Identifying Activity Centers in the San Diego Region

Kelly Luu k6luu@ucsd.edu
Kelvin Nguyen ktn015@ucsd.edu
Mike Smith mss002@ucsd.edu
Manav Dixit mudixit@ucsd.edu

Liang Tian
Liang.Tian@sandag.org

Abstract

In this study, a exploration of the San Diego region's economic dynamics through the lens of 'Activity Centers,' leveraging clustering techniques in geo-spatial analysis. San Diego contains a vibrant population and serves as a residency for people living and working. Traditional employment centers, focused on concentrated job opportunities, often lack a nuanced regional understanding. In response, 'Activity Centers' are introduced, rooted in San Diego businesses, offering essential employment details and generating a holistic comprehension of the region. Clustered geo-spatial analysis identifies different activity centers categorized by dominant business activity types. These centers serve as an important tools for summarizing statistics and generating data visualizations, showcasing the potential of clustering techniques to augment traditional methods for comprehensive regional insights. The study provides valuable insights for policy makers and planners navigating patterns between business activities and urban development.

Website: https://dixitmanav.github.io/DSC180B_Website/
Code: https://github.com/michaelmss/DSC180A-Activity-Centers

1 Introduction 2
2 Methods 3
3 Results 5
4 Discussion 8
5 Conclusion 10

1 Introduction

The San Diego region houses a population of 1.3 million which creates a dynamic hub for both working and living. Policy makers and regional planners at the San Diego Association of Governments (SANDAG) have utilized diverse data sources to create employment centers that inform future urban and transit planning endeavors. Traditionally, the employment centers have been defined as areas showcasing high concentration of employment opportunities. Employment centers offer valuable information on job opportunities, yet they may lack comprehensive insights. On the other hand, activity centers, based on San Diego businesses, not only supply pertinent employment details but also foster a holistic understanding of the region. Businesses can create an understanding of which specific industry or sector of work plays apart in the region. The broader perspective is beneficial in advancing urban and planning endeavors.

1.1 Definition and Purpose

Activity centers are defined to be locations of high business concentrations. The primary purpose of the study is to create improved clustering of activities to create a holistic understanding of the San Diego region for urban planning purposes.

1.2 Literature Review

The Brookings Metro report, "Mapping America's Activity Centers: Methodology Appendix," by Loh et al. (2022) describe the process in creating activity centers and important metrics that are considered when identifying good activity centers. The study demonstrates a vast array of data sources that are studied such as the Longitudinal Employer-Household Dynamics (LEHD), public library survey, DHS Homeland Infrastructure data, national register of historical places historical site data, and more. Through the combination of the data, the study was able to derive valuable insights. After their data was collected, the study started to determine activity centers. Their method of identifying activity included looking at community, tourism, consumption, institutional, and economic assets. After they have gathered information regarding the identifying features of activity centers, the study starts to use job density and job share, real estate value, and demographics to characterize the activity centers. Through the study, it identifies important characteristics and factors to think about when generating activity centers.

1.3 SANDAG's Previous Methodology

Previously, employment centers were able to be implemented by SANDAG. The employment centers displayed areas of employment grouped by density. Their steps in creating employment centers included finding the local maxima as the initial focal point. After identifying the local maxima, SANDAG would expand the radius of the area until it has reached a 2

mile radius. When the threshold is met, the employment center boundaries are then refined by accounting for the freeways or other barriers such as topography. SANDAG Master Geographic Reference Areas (MGRAs) are selected based on activity weight centroids based on the population and employment metrics. The outcome of the technique has lead to over 70 clusters of employment centers. The different employment centers are used for evaluating travel patterns, employment trends, and residential information. Using their previous methodology, this study builds upon that but adds upon business information. Through activity centers, more patterns and trends are able to be analyzed.

2 Methods

This study employs a geo-spatial analysis approach to identify Activity Centers across San Diego County. These centers are defined as geographical clusters representing areas of concentrated business activity, which are important for understanding regional economic patterns and aiding in urban planning and development. In our methodology, we discuss the two approaches that are developed through the process. Our methodology integrates business location data with census information to generate comprehensive mappings of business activity and extract summary statistics. This section outlines the data sources, processing steps, and analytical techniques utilized in our research.

2.1 Data Sources

Our primary dataset, SANGIS data, comprises the coordinates and categories of 98,163 individual businesses within San Diego County. This granular data allows for precise mapping and analysis of commercial activity at a detailed spatial level. This data contains several pieces of information about each individual business, including their name, category, and precise location, which will allow us to further categorize them in the future.

Additionally, we incorporate US Census data that provides insights into employee commuting patterns and a look of census tracts. These tracts are small geographical units used for demographic surveys, which serve as a basis for aggregating business data and analyzing spatial trends. This data included aggregated information on a tract granularity, which we used to generate summary statistics for each activity center.

Lastly, we included data from the SANDAG Open Data Portal. The open data portal allowed us to access public SANDAG data that was used for their employment center clustering creation. Within the open portal, information about employee patterns displayed which we merged with our data in order to create summary statistics. The Open Data Portal included a large amount of data that was cleaned and analyzed, which saved us a great amount of time in our data processing, including pre-cleaned boundaries for Census Blocks, Tracts, and Block Groups.

2.2 Data Categorization

With the 98,163 different businesses, allows for many different business types. Our data had more than 200 different business types. However, we wanted a more generalize category that would allow us to classify the different activity centers. Upon looking at all the different categories, we were able to group the different business types and find the top 10 maximum occurrence of business types. Through looking at this statistic, we generated 4 generalize categories in which we would fit the business types into.

The categories are: professional and financial, retail and services, construction and manufacturing, and health and welfare.

Professional and financial includes professions and services related to finance, accounting, and professional consultancy. Examples of such are accountants, attorneys, insurance brokers, and insurance companies. Retail and services encompasses a range of retail businesses and services provided directly to consumers. Examples of these include cafes, grocery stores, auto repair shops, hair salons, clothing stores, and restaurants. Construction and manufacturing involves businesses engaged in construction, manufacturing, and heavy equipment. Examples are machine shops, lumber retail, farming and heavy equipment retail. Lastly, health and welfare which include anything related to well being and providing for others. Examples include fitness centers, medical laboratories, hospitals, clinics, and spas.

These categories were then merged onto the SANGIS Business Sites data, which allowed for grouped analysis.

2.3 Data Processing

For our hex-bin approach, the process was greatly simplified by using a library called h3pandas, which allowed for the generation of hex-bins based on the granular business data. The hex-bins then were modified to contain counts of the amount of businesses inside them for each category. This hex-bin approach enables us to quantify business density across the county efficiently, facilitating the identification of areas with high concentrations of commercial activity. Each hex-bin is assigned a count representing the number of businesses it contains, with consideration for weighting these counts based on business type or relevant travel data to enhance the accuracy of activity representation.

In our census tract approach, the first step consists of aggregating business-level point data into clusters based on their geographic features (e.g. X and Y coordinates). After aggregating our business level data with cluster labels, we use these cluster labels to find the largest number of a cluster (e.g, cluster 6/7, etc.) within a certain tract to define our final number of clusters, assigning each tract a cluster based on the businesses inside of it. This allows us to have a final general shape that we can use to calculate our activity centers with the tract method.

2.4 Identifying Activity Centers

In the hex-bin approach, to determine activity centers, we employ a two-step clustering process. First, we identify the 70 hex-bins with the highest business density, ensuring these are spatially distributed to avoid over representation of central areas. These bins serve as preliminary centers around which activity centers are formed.

Subsequently, we apply distance clustering to assign the remaining bins to the nearest of the 70 selected centers, based on proximity. This clustering technique allows us to delineate geo-spatial boundaries around each center, effectively capturing areas of concentrated activity. The analysis specifically accounts for the spatial distribution to prevent the formation of disproportionately large clusters in less urbanized regions, addressing the challenge of representing diverse urban and rural dynamics across the county.

In the census tract approach, similar to the hex-bin approach, we identify the 70 largest clusters based on number of businesses closest to a centroid (e.g. k-means centroid). These centers serve, similar to the above process, as a foundation for the dissolving of geometries to create our final tract-like shapes. Dissolving geometries allows us to group neighboring tracts with the same cluster to create an explainable activity center. We dissolve the geometries and calculate the various types of centers based on activity. Given the shape of the geographic boundaries, we are able to generate summary statistics that can give us a better idea of what goes in a geographic shape based on what type of activity it is.

3 Results

The outcome of our methodology is the identification of approximately 70 Activity Centers across San Diego County, categorized by the dominant type of business activity. These centers enable the aggregation of summary statistics and the generation of detailed data visualizations, assisting in the exploration of business-employee interactions within the region. Our approach aligns with the objectives of the San Diego Association of Governments (SANDAG), providing a foundational tool for urban planning and the development of public-facing documents that illustrate the economic landscape of San Diego County.

3.1 Hex-Bins

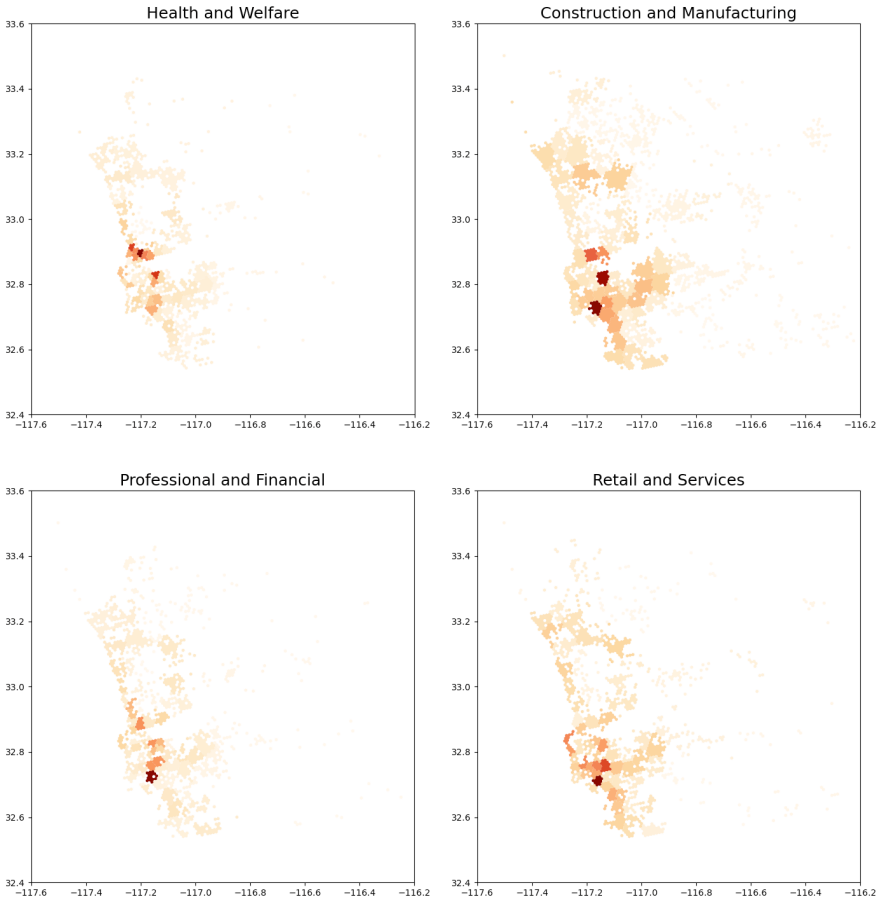

Figure 1: Hex-Bin Clusters based on the four different categories

3.2 Census Tract

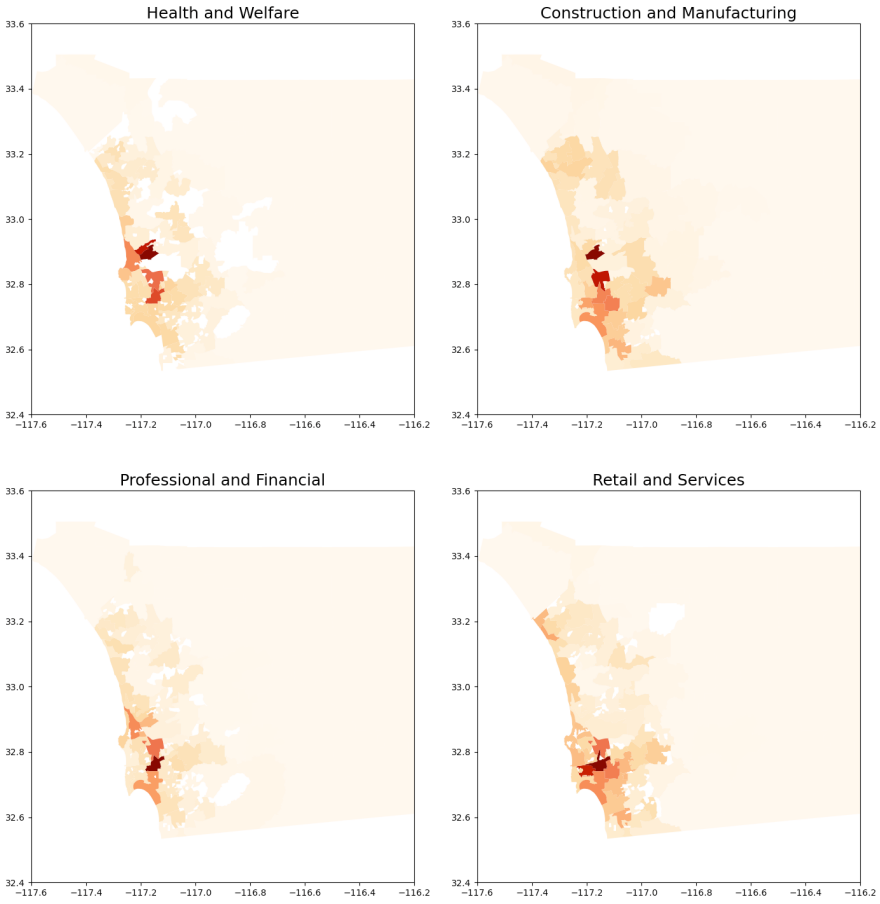

Figure 2: Census Tract Clusters based on the four different categories

3.3 Summary Statistic

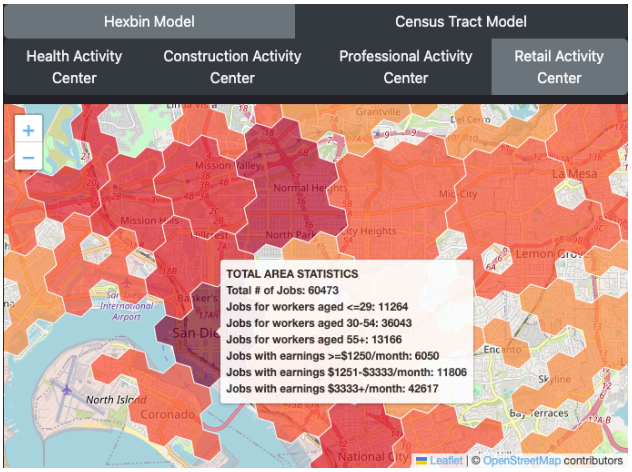

Figure 3: Summary Statistics for Hex-Bin Approach

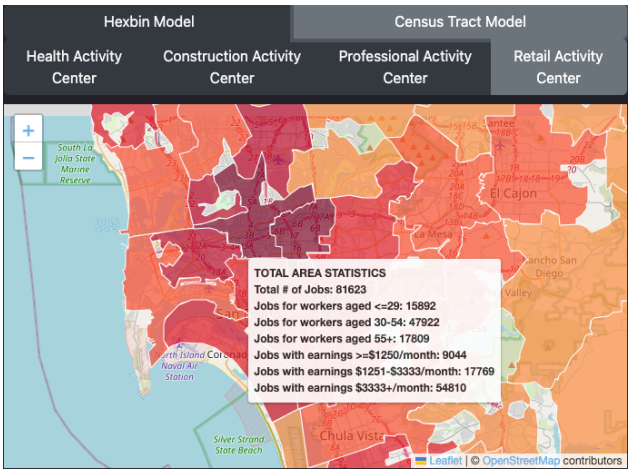

Figure 4: Summary Statistics for Census Tract Approach

4 Discussion

4.1 Interpretation of Results

The two approaches, hex-bin and census tract offer similar concentration of clusters across San Diego. To compare the different approaches, if we focus on the health and welfare activity center, the most concentrated area would be the mid-western part of San Diego. The area includes La Jolla, Torry Pines, and Sorrento Valley which are notably big contributes in location of biotech industries and locations of major hospitals. Although the two approaches are geo-spatially similar, both bring different disadvantages and limitations.

The hex-bin approach demonstrates fewer biases towards preexisting boundaries allowing for a more nuanced understanding of spatial patterns. By grouping the data into uniform hexagonal bins, the method allows for clustering to be discovered without any limitations of boundaries and solely based on the data. However, the lack of alignment towards boundaries can create unclear analysis and could hinder comparison between different hex-bins.

The census tract approach provides boundaries that align with the preexisting boundaries. These boundaries are aligned with freeways and physical boundaries that might not be found with the dataset. Since the clusters are derived from census tract, the clusters created can be compared with the census tract and blocks which contain employment data and could be merged with preexisting datasets. The clusters are able to be contextualized with socioeconomic details that are derived from the employment dataset. However, the census tract may result in unnecessarily large geometries because of the lack of data of businesses.

Upon looking at the summary statistics, the two clustering were chosen since they had one of the greater business densities. When comparing the census tract approach and the hex-bin approach, it is noted that the census tract for the area has a higher amount of jobs. The greater number is due to the larger surface area that the census tract block has. The hex-bin approach would have a smaller number because of the smaller surface area. Both of the methods provide similar summary statistics, however, it is a matter of where the boundaries are placed.

While both of the approaches offer insight and contribute to the purpose of the report, the hex-bins prioritizes unbiased representation while the census tract approach allows for appropriate alignment with administrative boundaries.

4.2 Limitations and Challenges

One notable limitation of the approaches is its failure to account for the varying weights of business sites based on their employee count. As a result, the model provides an equal representation of all businesses within the study area, regardless of their size or significance. By overlooking the differential contributions of businesses based on their employee count or economic output, the approaches may obscure important insights into the distribution and concentration of economic activity. Incorporating such weighting schemes would enable researchers and policymakers to obtain a more accurate representation of business activity.

Another limitation is from the underlying that used for the methods. The dataset provided by the San Diego Geographic Information Source (SANGIS) suffers from infrequent updates. The model can not be updated unless the the data is redownloaded from SANDGIS and inserted into the model. A more dynamic data would allow the model to become more robust and updated. By leveraging more frequently updated datasets and implementing systematic data validation procedures, we can ensure the accuracy of our analytical framework.

4.3 Future Research

To accurately represent the density of the businesses, incorporating weight into our methods can reflect the businesses relative importance. By assigning weight to specific business types or individual businesses, we can ensure a more accurate depiction of their spatial distribution. With our current uniform weight, we are showing that a small business with employment of 5 is equal to big corporate offices with 500 people. The approach of adding weight would create a more accurate depiction of the clusters. In addition, the weights would enhance interpretation of visualization and summary statistics to provide insight.

In addition to adding weight to our models, finding detailed, comprehensive regional insights are equally as valuable. The model can only be as good as the data source we use. While existing datasets may provide valuable information, they may not offer the level of detail necessary to accurately represent the area's economic dynamics. By obtaining more data sources that are more granular, we would be able to not only merge on census tract, but also the point in where we can derive summary statistics for the hex-bin approach as well. By leveraging more granular data sources, we can gain a deeper understanding of the connection between business activity, urban development, and socioeconomic trends, laying the groundwork for more effective strategies.

5 Conclusion

In conclusion, our research has uncovered a total of 70 clusters representing various categories of activity centers across the San Diego region. Through analysis, we have identified distinct geographical areas characterized by concentrated business activity. These activity centers offer valuable insights into the economic dynamics of the region, providing policymakers and planners with a comprehensive understanding of the distribution of businesses within San Diego.

Furthermore, we have derived summary statistics by aggregating business data towards census tract boundaries. By leveraging census tract data, we have gained valuable insights into the spatial distribution of businesses and their interactions with local communities. We were able to align the census tract data and merge the data with the employment center to compare and contrast the two different clustering.

Throughout our research, we have also gained a understanding of the benefits and drawbacks of aggregating data by hex-bin or census tract. Hex-bins offers flexibility whereas the census tract allow for boundaries that are limited by natural boundaries. In addition, hex-bins are unbiased whereas the census tract can be combined with other census data to create meaningful summary statistics.

In summary, our research demonstrates the importance of leveraging clustering techniques and spatial analysis tools to gain insights into regional economic dynamics. By carefully considering the benefits and drawbacks of different aggregation approaches, we can develop more robust analytical frameworks for urban planning and economic development goals in San Diego and beyond.

References

Loh, Tracy Hadden, DW Rowlands, Adie Tomer, Joseph Kane, , and Jennifer Vey. 2022. "MAPPING AMERICA'S ACTIVITY CENTERS:Methodology Appendix." [Link]

"San Diego Association of Governments Open Data Portal." [Link]