Estimating Intersection Pedestrian Volume from Built Environment Characteristics in Los Angeles, CA

Domain Background

In order for cities to plan transportation systems, they need to have robust data on the number of people using different modes of transport. This is especially important when the urban landscape changes, such as an increase in urban density, and cities need to plan for the changes in traffic that will arrive. Typically, cities collect data on the volume of vehicle, bicycle, and pedestrian traffic. Unfortunately, collecting these data is costly and time consuming, and in some cases impossible. Often transportation engineers and planners need to understand the consequences of different design decisions, such as the estimated effect of a traffic signal on safety. Since it is impossible to build out these alternatives and then perform volume counts, engineers and planners need a way to be able to estimate the effect of changes on pedestrian volume.

Although many cities have models to estimate the effect of the built environment on pedestrian volume, they are typically at the scale of traffic analysis zones, a geographic scale much larger compared to the intersection. Intersection-based models exist for the following locations: San Francisco, CA (1,2); Charlotte, NC (3); Alameda County, CA (4); San Diego County, CA (5); Santa Monica, CA (6); and Quebec (7). Most of these intersection-based models use either a linear or log-linear model, and the most common features found to significantly affect pedestrian volumes include population density, employment density, and transit accessibility.

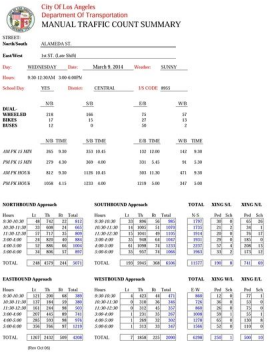
Despite general agreement on the most important features, there are differences among the models on other significant features from the built environment. For example, the City of Santa Monica found the distance from the ocean to be a significant variable in prediction (6); it is highly unlikely that a landlocked city would find the significance in that variable. Even when the models agree on which features of the built environment are significant predictors, they often disagree on the extent to which they influence pedestrian volume. As suggested by Schneider et al., this variation should be addressed by creating models that are sensitive to the context of the local environment (2). Since there currently does not exist a model to predict pedestrian intersection volumes for the City of Los Angeles, my machine learning project aims to fill that gap.

Problem Statement

For this project, my goal is to answer the following statement: What is the relationship between the built environment and the daily pedestrian volumes at intersections in the City of Los Angeles?

Datasets and Inputs

The Los Angeles Department of Transportation (LADOT) routinely collects volume data related to bicyclists, pedestrians, and motor vehicles for the purposes of transportation planning. Historically, these data have been stored in a PDF format, which makes it easy to digest a single traffic count, but prevents comparison and analysis of multiple counts. These PDF files are publicly available on the *Navigate LA* portal at http://navigatela.lacity.org/navigatela/.


City Of Los Angeles
Department Of Transportation
MANUAL TRAFFIC COUNT SUMMARY

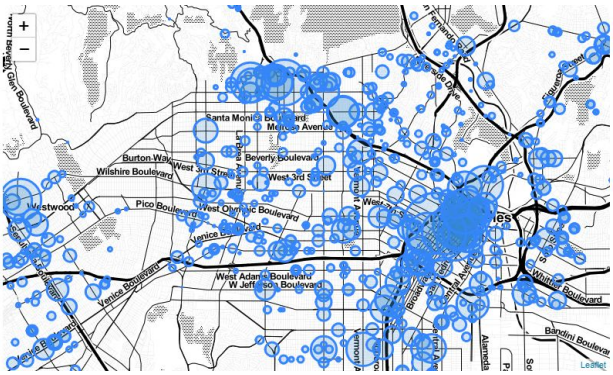


Figure 1. Example of LADOT Pedestrian Volume Count PDF

Figure 2. Map of Pedestrian Volumes, After Extracting from PDF

The first step to solving the proposed problem includes extracting pedestrian volume from these sheets and storing it in a format that could be used in a model. I developed a python script to read the pedestrian volume data from the PDF sheets and format them so they can be used in building the model. The daily volume (what I will be constructing a model to predict), is represented by the 'volume' attribute in the table. Each row is a separate count event ('count_id'), which occurs at an individual intersection ('ASSETID', 'cl_node_id').

After extracting pedestrian volume data from these sheets, I assembled data from the built environment that I thought could possibly be significant in predicting pedestrian volume at intersections in Los Angeles. These data are public, but were assembled for a previous project at LADOT. I looked to the literature review to inform the types of data to collect for evaluating. My built environment data (explanatory variables) include:

- Population within 0.25 mi. ('SUM_POPTTL')
- Employment within 0.25 mi. ('EMPTOT')
- Count of Schools within 0.25 mi. ('SCH_CT')
- Presence of a traffic signal ('SIG', 1 = yes, 0 = no)
- Count of transit stops within 100 ft. ('TRANSITSTOP')
- Transit Ridership ('RIDERSHIP')

Solution Statement

My proposed solution is to build a regression model that can take inputs from the built environment and predict the daily volume of pedestrians at an intersection. I will test the fit for both a linear model and log linear model, since both have been demonstrated to perform well for different cities. For my project, I prefer implementing a linear or log linear regression model because it is important to be able to describe the effect of each explanatory variable on the output. It is not just important to be able to accurately predict the pedestrian volume; it is also important to understand how the characteristics of the built environment affect the volume. I anticipate my resulting model to take one of two forms shown below:

(1) $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_j X_{ji}$

(2) $Y_i = exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_j X_{ji})$

where:

Y_i = weekday pedestrian volume at intersection

X_{ji} = value of explanatory variable j at intersection i

β_j = model coefficient for variable j

Benchmark Model

There have been a few attempts to build models estimating the effect of the built environment on pedestrian traffic volume, but none of them have focused on Los Angeles. In addition, most have also used a sample set smaller than my own. Below is a table with the results from a previous model for San Francisco (1):

Model Structure	Adjusted R^2 Value	F-Value (Test Value)
Log-Linear	0.804	34.4 ($p < .001$)

Evaluation Metrics

My proposed metric for evaluating my linear model is the r-squared score, the proportion of the variance in the dependent variable that is predictable from the independent variable. I will also compute the F-Value to ensure that the regression model provides a better fit to the data than a model that contains no independent variables.

Project Design

I've already completed much of the data preparation for this project. Since I am starting this project with a data table, I will begin by calculating some basic summary statistics, which may suggest whether any of the explanatory variables need to be transformed (such as rescaled, etc.). It can also give me a sense for how the intersections themselves are different. I will build at least three different models to predict the pedestrian volume from the explanatory variables: linear regression, log-linear regression, and decision tree regressor. I prefer to have a linear or log-linear regression model, because the output of the model is easily interpreted, but I also want to compare it to the performance of the decision tree regressor.

References

(1) Schneider, R. J., T. Henry, M. F. Mitman, L. Stonehill, and J. Koehler. Development and Application of a Pedestrian Volume Model in San Francisco. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2299, 2012, pp. 65–78.

(2) Liu, X., and J. Griswold. Pedestrian Volume Modeling: A Case Study of San Francisco. *Association of Pacific Coast Geographers Yearbook*, Vol. 71, 2009.

(3) Pulugurtha, S. S., and S. R. Repaka. Assessment of Models to Measure Pedestrian Activity at Signalized Intersections. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2073*, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 39-48.

(4) Schneider, R. J., L. S. Arnold, and D. R. Ragland. Pilot Model for Estimating Pedestrian Intersection Crossing Volumes. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2140*, Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 13-26.

(5) Jones, M. G., S. Ryan, J. Donlan, L. Ledbetter, L. Arnold, and D. Ragland. *Seamless Travel: Measuring Bicycle and Pedestrian Activity in San Diego County and Its Relationship to Land Use, Transportation, Safety, and Facility Type.* Alta Planning and Design and Safe Transportation Research and Education Center, University of California, Berkeley, 2010.

(6) Haynes, M., and S. Andrzejewski. *GIS Based Bicycle & Pedestrian Demand Forecasting Techniques.* Presentation to Travel Model Improvement Program, U.S. Department of Transportation. Fehr & Peers Transportation Consultants, San Francisco, Calif., 2010.

(7) Miranda-Moreno, L. F., and D. Fernandes. Modeling of Pedestrian Activity at Signalized Intersections: Land Use, Urban Form, Weather, and Spatiotemporal Patterns. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2264*, Transportation Research Board of the National Academies, Washington, D.C., 2011, pp. 74-82.