Decision Tree Based Station-Level Rail Transit Ridership Forecasting

Xin Li[1]; Yue Liu, A.M.ASCE[2]; Zhigang Gao[3]; and Daizong Liu[4]

Abstract: This paper presents a decision-tree based model to forecast rail transit ridership at the station level according to the surrounding land-use patterns. The canonical correlation analysis (CCA) method is used to identify key land use variables by evaluating their degrees of contribution to the rail transit station demand, which can effectively reduce dimensionality and complexity of the decision tree. A full month of Smart Card data and detailed regulatory land use plan from Chongqing, China are collected for model development and validation. The proposed model offers the capability of targeting key lane use patterns and associating them with rail transit station boarding and alighting demand at a high level of accuracy. The proposed model can reveal underlying rules between rail transit station demand and land use variables, and can be used to assist in developing the Transit Oriented Development (TOD) plans to improve land use and transit operational efficiency. **DOI: 10.1061/(ASCE)UP.1943-5444.0000331.**

Author keywords: Urban rail transit station; Demand forecasting; Land-use patterns; Decision tree; Canonical correlation analysis (CCA).

Introduction

Forecasting transit demand/ridership is critical to urban transport planning and policy making, and has been well addressed and widely applied in the transportation planning process over the past several decades. The traditional four-step travel forecasting model that considers trip generation, distribution, mode choice, and route assignment has been widely used to this regard (McNally 2000). However, some limitations have been indicated in the literature. For example, the four-step model typically functions to predict travel behavior at an aggregate level that is not able to estimate the travel impacts of neighborhood-level development. This presents an issue for regional travel demand models given the relatively low transit usage; even minor modeling imprecisions can cause significant changes in the location-specific ridership estimates yielding unreliable transit share forecasts (Liu et al. 2014; Cervero 2006; Duduta 2013; Fehr and Peers 2013). In addition, estimation of the four-step model mostly relies on the travel input data from household surveys that include only a small number of transit trips in the area of interest (Marshall and Grady 2006; Gutierrez et al. 2011). It cannot adequately reflect the built environment's impact on transit ridership, and is generally insensitive to land use (Cervero 2006, 2009; Duduta 2013). Finally, using the four-step model for transit ridership forecasting faces institutional and financial barriers as developing and maintaining them require staff resources and interagency or consultant involvement (Marshall and Grady 2006).

Realizing the limitations of the four-step model, researchers have developed simpler and quick response methods that are capable of generating demand estimates directly and economically, e.g., direct public transport demand models, which are widely employed in rail demand forecasting (Owen and Philips 1987; Preston 1991; Wardman and Tyler 2000; Blainey and Preston 2010; Dargay 2010). Such models based on multiple regression analysis comprise a complementary approach to estimating transit ridership as a function of built environment and services features (Kuby et al. 2004; Chu 2004; Cervero 2006). For example, Sung and Oh (2011) have developed multiple regression models to examine the relationship between characteristics of planning factors, such as transit supply service, land use, and transit ridership. Lee et al. (2013) have used automated fare collection (AFC) systems and automated passenger counter (APC) systems combined with geographic information system (GIS) tools to capture the temporal and spatial dimensions of the use of public transit and further to investigate how different land use patterns affect the spatial and temporal demand for public transit services. These models address combinations of transit alignments, station locations, and vehicle technology types, and conduct quick response evaluation of variations in parking, feeder bus service, station spacing, and transit speed and frequency. They can capture the effects of local land use characteristics, such as increased densities and improved walkability, within transit station areas and transit-served communities. These quick response models can be developed, calibrated, validated, and applied within an accelerated time schedule (Walters and Cervero 2003; Gutierrez et al. 2011; Liu et al. 2014). However, methodological problems associated with multiple regression models including nonlinearity, multicollinearity, function form misspecification, and heteroscedasticity (Larsen and Peterson 1988; Mark and Goldberg 1988; Do and Grudnitski 1992) have been neglected in the development of most of the regression-based models and may result in biased and inaccurate estimation.

At the station level transit ridership forecasting, there are an increasing number of studies highlighting the importance and the integration of public transport demand and land-use characteristics

[1]Ph.D. Candidate, Research Assistant, Dept. of Civil and Environmental Engineering, Univ. of Wisconsin at Milwaukee, P.O. Box 784, Milwaukee, WI 53201-0784. E-mail: li44@uwm.edu

[2]Associate Professor, Dept. of Civil and Environmental Engineering, Univ. of Wisconsin at Milwaukee, P.O. Box 784, Milwaukee, WI 53201-0784 (corresponding author). E-mail: liu28@uwm.edu

[3]Vice Chief, Chongqing Urban Transport Planning and Research Institute, No. 18, Yanghe'er Rd., Chongqing 400020, China. E-mail: gaozhigang1979@sina.com.cn

[4]Senior Engineer, China Sustainable Transportation Center, Energy Foundation US, Room 1903, CITIC Bldg., No. 19 Jianguomenwai Ave., Beijing 100004, China. E-mail: daizongliu@chinastc.org

Note. This manuscript was submitted on August 19, 2015; approved on December 23, 2015; published online on March 30, 2016. Discussion period open until August 30, 2016; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Urban Planning and Development***, © ASCE, ISSN 0733-9488.

(Tsai et al. 2013). Sohn and Shim (2010) have used both multiple linear regressions and the structure equation model (SEM) to investigate the factors affecting metro demand at a station level in Seoul City, Korea. In their final regression models, only commercial land was found to have a significant contribution to transit demand. For the SEM model, floor area in commercial and office buildings had a higher impact on boarding trips. They also clearly stated that some findings contradict with prior expectations because of the potential errors when aggregating raw data. Guerra et al. (2011) conducted a study to identify whether there is clear benchmark between distance and ridership that provides a norm for station-area planning and prediction. In their study, data on 832 heavy rail, 589 light rail, and 36 bus rapid transit stations and their surroundings from 20 American transit agencies were collected as input for a direct demand model to validate whether different catchment areas have different influences on a model's predictive power. Despite their announcement that planners and researchers used land use to predict direct ridership demand at the transit station level, they used the number of jobs and population to represent the land use pattern and failed to distinguish the different impacts brought by different types of land use. Similarly, Kuby et al. (2004) selected the number of employees and population as land use variables and found that all of those within walking distance are significantly associated with the average weekday boarding trips at light rail stations. Gutierrez et al. (2011) summarized that factors affecting ridership at the station level can be classified into three types: built environment, socioeconomic factors, and characteristics of the stations. Among those three factors, land use type was classified into factors of built environment. Similar to other existing studies, they focused on land mixed-use and subsequently introduced an indicator representing the mix-use of the different types of land patterns.

Although most of the aforementioned studies state that land use is closely associated with transit ridership at the station level, they used indirect variables such as population or the number of jobs instead of land use variables when developing the forecasting models. No clear interrelation between land use variables and transit demand at the station level has been established in the literature. In addition, transit demand forecasting at the station level by most previous studies has not differentiated boarding and alighting, resulting in their inability to capture imbalanced transit directional demand and their impact on the station capacity.

In view of such deficiencies, this paper contributes to developing a theoretically sound and practically accurate model for rail transit ridership forecasting at the station level associated with key land-use variables. In the proposed model, to efficiently target key land use variables, canonical correlation analysis (CCA) is used to analyze the contribution of each land use variable to the directional transit demand at the station. Following the preselection of key land-use variables, this study will further develop a station-level demand forecasting model based on a decision tree approach, which has been widely demonstrated to be effective in dealing with nonlinearity and multicollinearity problems between predictive and independent variables (Zurada et al. 2011; Esmeir and Markovitch 2007; Xie et al. 2003) faced by many traditional regression models. Tree-based models will be established for both boarding and alighting demand prediction to capture imbalanced directional transit load. The proposed decision-tree based framework can certainly accommodate other widely recognized factors affecting ridership (i.e., socioeconomic factors and transit operational factors) when they become available after the planning stage.

Data Source

The proposed model is developed and validated based on data collected in Chongqing metropolitan area, which is situated in the upper reaches of the Yangtze River at the confluence of the Yangtze and Jialing Rivers in southwest China. With an area of 82,400 km^2 (31,800 mi^2) and a population of 30 million people, Chongqing is the biggest municipality (in terms of area and population size) under the direct administration of the Chinese central government. The Chongqing Rail Transit (CRT) is a metro system that has been in operation since 2005. CRT serves the city's main business and entertainment downtown areas. As of May 2013, CRT consisted of four operating lines, with a total track length of 168 km in operation and 100 metro stations. Representative and well-developed urban rail network, mature urban system, and good data accessibility and quality are the major reasons to use Chongqing's data for model development. The general modeling framework developed in this study can be applied to other cities with local calibration.

Smart Card (IC Card) Data

IC card data set including approximately 32.23 million transactions in the full month of November 2014 have been collected from the CRT. Unlike the conventional demand forecasting models, the proposed model features to predict boarding and alighting demand separately. Hence, the archived IC card data is aggregated into boarding demand and alighting demand, respectively. In this study, only data in the morning peak (8:00 a.m.–9:00 a.m.) of weekdays are selected for model development as the capacity of the transit system often meets its operational bottleneck during that periods. Obviously, the proposed model is not limited to forecasting peak hour demand.

Regulatory Land Use Data

Since 2000 a series of regulatory land use plans have been prepared in order to support transit oriented development (TOD) along urban rail transit corridors and improve urban rail operational efficiency in Chongqing city. The plans cover the areas within a 500-meter radius of urban rail stations along each metro line, and were developed by the local planning bureau concurrent with planning of the local metro network. With the city's rapid growth, the land use plans adopted in this case have already been fully implemented. In this study, 49 major stations in Chongqing City have been selected for model development due to their high maturity in land use development and stable passenger demands. Table 1 lists the description statistical analysis for boarding and alighting data of 49 stations. In the plan, land use in the 500-meter circular buffer area has been classified into 17 categories (Table 2). The residential land is further divided into two types according to the floor area ratio (FAR) while the industrial land is also categorized into two types depending on the industry types. Specifically, the residential lands with FAR 1.2 or lower are classified into Type I, and the lands with FAR greater than 1.2 are considered as Type II. Regarding the classification of industry, Type I is defined as light and low-level pollution industry, such as electronics industry and clothing industry, while Type II is defined as moderate-level pollution industry, such as food industry and textile industry.

Identification of Key Land Use Variables

To avoid the impact of irrelevant land use variables, this study has employed the canonical correlation analysis (CCA) method

Table 1. Boarding and Alighting Demand at Selected Rail Transit Stations (in trips/h)

	Mean		Minimum		Maximum		SD	
Stations	Boarding	Alighting	Boarding	Alighting	Boarding	Alighting	Boarding	Alighting
XiaoShizi	1,215	5,802	1,151	5,687	1,251	6,006	25.54	93.34
JiaoChangKou	1,385	5,710	1,350	5,510	1,440	6,100	20.42	139.75
QiXingGang	1,748	2,437	1,696	2,318	1,816	2,596	35.19	83.55
Eling	1,219	1,326	1,184	1,271	1,277	1,426	26.03	39.36
ShiYouLu	2,647	2,906	2,585	2,732	2,768	3,127	47.22	119.95
XieTaiZi	2,555	1,817	2,436	1,717	2,666	1,950	55.47	58.05
ShiQiaoPu	5,214	6,159	5,093	5,703	5,430	6,546	86.67	219.14
GaoMiaoCun	3,142	793	3,064	715	3,334	843	61.26	36.62
XiaoLongKan	3,123	1,329	2,975	1,246	3,300	1,454	77.26	61.52
ShaPingBa	3,794	3,682	3,667	3,373	4,036	4,032	94.81	187.28
YangGongQiao	1,630	348	1,614	313	1,648	406	8.98	24.49
LieShiMu	2,046	622	1,767	585	2,345	659	150.83	18.70
CiQiKou	1,503	435	1,404	414	1,627	458	52.95	11.63
ShuangBei	2,571	729	2,440	694	2,700	760	75.53	20.61
DaXueCheng	2,935	3,158	2,779	2,980	3,308	3,578	136.89	164.68
LinJiangMen	499	4,594	472	4,459	524	4,741	12.94	84.30
ZengJiaYan	337	903	274	854	366	982	23.98	30.65
LiZiBa	1,020	655	986	642	1,035	666	12.54	6.97
FoTuGuan	176	168	150	158	211	180	16.90	5.76
YuanJiaGang	2,111	2,437	2,046	2,135	2,177	2,744	38.21	163.39
XieJiaWan	2,294	1,261	2,240	1,163	2,375	1,369	34.30	52.93
YangJiaPing	3,955	2,366	3,710	2,299	4,099	2,414	99.41	29.97
DongWuYuan	2,394	722	2,337	685	2,509	773	41.28	22.54
DaYanCun	1,126	492	1,076	447	1,210	564	37.78	33.37
MaWangChang	1,574	551	1,512	507	1,646	615	39.09	26.71
PingAn	2,053	809	1,959	767	2,196	825	54.53	15.08
DaDuKou	1,041	501	979	469	1,094	559	42.54	28.39
XinShanCun	2,961	1,274	2,790	1,224	3,107	1,343	89.34	38.02
LiuGongLi	2,075	527	2,019	473	2,124	581	28.86	34.03
5Gongli	2,125	1,145	1,999	993	2,196	1,429	53.84	121.80
4Gongli	3,463	1,376	3,365	1,307	3,667	1,478	87.05	43.47
NanPing	7,017	5,197	6,780	4,918	7,301	5,370	115.54	106.37
GongMao	3,195	4,238	3,141	4,109	3,270	4,415	30.27	79.09
TongYuanJu	1,100	206	1,076	190	1,113	221	8.82	9.99
LiangLuKou	1,825	5,495	1,713	5,155	2,008	5,659	82.15	127.83
HuaXinJie	1,944	1,013	1,856	970	2,040	1,045	41.08	17.30
GuanYinQiao	4,448	9,268	4,324	8,710	4,632	9,583	80.00	223.90
TangJiaYuanZi	860	927	840	854	896	988	14.14	37.05
ShiZiPing	2,088	1,303	1,993	1,252	2,219	1,348	60.14	25.73
ChongQingBei	1,613	2,571	1,484	2,463	1,849	2,736	95.60	75.51
TongJiaYuanZi	2,274	402	1,555	376	2,613	432	304.29	15.91
YuanYang	1,170	837	1,123	773	1,227	893	30.45	32.16
CuiYun	379	340	358	318	400	357	13.10	9.59
ShuangLong	1,649	529	1,587	521	1,788	543	47.93	5.82
HuangNiBang	1,702	1,411	1,643	1,401	1,763	1,426	36.44	7.35
HuaHuiYuan	4,446	1,239	4,341	1,206	4,658	1,268	90.30	18.04
DaLongShan	2,496	1,157	2,415	1,046	2,688	1,221	70.09	44.56
RanJiaBa	1,279	2,147	1,233	1,937	1,325	2,330	23.89	92.79
DaZhuLin	1,494	459	1,469	434	1,528	495	15.83	15.33

to investigate the relationship between all land use variables and the boarding/alighting demand variables in a multivariate framework. Land use variables with statistically significant contribution to the demand will be preselected through this process and used as input variables for the proposed decision tree model.

CCA was first proposed (Hotelling 1936) with the aim to evaluate correlations between multidimensional datasets. It offers the following advantages over other methods for the purpose of this study. Firstly, CCA limits the probability of a Type I error, which is related to the likelihood of finding a statistically significant result when one should not have (e.g., finding a relationship when it really does not exist in the population) (Sherry and Henson 2005). Secondly, CCA avoids the limit of examining singular causes and effects by taking into account the correlation within each multidimensional variable. Finally, CCA can find a pattern correlation, or a correlation in linear combinations of multiple variables, so that it can extract more information from the pattern of multiple variables but not from an individual variable (Misaki et al. 2012). In this study, because of the existence of two dependent variables, boarding and alighting demand, CCA is adopted to identify all land use variables' importance degree for both dependent variables simultaneously rather than individually.

In recent years, CCA has been widely applied in various research areas. In hydrology, Khalil et al. (2011) built up the functional relationship between the water quality mean values and the canonical attribute space. In genetic research, Hong et al. (2013)

Table 2. Regulatory Land Use Patterns Surrounding Rail Transit Stations

Code	Land use
C1	Administration and office
C2	Business and financial
C3	Cultural and entertainment
C4	Sports and physical training
C5	Medical and health
C6	Education and research
C7	Historical relic
G1	Public park
G	Public green land
M1	Type-I industrial
M2	Type-II industrial
R22	Primary and middle school
R2	Type-II residential
R	Type-I residential
S2	Public square
S	Public Parking
U	Public utilities land

has explored the observed expression variation in exons or in genomic position across the genes using CCA. Naylor et al. (2010) assessed the potential of applying canonical correlation analysis to partitioned genome wide data as a method for discovering regulatory variants and concluded that CCA outperformed pairwise univariate regression in simulation. However, CCA has not been applied to the transportation planning literature. In this study CCA is applied to find the importance degree of each land use variable to both boarding and alighting demands. The obtained value will be adopted as the criterion to select variable for tree-based forecasting models.

CCA Model

CCA is designed to seek the linear combinations of two sets of variables that maximize the correlation between them. To achieve this, the first pair of canonical variables (U_1, V_1) is defined as the linear combinations that have the maximum correlation. Then, the second pair of canonical variables (U_2, V_2) is identified as the linear combinations that have the largest correlation among all pairs uncorrelated to the first pair, and so on.

The general goal of CCA could be defined as the problem of finding pairs of linear combinations that are maximally correlated (Khalil et al. 2011), where the combinations are called canonical variables. The number of pairs of canonical variables is equal to the smallest dimensionality of X and Y (Tabachnick and Fidell 1996). CCA seeks vectors that maximize the correlation between the canonical variables (Fujikoshi et al. 2010). In this study, the introduction of CCA excludes those insignificantly correlated land use variables from ridership forecasting model and this function is named as the reduction of dimensionality. This reduction refines the model structure and further improves computational efficiency as well as estimation accuracy. The following concepts are needed to interpret the solution of CCA (Sherry and Henson 2005):

1. A structure coefficient (r_s) is the Pearson r between an observed variable (e.g., a land use variable) and the canonical variable for the variable's set (e.g., the canonical variable U_1 created from all the land use variables via linear equation).
2. A squared canonical structure coefficient (r_s^2) is the square of the structure coefficients. It indicates the proportion of the variance an observed variable linearly shares with the canonical variable generated from the observed variable's set.
3. A canonical communality coefficient (h^2) is computed as the sum of the r_s^2 across all canonical variables that are interpreted. It is the proportion of variance each variable is explained by the complete canonical solution. This statistic informs one about how useful the observed variable is for the entire analysis.

Application of CCA in Land Use Variables Selection

To apply the CCA, the dimensionality of variables needs to be reduced to the rank of the covariance matrix Σ_{11} to ensure nonsingularity (Naylor et al. 2010). The following requirements are recommended to satisfy when CCA is applied: (1) linearity; (2) multivariate normality; (3) homoscedasticity; and (4) low multicollinearity (Hair et al. 1998). Step 1 to Step 2 of the following procedure aim at reducing the dimensionality of land use variables by first ruling out unrelated variables and then reclassifying the land use variables. Then, Step 3 to Step 6 will be performed to check the requirements of CCA through a categorization transformation and a Box-Cox transformation.

The procedure of applying CCA in preselection of critical land use variables is detailed as follows:

Step 1: A preliminary correlation analysis is employed to measure the relationship between the transit demand variables and the land use variables. The variance/covariance matrix Σ_{12} is constructed, and the entries of the following eight land use variables C3, C4, C5, C7, G, M1, M2, and U (Table 2) are found to be zero, indicating that they are not linearly correlated with the boarding and alighting demand and therefore can be excluded from consideration. The remaining nine land use variables will be considered in the next step.

Step 2: The variance/covariance matrices Σ_{11} and Σ_{22} are suggested to be nonsingular (Naylor et al. 2010). The ranks are computed to be 5 and 2, respectively, indicating the singularity of Σ_{11} and nonsingularity of Σ_{22}. To ensure that the covariance matrix Σ_{11} is fully-ranked so that it can be inversed, the remaining nine land use variables are reclassified into five new sets of land use variables according to their characteristics, namely the entertainment-based land (denoted by A) combining public park (G1) and public square (S2); the education-based land (denoted by B), which is the combination of education and research (C6) and primary and middle

Table 3. Data Categorization of Reclassified Land Use Variables

Reclassified land use variables	Entertainment-based land (A)		Education-based land (B)		Work-based land (C)		Residence-based land (R)		Public parking land (S)	
Attribute	Category	Frequency	Category	Frequency	Category	Frequency	Category	Frequency	Category	Frequency
Values (m^2)	≤6,518	180	≤13,336	180	≤8,985	160	≤4,563	160	0	380
	6,518–45,533	240	13,336–32,961	200	8,985–35,380	240	87,458–1,85,053	180	1–2,509	160
	45,533–78,088	180	32,961–60,916	180	35,380–74,739	200	1,85,053–2,82,006	200	2,509–5,870	200
	78,088–1,51,601	200	60,916-1,03,975	220	74,739–1,80,206	180	2,82,006–4,32,346	240	≥5,870	240
	≥1,51,601	180	≥10,743	200	≥1,80,206	200	≥4,32,346	200	—	—

Table 4. Data Categorization of Transit Demand Variables

Transit demand variables	Boarding		Alighting	
Attribute	Category	Frequency	Category	Frequency
Values (trips)	≤1,071	132	≤478	152
	1,072–1,593	212	479–832	212
	1,594–2,095	212	833–1,309	193
	2,096–2,998	212	1,310–2,856	211
	≥2,999	212	≥2,857	212

Table 5. Results for Linearity Test through Multivariate Regressions

Regression	F-statistics	p value	R^2
Boarding demand	111.3	<0.001	0.8529
Alighting demand	127.4	<0.001	0.8691

school (R22); the work-based land (denoted by C), which combines business and financial (C2) and administration and office (C1); the residence-based land (denoted by R) that combines Type-I residential (R) and Type-II residential (R2); and the public parking land (S).

Step 3: To ensure the homoscedasticity of CCA, the continuous land use variables and boarding/alighting demand variables are discretized into categorical variables to eliminate the differences among variances. An equal depth categorization method is employed. Tables 3 and 4 show the categorized variables and the corresponding frequencies of each category.

Step 4: To test the linear relationship between the two variable sets, two multivariate regressions are performed with the land use variables as independent variables, and the boarding and alighting demands as the dependent variables. Table 5 shows the F-statistics and the p values for testing the general linear hypothesis, H_0:$\beta_1 = \cdots = \beta_5 = 0$. The coefficients of determination, R^2, for the two regressions are also given. The p values indicate that H_0 should be rejected and the requirement for linearity of CCA is satisfied.

Step 5: The authors have employed the eigenvalue analysis to check multicollinearity. A condition number is defined as $\lambda_{\max}/\lambda_{\min}$, where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues of the matrix $X'X$ matrix. A condition number greater than 1,000 is an indicator of strong multicollinearity (Chang and Mastrangelo 2011). The condition number in this study is found to be 93, which is far less than the suggested threshold. Hence, the requirement of low multicollinearity is satisfied for the adopted data set.

Step 6: A kurtosis test is finally used to assess the multivariate normality requirement (Mardia 1975). It is advised to reject the null hypothesis (H_0: The distribution is multivariate normal) because the p value for the kurtosis test equals 0.02 at 5% confidence level. It is found that the assumption of multivariate normality is violated in this study. To satisfy this requirement, all variables are further transformed by a Box-Cox transformation on Y for normality, $Y_i^{(\lambda)} = Y_i^{\lambda} - 1/\lambda$, with a selection of $\lambda = 0.65$ (Khalil et al. 2011). The value of λ is selected to maximize the log-likelihood function of the Box-Cox transformation. The p value for the kurtosis test after transformation is 0.053, indicating that the null hypothesis should not be rejected. Therefore, the requirement of multivariate normality is satisfied after the transformation.

Table 6. CCA Analysis Results

Dimensionality test	Wilks λ	F statistic	Df1	Df2	p value
Full model	0.02895025	92.66767	10	190	<0.001
(U_2, V_2)	0.29728416	56.73084	4	96	<0.001

Table 7. Canonical Solutions

	(U_1, V_1)		(U_2, V_2)		
Variables	r_s	r_s^2 (%)	r_s	r_s^2 (%)	h^2 (%)
A	0.252	0.063	—	0.691	0.754
B	0.503	0.253	—	0.490	0.743
C	—	0.564	0.432	0.187	0.751
R	—	0.152	—	0.480	0.631
S	0.880	0.774	0.225	0.051	0.85
Boarding	—	0.789	—	0.211	100
Alighting	—	0.801	0.374	0.199	100

Analysis Results

Following the procedure presented in the previous section, the CCA produces two pairs of the linear combinations of land use and demand variables, denoted by (U_1, V_1) and (U_2, V_2). According to the definition of CCA, the first pair (U_1, V_1) maximizes the correlation between X (land use factors) and Y (transit demand variables). The second pair (U_2, V_2) produces the second highest correlation. The two pairs of the linear combinations have been created to examine the importance degree of each land-use variable, which further helps to verify and validate the land-use information.

The canonical correlation coefficients (ρ) and the Pearson r relationship between the canonical variables are 0.9500 and 0.8383, respectively, suggesting that the two pairs of canonical variables are highly correlated. The squared canonical correlations (ρ^2) are 0.9026and 0.7027, indicating that U_1 and V_1 share 90.26% of the variance, and the proportion for the pair of U_2 and V_2 is 70.27%. The CCA result shows strong evidence that the reclassified categorized land use variables and demand variables are importantly correlated.

Table 6 shows the Wilks λ, F statistic and p values of the CCA analysis. The full model (both pairs of canonical variables) is statistically significant with $p < 0.001$. The second pair of canonical variables is also statistically significant. The test result suggests that both pairs of the canonical variables need to be considered in the study.

Table 7 shows the canonical solutions, including structure coefficients (r_s), the squared structure coefficients (r_s^2), and the canonical communality coefficients (h^2), which are normally used to identify the variables' importance degree and further help to select the key variables. Most of the existing studies focus on the interpretation of r_s, which only reflects the importance of one observed variable on one dimension of canonical variable. In this study, the authors employed the communality coefficient (h^2) as the contribution degree of each variable to both boarding and alighting demand variables because it reflects the proportion of variance in an observed variable captured by the CCA across all involved canonical variables. As indicated in Table 7, all reclassified land use variables have h^2 greater than 45%, which is the threshold of being useful (Sherry and Henson 2005). Considering the critical role each land use variable plays in CCA, all of them will be selected as the predictor variables in decision tree based model presented in the next section.

In this section, the CCA has been extensively adopted to target critical land use variables, which further helps reduce the dimensionality of the decision tree based demand forecasting model. The process of CCA application progressively eliminates the unnecessary variables by evaluating the relationship between land use patterns and boarding/alighting demand at rail transit stations. A general contribution of each variable is introduced as the selection criterion. Finally, five reclassified land use types have been selected as the input variables for decision tree model because of their significant contribution to rail transit demand at stations.

Forecasting Model

After the preselection of critical land use variables, a popular artificial intelligence approach, the decision tree, is adopted to develop the rail transit station demand forecasting model. The decision tree model is a rule-based decision support tool with a tree structure where a node represents a test on an attribute, a branch denotes an outcome of the test, and a leaf represents a class or class distribution. Each path from the root to a leaf represents a decision rule (Tang et al. 2014). The goal of the decision tree is to create a model that predicts the value of a target variable based on several input variables. Various studies have indicated that the decision tree model has significant benefits to implicitly perform variable screening or feature selection and to require relatively little efforts from users for data preparation (Esmeir and Markovitch 2007; Xie et al. 2003). In addition, unlike the regression model that is widely used in transit demand forecasting, nonlinear relationships between prediction parameters do not affect tree performance and the outcome of decision tree is easy to interpret.

The decision tree model has been widely applied in transportation research. Xie et al. (2003) conducted a thorough comparison of the predictive performances of decision trees, neural networks, and Logit model in terms of mode choice. Tang et al. (2014) proposed a decision tree approach to modeling travel model switching in a dynamic behavior process. Nejad et al. (2009) applied the decision tree model to learn traffic behavior and to predict new events. Wang et al. (2009) proposed a decision tree model to predict short term traffic flow conditions. The decision tree model has also been extensively used in traffic accident analysis (Zhang and Fan 2013; Olutayo and Eludire 2014). Although the decision tree model has been used in many transportation studies, very limited application has been found in transit demand forecasting associated with lane use patterns. In this study, the reclassified categorical land-use variables are treated as variable attributes in the decision tree while the boarding/alighting demands are considered as the target values.

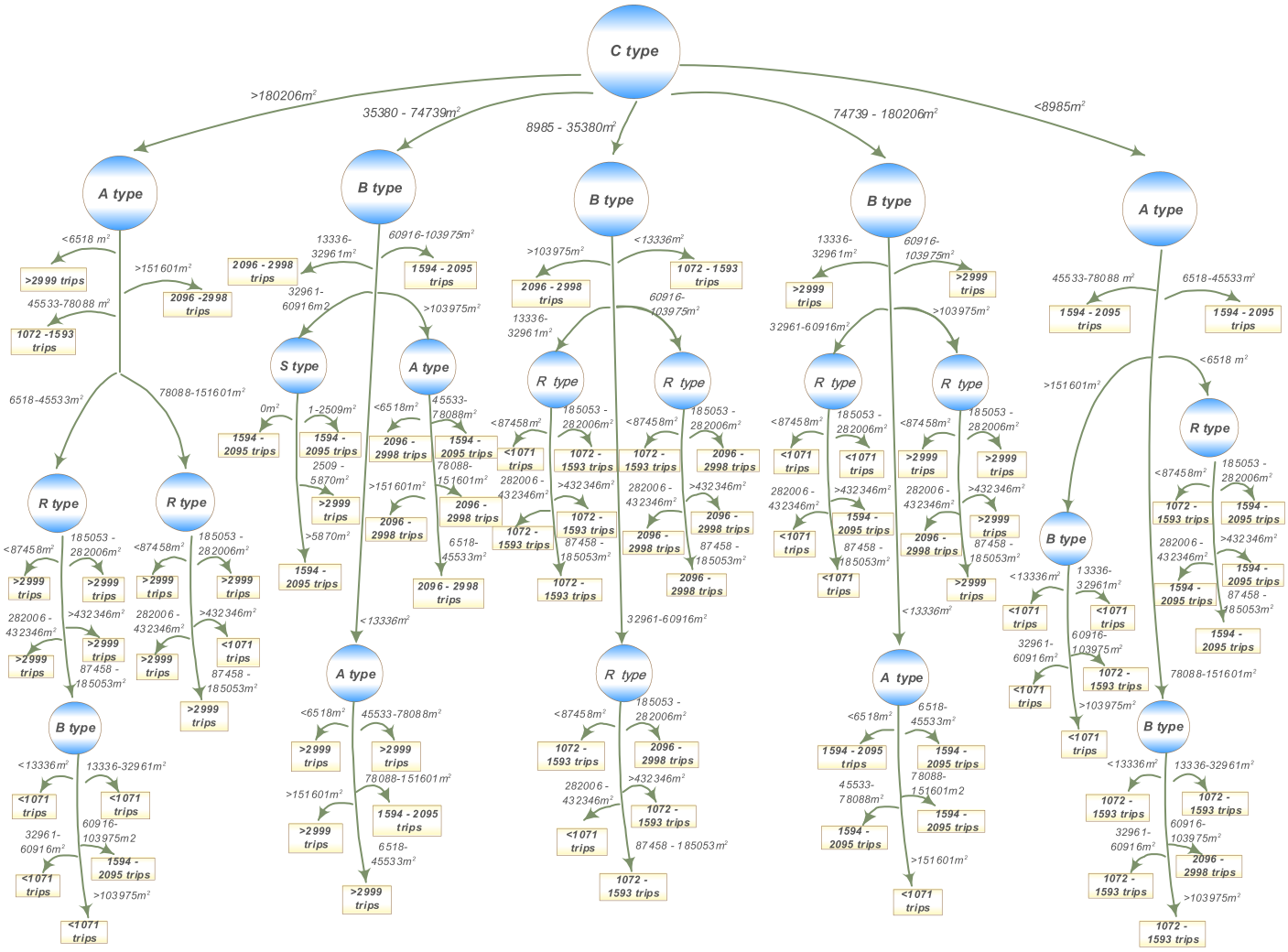

Fig. 1. Decision tree model for boarding demand forecasting

J. Urban Plann. Dev., 2016, 142(4): -1--1

Decision Tree Structure

In this study, the authors have employed the C4.5 algorithm to construct the decision tree model. The algorithm builds the decision tree top-down and prunes them. The tree is constructed by finding the best single feature test to conduct at the root node of the tree (Quinlan 1986; Kohavi 1995). The C4.5 algorithm employs the information gain in an information-based measure that takes into account different numbers (and different probabilities) of test outcomes to split each node, choosing the attribute at each node that produces the purest daughter node on which to split.

Two decision tree models have been constructed to forecast boarding and lighting demand at rail transit stations with the C4.5 algorithm. A total of 980 instances are used to train each tree model. In each decision tree model two critical parameters, the confidence factor of which is used to determine the pruning degree and the minimum number of instances per leaf, can influence the model's structure and performance. In this study, 10-folder cross validation, which is one of the most commonly used methods for parameter tuning and model selection, is used to find the optimal parameters (Kohavi 1995; Arlot and Celisse 2010). More specifically, each tree model is built to search the optimal value of confidence factor from 0.25 to 0.8 with an increment of 0.055, and look for the optimal number of instances per leaf from 1 to 10 at Step 1. Through the optimization process, the confidence factor is finally set to be 0.3 and the optimal number of instances per leaf is determined to be 2 in each tree model.

Figs. 1 and 2 show the constructed decision tree models for forecasting boarding and alighting demand at rail transit stations. More specifically, the tree of boarding demand is built into size 105 with 84 leaves whereas the tree of alighting demand is generated into size 109 with 87 leaves. By referring to the contracted trees, planners or engineers could easily associate land use plans with station-level demand. For example, a specific land use plan for new urban rail station development containing 70,000 m^2 education-based lands, 10,000 m^2 work-based lands, 2,00,000 m^2 residence-based lands, and 50,000 m^2 entertainment-based lands leads to the range of 1594–2095 trips of boarding and less than 478 trips of alighting in the peak hour, respectively.

Performance Evaluation

A very commonly used criterion to evaluate the performance of a decision tree (DT) is the predictive accuracy rate (i.e., correct classification rate). For DTs with binary target variables and a specified target event, various combinations of sensitivity (i.e., true positives/actual positives) and specificity (i.e., true negatives/actual negatives) have also been considered as measures of accuracy (Muata and Bryson 2004). In addition, some more complicated and comprehensive criteria, such as receiver operator characteristic (ROC)

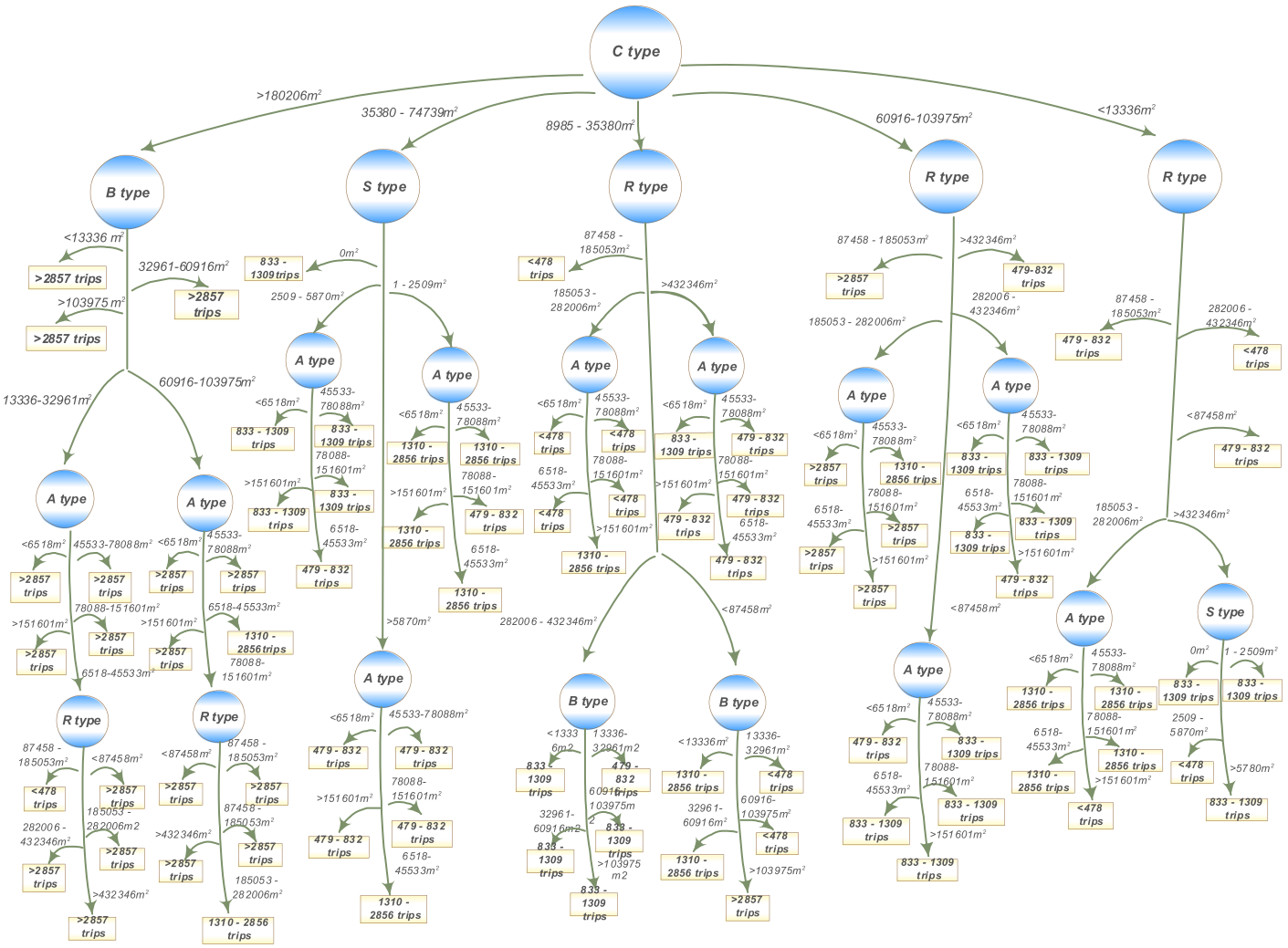

Fig. 2. Decision tree model for alighting demand forecasting

Table 8. Performance Evaluation of the Developed Decision Tree Models

DT model	Values (trips)	TP rate	Precision	Recall	F-measure	ROC area
Boarding	Class 1 = (< = 1071);	1.000	0.009	0.943	1.000	0.971
	Class 2 = (1072–1593);	0.925	0.022	0.92	0.925	0.922
	Class 3 = (1594–2095);	0.807	0.039	0.851	0.807	0.828
	Class 4 = (2096–2998);	0.887	0.049	0.832	0.887	0.858
	Class 5 = (>= 2999);	0.939	0.001	0.995	0.939	0.966
	Weighted average	0.904	0.025	0.905	0.904	0.904
Alighting	Class 1 = (< = 478);	0.914	0.001	0.993	0.914	0.952
	Class 2 = (479–832);	0.958	0.022	0.923	0.958	0.940
	Class 3 = (833–1309);	0.839	0.029	0.876	0.839	0.857
	Class 4 = (1310–2856);	0.891	0.035	0.874	0.891	0.883
	Class 5 = (>= 2857);	1.000	0.01	0.964	1.000	0.981
	Weighted average	0.922	0.021	0.923	0.922	0.922

Table 9. Statistical Performance of the Developed Decision Tree Models

DT model	Percentage of correctly classified instances (%)	Mean ABS error	Coverage (95% confidence level) (%)
Boarding	90.41	0.0519	99.59
Alighting	92.24	0.0413	99.59

precision, recall and F-measure (a combination of recall and precision), have also been developed for a more objective assessment on the classifier's performance (Bradley 1997; Davis and Goadrich 2006).

This study has employed five different criteria to evaluate the developed DTs performance, including the TP rate, precision, recall, F-measure, and the ROC area. In addition, other three statistical indicators, percentage of correctly classified instances, mean absolute error, and coverage of cases (at 95% confidence level), are also used to illustrate the model's performance. Evaluation results are summarized in Tables 8 and 9.

As shown in Tables 8 and 9, the developed DT models perform very well in forecasting both boarding and alighting demand at rail transit stations. The general performance of the alighting DT model is slightly better than the boarding DT model. Both models have correctly classified around 90% of the total instances, with 92.24% for alighting demand forecasting and 90.41% for boarding demand forecasting. The values of the ROC area for both models exceed 0.9 (0.904 for the boarding model and 0.922 for the alighting model), while the mean absolute errors of both models are less than 0.08, indicating very good predicting accuracy. Such prediction accuracy shows the models' applicability in real-world planning applications.

Conclusions

In this paper, decision tree based models are developed to quantify the underlying relations between the boarding/alighting demand and surrounding land use patterns at rail transit stations. IC card data and regulatory land use plans at 49 urban rail transit stations in Chongqing, China were used to train and validate the proposed models. Canonical correlation analysis (CCA) is applied in the process of preselecting key land use variables to reduce the dimensionality of the decision tree models. Five reclassified land use variables have been selected because of their significant contributions to demand. Two decision tree models are constructed and structurally optimized, which can yield around 80% of correct classification rate for both boarding and alighting demand forecasting. Statistical analysis of the models' performance also demonstrates its prediction accuracy and effectiveness.

The authors fully recognize that any data-oriented models (like the one developed in this study) intended for use in practice can achieve its best performance if calibrated properly with local data. The proposed study serves as a useful reference for any transit planning agencies in developing a similar model to reliably forecast rail transit ridership at the station level and to improve land use and transit operational efficiency.

The presented model and evaluation results, though preliminary, offer the advantage of computational convenience and operational flexibility, allowing potential users to customize its application depending on the data availability in the target region. Although the proposed model is calibrated from data collected at a limited number of rail transit stations, the analysis procedure and results clearly indicate different combinations of land use patterns surrounding the target transit development site can have critical impacts on its future demand level. Moreover, the proposed decision tree model with its simple structure and convenience for application offer the potential for its use in real-world planning applications.

The authors future research along this line is to extend the dataset from various types of rail transit stations and recalibrate the proposed model to make it more robust and applicable.

Notation

The following symbols are used in this paper:

U_i = ith canonical variable of $X(i = 1, 2)$;
V_i = ith canonical variable of $Y(i = 1, 2)$;
X = corresponding land use matrix;
Y = corresponding transit demand matrix;
Σ_{11} = $(q \times q)$ covariance matrix of the land use variables;
Σ_{22} = (2×2) covariance matrix of demand variables; and
Σ_{21} = symmetric covariance matrix of land use variable sand demand variables.

References

Arlot, S., and Celisse, A. (2010). "A survey of cross-validation procedures for model selection." *Stat. Surv.*, 4, 40–79.

Blainey, S. P., and Preston, J. M. (2010). "Modelling local rail demand in South Wales." *Transp. Plann. Technol.*, 33(1), 55–73.

Bradley, A. (1997). "The use of the area under the ROC curve in the evaluation of machine learning algorithms." *Pattern Recognit.*, 30(7), 1145–1159.

Cervero, R. (2006). "Alternative approaches to modeling the travel-demand impacts of smart growth." *J. Am. Plann. Assoc.*, 72(3), 285–295.

Cervero, R., Sarmiento, O. L., Jacoby, E., Gomez, L. F., and Neiman, A. (2009). "Influences of built environments on walking and cycling: Lessons from Bogotá." *Int. J. Sustainable Transp.*, 3(4), 203–226.
Chang, Y., and Mastrangelo, C. (2011). "Addressing multicollinearity in semiconductor manufacturing." *Qual. Reliab. Eng. Int.*, 27(6), 843–854.
Chu, X. (2004). "Ridership models at the stop level." National Center for Transit Research, Center for Urban Transportation Research, Univ. of South Florida, Tampa, FL.
Dargay, J. (2010). "A forecasting model for long distance travel in Great Britain." *European Transport Conf.*, Association for European Transport, Henley-in-Arden, U.K.
Davis, J., and Goadrich, M. (2006). "The relationship between precision-recall and ROC curves." *Proc., 23rd Int. Conf. on Machine Learning*, Association of Computing Machinery (ACM), New York, 233–240.
Do, A. Q., and Grudnitski, G. (1992). "A neural network approach to residential property appraisal." *Real Estate Appraiser*, 58(3), 38–45.
Duduta, N. (2013). "Direct ridership model of Mexico City's BRT and metro systems." *Transp. Res. Rec.*, 2394(1), 93–99.
Esmeir, S., and Markovitch, S. (2007). "Anytime learning of decision trees." *J. Mach. Learn. Res.*, 8, 891–933.
Fehr and Peers. (2013). "DRM (direct ridership models)." ⟨http://www.fehrandpeers.com/drm-direct-ridership-models/⟩ (Oct. 17, 2015).
Fujikoshi, Y., Ulyanov, V. V., and Shimizu, R. (2010). *Multivariate statistics: High-dimensional and large-sample approximations*, Wiley, Hoboken, NJ.
Guerra, E., Cervero, R., and Tischler, D. (2011). "The half-mile circle: Does it best represent transit station catchments?" Univ. of California, Berkeley, CA.
Gutierrez, J., Cardozo, O. D., and Garcia-Palomares, J. C. (2011). "Transit ridership forecasting at station level: An approach based on distance-decay weighted regression." *J. Transp. Geogr.*, 19(6), 1081–1092.
Hair, J. F., Anderson, R. E., Tatham, R. L., and Black, W. C. (1998). *Multivariate data analysis*, Prentice Hall, Upper Saddle River, NJ.
Hong, S., Chen, X., Jin, L., and Xiong, M. (2013). "Canonical correlation analysis for RNA-seq co-expression networks." *Nucleic Acids Res.*, 41(8), e95.
Hotelling, H. (1936). "Relations between two sets of variates." *Biometrika*, 28(3–4), 321–377.
Khalil, B., Ouarda, T. B. M. J., and St-Hilaire, A. (2011). "Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis." *J. Hydrol.*, 405(3–4), 277–287.
Kohavi, R. (1995). "Wrappers for performance enhancement and oblivious decision graphs." Ph.D. dissertation, Dept. of Computer Science, Stanford Univ., Standford, CA.
Kuby, M., Barranda, A., and Upchurch, C. (2004). "Factors influencing light-rail station boardings in the United States." *Transp. Res. Part A*, 38(3), 223–247.
Larsen, J. E., and Peterson, M. O. (1988). "Correcting for errors in statistical appraisal equations." *Real Estate Appraiser Anal.*, 54(3), 45–49.
Lee, S. G., Hickman, M., and Tong, D. Q. (2013). "Development of a temporal and spatial linkage between transit demand and landuse patterns." *J. Transport Land Use*, 6(2), 33–46.
Liu, C., Erdogan, S., Ma, T., and Ducca, F. W. (2014). "How to increase rail ridership in Maryland? Direct ridership models (DRM) for policy guidance." *TRB 93rd Annual Meeting Compendium of Papers*, Transportation Research Board, Washington, DC, 3414–3417.
Mardia, K. V. (1975). "Assessment of multinormality and the robustness of Hotelling's test." *Appl. Stat.*, 24(2), 163.
Mark, J., and Goldberg, M. (1988). "Multiple regression analysis and mass assessment: A review of the issues." *Appraisal J.*, 56(1), 89–109.
Marshall, N., and Grady, B. (2006). "Sketch transit modeling based on 2000 census data." *Transp. Res. Rec.*, 1986(1), 182–189.
McNally, M. G. (2000). "The four step model." Institute for Transportation Studies, Univ. of California, Irvine, CA.
Misaki, M., Wallace, G. L., Dankner, N., Martin, A., and Bandettini, P. A. (2012). "Characteristic cortical thickness patterns in adolescents with autism spectrum disorders: Interactions with age and intellectual ability revealed by canonical correlation analysis." *NeuroImage*, 60(3), 1890–1901.
Muata, K., and Bryson, O. (2004). "Evaluation of decision trees: A multi-criteria approach." *Comput. Oper. Res.*, 31(11), 1933–1945.
Naylor, M. G., Lin, X., Weiss, S. T., Raby, B. A., and Lange, C. (2010). "Using canonical correlation analysis to discover genetic regulatory variants." *PLoS One*, 5(5), e10395.
Nejad, S. K., Seifi, F., Ahmadi, H., and Seifi, N. (2009). "Applying data mining in prediction and classification of urban traffic." *Computer Science and Information Engineering, 2009 WRI World Congress*, IEEE Computer Society, Los Alamitos, CA, 674–678.
Olutayo, V. A, and Eludire, A. A. (2014). "Traffic accident analysis using decision trees and neural networks." *Int. J. Inf. Technol. Comput. Sci.*, 6(2), 22–28.
Owen, A. D., and Philips, G. D. A. P. (1987). "An econometric investigation into the characteristics of railway passenger demand." *J. Transp. Econ. Policy*, 21, 231–253.
Preston, J. (1991). "Demand forecasting for new local rail stations and services." *J. Transp. Econ. Policy*, 25, 183–202.
Quinlan, J. R. (1986). "Induction of decision trees." *Mach. Learn.*, 1(1), 81–106.
Sherry, A., and Henson, R. K. (2005). "Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer." *J. Pers. Assess.*, 84(1), 37–48.
Sohn, K., and Shim, H. (2010). "Factors generating boardings at metro stations in the Seoul metropolitan area." *Cities*, 27(5), 358–368.
Sung, H. J., and Oh, J. T. (2011). "Transit-oriented development in a high-density city: Identifying its association with transit ridership in Seoul, Korea." *Cities*, 28(1), 70–82.
Tabchnick, B. G., Fidell, L. S. (1996). *Using multivariate statistics*, Allyn and Bacon, London, 879.
Tang, L., Xiong, C., and Zhang, L. (2014). "Artificial intelligence approach to modeling travel mode switching in a dynamic behavioral process." *Transportation Research Board 93rd Annual Meeting*, Transportation Research Board, Washington, DC.
Tsai, C. H., Mulley, C., and Clifton, G. (2013). "Forecasting public transport demand for the Sydney greater metropolitan area: A comparison of univariate and multivariate methods." Australasian Transport Research Forum, Brisbane, Australia.
Walters, G., and Cervero, R. (2003). *Forecasting transit demand in a fast growing corridor: The direct-ridership model approach*, Fehr and Peers, Lafayette, CA.
Wang, J. J., Wang, J. F., Lu, F., Cao, Z. D., Liao, Y. L., and Deng, Y. (2009). "Comparison study on classification performance for short-term urban traffic flow condition using decision tree algorithms." *Software Eng.*, 4, 434–438.
Wardman, M., and Tyler, J. (2000). "Rail network accessibility and the demand for interurban rail travel." *Transp. Rev.*, 20(1), 3–24.
Xie, C., Lu, J., and Parkany, E. (2003). "Work travel mode choice modeling with data mining: Decision trees and neural networks." *Transp. Res. Rec.*, 1854, 50–61.
Zhang, X. F., and Fan, L. (2013). "A decision tree approach for traffic accident analysis of Saskatchewan highways." *Electrical and Computer Engineering (CCECE), 26th Annual IEEE Canadian Conf.*, IEEE, Piscataway, NJ, 1–4.
Zurada, J., Levitan, A., and Guan, J. (2011). "A comparison of regression and artificial intelligence methods in a mass appraisal context." *J. Real Estate Res.*, 33(3), 349–387.