throughout greater Los Angeles

USA Census [2]. The 2021 United States Census data included household income, housing cost, and other demographics for each county.

Greenspace [6]. Developed and undeveloped land is provided for each county.

2 Methodology:

The data sets were combined, analyzed, and graphed through the process below.

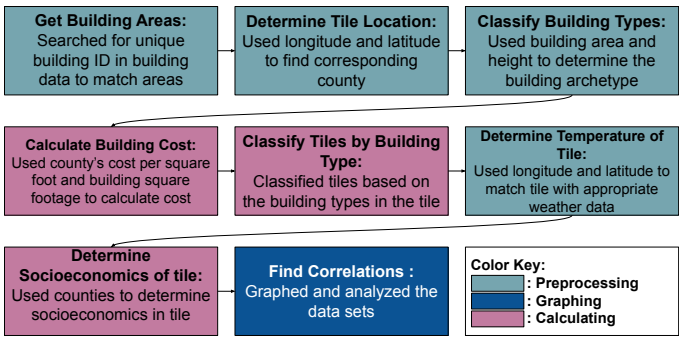

Fig. 2. A flowchart showing our data analysis procedure.

3 Challenges:

3.1 Challenge 1: Urban Density in Relation to Construction Quality and Value

Description: For challenge one, we analyzed how the number of buildings in a specific tile impacts the building quality and price in that specific tile [1].

Method: The square footage of a building can be used to calculate its value when multiplying it by the appropriate price per square foot constant. This concept was the basis of our method. We matched each building in the tile files with its corresponding building in the building data CSV file [1]. This allowed us to add each building's area to the tile files. Next, we determined what county each tile was in by calculating the distance between the tile center and the center of every city with over 100,000 people in the Greater Los Angeles Area. Using the closest city, we determined what county the tile was in. Knowing the area and county allowed us to calculate each building value using the price per square footage constants for Los Angeles, Orange, Riverside, San Bernardino, and Ventura Counties, which are 604, 570, 319, 323, 491 [4] dollars respectively. To determine construction quality, we looked at the building standard of each building [1]. We classified each building's type by comparing its area and height to the building archetypes within a 20-meter height variance. Once all the buildings were classified in their archetypes the average standard year was calculated. We created graphs to show the relationships between the number of buildings in a tile and the average value and standard.

Results:
In general, there is a negative correlation between the number of buildings and building value. The building value was found to decrease slightly as the number of buildings in a tile increases for all counties accept San Bernardino.

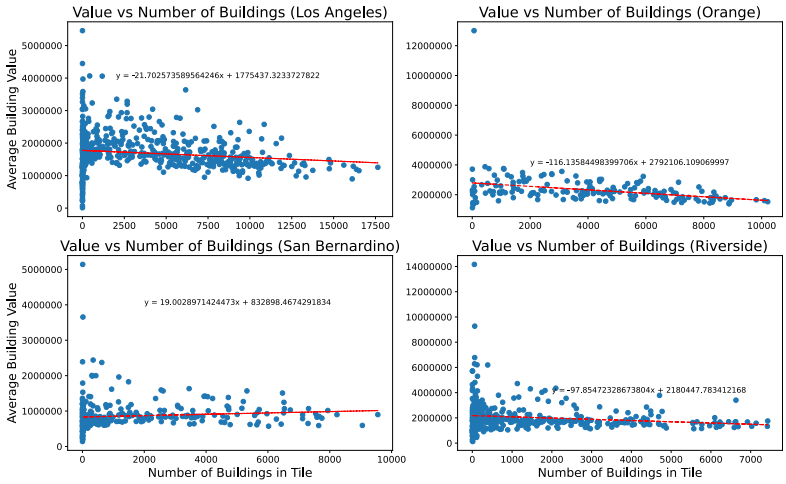

Fig. 3. Average building price of each tile for the counties in Greater Los Angeles.

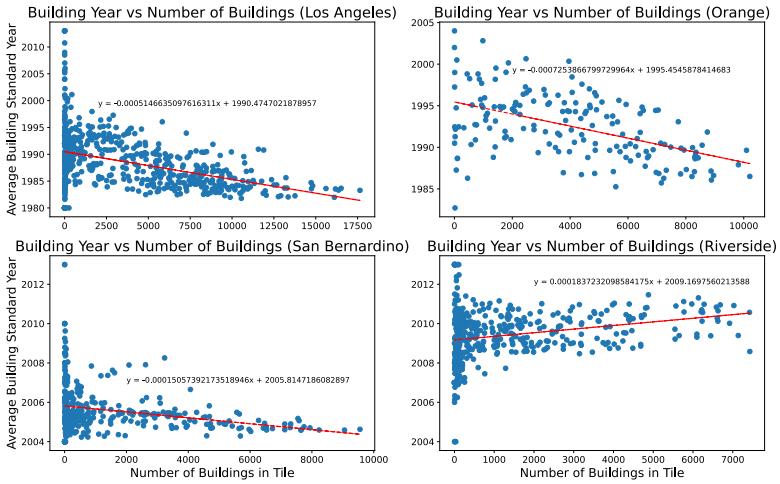

Fig. 4. Average building standard year of each tile for the counties in Greatest Los Angeles.

Construction quality can be measured by a building's building standard year. A new standard means the construction quality is better. Therefore, there is also a negative correlation between the number of buildings and construction quality. The building standard year was found to decrease when the number of buildings increased for all counties accept Riverside [1], which has had housing developments over recent years [5].

Discussion: A possible explanation for the negative correlation between building value and the number of buildings in an area is that people prefer to live in less congested areas. People like to have spaces to themselves. This idea is depicted by how there are

substantially more tiles with fewer buildings than there are with lots of buildings. The negative correlation between construction quality and the number of buildings can be explained by developments in transportation technology. With improvements in transportation, buildings (with newer standards) can be spread further apart but still be reachable for the public.

3.2 Challenge 2: Building Type Distribution and its Relation to Building Value, Year, and Area

Description: Challenge 2 asks us to see how buildings are distributed by type in each tile and see if building type is correlated to building value, year, and area [1].

Method: To determine the distribution of building types in each tile we defined categories to place each tile in. We defined dominant categories as categories that have a building type which makes up over 65% of buildings in the tile. Therefore, a retail/service, business, or residential tile has more than 65% of its respective buildings. Shared categories are where two building types make up 80% of the buildings, and one building type does not make up over 65%. The three shared categories we created were "Retail + Residential", "Retail + Business", and "Business + Residential". If the tile did not match any of the above categories or matched multiple shared categories, it was defined as "Mixed" [1]. By classifying each tile, we can illustrate the distribution of buildings in Greater Los Angeles tiles. Possible correlations that building type may have with building value, standard, and area can be analyzed [1]. We developed these thresholds.

Results.

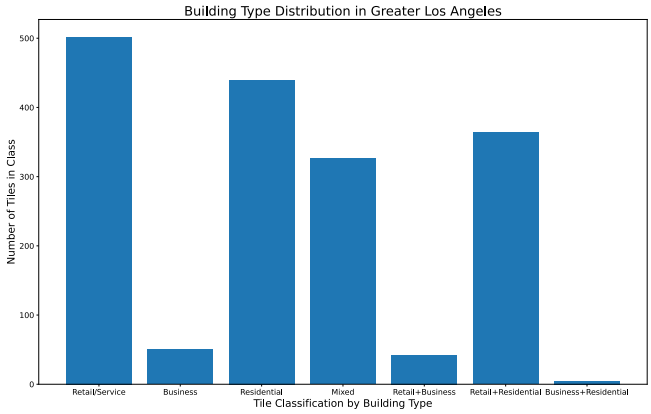

Fig. 5. The distribution of building types by tile in Greater Los Angeles.

Figure 5 clearly depicts that Greater Los Angeles is primary composed of Retail/Service and Residential Buildings.

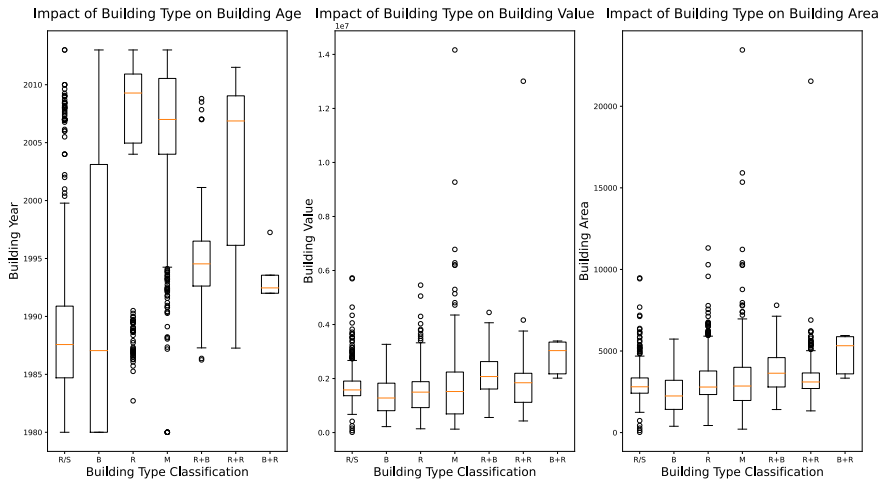

Fig. 6. Building type's relation with building year, value, and area in each tile.

For building area, the residential building tiles have the largest range of all the buildings with a similar median as the rest [1]. They also have the largest outliers amongst the three major categories. This leads us to believe that the residential buildings are overall the largest of the three major groupings. The retail and service tiles have one of the smallest ranges, yet it still contains some large outliers which makes sense because there are quite a few large shopping complexes in the area.

The effect of building type on building value is very similar to the area. We observe the same trends on building value as we did with building area, with the one exception being that there are fewer and not as drastic outliers in the residential tiles. The highest value tiles are the residential/business tiles with them having a significantly larger median than all the other tile types. This correlates with the residential/business size as well, with the median being almost as high as the other categories upper bounds. All the other categories follow suit with them being very similar. Residential tile building standards are much newer than other building category tiles. The last notable difference lies with the mixed tiles. With these tiles we see they have the largest outliers by far when compared to all other building types [1].

Discussion. The data shows that with the building size of the residential building tiles having so many high outliers and such a wide range with a similar median as the other three major categories, that this is the largest building type. On the other hand, when it comes to the building value the residential tiles are on par with the other sets. This leads to the conclusion that there is a big gap between the rich and poor in terms of housing size and quality. On the other hand, the retail/service and business tiles do not have as large of an outlier gap illustrating a smaller economic divide for these building types. The one tile type that has the largest area of the subgroups is the business/residential. This building type also has the highest average value, indicating that these tiles are in some of the nicest areas in the city. This leads us to believe that these are the luxury high-rises with nice office spaces and million-dollar penthouses.

3.3 Challenge 3: Temperature in relation to building density

Description: For challenge 3, the question asks us to use temperature data [3] from a source of our choosing to see if there is any relation between the density of buildings in each tile and the temperature in said tile.

Method: The data that we chose to use was the temperature data from NASA's Power project [3]. This data was great because it had exact latitude and longitude for each of the data points, which lets us pinpoint precisely each location. The first step to organizing this data with each of the tiles is finding which tile is closest to which temperature data point [1] [3]. To do this we used the distance function and rated each tile by their distance from each temperature point. Looking at these rankings we were able to tell the best temperature to use for each tile. After assigning each tile with a temperature we graphed the tiles according to their temperature and density as seen in the dot plot. Each dot (i.e. tile) was also colored to fit its respective county. Then we analyzed the dot plots to see if there was any positive correlation between the temperature and density.

Results:

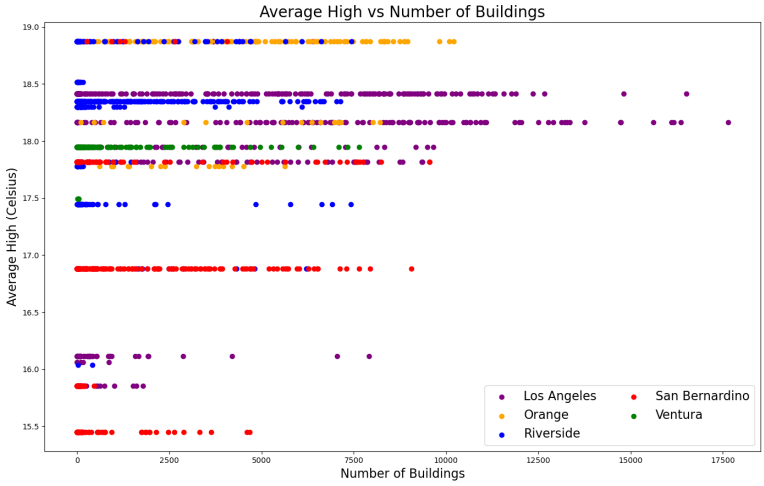

Fig. 7. Average high temperature from January to March in relation to the number of buildings for each tile in Greater Los Angeles.

In Figure 7, the average high dot plot, there is a correlation observed between the building density and temperature [1] [3]. The lower temperature tiles tend to have less buildings while the very dense tiles tend to have a higher temperature. This can also be seen because there are few to no tiles with a density of over 10,000 under 18.5 degrees Celsius. Within the same county the temperatures clearly vary with relation to the building density.

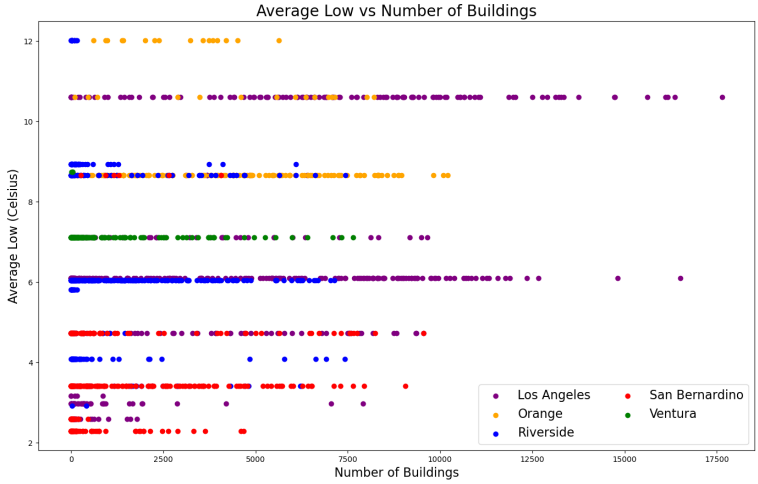

Fig. 8. Average low temperature from January to March in relation to the number of buildings for each tile in Greater Los Angeles.

In Figure 8, the average low dot plot, the data is a bit more muddled with some counties like Los Angeles not having a clear correlation between the density and the temperature aside from a few outliers but looking at the box plot Los Angeles tends to be a hotter county (Figure 9) so taking the lows as the main source of data is flawed when analyzing the true heat experienced by the population. This is because the low data will be rarer, with the majority being outliers to the temperature anyways. However, when looking at the San Bernardino County there is a very clear correlation between the density and temperature with the higher temperatures having a significant number of dense tiles and the lower temperatures possessing none of these dense tiles [1] [3].

Discussion: One thing that can be taken into question about the data is that some counties are bound to be hotter on average because they are closer to the equator than other counties. The urban heat island phenomenon can be seen clearly occurring here. For example, it can still be seen that even within one county, take Los Angeles' high temperature data as an example, the trend still holds true because the less dense tiles are more likely to land in lower temperatures as the trend shows. The other issue with the data is that there were only a few nodes from which the weather data was taken. One of these nodes being in San Bernardino, and therefore for the low temperature data set the correlation is so strong between the density and temperature. However, this leads to some counties having somewhat inaccurate data because during certain weather patterns only very specific areas could be targeted. Overall, taking these ideas into account there is still a clear correlation in a significant part of the data to lead us to the conclusion that the building density does influence the area's temperature.

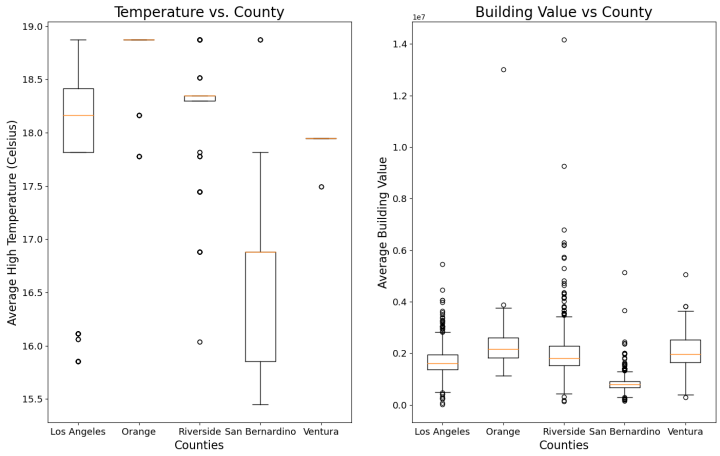

Fig. 9. This figure shows how temperature and value differs by county.

3.4 Challenge 4: How does the building environment and temperature correlate with socioeconomics and demographic data

Description. For challenge four, we are asked to see if there is any correlation with the built environment and the socioeconomic/demographic data that we chose.

Method. For this question we first had to find socioeconomic and demographic data. The data source that we ended up using was the US census data [2] because it was very recent, and we could divide it up by county which made the data processing much easier. Next, we found the individual price of each building by multiplying the square footage with the average square foot price in the county that the building resides in [1]. Then we filtered in the average home cost of each county from the census data. In addition to this we also filtered in the demographic data, the average rent cost, and the average household income for each county. We wanted to see the correlation between the average cost of a home in each county with our predicted price of each building. Once we graphed the box plot for each county's building price, we quickly saw a correlation between this and the average home and income of each county [1] [2]. Finally, we wanted to see if there was a correlation between the temperature and average building value. To do this we graphed a box plot of the relation between temperature and county. Using this we saw if there was any correlation between the value graph and the temperature graph to determine if there is a relation between the two [1] [2].

Results.

Looking at the building value vs county graph we can see that Orange County has the highest building value with the median and upper bound both being higher than the rest of the data points. We can also see that San Bernardino has the lowest building value [2] with a drastically lower mean than the other data points, and because there the upper and lower bounds are so close to the median there seems to be little variation in the main part of the data set. There are quite a few high outliers which shows that there

may be some disproportionately affluent areas in the San Bernardino area. Riverside and Los Angeles counties are very similar with both counties riding in the middle of the pack. The only significant observation to make here is the Riverside also has some very large outliers which could be due to some disproportionately affluent areas. In the temperature vs county graphs Orange County is the hottest county by far with very little variation in the temperature. The coldest county is San Bernardino, but the temperature fluctuates a lot as seen by the large range when compared to the other counties [1] [2]. Riverside and Los Angeles are very similar with the only notable difference being that Los Angeles has a larger range but not by too much [1] [2]. In addition, to this there is a larger population of Hispanic persons living in cooler climates in comparison to their population percent in hotter climates such as Orange County. On the other hand, there is a larger population of white persons living in warmer climates such as Orange County than in cooler climates such as San Bernardino [2].

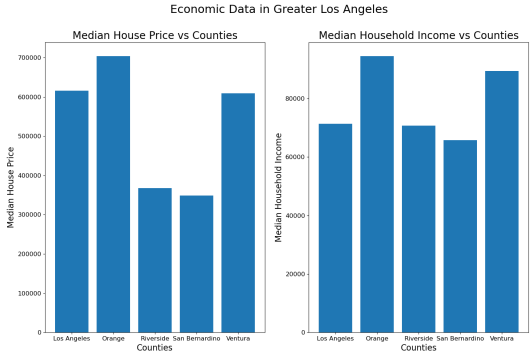

Fig. 10. This figure reflects the data collected from the 2021 US Census.

Discussion. Looking at Figure 9 it can be seen to be very accurate because it falls right into line with what the US census data shows [2]. This is because Orange County is the richest county with the mean home value according to the census data is 703,800 dollars and San Bernardino being the poorest with their mean around 348,500 dollars [2]. The only misrepresentation in the data is Riverside County with it having a median home value of slightly higher than San Bernardino yet on the box plot graph it is in line with Los Angeles. This could be because there are large shopping areas or large business sectors in Riverside while San Bernardino is mainly housing. This means that the numbers are skewed for Riverside because it considers these large buildings that are not homes. Another conclusion that can be made here is that San Bernardino is mainly just housing with not too much of a wealthy and large business sector. In the temperature data Orange County has the highest temperatures. This is in direct correlation with them being the richest county. This could be due to higher living costs in concern to cooling or people just move to California to get warmer weather. This relationship continues but is not as defined in Los Angeles and Riverside. This is because Riverside has higher temperatures than San Bernardino but still lower highs than Orange and Los Angeles. Los Angeles gets up there in temperature with its upper bound, but the area is still cooler

than Orange and almost ties Riverside with the median. The US census data [2] shows us that there is a larger population of Hispanic persons living in San Bernardino than in Orange County. This could be because of the cooler climate in San Bernardino but is more likely due to the wealth gap between the two counties.

3.5 Challenge 5: The effect of green space on temperature and demographic

Description. For this challenge we used land space data to determine if greenspace had any effect on the temperature of the county or the demographics

Method. We began by parsing through the land use data [6] to find the total square miles of each county. Then we parsed through the data throwing out all the nongreen space leaving us with only the square miles of green space. Next, we used the total square miles and green space square miles to calculate a percentage of green space for each county. Then we compared these results to the wealth of each county and that counties temperature. Finally, we saw if there was a correlation between the green space and the counties wealth/demographic.

Results.

County	Total Area (mi^2)	Green Area (mi^2)	Percent Green
Orange	947.98	313.22	33.04078145
San Bernardino	20106.24	19277.75	95.87943842
LA	4731.59	2712.53	57.32808633
Ventura	2208.17	1623.05	73.50204015
Riverside	7302.67	6487.45	88.83668576

Fig. 11. Orange County has 33 percent, San Bernardino has 95 percent, Los Angeles has 57 percent, Riverside has 88 percent, and Ventura has 73 percent greenspace [6].

Discussion. There is a clear correlation between the temperature and greenspace because Orange County only has 33 percent green space and has the highest temperatures by far. In addition to this, San Bernardino has the lowest overall temperature and the highest amount of greenspace. There is also a strong correlation between wealth and greenspace. This correlation is that the richer a community is the less green space they have. This can be seen in Orange County with only 33 percent green space while having the richest of all the communities from our data. On the other hand, San Bernardino is the poorest community and has by far the largest green space at 95 percent.

4 Conclusion

This data challenge was primarily focused on the infrastructure, socioeconomics, and urban morphology of Greater Los Angeles. For challenge question one, we saw a negative correlation between the density [1] and building value [4]. We determined that this was due to people not wanting to live in a congested environment. In challenge

question two, we found the building types in each tile. With this information we found that the main building types in tiles were a retail/service buildings and residential buildings [1]. The third question had us find how building density correlated with the given temperature. We found that in most cases there was a clear correlation between building density and temperature. For the fourth question we had to determine if the temperature in an area had any socioeconomic impacts [2] on the area. We concluded that there were socioeconomic impacts because the wealthier population was primary living in the warmer climate of Orange County, while the less earning people lived in cooler areas such as San Bernardino. For the fifth and final question we looked at land data [6] to determine if greenspace influences temperatures.

These conclusions and data sets that we came up with and created can be used in the future to better understand the Los Angeles area. This research can be broken down even further in the future by dividing Greater Los Angeles into cities or neighborhoods, and seeing how urban development, temperature, and socioeconomic differ in a more compact region. In addition, these data sets can potentially be used in city planning to better understand why some areas become more affluent than others. With this understanding city planners may be able to create developments that can be more inclusive to all groups of people.

5 References:

1. New, Joshua R., Bass, Brett, Adams, Mark, Berres, Anne, and Luo, Xuan (2021). “Los Angeles County Archetypes in Weather Research and Forecasting (WRF) Region from ORNL’s AutoBEM [Data set].” Zenodo, doi.org/10.5281/zenodo.4726136, Apr. 28, 2021. [Data]
2. 2021 US Census Bureau QuickFacts, https://www.census.gov/quickfacts/fact/table/venturacountycalifornia,riversidecountycalifornia,orangecountycalifornia,sanbernardinocountycalifornia,losangelescountycalifornia/PST045221, last accessed 2022/7/12
3. These data were obtained from the NASA Langley Research Center (LaRC) POWER Project funded through the NASA Earth Science/Applied Science Program.
4. Realtor.com (Los Angeles, Orange, Riverside, San Bernardino Pages), https://www.realtor.com/realestateandhomes-search/Los-Angeles-County_CA/overview, last accessed 2022/7/12
5. “Riverside Approves Plan For More Than 20,000 New Homes by 2029 – Press Enterprise”. *Pe.Com*, 2022, https://www.pe.com/2021/10/06/riverside-oks-plan-for-more-than-20000-new-homes-by-2029/
6. Barnes, Christopher A., National Land Cover Database Evaluation Visualization & Analysis (NLCD EVA) Tool [abs.], v. Conference Abstracts, p. 1–1, at https://www.asprs.org/wp-content/uploads/2014/11/MTSTC1-21.pdf

"Support for DOI 10.13139/ORNLNCCS/1854856 dataset is provided by the U.S. Department of Energy, project IM3 under Contract DE-AC05-00OR22725. Project IM3 used resources of the Oak Ridge Leadership Computing Facility at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725"