Centre for
Transport Studies
STOCKHOLM

Comparison of pedestrian trip generation models

Kim Nam Seok – Hanyang University (South Korea)
Yusak O. Susilo – Royal Institute of Technology (KTH)

CTS Working Paper 2013:22

Abstract
Using Poisson regression and negative binomial regression, this paper presents an empirical comparison of four different regression models for the estimation of pedestrian demand at the regional level and finds the most appropriate model with reference to the National Household Travel Survey (NHTS) 2001 data for the Baltimore (USA) region. The results show that Poisson regression seems to be more appropriate for pedestrian trip generation modeling in terms of x2 ratio test, Pseudo R2, and Akaike's information criterion (AIC). However, R2 based on deviance residuals and estimated log-likelihood value at convergence confirmed the empirical studies that negative binomial regression is more appropriate for the over-dispersed dependent variable than Poisson regression.

Keywords: pedestrian; trip generation; Poisson; negative binomial; regression

JEL Codes: R41, O18

Published as: Kim, N.S. and Susilo, Y.O. (2013) Comparison of pedestrian trip generation models. Journal of Advanced Transportation, 47, pp. 399–412.

Centre for Transport Studies
SE-100 44 Stockholm
Sweden
www.cts.kth.se

COMPARISON OF PEDESTRIAN TRIP GENERATION MODELS

Nam Seok Kim[1,2*] and Yusak O. Susilo[3]

[1] *The Korea Transport Institute, Goyang-si, Gyeonggi-do, Republic of Korea*
[2] *Department of Transport and Infrastructure, Delft University of Technology, Delft, The Netherlands*
[3]*Centre for Transport and Society, University of the West of England, Bristol, United Kingdom*

Published as: Kim, N.S. and Susilo, Y.O. (2013) Comparison of pedestrian trip generation models. Journal of Advanced Transportation, 47, pp. 399–412.

ABSTRACT

Using Poisson regression and Negative Binomial regression, this paper presents an empirical comparison of four different regression models for the estimation of pedestrian demand at the regional level and finds the most appropriate model with reference to the National Household Travel Survey 2001 data for the Baltimore (USA) region. The results show that Poisson regression seems to be more appropriate for pedestrian trip generation modeling in terms of Chi^2 ratio test, Pseudo R^2 and Akaike's Information Criterion (AIC). However, R^2 based on Deviance residuals and estimated Log-likelihood value at convergence confirmed the empirical studies that negative binomial regression is more appropriate for the over-dispersed dependent variable than Poisson regression.

KEY WORDS: Pedestrian, Trip generation, Poisson, Negative binomial, Regression

1. INTRODUCTION

Since non-motorized transportation has both social and individual benefits, governments, urban planners, and social environmental activists have been actively looking for appropriate ways of encouraging non-motorized travel (FHWA, 1994, FHWA, 1999a, FHWA, 1999b). One of the prime requirements in this field is a reliable method of identifying and estimating the demand for non-motorized travel. However, there is not a standard technique available for this purpose yet. This is likely a result of most research activities having been focused on motorized transportation modes such as private car and public transportation since transport modeling was first developed in the 1950s (Bate, 2000). Empirical studies of pedestrian demand modeling show various approaches depending on the

[*] Correspondence to: Nam Seok Kim, The Korea Transport Institute, Gyeonggi-do, Republic of Korea
E-mails: nskim@koti.re.kr; n.s.kim@tudelft.nl

size of research area (Behnam and Patel, 1997, Ercolano, et al., 1997, Kim, 2005, Pushkarev and Jeffrey, 1971, Targa and Clifton, 2005). Through reference to available literature, this paper will (a) trace the development of pedestrian demand modeling techniques in the United States, (b) construct various pedestrian trip generation models for the Baltimore region using National Household Travel Survey (NHTS) data base on different specifications, and (c) compare the results obtained in order to find the most appropriate model for the description of pedestrian trips. In particular, the merits of two regression techniques that have recently been applied to pedestrian demand modeling – the Poisson regression model (PRM) and the negative binomial regression model (NBRM) – are compared with reference to a review of the literature and evaluation of modeling results. A set of evaluation measures of goodness of fit was developed to facilitate comparison of the different models. The modeling results were used to estimate the influence of the built environment and socio-economic factors on “walking to work” behavior in Baltimore region.

The rest of this paper is organized as follows. In the next the section, a literature review of previous pedestrian trip generation studies in the United State is presented. Then, data and theories are described focusing on Poisson and Negative binomial regression model. In the section on results and analyses, four types of regression models for the same dependent variable (i.e. walking to work) and the same independent variables are developed and compared. Finally, conclusions are drawn.

2. LITERATURE REVIEW

The ‘Bicycle and Pedestrian Trip Generation Workshop’(FHWA, 1997) and the ‘Guidebook on Methods to Estimate Non-Motorized Travel’ (FHWA, 1999a, FHWA, 1999b) both describe the “sketch plan method” among several approaches to pedestrian volume estimation. The sketch plan method is originally described as “methods generally use pedestrian counts and regression analysis to predict pedestrian volumes as a function of adjacent land uses and indicators of transportation trip generation” (FHWA, 1999b). This method was developed to permit quick estimation of pedestrian demand under existing and future conditions. Table 1 summarizes the features of the selected studies in terms of level of study area, observation frequency, data requirements, and estimation methods used.

Pushkarev and Zupan (1971) and Behnam and Patel (1997), for example, estimated pedestrian demand in areas of high population density using existing land-use data and pedestrian counts. They counted the number of pedestrians and surveyed the characteristics of their trips, including trip times and distances. Pushkarev and Zupan (1971) used linear regression analysis to predict total pedestrian volumes per block. Explanatory variables included commercial land uses, distance to transit stops, and the presence or absence and condition of sidewalks. Behnam and Patel (1997) used a similar regression model. Both groups of researchers used pedestrian volume per hour per block as the dependent variable in the regression equation. The independent variables included commercial space, office space, cultural and entertainment space, manufacturing space, residential space, parking space, vacant space, and storage and maintenance space. Furthermore, based on future land use variables, Behnam and Patel (1997) predicted future pedestrian volumes in Milwaukee's central business district (CBD). Ercolano et al. (1997) used peak vehicles per hour, transit-ridership, and non-motorized mode share to estimate the pedestrian travel demand at the peak hour in suburban areas. Moreover, they used the estimated pedestrian travel demand at the peak hour to determine the location of pedestrian crossings, sidewalks, and signal re-timings. Matlick (1996) also modeled the pedestrian demand in terms of household population, transportation mode share, and activity center data. Matlick's model has been used to determine the priority areas or corridors for improvement of pedestrian facilities. The above-mentioned publications show that pedestrian demand at block and corridor level can be estimated using either linear regression or simple calculations. Other regression techniques have recently also been applied to pedestrian modeling. Kim (2005) studied the feasibility of the use of non-linear (Poisson) regression for estimating walking demand at the macroscopic level, while Cao et al. (2006) introduced negative binomial regression (NBRM) in pedestrian demand modeling.

Table 1 Features of Sketch-Plan Model

Researchers	Level of Study Area	Observation frequency	Data requirements		Estimation technique
			Pedestrian volume	Land-use and socio-economic data	
Pushkarev and Zupan, 1971	Block (Midtown Manhattan)	Hourly	Pedestrian counts (aerial photography)	Square mile of office, retail, and restaurant space	Linear regression
Behnam and Patel, 1977	Block (CBD of Milwaukee, WI)	Hourly (extrapolated from 6-minute counts)	Pedestrian counts (real counts)	Commercial space, Office space Cultural and entertainment space, Manufacturing space, Residential space, Parking space Vacant space, Storage and maintenance space	Linear regression
Davis et al., 1991	Crosswalk level (Washington D.C)	5- to 10-minute time segments during peak hours	Pedestrian counts (real counts)	Vehicle traffic counts	Relationship between vehicle and pedestrian counts
Matlick, 1996	Corridor-level (Seattle, WA)	Daily	Transportation mode share information (Census) / National Personal Travel Survey (NPTS)	Housing types, density, persons per household unit, and hotels Retail, recreation, social facilities, schools, employment, and churches.	Linear regression
Ercolano et al. 1997	City level (Plattsburgh, New York)	Hourly (peak hour)	Vehicles per hour from traffic counts and mode share from Census	Vehicle traffic counts	Computation using spreadsheets
Targa and Clifton, 2005	City level (Baltimore City, MD)	One day	Number of walk trips from NHTS 2001	Car ownership in household, type of housing unit, household income, age, sex, driver status, educational status, attitudes/ perceptions of pedestrians, household density, street connectivity, land-use diversity, proportion of commercial units	Poisson regression
Kim, 2005	Metropolitan Level (6 Counties and 1 City in Baltimore Metropolitan region)	One day	Number of walk trips from NHTS 2001	Age, driver Status, education level, income, race, percentage of adult drivers in household, non-residential density (tract level), road density within ¼ mile, mixed land use (tract level)	Poisson regression
Cao et al, 2006	Town level (6 neighborhoods in Austin, TX)	30 Days	Number of pedestrians derived from a self-administered survey mailed in 1995 (Handy et al., 1998)	Major stores within walking distance, traffic volume, pedestrian connections, perception of stores, perception of walk advantage, perception of walk comfort, perception of traffic, miles to the nearest store, sex, age, worker status, presence of children, household income	Negative Binomial regression

Table 1 (Cont.) Features of Sketch-Plan Model

Researchers	Level of Study Area	Observation frequency	Data requirements		Estimation technique
			Pedestrian volume	Land-use and socio-economic data	
(Shay, et al., 2006)	Town level (Southern Village in Chapel Hill, NC)	One day	Number of walking trips from travel diary	Sex, age, number of children, number of cars/household, number of licensed drivers per car, walking is enjoyable*, environmental protection is important*, value shops and services close by*, distance from home to activity center. * Scale based variable: 1 = strongly disagree, 5 = strongly agree.	Negative binomial regression
(Pulugurtha and Repaka, 2008)	Intersection level (Charlotte, NC)	12 hours (extrapolated from hourly volumes)	Pedestrian counts (real counts)	Household units, population, total employment, urban residential area, neighborhood business, mixed land use, transit stops, speed limit, vehicular volume: All variables above are captured within 1/4, 1/2, and 1 mile buffers.	Linear regressions
(Baran, et al., 2008)	Town level (a New Urbanist community and conventional suburban neighbourhood)	One day	Number of walking trips from NHTS 2001: either leisure or utilitarian walk trips	Age, gender, household size, vehicles per household and respondent's occupational status, two space syntax variables (global integration, local integration, and control variable)	Poisson Regression and Negative binomial regression
(Schneider, et al., 2009)	Intersection level (Alameda County, CA)	Weekly (extrapolated from 2 hours volumes with distinction of weekdays and Sundays)	Pedestrian counts (real counts)	Total population, total employment, proportion of housing units (either vacant or rented), number of housing units (either vacant or rented), number of commercial properties, number of elementary/middle/high schools& colleges, number of transit stations(bus, rail), sidewalk coverage, freeway presence, total street centerline distance, race (white), car ownership, income, age (categorical variable): All variables above are captured within both 1/10 and 1/4 mile buffers. Level of traffic, number of lanes, crosswalks, bicycle lanes, traffic signal, and curb radius	Linear regression

The above-mentioned studies, with the exception of Cao et al. (2006), may be regarded as applications of the sketch-plan method based on empirical examination of the relationship between pedestrian demand, the characteristics of the built environment, and socio-economic variables. This approach is not unlike that used in pedestrian trip generation. The purposes are slightly different in the two cases, however. The sketch-plan method is used to estimate the number of people walking either at present or in the future as mentioned previously, while the aim of pedestrian trip generation consists of identifying the relationship between pedestrian trip demand (i.e. trip generation rate) and other factors, finding critical factors both positively and negatively affecting pedestrian trips, and eventually estimating pedestrian trips. Thus, empirical studies can be a preliminary step towards a trip generation model. There have been several empirical studies examining the effect of land-use and socio-economic characteristics on pedestrian behavior. Levinson and Wynn (1963), for example, were among the first to investigate the impact of neighborhood characteristics on travel demand. They found that increasing neighborhood density is closely associated with a decrease in the frequency of private vehicle trips, and an increase in the frequency of the use of public transit and non-motorized trips. Ewing and Cervero (2001) summarized empirical findings and provided a synthesis of the relationship between travel and the characteristics of the built environment. The publications they reviewed show that walking trips are associated with transit-oriented neighborhoods, the distance between commercial districts and residential areas, higher population density, mixed land-use, and multi-story buildings. Targa and Clifton (2005) showed that lower vehicle ownership, college-dorm type accommodation, and lower household income are associated with higher walking frequency. In addition, higher urban population density, higher street connectivity, and more mixed land use generated more walking trips. They used the Poisson regression model (PRM), which assumes that the frequency of walking trips in a single day follows a Poisson distribution in their study and, as far as authors' aware of, it was the first attempt to use non-linear regression in trip generation study instead of the traditional linear regression. However, they do not sufficiently explain why PRM is preferred to linear regression and even compare their results with those from linear regression. The justification for using Poisson regression was to apply the nature of count-type data into pedestrian demand modeling and to show a reasonable result. As mentioned briefly above, Cao et al. (2006) used the negative binomial regression model (NBRM) to study the influence of the built environment and residential self-selection on pedestrian behavior. They argued that NBRM is much more suitable than PRM for studies of pedestrian behavior, since such behavior rarely satisfies the underlying assumption in PRM that the mean of the dependent variable is equal to its variance. Baran, et al.(2008)

confirmed that NBRM is superior to PRM in terms of explaining utilitarian walking behavior as well as the statistic goodness of fitness (i.e. Pseudo R^2). They, beyond adding one more empirical result that walking trips are related to the built environment, showed that Space syntax was useful to develop indicators to estimate pedestrian accessibility. Space syntax is a quantitative tool to extract characteristics of links such as sidewalks and streets, and furthermore to develop indicators such as levels of connectivity, of accessibility, and of integration (for more detailed information on Space syntax see Hillier (1998)). Recently, Pulugurtha and Repaka (2008) and Schneider et al. (2009) examines extensive buffer-based independent variables by using multiple liner regressions. Shay et al. (2006) included three attitude variables in a NBRM: ‘walking is enjoyable’, ‘environmental protection is important’, and ‘value shops and services close by’.

3. DATA AND METHODOLOGY

3.1. Data

As mentioned above, the sketch-plan model approach is based on the assumption that the walking frequency is a function of socio-economic factors and characteristics of the built environment. The National Household Travel Survey (NHTS) 2001 satisfies the data requirements of this study, since it provides not only information on trip frequency with trip purpose and transportation modes, but also socio-economic and land-use data at Census tract level. For the purposes of the survey, each member of the sample households records all trips for a designated 24-hour period known as the ‘Travel Day’. NHTS 2001 contains four kinds of data sets: household characteristics, personal characteristics, vehicle characteristics and travel information which are collected by interviews held from April 2001 through May 2002. It comprises 66,000 sample households, including 40,000 households from nine so-called “add-on areas”. These add-on areas include the Baltimore metropolitan region which was chosen as the study area because the sample size in this region permits modeling at the individual level. This area has approximately 2,500,000 population and is 2,256 square miles in size. NHTS 2001 Baltimore add-on includes 3,519 sample households generating 27,366 trips during the designated travel day. Since the data on built-environment factors in NHTS are too simple and limited for the purposes of our analysis, MD Property View 2001 and MD Transit View 2001 (both published and managed by Maryland Department of Planning, MDP) were used to generate more detailed land-use variables.

The frequencies of walking trips to work for respondents in a single day were taken from the travel dataset of NHTS 2001. Specifically, the home-based trips were classified in terms of the origin and destination of each trip and walking trips were then extracted in travel mode. Finally, the frequency of walking trips for the specified purposes (walks to work) was summarized with respect to the NHTS identification number of each person concerned and the provisos that all trips considered were home-based. The 3,915 persons in the sample made at least one commuting trip per day on average; less than 3% of these trips were on foot, however.

Socio-economic variables were also collected from NHTS 2001. Using personal identification numbers, the trip frequency for each person was assigned to personal and household characteristics given in NHTS 2001. The socio-economic variables used included household size, age, income, race, education, car ownership, and driver status. The break values (e.g. college graduation or lower educational level and more than US$ 40,000 annual income or less) were determined by simple correlation with the dependent variable: the chosen break points maximized the differences in walking trip frequency found between the two classes. As indicated above, land-use patterns might also affect walking trip frequencies. However, since the NHTS 2001 provides only very limited land-use variables such as population density and household size, more detailed information on land use and the built environment (e.g. floor space of single-family and multiple-family dwelling units, and extent of mixed land use) was taken from MD Property View 2001. The relationship between the sq. footage of dwellings/ non-dwellings and travel behavior has been demonstrated by empirical studies (Ewing and Cervero, 2001).

The road (sidewalk) density within ¼ mile of the home was calculated using GIS-based Census TIGER/Line 2001 (Matlick, 1996). More specifically, Census Feature Class Code (CFCC) (A10 to A18 in the code table) in TIGER/Line 2001 was used to distinguish the motorized traffic-only roads (i.e. interstate highways, arterial roads, local road without sidewalk) and remove them for calculation of sidewalk density. The above-mentioned variables are summarized in Table 2.

Table 2 Description of dependent and independent variables

Variable	Definition	Mean	Std. Dev.
Walk_Trips	Number of walking trips for commuting purposes per day (personal level)	0.06	0.36
Personal Characteristics			
Age	Age of respondents (years)	41.91	14.00
Sex	1 if male; otherwise 0	0.50	0.50
Driver	1 if driver; otherwise 0	0.92	0.27
Education	1, if less than college graduation; 0, otherwise	0.49	0.50
Household Characteristics			
Income	1 if less than US$ 40,000; otherwise 0	0.41	0.49
Race	1 if Caucasian; otherwise 0	0.76	0.40
Adult driver	Number of adult drivers in household	0.94	0.27
Vehicle number	Number of vehicles in household	2.08	1.18
Household size	Number of household members	2.73	1.28
No of driver per hh	Number of drivers in household	1.92	0.85
No of adult per hh	Number of adults in household	2.06	0.76
Employed hh	Number of household members in employment	1.87	0.76
Characteristics of built environment			
Residential density	Residential units / tract	23.71	19.75
Single family density	Single Family Dwelling Units per tract	2,119.00	3,262.00
Multiple family density	Multiple Family Dwelling Units per tract	21.23	64.97
Non-residential unit density	Non-residential units/ tract	220.48	487.28
Non-residential unit size	Total floor space of non-residential units per tract (sq. ft.)	7,334	27,110
Degree of urbanism	Non-residential units/ residential units	2.79	32.05
Bus stop density	# of bus stops per tract	29.08	44.99
Road density	Road length per tract (mi.)	23.71	19.75

3.2. Methodology

3.2.1. Theoretical Background

The trip generation phase in 4-step traditional transportation demand forecasting generally uses multiple linear regression. This approach assumes that the residuals follow a normal distribution. As shown in Figure 1 (a) and (b), however, the distributions of the walking frequency and the walking frequency to work are far from normal; it follows that the application of linear regression is not appropriate.

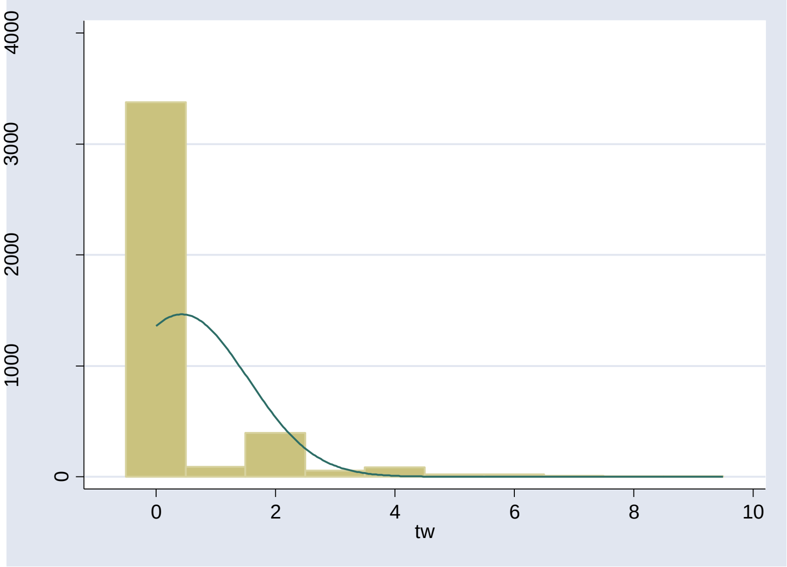

Figure 1(a) Walking Frequency regardless of trip purpose (mean 0.434, variance 1.117)

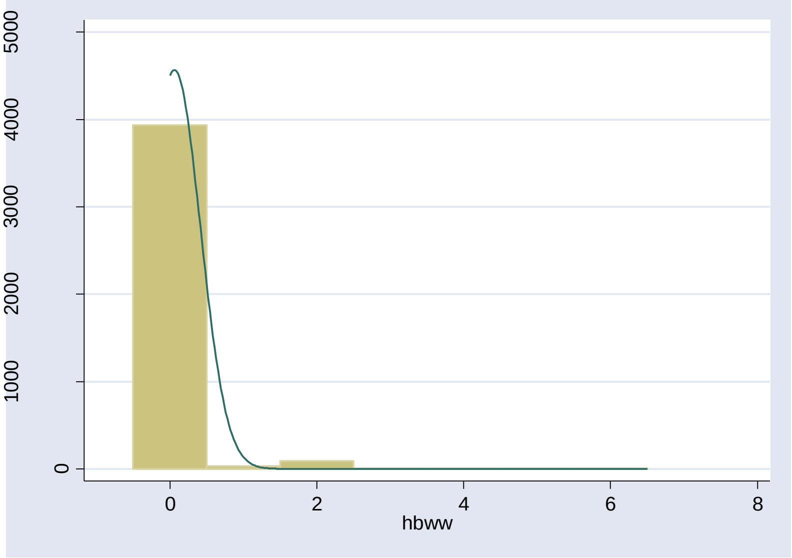

Figure 1(b) Walking Frequency to work (mean 0.060, variance 0.355)

The Poisson regression model (PRM) and the negative binomial regression model (NBRM) come into consideration as the common statistical techniques for the analysis of non-negative dependent variables (specifically, count data). The event in question should occur in a given observation period and/or a given observation space (DeMaris, 2004). In the present study, the time period is *1 day* and the space is *the Baltimore metropolitan region*. PRM assumes that the dependent variable, which follows a Poisson distribution with parameter μ_i, is controlled by independent variables (x_i). The density of the dependent variables is given by Equation (1) (Cameron and Trivedi, 1998).

$$f(y_i \mid x^i, \beta) = \frac{e^{-\mu_i}\mu_i^{\,y_i}}{y_i!} \tag{1}$$

where μ_i =E[y_i]= $\boldsymbol{e}^{\sum_{k=1}^{K}\beta_k X_{iK}}$, y_i = 0, 1, 2, …., n

μ_i is an exponential function of the covariates that is conditional on the covariates in each case. Equation (2) shows how μ_i can be written as a linear combination of the independent variables (X_i).

$$\mathrm{Ln}(\mu_i\ ) = \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} \tag{2}$$

It is further assumed that the mean of the dependent variable is equal to its variance (e.g. E[y_i]=VAR[y_i]= μ_i, where, E[y_i] is the expected value of dependent variable and VAR[y_i] is its variance). However, the data used in this study showed that the dependent variable ranged from 0 to 6 while the sample mean was 0.06 and the variance 0.355 – in other words, the dependent variable is over-dispersed (i.e. E[y_i]<VAR[y_i]). It would thus seem that Poisson regression is not the appropriate model for this purpose. Another seemingly inappropriate assumption is found in King (1989): "the independence assumption is that the probability of subsequence event is independent of the occurrence of a previous event"(DeMaris, 2004). The commuting trips are usually done by a transportation mode or a same set of combinations although it is not ensured. Since NBRM has no such restrictive properties (in fact, PRM may be regarded as a special case of NBRM (Cameron and Trivedi, 1998)), it would theoretically seem to be more appropriate for the explanation of walking behavior than Poisson regression. More specifically, the unobserved term in PRM is determined by independent variables while one in NBRM is not determined by independent variables but

allowed as just a disturbance term. The disturbance term has two densities: discrete and continuous. To specify the density of yi, the allowed random disturbance term should be described as in Equation (3).

$$\theta_i = e^{\sum_{k=1}^{K}\beta_k X_{iK} + \varepsilon_i} = \mu_i e^{\varepsilon_i} \tag{3}$$

Where, E[Yi]= θ_i , μ_i = $e^{\sum_{k=1}^{K}\beta_k X_{iK}}$, ε_i is disturbance terms (continuous)

To derive the likelihood function, the unobserved should be replaced to an appropriate distribution function. Thus, after assuming that e^{ε_i} is followed by Gamma distribution with parameter 1/ α , the marginal density of the dependent variables can be derived as shown in Equation (4) (Cameron and Trivedi, 1998)

$$f(y_i \mid x^i, \beta, \alpha) = \frac{\Gamma((1/\alpha) + y_i)}{\Gamma(1/\alpha)\Gamma(y_i!+1)} \left(\frac{1/\alpha}{(1/\alpha) + \mu_i} \right)^{1/\alpha} \left(\frac{\mu_i}{(1/\alpha) + \mu_i} \right)^{y_i} \tag{4}$$

Where, $\Gamma(.)$ is a gamma function with parameter 1/ α (α is called the over-dispersion parameter), μ_i =

$e^{\sum_{k=1}^{K}\beta_k X_{ik}}$ as in PRM, y_i = 0, 1, 2….,

While PRM only focuses on model estimation, two kinds of tests are required in NBRM. The first is to test over-dispersion of dependent variable: H0: α =0. Through this test, theoretically, NBRM is selected if the hypothesis is rejected (i.e. α >0) (DeMaris, 2004). Thus, the goodness of fit of NBRM for over-dispersed dependent variable should be greater than that of PRM. The second is to estimate coefficients in the model as PRM.

3.2.1. Evaluation measures of goodness of fit

The most general evaluation measure in count data regression models is the maximum likelihood (ML) estimate. The likelihood-ratio test is presented as follows (Hensher, et al., 2005a):

$$2(\mathrm{LL}_{\text{estimated model}} - \mathrm{LL}_{\text{base model}}) \sim \chi^2(_{\text{number of new parameters estimated in the estimated model}}) \tag{5}$$

Where, LL $_{\text{estimated model}}$ is the log-likelihood function of an estimated model and LL $_{\text{base model}}$ is the corresponding function for a base model (i.e. constant-only model). Equation (5) shows that 2 (LL $_{\text{estimated model}}$ –LL $_{\text{base mode}}$) is compared to a Chi-square statistic (critical value of Chi-square (χ^2)). The null hypothesis that the specified model is not statistically better than the base model is rejected if the LL ratio is larger than Chi-square value with degrees of freedom equal to the difference in the number of parameters between the two models.

Pseudo R^2 and adjusted Pseudo R^2, as defined in Equation (6) and (7) respectively, are other widely used measures of goodness of fit:

$$\text{Pseudo } R^2 = (1\text{- } LL_{\text{estimated model}} / LL_{\text{base model}}) \qquad (6)$$

$$\text{Adjusted Pseudo } R^2 = (1\text{- } (LL_{\text{estimated model}} - K)/ LL_{\text{base model}}) \qquad (7)$$

Where, K is the number of parameters estimated in the model. The interpretation of the two measures of goodness of fit is analogous to the R^2 in linear regression. Higher value close to 1 is better.

The expression -2*(LL $_{\text{base model}}$ –K), known as the Akaike Information Criterion(AIC), is also used as a goodness of fit measure; see Equation (8) (Akaike, 1973, Potoglou and Susilo, 2007)

$$AIC = -2\ (LL_{\text{estimated model}} - K) \qquad (8)$$

The related AIC c criterion has been suggested as a measure that can be used to correct for small sample size. This is defined as shown in Equation (9) (Hurvich and Tasai, 1989)

$$AICc = -2\ (LL_{\text{estimated model}} - K) + 2K(K+1)/(N-K+1) \qquad (9)$$

Where, K is the number of parameters and N is the number of observations.

The above-mentioned measures of goodness of fit are commonly used in non-linear regression analysis. However, the R-squared measures for PRM and NBRM have been specially considered by Cameron and Windmeijer (1995) as follows.

Equations (10) and (11) show the R^2 for PRM based on Pearson residuals and deviance residual, respectively.

$$R^2_{Pearson,P} = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i}{\sum_{i=1}^{N} (y_i - \bar{y})^2 / \bar{y}} \qquad (10)$$

$$R^2_{DEV,P} = 1 - \frac{\sum_{i=1}^{N} [y_i \log(y_i / \hat{\mu}_i) - (y_i / \hat{\mu}_i)]}{\sum_{i=1}^{N} y_i (\log(y_i / \bar{y})} \qquad (11)$$

Equations (12) and (13) show the R^2 for NBRM based on Pearson residuals and deviance residual, respectively.

$$R^2_{Pearson,N} = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{\mu}_i)^2 / (\hat{\mu}_i + \hat{\alpha}\hat{\mu}_i^2)}{\sum_{i=1}^{N} (y_i - \bar{y})^2 / (\bar{y} + \hat{\alpha}\bar{y}^2)} \qquad (12)$$

$$R^2_{DEV,N} = 1 - \frac{\sum_{i=1}^{N} [y_i \log(y_i / \hat{\mu}_i) - (y_i + \hat{\alpha}^{-1}) \log(y_i + \hat{\alpha}^{-1}) / (\hat{\mu}_i + \hat{\alpha}^{-1})]}{\sum_{i=1}^{N} [y_i \log(y_i / \bar{y}) - (y_i + \hat{\alpha}^{-1}) \log(y_i + \hat{\alpha}^{-1}) / (\bar{y} + \hat{\alpha}^{-1})]} \qquad (13)$$

It is notable that Negbin 2 variance function is used for calculation. Negbin 2 variance is defined as $\mu_i + \alpha\mu_i{}^2$ in Cameron and Trivedi (1998). The interpretation of four measurements is also analogous to R^2 in linear regression.

4. ANALYSIS AND RESULTS

4.1. Analysis

Eight independent variables out of the original 23 were selected and considered for analysis in this study through (a) correlation tests with the dependent variable ($|p| < 0.2$) and (b) correlations between the independent variables ($|p| > 0.7$). These are listed as follows, together with the abbreviations by which they were designated in the course of the investigation: age of respondent (Age), driver license holder (Driver), college graduation status (Education), $40,000 income level (Income), the percentage of adult drivers in the household (Adult driver), residential density (Residential density), non-residential floor space (Non-residential unit size), and degree of urbanism (degree of urbanism). The software package STATA 8.1 was used for statistical analysis.

One may wonder why the bus stop is not selected as a variable (i.e. bus stop density shown in Table 2) was extracted from MD Transit view and bus stops seem to increase rates of walking to work. It would be understandable when the characteristic of dependent variable (the number of walking trip to walk) is examined as follows. In NHTS (National Household Travel Survey), pure walking trips and the walking trips to access/egress transit stops are separately identified. The variable extracted from MD Transit (Bus stop density) could be selected if either residential density or non-residential unit size is not selected in the final analysis. As expected, the bus stop density was not more significant than residential density and non-residential unit size and was correlated with the other independent variables (i.e. higher bust stop density is observed in the higher residential density and higher non-residential unit size). Thus, it was removed from the set of independent variables.

Inspection of Table 3 shows that both PRM and NBRM give a statistically significant measure of goodness of fit with the observed data. In the case of PRM, the LL function for the estimated model (-780.27) is statistically closer to zero than that for the base model (constant only, -975.21), implying that the former is statistically a better fit. This result is confirmed by the likelihood ratio test (Prob>Chi^2 = 0.000). Pseudo R^2 and adjusted Pseudo R^2 are 0.200 and 0.192, respectively. Also, AIC and AICc support the conclusion that PRM gives a statistically significant fit with the observed data.

Two sets of LL ratio tests are presented for NBRM in Table 3. The NBRM model is compared with the base model under the heading '*Estimate A*', and with the Poisson model under the heading '*Estimate B*'. In particular, LL for the base model (-632.24) is used as the basic level for the likelihood ratio test in *Estimate A*, while LL for the estimated Poisson model (780.27) is used in *Estimate B*. With reference to the interpretation of PRM given above, we see that in *Estimate A* the value of LL for the estimated model (-632.24) is closer to zero than that of the base model (-727.88). This means that the null hypothesis (the estimated model is no better than the base model) must be rejected (191.28 > 15.507, 8 df (degree of freedom)). Other measures such as Pseudo R^2, adjusted Pseudo R^2, AIC, and AICc show that NBRM also gives a statistically significant good fit.

Estimate B is used to examine the extent of over-dispersion of the walking frequency. LL for the estimated model (-632.24) is still statistically closer to zero here than that of the Poisson model (-780.27). The null hypothesis of $\alpha = 0$ is rejected by the likelihood ratio test. This result is confirmed by the fact that the dispersion parameter (α) is significantly greater than 0, indicating the statistical significance of over-dispersion of the walking frequency.

Both regression models confirmed the results of previous studies (e.g. Cao, et al., 2006, FHWA, 1999b, Kim, 2005, Targa and Clifton, 2005), showing that higher density, mixed land use, and residential density are positively associated with higher frequencies of pedestrian trips; old age, drivers, and highly educated persons on the other hand are negatively associated with walking frequency. However, it is notable that the old age, one of the variables, does not include those of retirement age, in that the older commuters are associated with walking frequency. It is interesting that the highly educated persons are associated with lower walk trips. This is contrary to the result estimated by Tagar and Clifton (2005). While they considered pedestrian trip frequency regardless of trip purpose, this study focused on commuting trips. The result, at least estimated in PRM and NBRM, indicated that highly educated persons are less likely to walk to work.

4.2. Discussion

While as mentioned above, NBRM is theoretically superior to PRM for the modeling of over-dispersed variables, our results actually showed PRM to give better model improvement between base model and estimated model: Chi^2(8) = 389.88 as compared with 191.28 for NBRM. In addition, the other measures of goodness of fit supports that PRM seems to give a better result than NBRM (e.g. Pseudo R^2 (0.20 for PRM as compared with 0.13 for NBRM) and AIC (1577 for PRM as compared with 1281 for NBRM)).

Table 3 Comparison of different models of the observed data

<table>
<tr><td>Walk Trips</td><td colspan="2">PRM</td><td colspan="2">NBRM</td></tr>
<tr><td></td><td>Coeff.</td><td>P>z</td><td>Coeff.</td><td>P>z</td></tr>
<tr><td>Age</td><td>-0.015</td><td>0.005</td><td>-0.015</td><td>0.062</td></tr>
<tr><td>Driver</td><td>-0.374</td><td>0.047</td><td>-0.875</td><td>0.010</td></tr>
<tr><td>Education</td><td>-0.424</td><td>0.005</td><td>-0.366</td><td>0.155</td></tr>
<tr><td>Income</td><td>0.370</td><td>0.018</td><td>0.241</td><td>0.339</td></tr>
<tr><td>Adult driver</td><td>-1.164</td><td>0.000</td><td>-1.340</td><td>0.000</td></tr>
<tr><td>Residential density</td><td>0.030</td><td>0.000</td><td>0.033</td><td>0.000</td></tr>
<tr><td>Non-residential unit size</td><td>0.000</td><td>0.000</td><td>0.000</td><td>0.000</td></tr>
<tr><td>Degree of urbanism</td><td>0.045</td><td>0.000</td><td>0.131</td><td>0.002</td></tr>
<tr><td>Constants</td><td>-2.230</td><td>0.000</td><td>-1.912</td><td>0.000</td></tr>
<tr><td></td><td colspan="2">Chi^2(8) = 389.88
Prob>chi^2 = 0.000
LL $_{estimated}$ = - 780.27
LL $_{base}$ = -975.21
α (alpha)= 0
Pseudo R^2 = 0.200
Adjusted Pseudo R^2 =0.192
AIC = 1576.54
AICc= 1576.58
R$^2_{Pearson,P}$ = 0.322</td><td>Estimate A
Chi^2(8) = 191.28
Prob>chi^2 = 0.0000
LL $_{estimated}$ = - 632.24
LL $_{base}$ = -727.88
Pseudo R^2 = 0.1314
Adjusted Pseudo R^2 =0.120
AIC =1280.48
AICc=1280.52
R$^2_{Pearson,N}$ = 0.129</td><td>Estimate B
Chi^2(1) = 494.65
Prob>chi^2= 0.000
LL $_{estimated}$ = -632.24
LL $_{base}$ = -780.27
Dispersion parameter α (alpha) = 11.534</td></tr>
</table>

	$R^2_{DEV,P}$ = 0.238	$R^2_{DEV,N}$ = 0.353	
Number of observations = 3,915			

However, *Estimate B*, which is designed to examine over-dispersion, indicated that NBRM is definitely better than PRM (Chi^2 (1) = 494.65, dispersion parameter α (alpha) = 11.534 >0). The most recommended measure of goodness of fit by Cameron and Windmeijer (1995), $R^2_{DEV,N}$ (0.353) supports *Estimate B*. In addition, the closer LL $_{estimated}$ for NBRM(-632.24) to 0 than the LL $_{estimated}$ for PRM (-780.27) is another evidence in that NBRM is more appropriate for an over-dispersed dependent variable than PRM.

There would seem to be two possible explanations of these conflicting results. Firstly, the selection of independent variables between the two models may affect the measure of goodness of fit. For example, if the researcher selects all variables that show significant correlation for PRM, the NBRM model using the same variables will probably show a poorer fit than PRM. Since the variables in this study were selected with reference to PRM, it is hardly surprising that the PRM model showed a better fit. This consideration is borne out by the fact that the coefficients estimated for all variables in PRM were significant at the 95% confidence level. In NBRM, three variables (Income, Education, and Age) were not found to be significant at the same level. This could be the reason why the PRM model gave a better fit with the observed variables than NBRM. On the other hand, if the researcher had chosen NBRM as a basis for fitting the model based on NBRM, it might be expected that NBRM would have given better results. In other words, the theoretical advantage of NBRM over PRM does not always show up in practice. Secondly, neither PRM nor NBRM might be the best model for studying the demand for pedestrian trips at the regional level. In order to check these suppositions, the traditional trip generation technique, linear regression, was also used to model the data. The characteristics of this model are shown in Table 4.

Table 4 Linear Regression model of pedestrian trip data

Walk Trips	Coefficient	Std. Err	T	P>t
Age	-0.001	0.000	-1.600	0.110
Driver	-0.101	0.023	-4.350	0.000
Education	-0.019	0.012	-1.610	0.108
Income	0.019	0.012	1.540	0.123
Adult driver	-0.043	0.013	-3.290	0.001
Residential density	0.002	0.000	5.840	0.000
Non-residential unit size	0.000	0.000	5.580	0.000
Degree of urbanism	0.017	0.003	6.300	0.000
Constants	0.162	0.032	5.190	0.000
R^2 = 0.0656				

Adjusted R^2=0.0637 F(8, 3906) = 34.29 Prob>F = 0.000

The results of the F test show that this model gives a statistically significant fit. However, the value of R^2 is not very good (0.07) even though this R^2 is not analogous to the measures of goodness of fit such as Pseudo R^2 and $R^2_{DEV,N}$ estimated for PRM and NBRM (Hensher, et al., 2005b). The estimated coefficients have the right signs (the same as those for PRM and NBRM), as is to be expected on the basis of the empirical studies. However, the education level (Education) and the $40,000 income level (Income) variable are not statistically significant (as was also the case in NBRM). Thus, since the basic statistical features of linear regression are acceptable, it cannot be said that this model is not appropriate for estimating pedestrian demand despite the reservations expressed above (see Methodology: Theoretical background) concerning the unsuitability of linear regression for modeling distributions that are far from normal as in the present case.

An alternative approach is to recode the trip frequency as a dummy variable (1 if respondent walks to work; 0 otherwise) and use logistic regression. This model can be a very reasonable alternative because the dependent variable in Figure 1(b), the number of ‘walk to work’ trips, showed relatively negligible walking events. As shown in Table 5, the Chi^2 test again rejects the null hypothesis. The value of Pseudo R^2 calculated from log likelihood values was higher in this logistic regression model than in any of the other maximum-likelihood models considered in this paper. Nonetheless, as indicated in DeMaris (2004), using logistic regression for countable data (i.e. the number of walking trips to work in this study) would be waste of information. This is the main reason that logistic regression is not selected as the best model. In other words, the outcome of logistic regression is not the substantial number of pedestrian trips but ‘trip generated or not’ regardless of the actual quantity. The income level is still not significant, as was the case in the NBRM and linear regression models.

Table 5 Logistic Regression model of pedestrian trip data

Walk Trips	Coefficient	Std. Err	T	P>t
Age	-0.016	0.008	-2.140	0.032
Driver	-0.725	0.275	-2.640	0.008
Education	-0.474	0.224	-2.110	0.035
Income	0.277	0.223	1.240	0.215
Adult driver	-1.337	0.289	-4.630	0.000
Residential density	0.028	0.005	5.560	0.000
Non-residential unit size	0.000	0.000	5.320	0.000
Degree of urbanism	0.066	0.024	2.770	0.006
Constants	-2.208	0.503	-4.390	0.000
$Chi^2(8)$ = 226.77				

Prob>chi^2 = 0.00
Log likelihood = -440.13
Pseudo R^2 = 0.21

5. CONCLUSIONS

The aim of this study was to find the best model for pedestrian commuter trips in a metropolitan area. Since Poisson regression (PRM) and Negative Binomial regression (NBRM) have recently been used to model walking frequencies, these two emerging models formed the main focus of the comparison here. The results of this research confirmed on the one hand the conclusion stated in previous publications that the frequency of pedestrian trips is associated with features of the built environment and socio-economic parameters. On the other hand, the results for this dataset confirmed previous findings that PRM can be a better modeling technique than NBRM in practice despite the theoretical advantages of NBRM for dealing with over-dispersed data. The measure of goodness of fit of the PRM model was confirmed by several evaluation measures. It follows that over-dispersion of the data set under investigation should not be used uncritically as a criterion for rejecting models based on certain types of regression, but that the suitability or unsuitability of the various models needs to be demonstrated empirically. In other words, caution should be exercised in the selection of the best regression technique for the modeling of the frequency of pedestrian trips in any given case. This paper further describes the testing of two other regression techniques for this purpose: linear regression and logistic regression. The parameter estimates and measure of goodness of fit found confirmed that PRM and NBRM are the most appropriate models for pedestrian trip generation of the regression techniques considered here. The suitability of zero-inflated PRM and zero-inflated NBRM for modeling the frequency of pedestrian trips to work would be an appropriate topic for future research, since the frequency of such trips is usually zero-inflated - at least in the United States.

ACKNOWLEGDEMENTS

The authors would like to thank the Maryland State Highway Administration for their assistance and Associate Professor Kelly J. Clifton at Portland State University encouraged and supported this study. The authors also would like to acknowledge three anonymous reviewers. The opinions presented are those of the authors and not necessarily

of the Maryland Department of Transportation and the National Center for Smart Growth. This paper has been presented at the 87th Transportation Research Board Annual Meeting at Washington DC, USA.

REFERENCES

Akaike, H., 2nd International Symposium on Information Theory, 267-281,1973.

Baran, P. K., Rodríguez, D. A. and Khattak, A. J., "*Space Syntax and Walking in a New Urbanist and Suburban Neighbourhoods*", Journal of Urban Design, **13-**1, 5 - 28,2008.

Bate, J., "*History of Demand Modelling* ", In D. A. Hensher and K. J. Button (eds), *Handbook of Transport Modelling Volume 1,* Elsevier Science Ltd.,2000.

Behnam, J. and Patel, B., "*A Method for Estimating Pedestrian Volume in a Central Business District*", Transportation Research Record **629-**,1997.

Cameron, A. C. and Windmeijer, F. A. G., "*R-Squared Measures for Count Data Regression Models With Applications to Health Care Utilization*", Journal of Business and Economic Statistics, **14-**2,1995.

Cameron, A. C. and Trivedi, P. K., "*Regression analysis of Count Data*", Cambridge: Cambridge University Press,1998.

Cao, X., Handy, S. L. and Mokhtarian, P. L., "*The influences of the built environment and residential self-selection on pedestrian behavior: evidence from Austin, TX*", Transportation, **33-**, 1-20,2006.

DeMaris, A., "*Regression with Social Data: Modeling Continuous and Limited Response Variables*", John Wiley & Sons, Inc.,2004.

Ercolano, J. M., Olson, J. S. and Spring, D. M., "*Sketch-Plan Method for Estimating Pedestrian Traffic for Central Business Districts and Suburban Growth Corridors*", Transportation Research Record **1578-**,1997.

Ewing, R. and Cervero, R., "*Travel and the Built Environment: A Synthesis*", Transportation Research Record, **1780-**,2001.

FHWA, "*Final Report: The National Bicycling and Walking Study*",1994.

FHWA, "*Bicycle and Pedestrian Trip Generation workshop*",1997.

FHWA, "*Guidebook on Methods to estimate Non-motorized Travel: Overview of Methods*",1999a.

FHWA, "*Guidebook on Methods to estimate Non-motorized Travel: Supporting Documentation*",1999b.

Hensher, D. A., Rose, J. M. and Greene, W. H., "*Applied Choice Anlaysis*", Cambridge University Press,2005a.

Hensher, D. A., Rose, J. M. and Greene, W. H., "*Applied Choice Analysis*", Cambridge University Press, New York,2005b.

Hillier, B., "*Space is the Machine: A Configurational Theory of Architecture*", Cambridge: Cambridge University Press,1998.

Hurvich, C. M. and Tasai, C., "*Regression and Time Series Model Selection in Small Samples*", Biometrika, **76-**2, 297-307,1989.

Kim, N. S., "*Trip Generation Model for Pedestrians based on NHTS 2001*", Civil Engineering, University of Maryland, Master of Science, 2005.

King, G., "*A SEEMINGLY UNRELATED POISSON REGRESSION-MODEL*", Sociological Methods & Research, **17-**3, 235-255,1989.

Levinson, H. S. and Whnn, F. H., "*Effects of Density on Urban Transportation Requirements*", Highway Research Record 2,, N. R. C. Highway Research Board, Washington, D.C. 38–64,1963.

Matlick, J. M., "*If We Build it, Will They Come? (Forecasting Pedestrian Use and Flows)*",
, pp. 315-319,1996.

Potoglou, D. and Susilo, Y. Comparison of Disaggregate Car Ownership Models. In *9th NECTAR conference*. Porto, Portugal.2007.

Pulugurtha, S. and Repaka, S., "*Assessment of Models to Measure Pedestrian Activity at Signalized Intersections*", Transportation Research Record: Journal of the Transportation Research Board, **2073-**-1, 39-48,2008.

Pushkarev, B. and Jeffrey, M. Z., "*Pedestrian Travel Demand*", Highway Research Record, **355-**,1971.

Schneider, R. J., Arnold, L. S. and Ragland, D. R., "*Pilot Model for Estimating Pedestrian Intersection Crossing Volumes*", Transportation Resesearch Record: Journal of the Transportation Research Board, **2140-**, 13-26,2009.

Shay, E., Fan, Y., Rodríguez, D. and Khattak, A., "*Drive or Walk?: Utilitarian Trips Within a Neotraditional Neighborhood*", Transportation Research Record: Journal of the Transportation Research Board, **1985-**-1, 154-161,2006.

Targa, F. and Clifton, K., "*Bulit Environment and Trip Generation for Non-motorized Travel*", Journal of Transportation and Statistics, **Special Edition (NHTS)-**,2005.