DRISI
CALTRANS DIVISION OF RESEARCH,
INNOVATION AND SYSTEM INFORMATION
TRANSFORMING IDEAS INTO SOLUTIONS

Research Notes

Developing a data fusion framework to map active transportation usage patterns in Orange County

To better understand active transportation (AT) usage with corrected bias in crowdsourced data collected through a GPS-enabled smartphone applications (apps)

Planning, Policy and Programming

NOVEMBER 2023

Project Title:
Developing a data fusion framework to map active transportation usage patterns in Orange County

Task Number: 3469

Start Date: July 1, 2023

Completion Date: June 30, 2024

Task Manager:
Frank Law
Senior Transportation Planner
frank.law@dot.ca.gov

WHAT IS THE NEED?

The recent emergence of GPS-enabled smartphone applications (apps) has facilitated a vast increase in capturing human movement data that is changing the landscape of Active Transportation (AT) research. These types of smart device captured data are emerging as an important data source for evidence-based city planning, decision-making, and more efficient provision of urban services.

However, the bias in crowdsourced data collected through apps leads to a major hindrance in utilizing the data directly into AI-driven transportation planning research. It also can eventually lead to inequitable infrastructure/policy decisions made by practitioners.

WHAT ARE WE DOING?

This research aims to directly address the fundamental aspects related to sampling bias in crowdsourced data at all stages of the Artificial Intelligence(AI) pipeline. The principal investigator will build on the research approach developed for Tempe, AZ. The investigative approach is comprised of three phases that align with the overall goals: 1) Using GIS and statistical models, to build a geographically-informed model to correct bias in Strava data; 2) Predict total ridership on all street segments for Irvine, Newport Beach, Anaheim, Santa Ana and Costa Mesa from ground truth data (i.e., official counts provided by the cities) and then evaluate model prediction accuracy using a separate test dataset; and 3) The model will generalize trends in bias correction factors for street segments in different geographic regions with varying built-environment, spatial, and socio-economic characteristics.

Caltrans
DRISI provides solutions and knowledge that improves California's transportation system

WHAT IS OUR GOAL?

The goal of the research is to create a set of adjustment factors accounting for the built environment, socio-economic, and land-use characteristics which can be applied to crowdsourced data so that policymakers and transportation practitioners across the United States can begin to incorporate exposure estimates more reliably and consistently into their safety, infrastructure planning, and decision-making analysis.

WHAT IS THE BENEFIT?

This research will provide insights into bicycling patterns that may be more broadly applicable, such as geographic and sociodemographic variables that consistently impact bicycling volumes in certain contexts or on certain street types regardless of context. These insights will be useful irrespective of whether a community has crowdsourced data or an established counting program. It will also highlight aspects of transportation equity that are often not well captured in count programs conducted by local authorities. The underrepresented communities which are often left out of planning decisions will be accounted for in the modeling framework by means of additional data acquired from US Census Bureaus' American Community Survey.

WHAT IS THE PROGRESS TO DATE?

The research team completed data preparation, including, but not limited to, data collection, data cleaning, and data visualization. The research team also finalized the bias-related modeling for Orange County using Strava and Orange County Transportation Authority 2018 Annual Bicycle Ridership Counts based on a regression model in which Annual Ridership Counts reached the highest adjusted R-square compared to hours, minutes, and days.

Preliminary results from the regression analysis indicate that the use of selected built environments and socio-economic factors have a sizable ability to correct bias, where the mean absolute error for the test dataset is 35.91 and the R-squared for the test set is 0.82.

The following are planned for the next quarter:

More variables will be added to the bias-related model next quarter. The selection of variables that are critical to Orange County is important to improve the model. The impact of spatial effects will also be investigated to see if it affects model performance.

The contents of this document reflect the views of the authors, who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the California Department of Transportation, the State of California, or the Federal Highway Administration. This document does not constitute a standard, specification, or regulation. No part of this publication should be construed as an endorsement for a commercial product, manufacturer, contractor, or consultant. Any trade names or photos of commercial products appearing in this document are for clarity only.