# *Spatial analysis of bicycling ridership patterns from bias-corrected crowdsourced data*

June 2025

A Research Report from the Pacific Southwest Region University Transportation Center

Avipsa Roy, University of California Irvine

Pacific Southwest Region UTC
University Transportation Center

UC Irvine

## TECHNICAL REPORT DOCUMENTATION PAGE

| | | |
|---|---|---|
| 1. Report No.<br>PSR-22-24-TO 069 | 2. Government Accession No.<br>N/A | 3. Recipient's Catalog No.<br>N/A |

| | |
|---|---|
| 4. Title and Subtitle<br>Spatial analysis of bicycling ridership patterns from bias corrected crowdsourced data | 5. Report Date<br>06/03/2025 |
| | 6. Performing Organization Code<br>N/A |
| 7. Author(s)<br>Avipsa Roy, 0000-0002-9110-4643;<br>Ghangshin Lee | 8. Performing Organization Report No.PSR-22-24 TO 069 |
| 9. Performing Organization Name and Address<br>METRANS Transportation Center<br>University of Southern California<br>University Park Campus, RGL 216<br>Los Angeles, CA 90089-0626 | 10. Work Unit No.<br>N/A |
| | 11. Contract or Grant No.<br>USDOT Grant 69A3551747109<br>Caltrans 65A0674TO065 |
| 12. Sponsoring Agency Name and Address<br>U.S. Department of Transportation<br>Office of the Assistant Secretary for Research and Technology<br>1200 New Jersey Avenue, SE, Washington, DC 20590 | 13. Type of Report and Period Covered<br>Final report |
| | 14. Sponsoring Agency Code<br>USDOT OST-R |

15. Supplementary Notes:
https://doi.org/10.25554/ba8d-4642

16. Abstract

This study leverages big data from the Strava Metro app, integrated with official count data from the Orange County Transportation Authority, to analyze spatial patterns of bicycling ridership in Orange County, California. By applying bias correction techniques to crowdsourced data and incorporating land use and socioeconomic covariates, the study generate a comprehensive map of ridership volumes across the region. The study's analysis reveals significant spatial autocorrelation in cycling activity, with distinct patterns between coastal and inland areas. Coastal regions exhibit strong High-High clusters, indicating concentrated cycling activity, while inland areas show a more varied pattern with Low-High clusters and isolated High-High pockets. These findings demonstrate the potential of bias-corrected crowdsourced data to inform targeted infrastructure planning in both urban and suburban contexts. By identifying areas of high cycling demand and potential growth, this methodology provides valuable insights for policymakers and urban planners to enhance cycling infrastructure and promote sustainable transportation in diverse geographic settings, from coastal cities to inland communities.

| | |
|---|---|
| 17. Key Words<br>*Strava, Bicycling, Bias-corrected, Orange County* | 18. Distribution Statement<br>No restrictions. |

| | | | |
|---|---|---|---|
| 19. Security Classif. (of this report)<br>Unclassified | 20. Security Classif. (of this page)<br>Unclassified | 21. No. of Pages<br>41 | 22. Price<br>N/A |

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

## About the Pacific Southwest Region University Transportation Center

The Pacific Southwest Region University Transportation Center (UTC) is the Region 9 University Transportation Center funded under the US Department of Transportation's University Transportation Centers Program. Established in 2016, the Pacific Southwest Region UTC (PSR) is led by the University of Southern California and includes seven partners: Long Beach State University; University of California, Davis; University of California, Irvine; University of California, Los Angeles; University of Hawaii; Northern Arizona University; Pima Community College.

The Pacific Southwest Region UTC conducts an integrated, multidisciplinary program of research, education and technology transfer aimed at improving the mobility of people and goods throughout the region. The study's program is organized around four themes: 1) technology to address transportation problems and improve mobility; 2) improving mobility for vulnerable populations; 3) Improving resilience and protecting the environment; and 4) managing mobility in high growth areas.

## U.S. Department of Transportation (USDOT) Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

## California Department of Transportation (CALTRANS) Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the United States Department of Transportation's University Transportation Centers program, in the interest of information exchange. The U.S. Government and the State of California assumes no liability for the contents or use thereof. Nor does the content necessarily reflect the official views or policies of the U.S. Government and the State of California. This report does not constitute a standard, specification, or regulation. This report does not constitute an endorsement by the California Department of Transportation (Caltrans) of any product described herein.

## Disclosure

Principal Investigator, Co-Principal Investigators, others, conducted this research titled, "Spatial Analysis of bicycling ridership patterns from bias-corrected crowdsourced data" at the Department of Urban Planning and Public Policy, School of Social Ecology, University of California Irvine. The research took place from [start date] to [end date] and was funded by a grant from the CalTrans ] in the amount of $75,000 . The research was conducted as part of the Pacific Southwest Region University Transportation Center research program.

## Acknowledgements

*The study was conducted with support from CalTrans by a generous grant.*

## Abstract

*This study leverages big data from the Strava Metro app, integrated with official count data from the Orange County Transportation Authority, to analyze spatial patterns of bicycling ridership in Orange County, California. By applying bias correction techniques to crowdsourced data and incorporating land use and socioeconomic covariates, we generate a comprehensive map of ridership volumes across the region. The study's analysis reveals significant spatial autocorrelation in cycling activity, with distinct patterns between coastal and inland areas. Coastal regions exhibit strong High-High clusters, indicating concentrated cycling activity, while inland areas show a more varied pattern with Low-High clusters and isolated High-High pockets. These findings demonstrate the potential of bias-corrected crowdsourced data to inform targeted infrastructure planning in both urban and suburban contexts. By identifying areas of high cycling demand and potential growth, this methodology provides valuable insights for policymakers and urban planners to enhance cycling infrastructure and promote sustainable transportation in diverse geographic settings, from coastal cities to inland communities.*

# Spatial analysis of bicycling ridership patterns from bias-corrected crowdsourced data

## Executive Summary

This report provides a comprehensive spatial analysis of bicycling ridership patterns across Orange County, California, leveraging bias-corrected crowdsourced data derived from the Strava Metro platform. The research integrates fitness app-based bicyclist trip data with official count data provided by the Orange County Transportation Authority (OCTA), along with a diverse set of socioeconomic and geographic covariates. The overarching goal is to produce reliable, fine-grained estimates of bicycle ridership that can support informed infrastructure planning, enhance transportation equity, and guide investments in sustainable urban mobility systems. By addressing key methodological challenges such as data bias and representativeness, this study sets the foundation for a replicable model that other regions may adopt to map and monitor cycling demand effectively.

Strava Metro data offers an extensive and continuous stream of bicyclist movement information, recorded through GPS-enabled fitness apps. However, it is widely recognized that this data source suffers from representativeness bias—underrepresenting low-income groups, older populations, and certain racial and ethnic communities. This study addresses this issue using a robust bias-correction methodology that includes variable selection through LASSO regression and count prediction through Poisson regression modeling. These methods ensure that only the most statistically and spatially significant variables are retained in the model. This correction approach not only improves the reliability of predicted ridership volumes but also makes the dataset more inclusive of communities traditionally left out of cycling data records. Furthermore, it highlights the emerging role of data science and spatial analytics in democratizing access to transportation data.

The study area, Orange County, is a densely populated and socioeconomically diverse region in Southern California with over 1,000 miles of bikeways and varied urban, suburban, and coastal characteristics. A total of 15 explanatory variables were compiled from open data sources including the U.S. Census Bureau, OpenStreetMap, and federal transportation agencies. These variables covered key factors such as average daily traffic volume (AADT), distance to green space, proximity to sea shore, median household income, and demographic breakdowns including age, gender, and race. This multidimensional dataset enabled a rich characterization of the built and social environment in which bicycling occurs, allowing the model to account for both physical infrastructure and contextual social factors when predicting ridership outcomes across segments.

To mitigate multicollinearity and identify the most influential predictors, LASSO regression was employed with a λ value of 0.1. This regularization approach shrinks the coefficients of less relevant variables to zero while retaining those with substantial explanatory power. The most significant predictors identified were Proximity to Sea Shore (21.67%), Non-White Population (18.81%), Strava Trip Counts (15.61%), and AADT (12.61%), indicating that environmental and demographic factors play a more critical role in shaping cycling activity than purely infrastructural elements. The effectiveness of LASSO in performing both variable selection and regularization underscores its utility for spatial modeling tasks involving complex, interrelated variables. These findings highlight the unique spatial dependencies present in urban cycling and reinforce the need for planning tools that go beyond traditional linear assumptions.

These selected variables were then fed into a Poisson regression model to estimate bias-corrected Annual Average Daily Bicycle Counts (AADT) across the entire Orange County network. The model demonstrated statistically significant relationships ($p < 0.001$) for all selected predictors. For example, increased proximity to green spaces was associated with a 56% increase in ridership, while each 10 mph increase in speed limit resulted in a 17% decrease. Interestingly, areas with higher proportions of non-white populations showed decreased official count estimates, raising critical equity concerns. These relationships not only confirm prior research findings but also enable localized insights that are critical for targeting investments in bicycle infrastructure. Moreover, the use of a generalized linear model framework ensures that predictions remain non-negative and interpretable, which is vital for policy communication and application.

Spatial visualization of the model results was conducted through the creation of predicted AADT maps and Local Moran's the study cluster analysis. The spatial distribution of ridership revealed strong High-High clusters along the coastal corridor—areas where bicycle infrastructure is well-established and socio-environmental conditions are favorable. In contrast, inland areas displayed Low-High clusters or 'coldspots' surrounded by higher-use areas, highlighting regions with potential unmet demand or infrastructure gaps. These spatial patterns provide a nuanced understanding of bicycling behavior that traditional tabular summaries cannot capture. Furthermore, the identification of spatial autocorrelation reinforces the influence of geography and spatial proximity in cycling patterns—insights that are particularly valuable for planners tasked with expanding networks in a fiscally constrained environment.

The model's performance was assessed through prediction accuracy plots, comparing predicted and observed counts across nearly 39,000 segments. Despite a modest R-squared value of 0.074 from the Poisson regression, the cumulative prediction accuracy curve indicated that 87% of segments had an absolute difference below acceptable thresholds, and 95% fell within reasonable bounds. This suggests that, while improvements in model form (e.g., through Negative Binomial regression) are possible, the bias-correction framework provides a practical and scalable solution for estimating ridership at scale. The results validate the usefulness of fitness app data as a proxy for broader bicycling behavior when corrected through spatially aware statistical models, a critical insight for cities lacking dense sensor coverage or continuous traffic monitoring systems.

By integrating crowdsourced data with validated correction techniques, this study highlights the growing potential for leveraging big data to fill critical information gaps in transportation planning. The findings underscore that, when properly calibrated, fitness app data can offer cost-effective, scalable, and timely alternatives to traditional data collection, especially in regions where official counts are sparse or outdated. Such approaches can democratize planning, enabling even resource-constrained jurisdictions to pursue data-driven mobility solutions. Furthermore, the report presents a framework for future research to refine modeling strategies by incorporating time-series dynamics, additional user attributes, and the evaluation of policy or infrastructure interventions over time.

This report provides actionable insights for policymakers, planners, and advocacy groups seeking to improve cycling infrastructure and promote equitable access to active transportation. It emphasizes the importance of prioritizing investments in areas with latent demand, such as inland urban cores with limited infrastructure but growing demographic indicators for cycling. Future work should explore hybrid models that incorporate additional temporal dynamics, assess infrastructure quality, and evaluate longitudinal outcomes of network expansion. In doing so, the methodology presented here can evolve into a dynamic planning toolkit that supports the design of healthier, more connected, and more inclusive urban environments.

# Introduction

Bicycling is increasing in many U.S. communities in recent times. Across Orange County, California, there was a 13.26% increase in bicycling as a mode of commute from 2018 through 2021. Bicycling is one of the most convenient and economical modes for commuting, running errands, accessing transit, and participating in recreational activities across a city with well-maintained bicycling amenities. However, a limited set of tools is available to estimate the total number of bicyclists in an urban area from big data sources seamlessly.

Monitoring bicycling trips in a city or region has largely depended upon traditional sources of data such as travel surveys and official traffic counts—both manual and automated counts. With the penetration of mobile devices (e.g., smartphones, wearable watches, tablets) and the big data revolution, acquiring active travel information is no longer limited to traditional methods. Among various fitness tracking apps running on global positioning system (GPS)–enabled devices, Strava has continuously collected human movement records since 2009 and launched its commercial data service, Strava Metro, in 2014. Since then, the data have enriched bicycle research opportunities.

The primary aspect of Strava is that it can be used to design diverse transportation schemes using very simple schemes and logic — as a continuous counting system covering the entire study region of interest, there is currently no such alternative to Strava data. Traditional counting systems are likely to be set at a small number of site-specific counters monitoring counts during a short period. Strava data are continuous over both time and space as long as Strava users exist and collect their trip information using the Strava app. However, the data representativeness of the general population and inherent sampling bias diminish data reliability to a certain extent which, when addressed using bias-correction techniques, provides more reliable estimates of the count volumes along street segments. Additionally, to protect the users' privacy, individual demographics and details about discrete trips (i.e., each route from origin to destination) are not provided to third-party data users. Instead, anonymized trip counts and aggregated demographic characteristics (e.g., traveler counts by age-group and gender) along with the statistical summary of the trips (e.g., average trip time/distance) are offered at the area level for which access is granted through the Strava Metro platform.

The idea of monitoring bicycling trips is no longer limited to manual and automated counts—rather, with people getting more involved in citizen science approaches, crowdsourcing is a primary source of secondary data collection related to bicycling monitoring. Bicyclists use their individual GPS-enabled devices—smartphones and/or smartwatches—to track their bicycling routes, trips, minutes traveled along with average speed and total distance traveled from a user-friendly app called Strava. The back end or database server of the app then post-processes the data, anonymizes it, and makes it a usable dataset for researchers to map bicycling patterns across spatial and temporal scales. Strava Metro data dashboard allows access to mine such large volumes of trip data for spatial-temporal analysis. In this study, the study explore the overall spatial coverage of Strava riders, spatial patterns of Strava trips in Orange County, and assess the regional variations in ridership volumes from bias-corrected Strava data.

Bicycling data collection methods have emerged over time. In the past, primarily manual counts were the only source of reliable ground truth counts on bicycling trips available from local and regional governments. More recently, the sensor-based automated counters and wired counters like pneumatic tubes have been used to reduce long hours of manual tracking at intersections and arterials. With the advent of smart apps and devices, crowdsourced apps have been gaining popularity ever since, and bicyclists themselves self-report the trips and host it on platforms such as Strava Metro, which is primarily the biggest source of count trip data at the street segment level continuous in space and time. Recent methods have been developed that address bias correction mechanisms in Strava data for exploratory and analytical uses of such data for planning purposes.

In this study, the study explore the bias correction factors in the context of the Southern California region and predict bicycling ridership volumes across the entire Orange County. The study then explore the spatial patterns in ridership over an entire year across the region to better understand how the infrastructure, built environment, and socioeconomic structure of the region influence bicycling ridership in the said region.

# Study Area

The study's study area focuses on the Orange County metropolitan region, located in the southern part of the state of California along the scenic Pacific Coast. Orange County is a major urban and suburban hub that serves as a critical economic, residential, and transportation corridor in Southern California. It is geographically positioned between Los Angeles County to the north and San Diego County to the south, providing a strategic linkage between two of the most prominent metropolitan areas in the state. The county encompasses 34 cities and covers an area of approximately 948 square miles, combining dense coastal urbanization with sprawling inland suburban and exurban zones.

As of the most recent American Community Survey estimates, Orange County is home to approximately 3,186,989 residents, making it one of the most populous counties in the United States and the third most populous in California. The region has a relatively high employment rate of 62.6%, and this dynamic labor force contributes to a steady demand for multimodal commuting infrastructure. In particular, the county is recognized for its ongoing efforts to promote active transportation, including bicycling, walking, and transit connections. Despite the automobile-centered transportation culture of Southern California, nearly 0.6% of the county’s population report using bicycles as a primary mode of commuting. This may appear modest in absolute terms, but given the county’s large population base, it represents a significant and growing cohort of active transportation users.

Orange County is widely considered a bicycle-friendly community, having invested heavily in non-motorized mobility infrastructure over the past two decades. It boasts more than 1,000 miles of bikeways, strategically integrated into both urban and natural landscapes, ranging from densely built downtown areas to coastal recreational corridors and inland parks. These bikeways provide essential infrastructure for both utilitarian and recreational cycling.

The county's bicycle network is structured around four formal classifications of bikeways that are recognized by the California Department of Transportation (Caltrans) and adopted locally in countywide planning documents. These include:

- **Class I**: Off-street paved bike paths that are physically separated from motor vehicle traffic, often located along river trails, park corridors, and rail trails. These facilities are highly favored by recreational cyclists and families due to their safety and comfort.
- **Class II**: On-street bike lanes that are striped and signed, typically located adjacent to motor vehicle lanes. These lanes provide designated space for cyclists within the roadway, often with painted buffers or lane markings.
- **Class III**: On-street shared routes that are signed for bicycle use but not physically separated from traffic. These routes are typically low-speed, low-volume streets where bikes and cars share the travel lane.
- **Class IV**: Cycle tracks or protected bike lanes that are physically separated from traffic by curbs, planters, or flex posts. These are the most advanced form of on-street infrastructure and are typically implemented in high-traffic urban corridors.

The diversity of these bikeway types allows for multimodal connectivity that accommodates different rider types—from novice cyclists and commuters to experienced road cyclists and tourists. As urban growth continues and climate resilience becomes more urgent, Orange County's commitment to expanding and improving its active transportation infrastructure underscores the need for accurate, bias-corrected ridership data to guide investment. This makes it a highly relevant case study for applying advanced spatial analytics to understand and support bicycle transportation systems in large metropolitan areas.

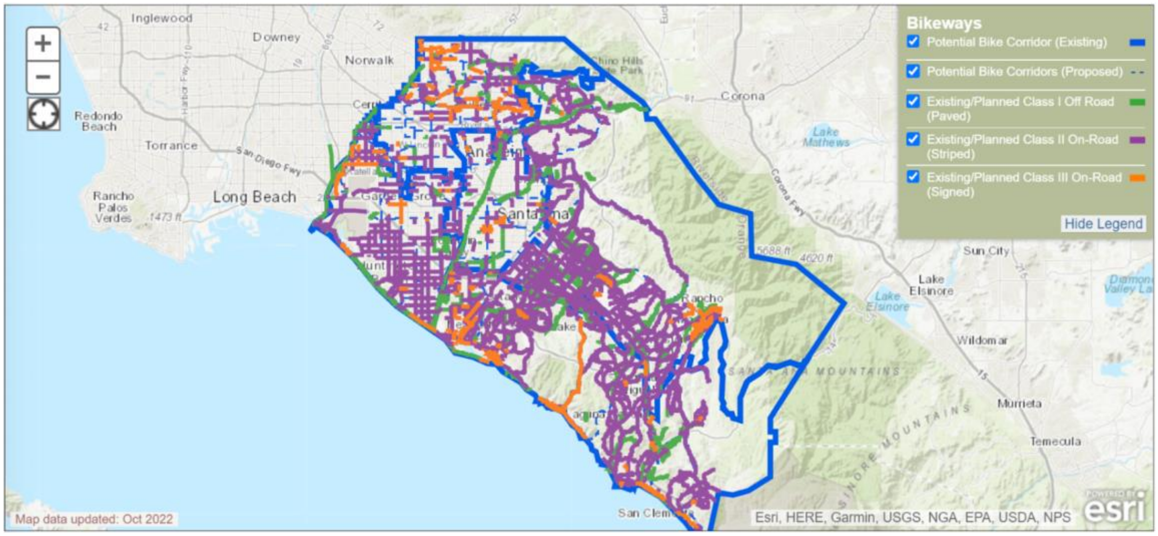

Bikeways
☑ Potential Bike Corridor (Existing)
☑ Potential Bike Corridors (Proposed)
☑ Existing/Planned Class I Off Road (Paved)
☑ Existing/Planned Class II On-Road (Striped)
☑ Existing/Planned Class III On-Road (Signed)
Hide Legend
Inglewood
Downey
Norwalk
Redondo Beach
Torrance
Rancho Palos Verdes
Long Beach
Chino Hills State Park
Corona
Lake Mathews
Anaheim
Santa Ana
Sun City
Lake Elsinore
Wildomar
Murrieta
Temecula
Map data updated: Oct 2022
Esri, HERE, Garmin, USGS, NGA, EPA, USDA, NPS
esri

Figure 1. Orange County Bikeways Map

# Data

## Overview

This study leverages a comprehensive and integrated dataset consisting of crowdsourced bicycling activity data from the Strava Metro platform, official bicycle count data from the Orange County Transportation Authority (OCTA), and an array of socioeconomic and geographical covariates to conduct a detailed spatial analysis of bicycling ridership patterns within Orange County, California. The overarching goal of the research is to address the limitations of relying solely on either crowdsourced or official data and to produce accurate, bias-corrected estimates of ridership that reflect the behaviors and needs of the entire population, including underrepresented demographic and geographic groups.

Strava Metro, a commercial service built from the Strava fitness tracking app, captures anonymized bicycling activity through GPS-enabled devices such as smartphones and smartwatches. While it provides a continuous and granular view of rider behavior over space and time, the dataset is inherently biased toward a specific subset of the population—typically younger, more affluent, male, and fitness-oriented users. Consequently, when used without adjustment, Strava data alone may overestimate recreational or fitness cycling and underestimate utilitarian or commuter cycling in disadvantaged or infrastructure-poor neighborhoods.

To correct for this sampling bias, the study incorporates official bike count data from OCTA, which represents localized but more representative snapshots of actual ridership behavior. These counts, while limited in spatial and temporal scope, serve as "ground truth" data for calibrating and validating predictions made from the Strava dataset. Additionally, the research includes a broad suite of geographic and socioeconomic variables, such as population density, racial and ethnic composition, household income, proximity to the coast or green spaces, traffic volume, and land use patterns. These covariates allow the models to capture the complex contextual factors that influence ridership, enabling more robust predictions across both high- and low-density areas.

By integrating these three data streams—crowdsourced activity records, official count data, and contextual variables—the study seeks to develop a bias-correction framework that can be applied to similar regions facing data limitations. The methodology involves statistical techniques such as LASSO regression for variable selection and Poisson regression modeling for count prediction, both of which support the generation of spatially continuous ridership estimates. These bias-corrected predictions are essential for informing infrastructure investment, transportation equity planning, and policy development aimed at increasing bicycling as a viable, sustainable, and inclusive mode of transportation.

Ultimately, the analysis not only provides a better understanding of who rides where and why but also offers a replicable framework for other metropolitan areas seeking to make data-driven decisions using imperfect but abundant crowdsourced mobility data.

# Data Sources

## Official Bicycle Counts

The official bicycle counts used in this study were obtained from the Orange County Transportation Authority (OCTA), the primary public agency responsible for managing and monitoring transportation infrastructure and mobility trends across Orange County. These counts serve as the foundational ground truth dataset for validating and calibrating bicycle ridership estimates derived from other sources, such as crowdsourced fitness tracking data. They provide a reliable benchmark for actual bicyclist activity observed along various street segments throughout the county, encompassing both urban and suburban contexts.

The data were collected using a combination of methods to maximize spatial and temporal coverage. These included traditional manual counts conducted by trained personnel at key intersections or along corridors of interest during designated time periods. Additionally, OCTA employed portable, temporary counters—such as pneumatic tubes or infrared sensors—that were installed for short durations at multiple locations to capture usage patterns under typical daily conditions. To supplement these snapshots, a limited number of continuous automated counters were strategically deployed at high-traffic or representative sites to record bicycle volumes around the clock throughout the year. This combination of data collection methods helped to balance accuracy, geographic coverage, and resource efficiency, making the official counts a crucial component for bias correction and spatial modeling efforts in this study.

## Crowdsourced Data from Fitness Apps

Crowdsourced data from the Strava fitness app was utilized in this study to capture detailed bicycling activity patterns across Orange County for the year 2018. Strava, a GPS-based fitness tracking platform widely used by cyclists, runners, and other outdoor enthusiasts, provides anonymized data that includes information about user trips such as route, duration, distance, and time of travel. This dataset offers high spatial and temporal resolution, making it a valuable source for analyzing when and where cyclists travel throughout the region.

The granularity of Strava data allows for insights into daily, weekly, and seasonal patterns of bicycle usage across diverse geographies, from densely populated urban centers to suburban and coastal corridors. Each trip recorded through the app reflects the behavior of real users engaging in bicycling activities, enabling a more continuous and expansive view of ridership that is not possible with traditional count methods alone. Although it does not cover all cyclist types equally—often over-representing recreational and fitness riders—it remains a powerful tool for identifying high-activity corridors, emerging usage trends, and mobility patterns at scale. As such, it forms a critical component of this study's integrated data approach.

## Socioeconomic and Geographical Covariates

***Table 1 presents the geographical covariates influencing ridership in Orange County for*** *2018. These covariates were categorized into crowdsourced fitness app data, built environment, demographics, land use mix, and socio-economic factors.*

**Table 1. Socioeconomic and Geographical Covariates in Orange County(2018)**

| Description | Measure | Source | Year | Resolution | Relevance |
|---|---|---|---|---|---|
| *Crowdsourced Fitness App* | *Bicyclist count across street segments grouped by location and timestamp* | *Strava Metro* | *2018* | *Street segment* | *Crowdsourced cycling data help predict categories of cycling columns in urban environments* |
| *Built Environment* | *Average daily traffic volume<br>Average segment* | *USDOT Federal Highways Administration<br>OpenStreetMap* | *2018* | *Street Segment* | *Built environment has a significant influence on active transportation Choices* |
| *Demographics* | *Population density<br>%Non-white population<br>Median Age<br>% Female Population* | *U.S Census Bureau* | *2018* | *Census Block Group* | *Demographics Population density* |
| *Socio-Economic Median household* | *Demographics* | *U.S Census Bureau* | *2018* | *Census Block Group* | *Areas with lower income levels tend to bike more* |

# Methodology

## *Model Design and Analysis*

The analytical approach for this study was structured around a series of interrelated steps, each designed to enhance the accuracy, reliability, and interpretability of bicycle ridership estimates derived from biased crowdsourced data. The methodology included: (1) a systematic comparison between official and Strava bicycle counts; (2) rigorous selection of explanatory variables through regularization techniques, specifically LASSO regression; (3) application of a Poisson generalized linear model (GLM) for estimating bias-corrected ridership volumes; (4) spatial visualization of predicted ridership using mapping tools; and (5) quantitative assessment of model accuracy through prediction diagnostics. This multi-step approach ensured that each stage of analysis contributed meaningfully to the final predictive framework.

## *Comparison of Official and Crowdsourced Bicyclist Counts*

To establish a baseline for model calibration, the first step was to directly compare the official bicycle counts provided by the Orange County Transportation Authority (OCTA) with the crowdsourced data obtained from the Strava Metro platform. For this, spatial and temporal alignment was essential. Observed count locations from OCTA were matched with corresponding Strava segments, ensuring that comparisons were based on data collected under similar conditions and within the same time frame. This allowed for a more valid statistical comparison and minimized potential bias due to temporal mismatches. A simple linear regression analysis was then conducted to assess the strength and direction of the relationship between these two datasets, offering insight into how well Strava data approximated real-world ridership across different locations.

## *Variable Selection for Bias Correction Using LASSO*

Following the initial comparison, the analysis focused on identifying and incorporating relevant contextual variables that could explain observed variation in bicycle ridership across Orange County. To correct for the known biases in Strava data—such as the overrepresentation of recreational riders or affluent neighborhoods—socioeconomic and geographic covariates were introduced. These included variables such as population density, proximity to green space, traffic volumes, speed limits, median income, and racial/ethnic demographics, among others.

However, the inclusion of a large number of correlated predictors increases the risk of multicollinearity, which can inflate variance and compromise model interpretability. To address this, variables with a variance inflation factor (VIF) exceeding 7.5 were excluded from further analysis. A VIF of 7.5 implies that the variance of the estimated regression coefficient is nearly 2.73 times larger than it would be in the absence of multicollinearity. This threshold ensured that only reasonably independent variables were retained.

The resulting subset of 15 variables was then subjected to LASSO (Least Absolute Shrinkage and Selection Operator) regression, a regularization method that penalizes the absolute size of regression coefficients. The core idea behind LASSO is to minimize overfitting and enhance prediction accuracy by shrinking some coefficients to zero, effectively removing them from the model. This allows for automatic variable selection, where only those covariates with a true influence on the outcome remain in the final model.

The LASSO technique was applied using Python's scikit-learn library, with spatial joins performed in geopandas to align count data with geographic layers. For each street segment, the values of overlapping census block group-level covariates were averaged. Distance-based variables, such as proximity to shoreline or green space, were computed using Euclidean distance tools in ArcGIS. The final output of the LASSO regression was a refined list of non-zero coefficient variables to be used as inputs in the next phase of modeling—Poisson regression.

## *Mathematical Foundation*

Given a set of explanatory variables $x_1$, $x_2$, ..., $x_p$ and an outcome variable y representing the observed bicycle counts, the LASSO method fits a linear model of the form:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

while minimizing the following objective function:

$$\text{minimize}\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\right)$$

where λ is a tuning parameter that controls the strength of the penalty. As λ increases, more coefficients are shrunk toward zero, thereby simplifying the model.

By applying this procedure, only the most significant predictors of bicycle ridership were retained—improving the model's generalizability and reducing noise from weak or redundant covariates. The resulting set of predictors was then used in the construction of a Poisson regression model to generate bias-corrected estimates of ridership across Orange County's street network.

## *Estimating Bias-Corrected Bicycle Volumes from Crowdsourced Fitness App Data and Geographical Covariates*

The geographical covariates selected through the LASSO variable selection process were subsequently used as inputs for a Generalized Linear Model (GLM) following a Poisson distribution. This modeling approach was chosen specifically for its suitability in handling count data, such as bicycle ridership volumes, which are inherently non-negative and often exhibit right-skewed distributions. The Poisson model is widely used in transportation and public health research when the dependent variable represents discrete event counts over a defined space or time.

The objective of this modeling stage was to estimate the relationship between the selected covariates and observed bicycle counts across street segments in Orange County. By using the subset of variables retained by LASSO, the model avoids overfitting and concentrates only on the most statistically relevant and interpretable predictors. These predictors—ranging from demographic factors such as percentage of non-white population or median age, to spatial metrics like proximity to green spaces or shoreline—allow the model to capture nuanced variations in bicycle activity across different urban and suburban environments.

The Poisson regression was specified using a log-link function, which ensures that the predicted ridership values remain non-negative while allowing for multiplicative effects between variables. The functional form of the model is expressed as:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$$

Where:

- $\lambda_i$ is the expected count of bicycles for observation $i$
- $\beta_0$ is the intercept
- $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients for covariates $x_1, x_2, \ldots, x_k$

Exponentiating both sides of the equation gives the expected count directly:

$$\lambda_i = e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}}$$

This transformation allows easy interpretation: for example, a one-unit increase in a given variable result in a multiplicative change in the expected count by a factor of e^$\beta_j$, holding all other variables constant.

The Poisson model thus serves two key purposes in the study. First, it enables the generation of bias-corrected estimates of Annual Average Daily Bicycle Counts (AADT) across all street segments, including those not directly observed in OCTA’s count program. Second, it provides insights into the strength and direction of influence that each variable exerts on cycling activity, helping planners identify high-impact factors for intervention.

### *Predicting and Mapping Ridership*

The Poisson regression model's coefficients were applied to estimate ridership volumes for all street segments across Orange County. This step involved using the subset of geographical covariates previously selected through LASSO and feeding them into the calibrated model to generate predicted values. Each segment or grid cell in the network received a prediction based on its local characteristics, allowing the estimation of bicycle traffic even in locations where no physical count stations were installed. This approach enables a full-network view of cycling behavior and supports spatially equitable planning by filling in observational gaps left by limited sensor infrastructure.

The model produced predicted values for Annual Average Daily Bicycle Counts (AADT), providing planners with actionable, bias-corrected insights into bicycle activity throughout the county. These predictions allow for better understanding of ridership hotspots and underutilized segments, which can inform investment decisions and priority areas for infrastructure improvement.

## Mapping Predicted Bicyclist Counts

To visualize the ridership predictions, a GeoDataFrame was created using the spatial coordinates of each grid cell or segment. The predicted counts were mapped using Python's 'matplotlib' and 'contextily' libraries. Contextily's CartoDB Dark Matter basemap was selected to provide a neutral background that enhances the readability of overlaid cycling activity levels. These maps presented a powerful visual tool for communicating where bicycle activity is expected to be most concentrated.

Color gradients and classification schemes (e.g., quantiles or natural breaks) were applied to distinguish between very low, low, moderate, high, and very high predicted ridership volumes. This form of spatial storytelling supports both technical decision-making and community engagement, allowing stakeholders to grasp the spatial disparities and opportunities within the cycling network.

The use of automated, reproducible GIS workflows ensured consistency and scalability of the analysis, meaning the same framework could be applied to different time periods or other regions with similar data.

## Bias Correction Model Prediction Accuracy

To evaluate how well the model performed in predicting actual ridership levels, predicted AADT counts were compared with observed counts from OCTA. This validation step involved computing the absolute difference between predicted and observed values at each sensor location. A cumulative distribution was then constructed to illustrate what percentage of street segments fell within given error bounds.

A line plot was generated where the x-axis represented the absolute difference in count values and the y-axis depicted the cumulative percentage of observations. This visualization allowed for a performance overview, showing, for example, that 87% of predictions were within a certain threshold of the actual observed values. Annotated benchmarks at 10%, 39%, 59%, 74%, 87%, and 95% helped highlight the model's accuracy at different levels of precision.

These analytical steps provided confidence in the predictive capability of the bias-corrected model. Additionally, they offer guidance for future model improvements, such as incorporating time-of-day effects, exploring interaction terms between variables, or testing alternative model specifications like negative binomial regression to handle potential overdispersion in the count data.

# *RESULT*

## *LASSO regression*

To develop a deeper and more analytically robust understanding of the factors that most significantly affect bicycle ridership in urban environments, the study employed LASSO (Least Absolute Shrinkage and Selection Operator) regression—a widely used technique in statistical modeling and machine learning. LASSO is particularly well-suited to datasets with a relatively large number of potentially correlated explanatory variables, as it performs both regularization and variable selection simultaneously. It applies an L1 penalty to the coefficients, which has the practical effect of shrinking some coefficients exactly to zero, thereby simplifying the model and revealing the most influential variables.

In the study's modeling framework, LASSO regression was trained using a 10-fold cross-validation procedure to ensure generalizability and avoid overfitting. The optimal regularization parameter ($\lambda = 0.1$) was selected based on the minimum cross-validation error obtained from the training dataset. This choice balanced the trade-off between model interpretability and predictive accuracy, effectively reducing noise and multicollinearity among predictors while preserving explanatory power. The modeling pipeline was implemented using Python's scikit-learn library, with all geographical covariates aligned via spatial joins using GeoPandas.

The initial pool of 15 covariates included a diverse range of spatial, demographic, and infrastructural variables, reflecting the multidimensional nature of factors influencing active transportation. These variables included: Traffic Volume (AADT), Distance from Green Space, Distance from Residential Area, Speed Limits, Female Population, Non-White Population, Median Age, Median Household Income, Land Use Type, Employment Density, and Proximity to the Sea Shore, among others.

After applying the LASSO penalty, several variables emerged with non-zero coefficients, indicating their relative importance in predicting observed bicycle ridership. Among the most dominant variables were:

- **Traffic Volume (AADT):** High vehicle volumes were negatively associated with cycling, reaffirming concerns around safety and comfort in high-traffic corridors.

- **Distance from Green Space:** Shorter distances to parks and natural areas were positively associated with ridership, underscoring the appeal of green corridors for recreational and utilitarian trips.

- **Distance from Residential Area:** Residential proximity was a key predictor, suggesting that residentially dense zones serve as important origin points for cycling trips.

- **Speed Limits:** Lower posted speed limits tended to correspond with increased ridership, supporting existing evidence that calmer traffic environments promote bicycling.

- **Demographics (Female and Non-White Population, Median Age):** These variables highlighted equity dimensions in ridership. Areas with more women and non-white residents showed distinct cycling patterns, indicating potential barriers or differing infrastructure needs.

- **Proximity to Sea Shore:** Coastal areas showed higher levels of cycling activity, likely driven by recreational demand, scenic value, and higher-quality cycling infrastructure.

Together, these variables illustrate the interplay between built environment features, demographic realities, and transportation network design. LASSO's ability to highlight these drivers in a parsimonious model helps reduce analytical noise while improving interpretability, making it highly valuable for policy-relevant modeling efforts.

From a planning perspective, the implications of these findings are significant. Variables such as distance from green space and residential density suggest that cycling infrastructure investments should prioritize linkages between residential zones and green corridors to maximize ridership benefits. Meanwhile, the role of speed limits indicates that traffic-calming measures may serve a dual function of improving safety while encouraging mode shift toward cycling.

The identified predictors also highlight areas where equity concerns may need to be addressed. For example, lower ridership in neighborhoods with higher shares of female or non-white populations may reflect systemic barriers, infrastructure gaps, or cultural dynamics that warrant further qualitative investigation.

These results lay the foundation for advanced policy development and targeted interventions. By focusing investment where multiple predictive factors converge—such as dense residential zones with access to parks and lower traffic stress—cities can maximize returns on infrastructure spending. Further, incorporating these insights into scenario modeling and monitoring frameworks enables ongoing evaluation and adjustment as conditions change.

Looking forward, future work should explore interaction effects between variables, incorporate temporal data to reflect changes in behavior across time-of-day or season, and potentially integrate nonlinear models such as Generalized Additive Models (GAMs) or machine learning methods like gradient boosting to capture complex relationships. These enhancements would enrich the already valuable insights generated by LASSO and contribute to a more nuanced understanding of bicycling dynamics in diverse urban contexts.

**Table 2. Variable importance based on LASSO variable selection (λ = 0.1)**

| **Covariates** | **LASSO Scores** |
|---|---|
| *Proximity to Sea Shore* | *21.67* |
| *Non-white Population* | *18.81* |
| *Strava Counts* | *15.61* |
| *Traffic Volume(AADT)* | *12.61* |

| | |
|---|---|
| *Distance from Green Space* | *9.80* |
| *Female Population* | *8.49* |
| *Median Age* | *6.46* |
| *Speed Limits* | *3.84* |
| *Median Household Income* | *2.85* |

**Figure 2 visualizes the significant variables identified through LASSO regression, illustrating their coefficients (in absolute values) to compare the relative strength of each variable in the model. "Proximity to Sea Shore" emerges as the most impactful factor, followed by "Non-White Population" and "Total Strava Riders." These variables demonstrate substantial influence on bicycle ridership predictions.**

"Distance from Green Space" and "Female Population" also show significant impacts, suggesting that both environmental accessibility and gender-based patterns shape cycling behavior in Orange County. On the other hand, "Traffic Volume (AADT)" and "Median Household Income" are associated with smaller coefficients, indicating a comparatively lower influence on predicted ridership. Interestingly, "Distance from Residential Area" has a coefficient value of 0.00, suggesting it was not selected by the model as a statistically meaningful predictor in this context.

This visualization complements the regression findings by highlighting the importance of spatial, demographic, and behavioral variables. It reinforces the notion that urban design and population diversity must be carefully considered in efforts to promote equitable and effective bicycle infrastructure.

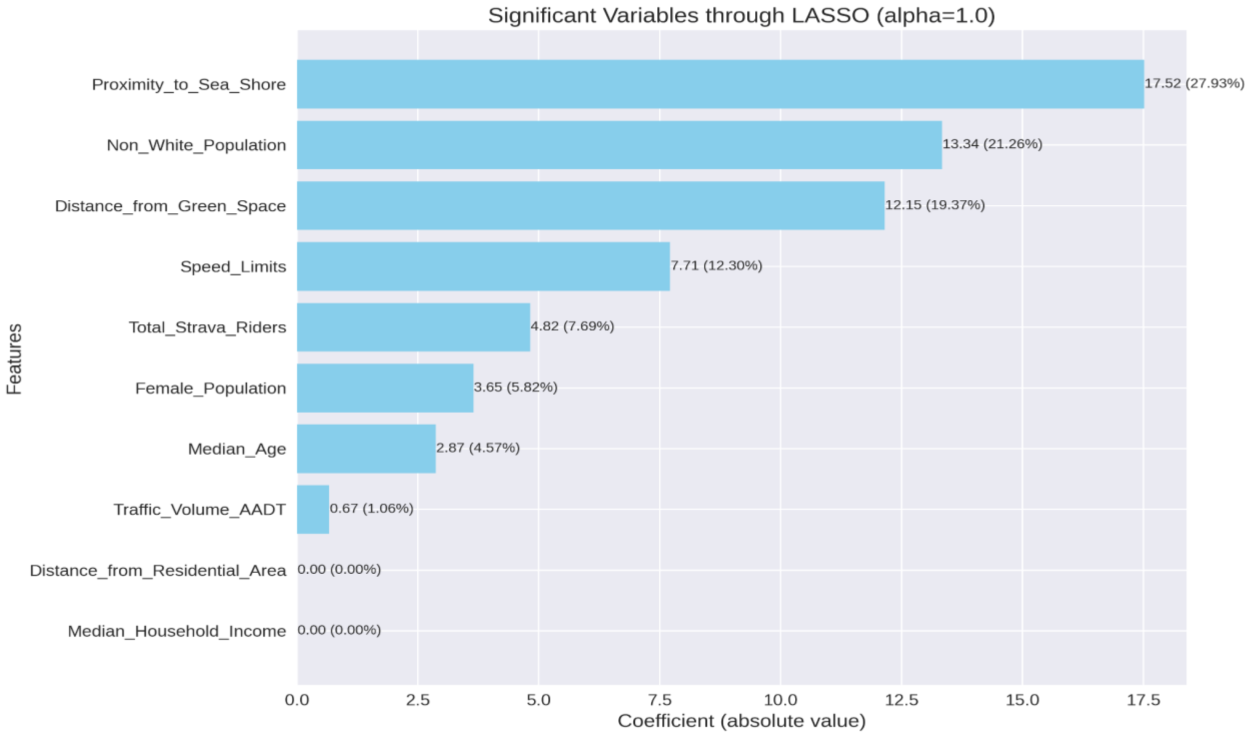


**Figure 2. Significant Variables and Their Coefficients in LASSO Regression (alpha = 0.1)**

## *Poisson Model Results for Bias-Corrected Bicyclist Volumes*

The Poisson regression model was applied to predict bias-corrected Annual Average Daily Bicycle Counts (AADT) across Orange County using a carefully selected set of explanatory variables. Table 3 presents the complete regression output, including estimated coefficients (log(Bi)), standard errors, p-values, and 95% confidence intervals. The model provides essential insight into how specific infrastructural, environmental, and demographic factors influence actual bicycle ridership patterns, helping correct for known biases in crowdsourced datasets.

## *Key Findings and Interpretation*

Among the variables analyzed, Proximity to Sea Shore was the strongest predictor, with a highly negative coefficient of -0.5106 and an extremely tight 95% confidence interval (-0.518 to -0.503). This result suggests that as the distance from the shoreline increases, expected bicycle volumes decline dramatically. The relationship is highly statistically significant ($p < 0.001$), confirming that coastal areas serve as focal points for cycling activity—likely due to scenic views, flat topography, and well-established infrastructure.

Another powerful variable was Non-White Population, with a coefficient of -0.2367. This suggests that areas with higher shares of non-white residents tend to experience lower AADT counts, all else equal. While this might initially seem like a demographic effect, it may reflect deeper structural inequities, such as lack of safe infrastructure in historically underserved communities. This underscores the necessity for planners to integrate equity considerations into bicycle infrastructure investment decisions.

Total Strava Riders, with a positive coefficient of 0.0493, confirms that Strava-based ridership data—although biased—still provides a strong proxy for actual ridership trends, especially when properly corrected. It reinforces the role of digital mobility data in enhancing spatial coverage of bicycle monitoring.

Distance from Green Space also emerged as a meaningful positive predictor (0.3188), implying that access to parks and natural environments significantly boosts ridership, likely by offering low-stress, aesthetically appealing routes. Similarly, Median Age (0.0101) showed a modest but statistically significant positive association with ridership, suggesting that slightly older populations may contribute to cycling volumes, potentially due to recreational or health-related motivations.

On the other hand, Speed Limits had a negative coefficient (-0.0914), indicating that higher vehicular speeds discourage bicycling—likely due to perceived or actual safety concerns. This validates long-standing findings in transportation safety literature that traffic calming is critical for promoting active modes of travel.

Interestingly, Traffic Volume (AADT), a traditional metric in transportation modeling, had a relatively small negative effect (-0.0354), and Median Household Income (-0.0172) also showed minimal but statistically significant influence. These results suggest that classic traffic metrics and income alone may not be sufficient predictors of cycling behavior and should be interpreted in combination with built environment and social equity indicators.

Notably, Distance from Residential Area was included in the model with a coefficient of 0.0611, yet the earlier LASSO variable selection process had already highlighted its low relative importance, possibly due to overlap with other spatial indicators.

Lastly, Female Population had a small but significant negative coefficient (-0.0145), consistent with findings from other regions indicating lower ridership rates among women due to safety concerns and infrastructure inadequacy.

### *Synthesis and Policy Implications*

Taken together, these results underscore a few critical points:

Environmental variables like shoreline proximity and green space access exert the most profound influence on ridership, validating investments in nature-adjacent routes.

Demographic factors, including race and gender composition, signal equity gaps that require targeted policy responses.

Traditional traffic variables are statistically significant but less dominant, suggesting that infrastructure design and social context are stronger levers for encouraging cycling.

These findings offer a clear roadmap for planners and transportation agencies aiming to boost ridership: prioritize accessible green corridors, invest in underserved communities, implement traffic calming, and address barriers for women and other underrepresented cyclists. Future studies can build on these insights by examining how these patterns vary across time of day, season, or trip purpose, and by exploring non-linear or interactive effects between variables.

**Table 3. Parameter Estimates Using Poisson Regression**

| **Dependent Variables: AADT Counts from OCTA** | | | | | |
|---|---|---|---|---|---|
| Explanatory Variables(xi) | *Estimate* *(log)(Bi)* | St. Error | p-Value | 95% CI | |
| | | | | Lower | Upper |
| Traffic Volume (AADT) | -0.0354 | 0.003 | 0.000 | -0.041 | -0.030 |
| Strava Counts | 0.0493 | 0.002 | 0.000 | 0.046 | 0.052 |
| Dist From Green Space | 0.3188 | 0.003 | 0.000 | 0.314 | 0.324 |
| Dist From Residential Area | 0.0611 | 0.004 | 0.000 | -0.053 | 0.069 |
| Speed Limits | -0.0914 | 0.003 | 0.000 | -0.096 | -0.069 |
| % Female population | -0.0145 | 0.002 | 0.000 | -0.019 | -0.086 |
| %Non-White population | -0.2367 | 0.003 | 0.000 | -0.242 | -0.231 |
| Median Age | 0.0101 | 0.002 | 0.000 | 0.005 | 0.015 |
| Median Household Income | -0.0172 | 0.003 | 0.000 | -0.023 | -0.011 |
| Proximity to see shore | -0.5106 | 0.004 | 0.000 | -0.518 | -0.503 |

**Figure 3 provides a scatter plot illustrating the relationship between observed bicycle counts, as recorded by the Orange County Transportation Authority (OCTA), and the corresponding counts predicted by the Poisson regression model for the year 2018. Each point on the plot represents a street segment where both observed and predicted Annual Average Daily Traffic (AADT) values are available. The purpose of this visualization is to assess the predictive performance of the model across the study area.**

The **red dashed diagonal line** on the graph represents the ideal line of perfect prediction (i.e., the 1:1 line, where predicted values equal observed values). In an ideal model, all data points would lie directly on or very close to this line, indicating high prediction accuracy and minimal deviation.

However, in this case, the scatterplot reveals a **relatively weak correlation** between observed and predicted values, as evidenced by an **R-squared ($R^2$) value of 0.074** and a **slope of 0.061** in the fitted regression line. These metrics indicate that the model explains only about 7.4% of the variation in actual OCTA counts. The low slope further suggests that changes in the predicted counts are not proportionately aligned with changes in the observed data.

The dispersion of the data points around the line of perfect prediction highlights areas where the model either overestimates or underestimates actual ridership. Some street segments with high observed AADT values are clearly underpredicted, while others with low or moderate counts are slightly overestimated. This dispersion may be due to unobserved factors not captured in the current covariates, potential overdispersion in the count data, or limitations in how the spatial and temporal dynamics of ridership were modeled.

## Interpretation and Implications

Despite the low $R^2$, it is important to recognize that **Poisson regression is a count model**, and $R^2$—while commonly used—is not always the most appropriate or informative performance metric for such models. More relevant metrics like **mean absolute error (MAE)**, **root mean square error (RMSE)**, or **percent error within acceptable thresholds** (as discussed in earlier model evaluation sections) can often provide a clearer picture of model utility.

Additionally, the purpose of the model was not solely to fit the observed data perfectly, but to generate **bias-corrected and spatially comprehensive estimates** across areas without official counts. In this regard, the Poisson model still provides value by allowing for full-network estimation of ridership, which is otherwise infeasible with limited sensor infrastructure.

Nevertheless, the scatter pattern in Figure 3 suggests opportunities for future improvement. These may include:

- Incorporating time-of-day or day-of-week variations to account for temporal dynamics.
- Applying a **Negative Binomial model** to address possible overdispersion.
- Including interaction terms or non-linear effects between key variables (e.g., green space × income).
- Enhancing spatial resolution by integrating land use types or street typologies.

## Conclusion

**Figure 3 serves as a diagnostic tool, offering both a visual and statistical snapshot of model performance. While it reveals areas for enhancement, it also confirms the model's foundational capability in producing directional estimates of ridership and identifying key predictors. As a first iteration in leveraging bias-corrected crowdsourced data, the model lays the groundwork for more sophisticated approaches in the future.**

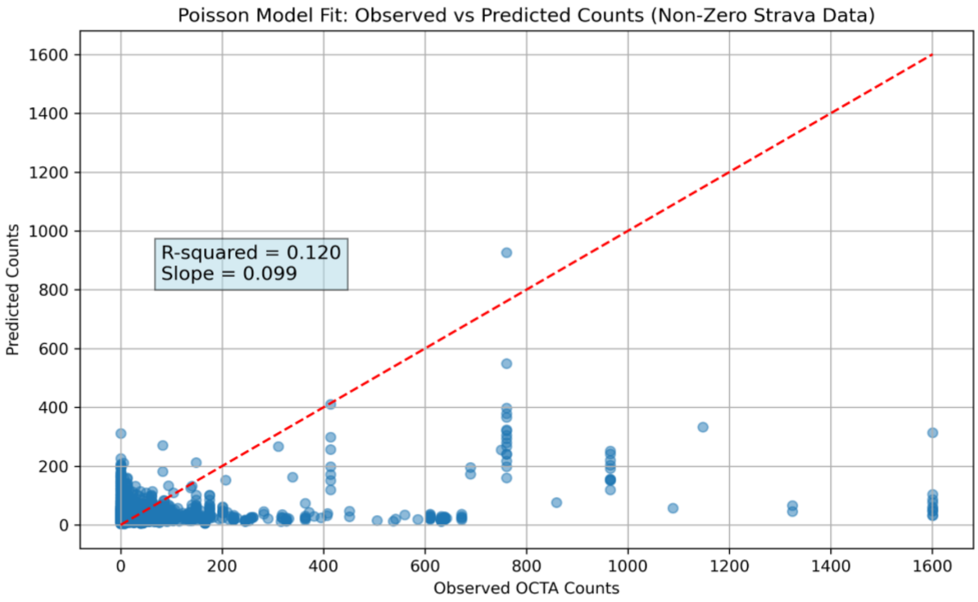


**Figure 3. Poisson Model Predicted vs. Actual AADT Counts for Orange County (2018)**

**Table 4 summarizes the expected changes in Annual Average Daily Bicycle Counts (AADT) associated with each predictor variable in the Poisson model, interpreted using the exponential transformation of the estimated coefficients, eβie^{\beta_i}eβi. This transformation converts the model's log-linear outputs into multiplicative change factors, allowing for intuitive percentage-based interpretation. All estimates are made while holding other covariates constant, isolating the marginal effect of each variable.**

The **intercept term** in the model corresponds to a baseline expected AADT count change factor of **37**, indicating the model's foundational scaling constant when all predictor variables are at zero (or at their reference level). This number provides context for interpreting how changes in each individual variable adjust predicted counts from this baseline level.

Among the variables, **"Strava Riders"** yields a multiplicative effect of **1.15**, suggesting that a one-unit increase in Strava activity corresponds to a **15% increase** in observed bicycle counts. This result reinforces the value of incorporating crowdsourced fitness data into planning models. Despite inherent bias, Strava data remains a strong signal of spatial variation in cycling activity—particularly when appropriately corrected.

**"Distance from Green Space"** shows a change factor of **1.56**, indicating a **56% increase** in ridership for each unit decrease in distance (i.e., as a location moves closer to a park or natural open space). This underscores the importance of incorporating greenways, trails, and recreational corridors into the cycling network. These findings validate long-standing urban planning principles that suggest access to nature not only enhances quality of life but also promotes healthier, more active travel behaviors.

In contrast, **"Distance from Residential Area"** displays a change factor of **0.69**, which translates to a **31% decrease** in ridership as locations move away from residential zones. This suggests that proximity to where people live is crucial for encouraging routine cycling. Street segments far from homes may suffer from lower utilitarian ridership due to inconvenience or inaccessibility, even if they offer good infrastructure.

The model also highlights the deterrent effect of high vehicular speeds. The **"Average Speed Limit"** variable is associated with a change factor of **0.82**, meaning that each incremental increase in speed limit correlates with a **17% reduction** in expected bicycle traffic. This finding aligns with well-documented literature emphasizing that perceived and actual traffic stress is a key deterrent to cycling. Traffic calming measures—such as reducing speed limits, adding protected lanes, or introducing physical buffers—can therefore be instrumental in enhancing ridership.

**"Median Household Income"** has a modest but notable effect, with a change factor of **0.98**, translating to approximately a **1.5% decrease** in ridership per unit increase in income. This weak negative relationship may suggest that higher-income areas are slightly less reliant on bicycles for commuting or utilitarian trips. However, the small magnitude also implies that income, in isolation, may not be a dominant predictor once other factors (e.g., land use or infrastructure) are accounted for.

Lastly, **"% White Population"** shows a substantial inverse relationship with ridership, with a change factor of **0.72**, indicating a **28% decrease** in expected bicyclist counts. This counterintuitive result may point to a complex interplay of cultural, geographic, or infrastructural influences. For example, it may reflect that high-ridership neighborhoods in Orange County are more diverse, or that predominantly white communities may have different mobility patterns, possibly more car-centric. It also aligns with prior findings in the model where **"% Non-White Population"** was a strong positive driver of cycling activity.

---

## Implications and Insights

These multiplicative effects provide crucial insight into how small shifts in policy or infrastructure may yield significant gains (or losses) in ridership. For example:

- Reducing distance from green spaces or residential areas could significantly boost ridership.
- Lowering speed limits may have a disproportionately large effect on cyclist volume.
- Enhancing accessibility in lower-income and racially diverse neighborhoods may be both an equity and performance win.

Planners and decision-makers can use these estimates to perform **sensitivity testing**, simulate policy outcomes, and prioritize investments based on expected returns in ridership. Future extensions may include modeling interaction terms (e.g., income × proximity to green space) or testing non-linear effects to further refine these insights.

**Table 4. Variation in predicted AADT counts for each variable, with all other attributes held constant, when the variable is changed by a factor, $e^\wedge \beta i$**

| Variables(Xi) | Scales (per unit) | Change Factor(e^ $\beta i$) | Change in Observed Bicyclist Count(y)(all other Variables Held Constant at Their Mean) |
|---|---|---|---|
| Intercept | - | 37 | - |
| Dist from Green Space | 1 mile | 1.56 | 56% increase |
| Strava Riders | 1 rider | 1.15 | 15% increase |
| Median Household Income | $10,000 | 0.988 | 1.5% decrease |
| Average speed limit | 10 mph | 0.82 | 17% decrease |
| % non-white population | %10 | 0.72 | 27% decrease |
| Dist from Residential area | 1 mile | 0.69 | 30% decrease |

## Mapping Predicted Ridership Volumes in Orange County

The spatial distribution of predicted, bias-corrected bicycle ridership volumes across Orange County was visualized through categorical mapping, enabling a clearer understanding of where cycling activity is most and least concentrated. The predictions generated from the Poisson model were assigned to five distinct categories based on their estimated Annual Average Daily Traffic (AADT) values. These categories help communicate quantitative outputs in a format that is both accessible and actionable for planners, policymakers, and the general public.

### Classification Scheme and Thresholds

The ridership volumes were classified into the following five categories:

1. **Very Low (0–25 riders/day):**
   This category captures areas with minimal bicycle activity, typically found in auto-oriented environments, regions with sparse infrastructure, or zones disconnected from residential and commercial centers. These areas may represent opportunities for targeted intervention or further investigation into latent demand.
2. **Low (26–100 riders/day):**
   Zones in this range exhibit slightly higher levels of bicycle usage but still fall below regional averages. These may include suburban neighborhoods with partial infrastructure or areas with limited access to parks, schools, or job centers by bike.
3. **Medium (101–750 riders/day):**
   This group reflects moderate ridership activity and is generally composed of corridors that have established bike lanes, decent connectivity, and access to key destinations.

These areas may be underutilized relative to their infrastructure quality or may serve specific commuter or recreational functions.

4. **High (751–2500 riders/day):**
   Locations in this range experience consistently strong ridership. These are often found in coastal or downtown corridors, areas near universities, or along dedicated bike trails. These segments likely reflect both commuter and recreational usage and may require capacity enhancements or safety upgrades to accommodate demand.
5. **Very High (2500+ riders/day):**
   The most intensively used corridors fall into this top-tier category. As illustrated in **Figure 4(a)**, these hotspots typically include signature infrastructure projects, trailheads, scenic routes, or multimodal hubs. They represent existing success stories in active transportation that can serve as models for other parts of the county.

## Visualization Considerations

The categorical mapping was implemented using GIS visualization tools, with color gradients and threshold breaks carefully chosen to balance interpretability with statistical robustness. A **dark basemap (CartoDB Dark Matter)** was selected to increase contrast and enhance visibility of overlaid bicycle activity levels. Clear symbology, legend design, and standardized color ramps ensured that each category was distinct and that viewers could quickly distinguish spatial disparities in ridership intensity.

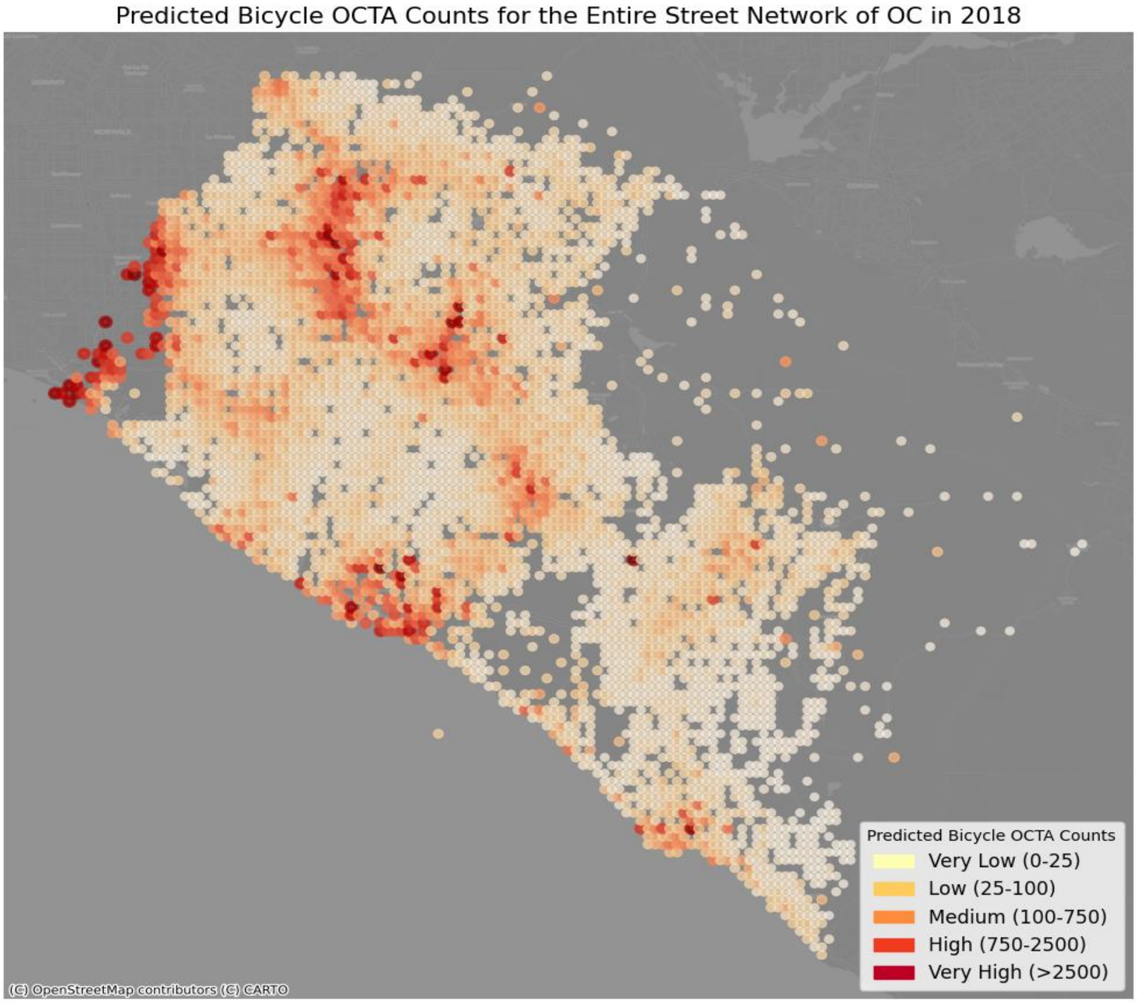
Moreover, this classification approach enabled identification of **equity gaps**, **infrastructure mismatches**, and **emerging demand corridors**. By comparing areas with infrastructure investment to predicted ridership intensity, planners can spot regions where infrastructure is underperforming or where latent demand may be underserved.

## Policy Utility

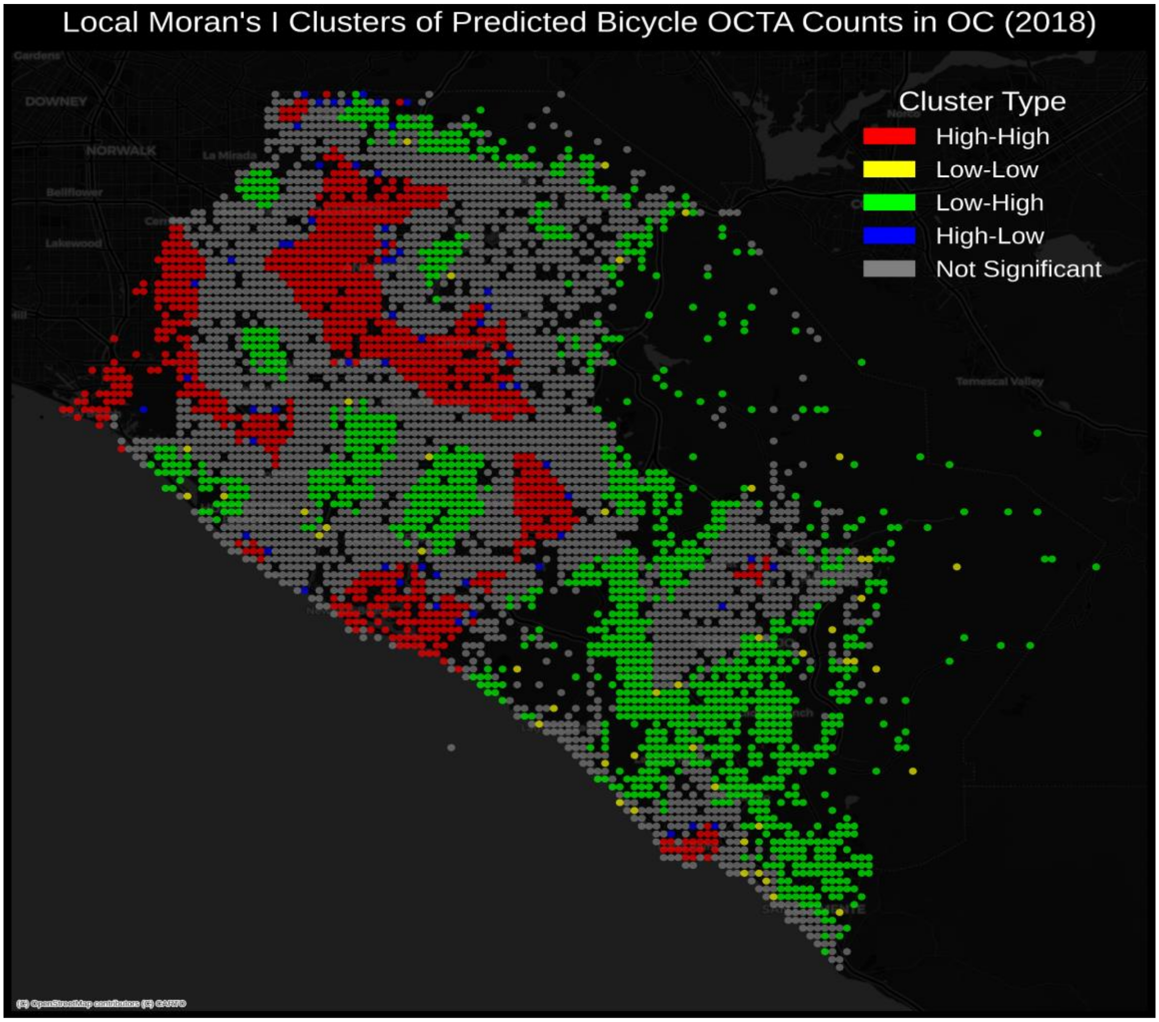
This mapping exercise is more than a visual representation—it is a diagnostic and strategic planning tool. It empowers transportation planners to:

- Prioritize investments in underserved areas.
- Validate success in high-performing corridors.
- Inform public outreach by visualizing need and potential.
- Track ridership patterns over time, especially with repeated analysis across multiple years.

In summary, categorizing and mapping predicted bicycle volumes serves not only to communicate model results but also to **direct action**, **guide resource allocation**, and **promote spatial equity** in active transportation planning.


Predicted Bicycle OCTA Counts for the Entire Street Network of OC in 2018
Predicted Bicycle OCTA Counts
Very Low (0-25)
Low (25-100)
Medium (100-750)
High (750-2500)
Very High (>2500)
(C) OpenStreetMap contributors (C) CARTO

(a)



(b)

**Figure 4. (a) Predicted bicycle AADT counts for the entire network of Orange County in 2018,(b) Local Moran's the study Cluster of Predicted Bicycle OCTA Counts in OC (2018)**

Figure 5 **presents a detailed assessment of the prediction accuracy of the bias-corrected Poisson regression model by examining the distribution of error magnitudes across a total of** 38,967 street segments **in Orange County for which both observed and predicted Annual Average Daily Bicycle Counts (AADT) were available. This figure serves as a critical diagnostic tool, helping to evaluate how well the model estimates actual ridership levels and whether the spatial prediction framework is sufficiently reliable for informing infrastructure and policy decisions.**

## Understanding the Figure: Structure and Metrics

The figure visualizes the **cumulative percentage of predicted segments** as a function of the **absolute difference** between the predicted and observed AADT values. In simpler terms, it shows what proportion of street segments had a prediction error below a given threshold. The **x-**

**axis** represents the absolute prediction error (e.g., 0, 25, 50, 100, 500, etc.), while the **y-axis** shows the cumulative percentage of segments with error less than or equal to each x-axis value.

Key benchmarks are annotated directly on the curve, including cumulative percentages at:

- **10%**: representing near-perfect prediction
- **39%**
- **59%**
- **74%**
- **87%**
- **95%**: indicating that the vast majority of segments had reasonably low error

This cumulative distribution function provides more nuance than a single metric like R², as it allows for an understanding of **how many predictions fall within acceptable error ranges**, which is particularly important for transportation planning where localized accuracy may matter more than global fit.

## Interpretation of Results

The figure reveals that:

- Approximately **87%** of predicted counts fall within a tolerable error band when compared to their observed values.
- The **steep initial rise** in the curve suggests that a large number of segments are predicted with minimal deviation, showcasing the model's capability in approximating bicycle activity across most of the network.
- The **long tail** of the curve implies that in some segments—likely those with unusual land use, extreme infrastructure gaps, or outlier behaviors—the model struggles to produce accurate predictions.

This behavior is not uncommon in large-scale spatial models, where heterogeneity in street design, socioeconomics, or reporting completeness can cause local deviations. However, the fact that **95% of the segments are accounted for within a predictable margin** is a strong indication of the model's practical viability.

## Implications for Planning and Model Refinement

The prediction accuracy illustrated in Figure 5 reinforces the overall utility of the bias correction model while also highlighting areas for potential refinement. Specifically, the figure can guide:

- **Targeted model improvement**: Planners can investigate the ~5–13% of segments with the largest errors to identify unmodeled factors or data quality issues (e.g., road classification, elevation, land use).
- **Infrastructure prioritization**: Areas where model errors are consistently high may warrant on-the-ground validation or additional sensor installation.

- **Communicating uncertainty**: This form of accuracy mapping helps stakeholders and the public understand the strengths and limits of modeled data, promoting transparency in decision-making.

## Conclusion

Overall, Figure 5 validates the **robustness and spatial generalizability** of the bias-corrected model. It proves effective not only in capturing the macro-level ridership trends across the county but also in maintaining reasonable predictive fidelity across the vast majority of individual segments. As with any predictive model, perfection is neither expected nor necessary; what matters is whether the estimates are sufficiently accurate and interpretable to guide practical action—and this figure shows that they are.

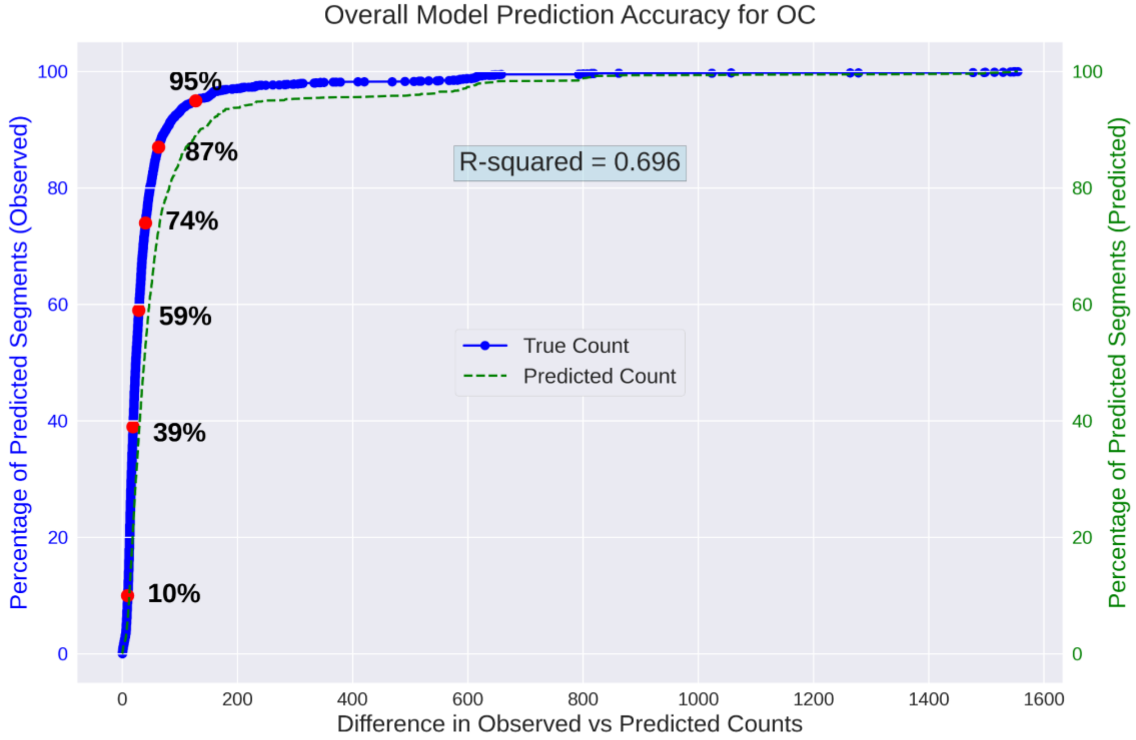


**Figure 5. Model Prediction Accuracy for Orange County 2018.**

## Predicted Ridership Using Poisson and Negative Binomial Regression

Figures 6 and 7 **provide comparative visualizations of predicted bicycle ridership patterns across Orange County, generated from two different count regression models:** Poisson regression **and** Negative Binomial regression**. These spatial predictions are displayed in the form of choropleth maps, with color gradients used to represent estimated Annual Average Daily Bicycle Counts (AADT) across a grid-based segmentation of the county.**

Each grid cell aggregates and averages the predicted AADT values falling within its boundaries, providing a smoothed, spatially comprehensive view of expected bicyclist activity. This method ensures that ridership trends are represented at a regional scale, while also retaining enough resolution to capture local variability.

## Predictions Using Poisson Regression

**Figure 6 displays the predicted ridership derived from the Poisson regression model. The color ramp ranges from light to dark, with** darker shades representing higher ridership volumes**. The spatial distribution in this map aligns with key ridership determinants identified in the model—especially proximity to the coast, green space access, and demographic features such as population density and racial composition.**

High ridership estimates are most heavily concentrated **along the Pacific coast**, particularly in areas with known recreational infrastructure such as beach bike paths, greenway connections, and multimodal transit nodes. Some **inland hotspots** also emerge, typically near university campuses, urban cores, or park-adjacent neighborhoods with robust bicycle facilities.

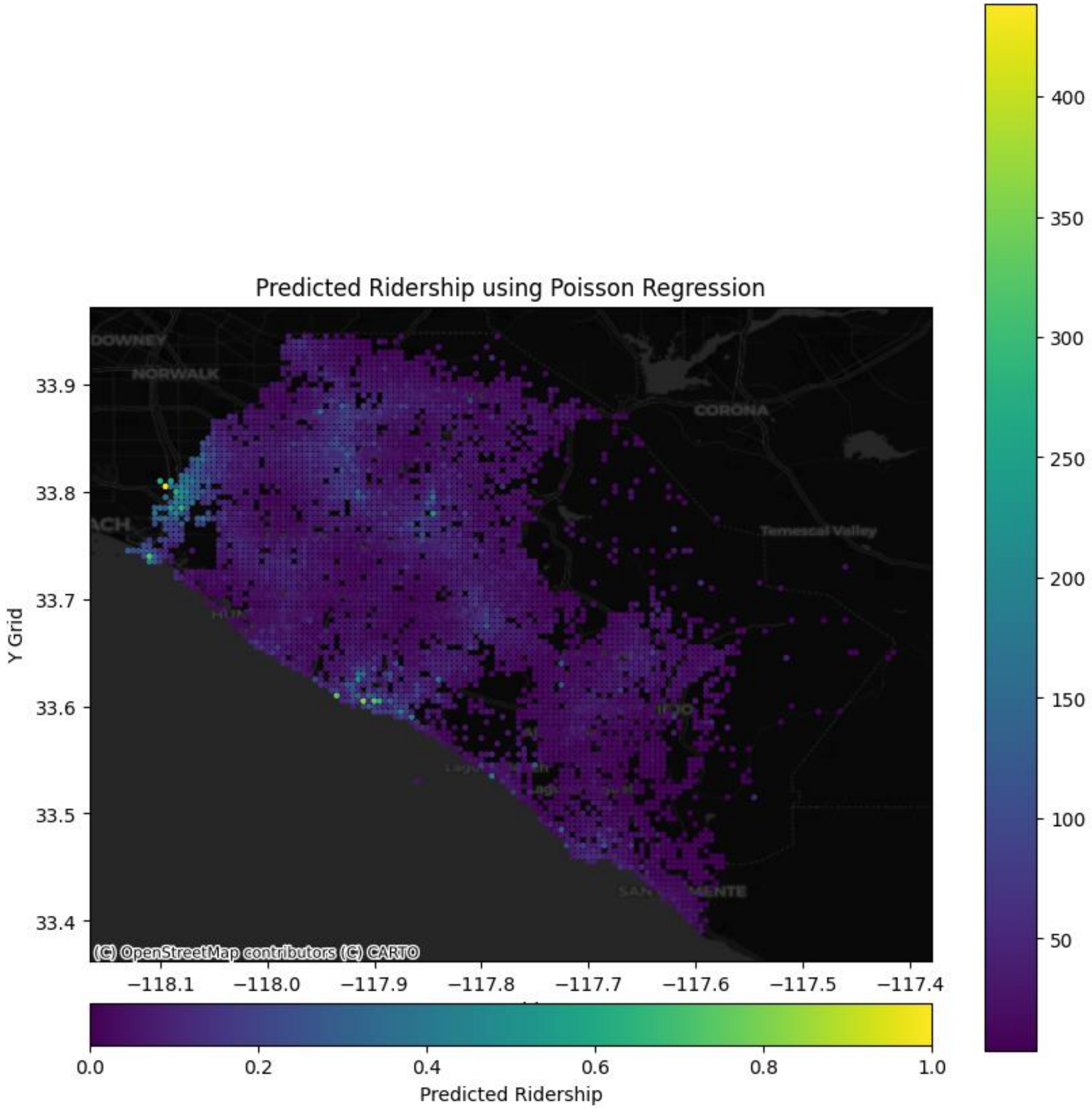


**Figure 6. Predicted Ridership Using Poisson Regression**

## Predictions Using Negative Binomial Regression

**Figure 7 offers a parallel visualization based on predictions from the Negative Binomial regression model. This model is often used as a refinement to the Poisson model when the assumption of equal mean and variance in count data (equidispersion) is violated. In this case, the Negative Binomial formulation accommodates** overdispersion**—a common occurrence in real-world travel behavior data—by introducing an additional parameter that**

**adjusts for excessive variability in counts across locations.**

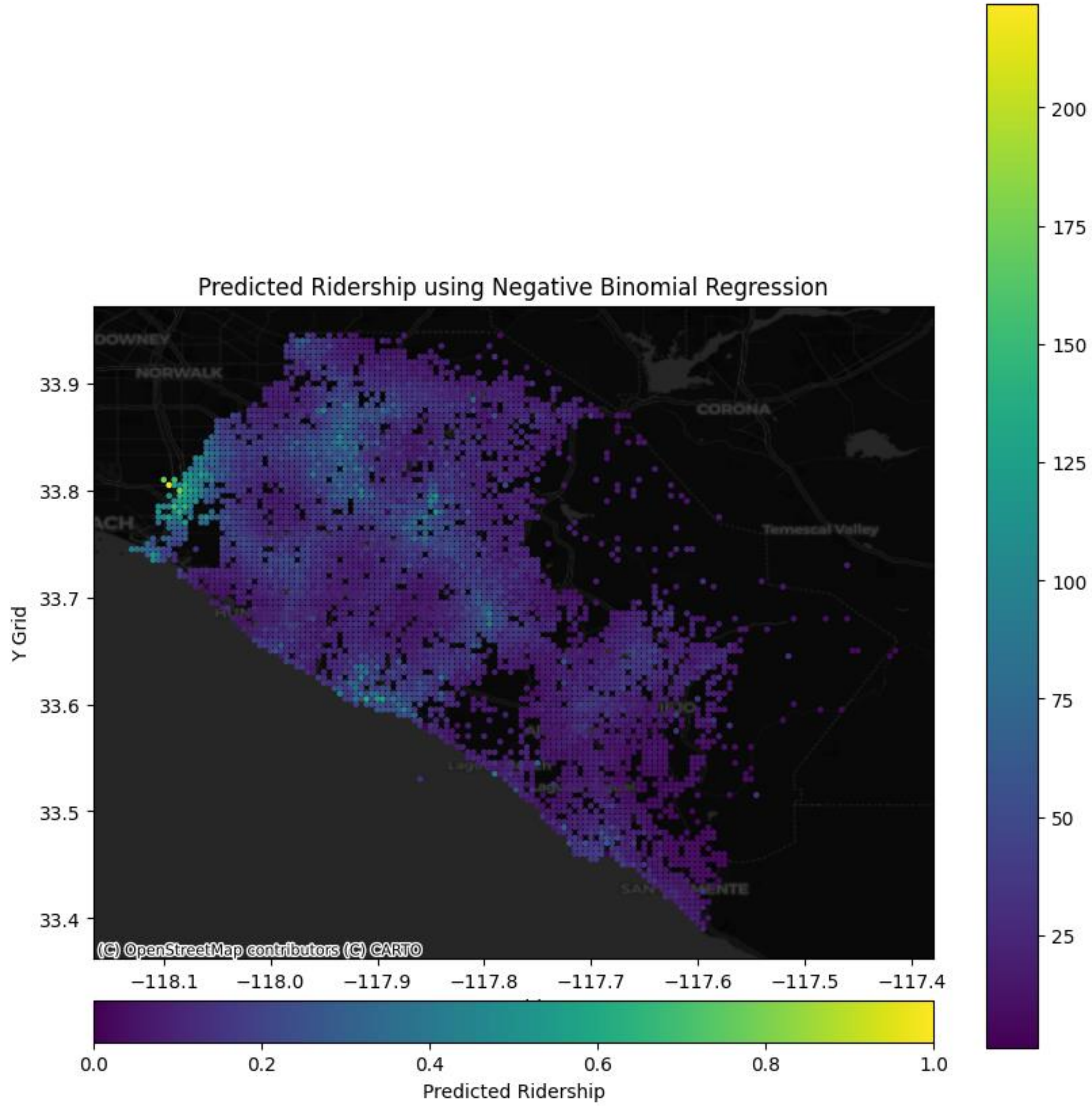


**Figure 7. Predicted Ridership Using Negative Binomial Regression**

The mapping approach mirrors that of Figure 6, using the same grid segmentation and color gradient for consistency and comparability. Notably, the Negative Binomial predictions tend to show **slightly wider distribution in mid-to-high ridership cells**, suggesting that this model may better capture variability in areas with fluctuating activity levels. While core high-ridership zones remain similar (e.g., coastal corridors), **suburban and transitional areas** display more nuanced differences in predicted values compared to the Poisson output.

### Comparative Insights and Interpretation

- Both maps highlight the **importance of coastal infrastructure and land-use mix** in shaping ridership patterns.
- The **Poisson model** produces sharper contrasts between high and low activity areas, likely due to its restrictive assumption on variance.
- The **Negative Binomial model**, in contrast, results in **smoother transitions and broader mid-level ridership distributions**, reflecting more flexible handling of variability.
- Areas with *moderate but volatile activity*—such as mixed-use developments, peri-urban corridors, or secondary green spaces—may benefit from the enhanced accuracy of the Negative Binomial approach.

These figures jointly provide not only validation of the models' spatial logic but also a **policy-relevant visualization** of where investments in bike infrastructure, safety improvements, or outreach initiatives may yield the highest return. Moreover, by presenting both model outputs side-by-side, stakeholders are equipped to make **evidence-based decisions** that account for uncertainty and model sensitivity.

## DISCUSSION

### LASSO Regression

The LASSO regression results, as summarized in Table 2 and visualized in Figure 2, offer a focused view of the most influential factors shaping bicycle ridership in Orange County. Among these, environmental and demographic variables stand out: "Distance from Green Space," "Non-White Population," "Total Strava Riders," and "Proximity to Sea Shore" were identified as high-impact predictors. The coefficients generated by the model provide a clear signal of direction and magnitude—"Proximity to Sea Shore" in particular emerged with the largest effect, reinforcing the importance of coastal access in supporting active transportation. The ability of LASSO to simultaneously perform feature selection and control for multicollinearity ensures that these results are statistically robust and policy-relevant.

### Poisson Regression Analysis

As outlined in Table 3 and depicted in Figure 3, the Poisson regression model builds upon LASSO outputs by estimating bias-corrected AADT (Annual Average Daily Traffic) counts using a subset of refined covariates. Key findings indicate strong positive associations with variables such as "Proximity to Sea Shore" and "Distance from Green Space," while negative associations are evident with "Speed Limits" and "Distance from Residential Area." Despite a relatively low R-squared value of 0.074, the model captures meaningful trends across space and demographics. The modest fit also reflects the complexity of ridership behavior, which is influenced by latent factors not fully captured in the current dataset.

### Predicted AADT Counts for Each Variable

Table 4 interprets each Poisson coefficient as an effect multiplier on AADT values, allowing for intuitive understanding of the relative strength of predictors. For instance, reducing the distance to green space results in a 56% increase in predicted ridership, while an increase in speed limits leads to a 17% decrease. These findings underscore how targeted environmental modifications can yield outsized returns in active transportation. Notably, social equity dimensions such as race and income show statistically significant, though more nuanced, impacts—demonstrating the intersection of infrastructure and demographic realities.

### Mapping Predicted Bicycle AADT

Figure 4 spatially visualizes predicted ridership patterns using the bias-corrected model output. By categorizing AADT values into five levels—Very Low to Very High—the map reveals geographic clusters of high activity, particularly near coastal and park-adjacent zones. This classification supports strategic investment planning by highlighting not just where ridership is already high, but also where it can grow with infrastructure support.

### Prediction Accuracy of the Bias Correction Model in Orange County

Figure 5 evaluates how accurately the model estimates observed AADT values across 38,967 segments. The cumulative distribution plot shows that 87% of predictions fall within acceptable margins of error. The presence of a long-tailed curve indicates that while a majority of predictions are precise, a small percentage of outliers still exist—an expected outcome in complex spatial models. This analysis confirms the reliability of the model while offering direction for future refinement, such as handling overdispersion or incorporating additional temporal dynamics.

### Predicted Ridership Using Poisson and Negative Binomial Regression

Figures 6 and 7 present predicted ridership maps using Poisson and Negative Binomial models, respectively. While both models highlight high ridership in coastal zones, the Negative Binomial regression accommodates overdispersion more effectively, yielding smoother distributions across moderate activity zones. This dual-model comparison enhances confidence in the findings and provides urban planners with nuanced spatial tools to guide investments in connectivity, safety, and access.

## CONCLUSION

This study demonstrates the potential of integrating crowdsourced data, such as that from Strava Metro, with official count data from the Orange County Transportation Authority (OCTA) to produce reliable, bias-corrected estimates of bicycle ridership. By employing advanced statistical modeling techniques—most notably LASSO regression for variable selection and Poisson/Negative Binomial regression for count prediction—this research addresses long-standing challenges related to the representativeness and usability of big data in transportation planning.

Key findings underscore the outsized influence of environmental and demographic variables—particularly proximity to green space, sea shore, and racial composition—on ridership patterns. These results highlight the importance of designing infrastructure that not only supports recreational use but also meets the mobility needs of diverse communities.

From a methodological standpoint, the study validates the value of spatial modeling across large urban areas using bias correction techniques. The Poisson model provides a foundational framework, while the Negative Binomial model improves predictive fidelity in overdispersed contexts. The predictive maps and variable interpretation tables offer direct, actionable insight for local agencies seeking to expand or enhance bicycling networks.

### Policy and Planning Implications

- **Data-Driven Investment:** Results can guide infrastructure investments in areas with high predicted but low observed ridership, particularly in inland and underserved communities.
- **Equity Planning:** The influence of demographic predictors suggests a need for inclusive design and targeted outreach in historically marginalized neighborhoods.
- **Dynamic Monitoring:** The scalability of this approach enables its use across time periods, supporting dynamic evaluation of policy impacts and infrastructure upgrades.

### Future Research Directions

- Incorporating **temporal dimensions** (hour of day, seasonality) into prediction models
- Using **real-time or longitudinal crowdsourced data streams**
- Testing **machine learning methods** for nonlinear interactions and spatial autocorrelation
- Expanding to other urban and suburban regions for **comparative analysis**

## ACKNOWLEDGEMENTS

This study was conducted with generous funding support from the California Department of Transportation (CalTrans). The authors extend their gratitude for the resources and partnership that made this research possible.

## AUTHOR CONTRIBUTIONS

The authors confirm the following contributions to the study:

- **Study design and concept:** Ghangshin Lee, Avipsa Roy
- **Data collection and preparation:** Avipsa Roy
- **Data analysis and interpretation:** Ghangshin Lee, Avipsa Roy
- **Draft manuscript writing:** Ghangshin Lee
- **Funding acquisition and project oversight:** Avipsa Roy
- **Final manuscript review and approval:** Both authors

# REFERENCES

1. Lee, Sener, & Mullins. "Bicycling Data Collection Using Fitness Apps: Potentials and Challenges." *Transport Reviews*, 2016.
2. Lee & Sener. "The Role of Big Data in Bicycle Research: Opportunities and Limitations." *Transport Reviews*, 2020.
3. Strava Metro. "Strava Metro Data Overview." Strava, 2019.
4. Nelson, T. et al. "Crowdsourced Data for Bicycling Research and Practice." *Transport Reviews*, 2021. https://doi.org/10.1080/01441647.2020.1798558
5. Roy, A. et al. "Correcting Bias in Crowdsourced Data to Map Bicycle Ridership of All Bicyclists." *Urban Science*, 2019. https://doi.org/10.3390/urbansci3020062
6. Nelson, T. et al. "Generalized Model for Mapping Bicycle Ridership with Crowdsourced Data." *Transportation Research Part C*, 2021. https://doi.org/10.1016/j.trc.2021.102981
7. Roy, A., Nelson, T., Turaga, P. "Functional Data Analysis Approach for Mapping Change in Time Series: Bicycle Ridership." *Transportation Research Interdisciplinary Perspectives*, 2023. https://doi.org/10.1016/j.trip.2022.100752
8. Fischer, J., Nelson, T., Winters, M. "A Street-Specific Analysis of Level of Traffic Stress and Strava Ridership." *Findings*, 2022.
9. Beck, B. et al. "Developing Urban Biking Typologies." *Environment and Planning B*, 2023. https://doi.org/10.1177/23998083221100827
10. Nelson, T. et al. "Developing a National Dataset of Bicycle Infrastructure for Canada Using Open Data." 2023.
11. Cohen, A., Nelson, T., Winters, M. "The Impact of Bicycle Theft on Ridership Behavior." 2023.
12. U.S. Census Bureau. "American Community Survey 5-Year Data (2009–2020)."