Geophysical Research Abstracts
Vol. 20, EGU2018-10983-3, 2018
EGU General Assembly 2018

Bias correction in geolocated crowdsourced data from Strava using Machine Learning based linear models

Avipsa Roy and Trisalyn Nelson
Arizona State University, Tempe, USA

The recent upheaval in geolocated crowdsourced data from mobile applications used by cyclists has opened a new horizon for active transportation research. It has, therefore, become essential to make sense of such high volumes of crowdsourced data with reference to active transportation. The challenge with crowdsourced data being incorporated into statistical models is its quality and accuracy in terms of predicting volumes of bike ridership at peak hours during the week.There have been previous studies [Jestico et al.] to determine if crowdsourced data obtained from fitness mobile applications are convenient for mapping bike ridership in urban areas across consistent spatial and temporal scales. However, the predicted ridership was categorized into very coarse scales based on cycling volumes. Our research goal is to quantify the bias if any in crowdsourced geolocated data and further refine the classification categories of bike ridership spanning across Maricopa County in the state of Arizona, USA. We find a suitable attribute resolution for incorporating big crowdsourced data in predictive machine learning models to estimate bike ridership volumes in urban areas. Using big data provided by Strava.com of exactly 99,784,569 riders at intersections and 68,469,637 riders along streets spanning the whole of Maricopa County, we try to build a multi-level classification model for predicting bike volumes across three cities Phoenix, Tempe, and Mesa at peak hours during the day. We use official bike counts from 44 locations across Maricopa County provided by the Maricopa Association of Governments as our ground truth data for comparison. Initial results have shown the R-squared values are in the range of 0.4 to 0.6 depending upon the test and training ratio variations between 20 to 60 percent. We also combine our analyses with socio-economic data from the Maricopa county to study how income groups, ethnicity, and proximity to urban facilities (eg: schools, convenience stores, markets etc.) affect the ridership patterns. The results would prove effective for urban planners and transportation agencies in terms of building and by expanding effective bicycle infrastructure network, identifying the high intensity of active commuters within specific regions of the cities to promote safe bike ridership and also prevent accidents with adequate monitoring at street intersections.

Reference:
Jestico, Ben, Trisalyn Nelson, and Meghan Winters. "Mapping ridership using crowdsourced cycling data." Journal of transport geography 52 (2016): 90-97.