Impacts of land use and amenities on public transport use, urban planning and design

Article *in* Land Use Policy · November 2016

DOI: 10.1016/j.landusepol.2016.06.004

Nan Hu
2 PUBLICATIONS **1** CITATION
SEE PROFILE

Erika Fille Tupas Legara
Institute Of High Performance Computing
26 PUBLICATIONS **76** CITATIONS
SEE PROFILE

Kee Khoon Lee
Agency for Science, Technology and Researc…
32 PUBLICATIONS **363** CITATIONS
SEE PROFILE

Land Use Policy 57 (2016) 356–367

Contents lists available at ScienceDirect

Land Use Policy

journal homepage: www.elsevier.com/locate/landusepol

Impacts of land use and amenities on public transport use, urban planning and design

Nan Hu[a], Erika Fille Legara[a], Kee Khoon Lee[a], Gih Guang Hung[a,b], Christopher Monterola[a,c,*]

[a] Institute of High Performance Computing, 1 Fusionopolis Way, #16-16 Connexis North, Singapore 138632, Singapore
[b] Rolls-Royce Singapore Pte. Ltd., Advanced Technology Centre, 6 Seletar Aerospace Rise, Singapore 797575, Singapore
[c] Complexity Institute, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore

ARTICLE INFO

Article history:
Received 8 September 2015
Received in revised form 24 May 2016
Accepted 1 June 2016

Keywords:
Land use
Amenities
Public transit
Urban planning

ABSTRACT

Various land-use configurations are known to have wide-ranging effects on the dynamics of and within other city components including the transportation system. In this work, we particularly focus on the complex relationship between land-use and transport offering an innovative approach to the problem by using land-use features at two differing levels of granularity (the more general *land-use sector types* and the more granular *amenity structures*) to evaluate their impact on public transit ridership in both time and space. To quantify the interdependencies, we explored three machine learning models and demonstrate that the decision tree model performs best in terms of overall performance—good predictive accuracy, generality, computational efficiency, and “interpretability”. Results also reveal that amenity-related features are better predictors than the more general ones, which suggests that high-resolution geo-information can provide more insights into the dependence of transit ridership on land-use. We then demonstrate how the developed framework can be applied to urban planning for transit-oriented development by exploring practicable scenarios based on Singapore’s urban plan toward 2030, which includes the development of “regional centers” (*RCs*) across the city-state. Results show an initial increase in transit ridership as the amount of amenities is increased. This trend, on the other hand, eventually reverses (particularly during peak hours) with continued strategic increase in amenities; a tipping point at 55% increase is identified where ridership begins to decline and at 110%, the predicted ridership begins to fall below current levels. Our *in-silico* experiments support one of the medium-term land-use transport goals of stakeholders—to alleviate future strains in the transportation system of Singapore through the development of *RCs*. The model put forward can serve as a good foundation in building decision-support tools that can assist planners in better strategizing and planning land-use configurations, in particular the amenity resource distribution, to influence and shape public transportation demand.

1. Introduction

With rapid urbanization, issues on urban sustainability and resilience have become more and more challenging to synergize with multiple, and sometimes opposing, objectives (Filion and McSpurren, 2007). As a consequence, it is essential to probe multiple urban indicators in a more systematic and holistic manner to capture various urban-related phenomena such as transport ridership and road traffic flow—those that are known to be influenced by the interactions of the different components of an urban system (e.g., land-use, transport, population, etc.). Understanding the interplay of these factors is vital to accurately evaluate existing conditions “on-ground” and to effectively determine how to plan and design urban systems including the identification of which public infrastructures, services, and resources need to be built and deployed (Batty, 2007).

Quantifying public transit ridership involves complex processes as it is affected by various factors: socio-economic (Wang, 2011; Greer and van Campen, 2011), geographic and spatial (land-use related) (Litman, 2007; Putman, 2013; Beimborn et al., 1999; Choi et al., 2012; Engelen, 1988), financial (public) (Neil et al., 2006; Bettencourt, 2013; Lee et al., 2015), and other more qualitative factors such as comfort and convenience (Chen et al., 2011; Litman, 2008). Most of these factors, however, can be quite laborious to

* Corresponding author.
E-mail addresses: hun@ihpc.a-star.edu.sg (N. Hu), legaraeft@ihpc.a-star.edu.sg (E.F. Legara), leekk@ihpc.a-star.edu.sg (K.K. Lee), terence.hung@rolls-royce.com (G.G. Hung), monterolac@ihpc.a-star.edu.sg (C. Monterola).

http://dx.doi.org/10.1016/j.landusepol.2016.06.004
0264-8377/© 2016 Elsevier Ltd. All rights reserved.

measure, more so to control vis-à-vis urban planning. Hence, in this work, we particularly zoom into the more tangible land-use features to quantify ridership. Although these physical features are less challenging to appraise, the relationship of land-use and transport is not any less complex. On one hand, land-use design and characteristics have been used to build various travel demand models (TDM) (e.g., the Four Step TDM (Manheim, 1979; Florian et al., 1988)) as land-use features impose specific spatial constraints for most, if not all, activities. On the other, the demand for transport within an area influences the price of land and the distribution of amenities within a locality, which then affects the future evolution of land-use design (Beimborn et al., 1999; Still et al., 1999; Decraene et al., 2013).

Traditionally, at the planning level, studies have been more macroscopic in nature—without detailed spatiotemporal analysis of the ridership at individual localities (Choi et al., 2012; Chakraborty and Mishra, 2013). Here, we offer a fresh approach by zooming in on certain planner controllable land-use features at higher resolutions and by analyzing their respective spatiotemporal impacts on ridership from both the aggregated and more localized levels. We believe that from an urban planning perspective, it is more suitable to focus on controllable features such as total population, transportation infrastructure, and land-use than factors at lower operational levels (e.g., people's perception on the use of alternative transport mode (Taylor and Fink, 2002)). Our approach makes use of travel data collected from an automated fare collection system, which is used together with certain geo-datasets such as amenity types and distributions obtained from OpenStreetMap (OSM, 2015)—an open source map database. Using open source data, the methods developed and proposed can be easily extended and applied to other countries and/or cities.

We then introduce three multivariate analytical models to quantify the relationship between a set of land-use features and public transit ridership. In multivariate analysis, collinearity is a big challenge (Gomez-Ibanez, 1996; Crane, 2000) as variables can be inter-correlated (e.g., residential and industrial estate can be separated naturally as the urban system evolves (Decraene et al., 2013)); thus, we implement different multivariate analytical models in the estimation of ridership. Our goal is to not only find the most suitable model that would result to the highest accuracy rates, but to also identify the relative importance of the different variables and their critical values in the forecasting process. The machine learning methods, therefore, are evaluated based on four criteria: (1) predictive accuracy; (2) generality in handling different constraints and assumptions of unknown variable values; (3) computational cost/complexity; and (4) "interpretability". To demonstrate the use of our best model, we apply it to tangible urban development scenarios, including the development of Regional Centers (*RC*s), to quantify the effects of the plans to the ridership.

The rest of the paper is organized as follows. In the next section, we discuss the different multivariate analytical models used to quantify the interrelation between various land-use features and public transit ridership. Section 3 then evaluates the predictive accuracy, generality, and ease of interpretation of these methods. Scenario studies on hypothetical plans that include amenity resource increments around *RC*s are implemented and discussed in Section 4. Finally, in Section 5, we conclude the paper and provide helpful insights into our work's usability to urban planners and other stakeholders.

2. Multivariate models and evaluations

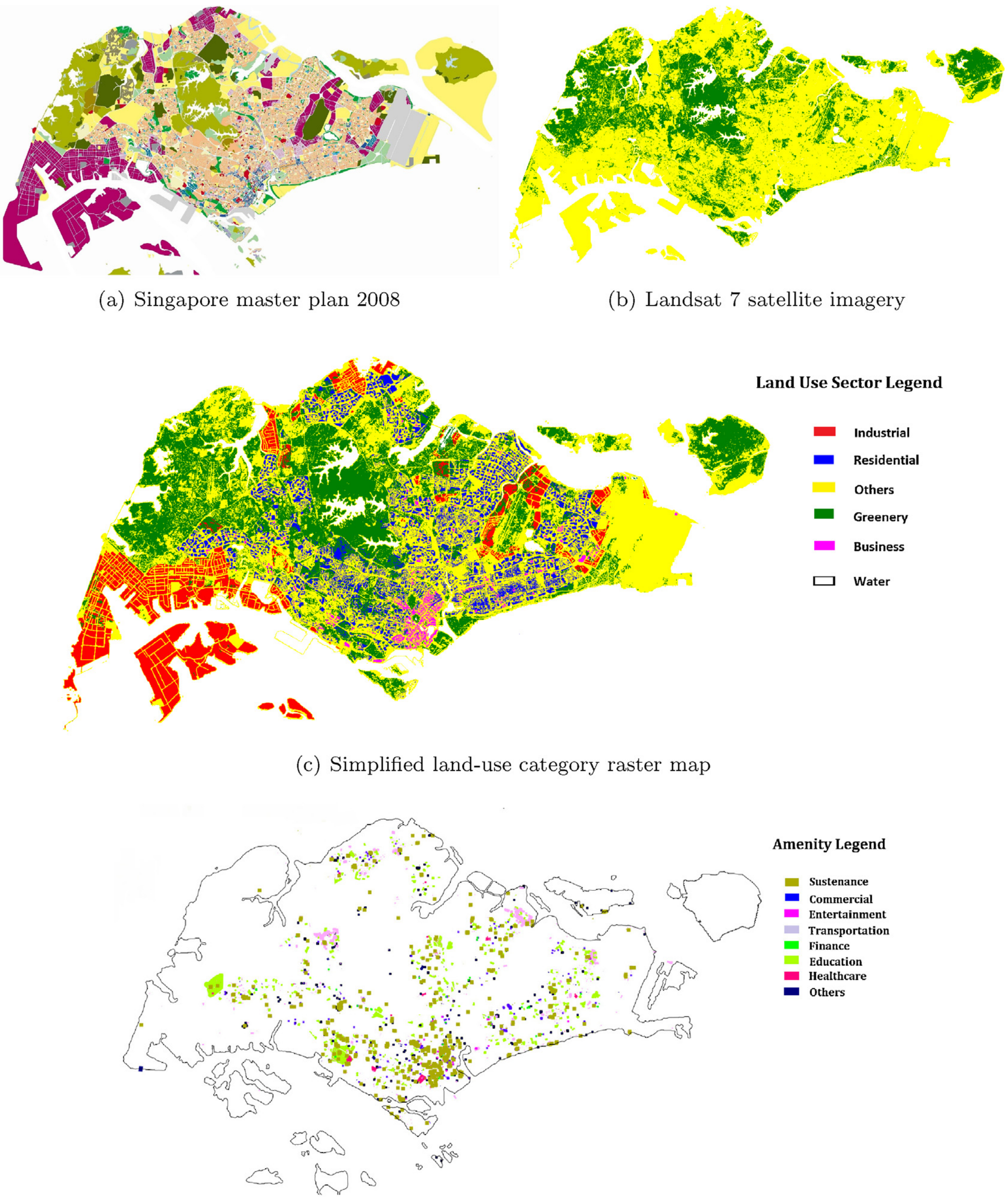
First, we define a *locality* as a surrounding area centered at a public transport station (bus or train). The ridership within a locality is the number of individuals going in and out of the station as recorded in the anonymized electronic ticketing cards. The transport demand at a given station (or stop) over time within a day gives us an idea on the role of the transport point in the entire system. We then infer that the utilization of land-use entities surrounding a station is linked to its ridership as what we have described in our previous work on city characterization based on different land-use category data (Decraene et al., 2013). We then implement three (3) machine learning models to reinforce this connection between land-use features and ridership.

2.1. Urban data in Singapore

The primary data utilized in this study are as follows:

- **Singapore URA Master Plan 2008 (MP2008)**. Land-use categories were extracted from the government's master plan on land-use allocation (Fig. 1(a)). In the original MP2008, the land-use categories include more detailed categories, where we aggregated similar categories (such as different types of business sector) into five broader sectors[1]: business, industrial, residential, water, and others. The original resolution of the image map is 9.7 × 9.7 sqms per pixel; the map was then merged and scaled down to 31.25 × 31.25 sqms per pixel (approximately 32 pixels per 1 km) to make it consistent with other complimentary datasets.
- **Greenery**. The greeneries dataset was extracted from Landsat 7 satellite multi-spectral imagery dated 2002 with less than 5% cloud coverage. The original resolution is 31.25 × 31.25 sqms per pixel. The density of greeneries is not discriminated in this dataset. This dataset complements the MP2008 such that the intersections between the greenery area in this map and the "others" land-use category is considered as a "greenery" (land-use type). Consequently, we now have six sectors of land-use: *residential*, *business*, *industrial*, *greenery*, *water*, and *others*.
- **Amenities**. The amenities of a locality are used to add more granularity to its land-use features, in addition to the land-use category data, see Fig. 1(d). Amenity information was retrieved from Open Street Map (OSM)—an open source geo-data platform that provides a free map of the world. At least for Singapore, we have verified that the geoinformation in OSM is accurate. Based on the usage of the amenities and their tag information in OSM, the amenities are grouped into eight categories: *sustenance*, *education*, *transportation*, *healthcare*, *entertainment*, *finance*, *commerce* and *others*. The same scale of 31.25 × 31.25 sqms per pixel was applied in calculating the amenity densities.
- **Transport data**. The smart fare card dataset used in this work was collected from a centralized automated fare collection (AFC) system and was provided by the Singapore Land Transport Authority (LTA). In recent years, "contactless" smart fare cards and the travel information they generate have particularly helped advance research in the field of human mobility, transport, and urban planning, among others, allowing researchers to probe various aspects of the dynamic spatiotemporal patterns that commuters generate within a given territory in a less invasive manner (Legara and Monterola, 2015; Hasan et al., 2013; Pelletier et al., 2011; Bagchi and White, 2005). In this work, for example, we utilize travel data to inform and train our supervised machine learning models with ridership demand as model output. For the purpose of our research, we only considered the following travel information generated from the AFC system: total number of commuters

[1] More details on the land sector aggregation process are reported in our previous work (Decraene et al., 2013).



(a) Singapore master plan 2008

(b) Landsat 7 satellite imagery

(c) Simplified land-use category raster map

(d) Singapore amenities extracted from OSM

Fig. 1. The urban data of Singapore on the land-use category and amenity distribution: six main land-use categories are merged and extracted. Eight amenity categories are processed as different vector layers in the OSM to remain the data completeness.

boarding and alighting at each (subway or bus) station for a given time interval (here, 10 min). Station information such as the longitude and latitude coordinates for 4544 bus stops and 90 MRT stations are also considered. The anonymized dataset used here runs for a week.

Given these urban data, we initially extracted and analyzed land parcels around each transportation point using three different radii (r_1, r_2, and r_3). The radii are set to 125 m, 250 m, 500 m for the bus stations and 250 m, 500 m, and 1 km for the subway stations, respectively (1 km approximates to 32 pixels in the raster map). Different radii are implemented in the analysis to address the differing density distributions of the bus and subway stations. For each land parcel, the densities of the different land-use categories based on the raster map of the MP2008 are then computed. Densities of each amenity category within the parcel are also calculated based on the intersection of the land parcel and the shape files of buildings where the amenities are situated. In the case where no shape file is found for an amenity resource (e.g., a foodcourt by the street side), a point with unit pixel size (31.25 × 31.25 sqm) is assigned to the OSM database. In summary, three categories of variables are used in this study for the public transit ridership prediction: (1) time (t), i.e., the time instances for ridership aggregation derived from the transport data; (2) land-use category density (L) calculated based the six categories of land-use category from the masterplan and the satellite greenery data; (3) amenities density (A) derived from the OSM dataset. The summary statistics for all variables is listed in Table 1.

Table 1
Descriptive statistics of predictor variables: both land-use category features (L) and amenity densities (A) are used. The minimum values for all features are 0 except the "other" type of L for MRT, which has a value of 0.2574, 0.2814 and 0.3018 for radius r_1, r_2 and r_3 respectively.

Locality	Cat.	Predictor	Mean			S.D.			Max		
			r_1	r_2	r_3	r_1	r_2	r_3	r_1	r_2	r3
MRT	L	Residential	0.1598	0.1816	0.1759	0.1335	0.1233	0.1038	0.4925	0.4348	0.4103
		Business	0.1154	0.0871	0.0700	0.1645	0.1248	0.1057	0.5600	0.4584	0.3933
		Industrial	0.0088	0.0114	0.0162	0.0557	0.0477	0.0470	0.4752	0.3834	0.3419
		Greenery	0.0763	0.1026	0.1230	0.0942	0.0965	0.0892	0.5050	0.4218	0.4491
		Water	0.0095	0.0198	0.0351	0.0309	0.0515	0.0741	0.1841	0.2363	0.3767
		Other	0.6301	0.5975	0.5798	0.1323	0.1138	0.1049	1	1	1
	A	Sustenance	0.0276	0.0138	0.0086	0.0585	0.0246	0.0103	0.3382	0.1542	0.0457
		Education	0.0324	0.0489	0.0540	0.0939	0.0807	0.0573	0.6167	0.4551	0.3435
		Transit	0.0555	0.0484	0.0318	0.0919	0.0589	0.0288	0.5570	0.3544	0.1253
		Finance	0.0012	0.0007	0.0005	0.0033	0.0016	0.0012	0.0149	0.0087	0.0093
		Health care	0.0166	0.0106	0.0077	0.0809	0.0461	0.0185	0.5521	0.3133	0.1262
		Entertainment	0.0025	0.0035	0.0037	0.0116	0.0117	0.0095	0.0846	0.0970	0.0426
		Commercial	0.0784	0.0331	0.0198	0.1557	0.0672	0.0334	1	0.4874	0.1663
		Other	0.0213	0.0175	0.0138	0.0751	0.0373	0.0206	0.6018	0.2325	0.0933
Bus	L	Residential	0.1594	0.1750	0.1704	0.1786	0.1567	0.1315	0.9000	0.7612	0.7419
		Business	0.0171	0.0168	0.0146	0.0597	0.0435	0.0276	0.6809	0.4700	0.2578
		Industrial	0.0726	0.0783	0.0772	0.1834	0.1870	0.1686	1	0.8450	0.7811
		Greenery	0.1122	0.1208	0.1338	0.1946	0.1706	0.1522	1	0.9500	0.9006
		Water	0.0152	0.0196	0.0302	0.0933	0.0979	0.1070	1	1	1
		Other	0.6234	0.5894	0.5736	0.2212	0.1864	0.1599	1	1	0.9826
	A	Sustenance	0.0074	0.0066	0.0058	0.0426	0.0252	0.0140	0.8555	0.3432	0.1591
		Education	0.0575	0.0596	0.0581	0.1598	0.1302	0.1016	1	1	1
		Transit	0.0361	0.0338	0.0298	0.0964	0.0713	0.0466	1	0.8007	0.3631
		Finance	0.0005	0.0004	0.0003	0.0052	0.0021	0.0012	0.1592	0.0398	0.0137
		Health care	0.0049	0.0045	0.0048	0.0448	0.0366	0.0266	1	0.7261	0.4700
		Entertainment	0.0013	0.0012	0.0010	0.0241	0.0124	0.0065	1	0.5371	0.1579
		Commercial	0.0153	0.0126	0.0108	0.0831	0.0521	0.0322	1	1	0.5247
		Other	0.0095	0.0091	0.0086	0.0640	0.0363	0.0232	1	0.6963	0.3158

2.2. Predictive models

Without loss of generality, we implemented and tested three (3) different multivariate analytical methods in the prediction of public transit ridership. These are: (1) decision tree (DT), (2) support vector regression (SVR), and (3) item-based collaborative filtering method based on cosine similarity (CF). The primary objective is to illustrate the ability of predicting public transport ridership across the day using various land-use features and amenities. It is worth noting that since ridership behaves nonlinearly in time, linear models are generally expected to fail to achieve good fits, and are thus excluded in the analysis.

The DT method predicts the ridership through a collection of rules that are constructed in a tree-like structure. At each tree node, a variable with a critical value is set to split the data sample into branches. The splitting criterion used in this study is the mean-squared error since the values predicted (ridership) are continuous numbers. On the other hand, the SVR model is a kernel-based regression method that "learns" a non-linear function by mapping variables into high-dimensional kernel-induced feature space. We use the epsilon-SVR model on the LIBSVM (Chang and Lin, 2011) with a radial basis function (RBF) kernel. Finally, the CF model is inspired by the nearest neighbor and recommender systems (Sarwar et al., 2001). A similarity level (cosine-based) α is measured between each pair of stations based on the following:

$$\alpha = \frac{U\cdot V}{\|U\|\|V\|} = \frac{\sum U_iV_i}{\sqrt{\sum U_i^2}\sqrt{\sum V_i^2}} \tag{1}$$

where U, V are two vectors of predictor variables of the comparing localities.

A threshold similarity value ϕ is used to set all low similarity levels to 0—this value is tuned to distinguish different stations while remaining robust to include stations that are similar to each other. The autocorrelation of ridership along time is modeled through a radial basis kernel density function β. To predict the ridership $y_j^{\tilde{t}}$ of a station j at time $\tilde{t}$, we have:

$$y_j^{\tilde{t}} = \frac{\sum_i^M \alpha_{ij}\left(\sum_k^K \beta_k y_i^k/\sum_k^K \beta_k\right)}{\sum_i^M \alpha_{ij}} \tag{2}$$

$$\beta_k = \frac{e^{-(\tilde{t}-k)^2}}{\gamma^2} \tag{3}$$

M, stations in the training set; α_{ij}, similarity level between the predicted station j and station i; y_i^k, ridership of station i at time instance k; γ is a kernel parameter to control the influential range of a time instance to the others.

3. Model comparison and results

We compare the multivariate models based on four criteria: accuracy, generality, computational cost, and "interpretability." Accuracy refers to how precise and exact the models are in reconstructing and/or predicting current ground statistics in public transit ridership. Generality describes how a method accurately captures the dynamics of the "unseen data." Computational cost looks at the efficiency of the algorithms; and finally, interpretability, in this context, looks at the extent at which model parameters can be translated into the language of urban planners—very abstract concepts and variables may prove to be more difficult to use in more practical situations.

To compare the predictive accuracies of the models, we conduct a set of experiments using some sample data and employ cross-validations within subsets of the transportation data. Following common heuristics for machine learning algorithms (Hair et al., 2009), the dataset of land-use features and the ridership of stations are split into training and test sets with a 2:1 ratio. Moreover, based on the observed daily ridership across stations, the time intervals of a day is divided into 144 of 10 min intervals. The total dataset including all the "station-time" instances is of size 12,960 (144 10-min intervals × 90 stations) for the subway and 654,336 (144 10-min intervals × 4544 stations) for the bus stations.

The predictive accuracy is then measured using Linfoot's three criteria (*FCQ*) of wavelet quantization (Wetherell, 2012). Linfoot's criteria are suitable here for a dependence measurement of the continuous variables (ridership); the *FCQ* values essentially quantify the "closeness" of two curves, borrowing the idea of wave optics comparison. Specifically, *F* quantifies similarity of values, *C* the alignment of peaks, and *Q* looks at consistency of trend/form. The equations for these measures are:

$$R^2 = 1 - \frac{\sum (Y' - Y)^2}{\sum (Y - \bar{Y})^2} \tag{4}$$

$$F = 1 - \frac{\sum (Y' - Y)^2}{\sum Y^2} \tag{5}$$

$$C = \frac{\sum Y'^2}{\sum Y^2} \tag{6}$$

$$Q = \frac{\sum (Y \times Y')}{\sum Y^2} \tag{7}$$

where *Y* and *Y*′ indicate the actual and predicted values at each time interval of the test set. For both measures, the closer the values to 1, the 'closer' the two curves (i.e., the actual and predicted ridership) are, which indicate higher prediction accuracy in the multivariate analysis. For each sample data, we then calculate the mean and standard deviation of the measures for each model, and aggregate such results over 10 iterations. Fig. 2 shows the prediction accuracies of the three models using *FCQ*.

In addition to Linfoot's criteria, in Fig. 3, we plot the probability density function of the residuals generated from the three models for different radii of land-use coverage around localities (1 km for subway stations and 500 m for bus stations). Finally, we performed sensitivity analyses between the predictor variables and the model residuals. For each feature in the cases considered, results display random normal distributions of residual values with mean values around 0, indicating that the model residuals are only due to random noise. This reinforces the idea that the multivariate models implemented are appropriate for the purpose of this research.

3.1. Model comparisons

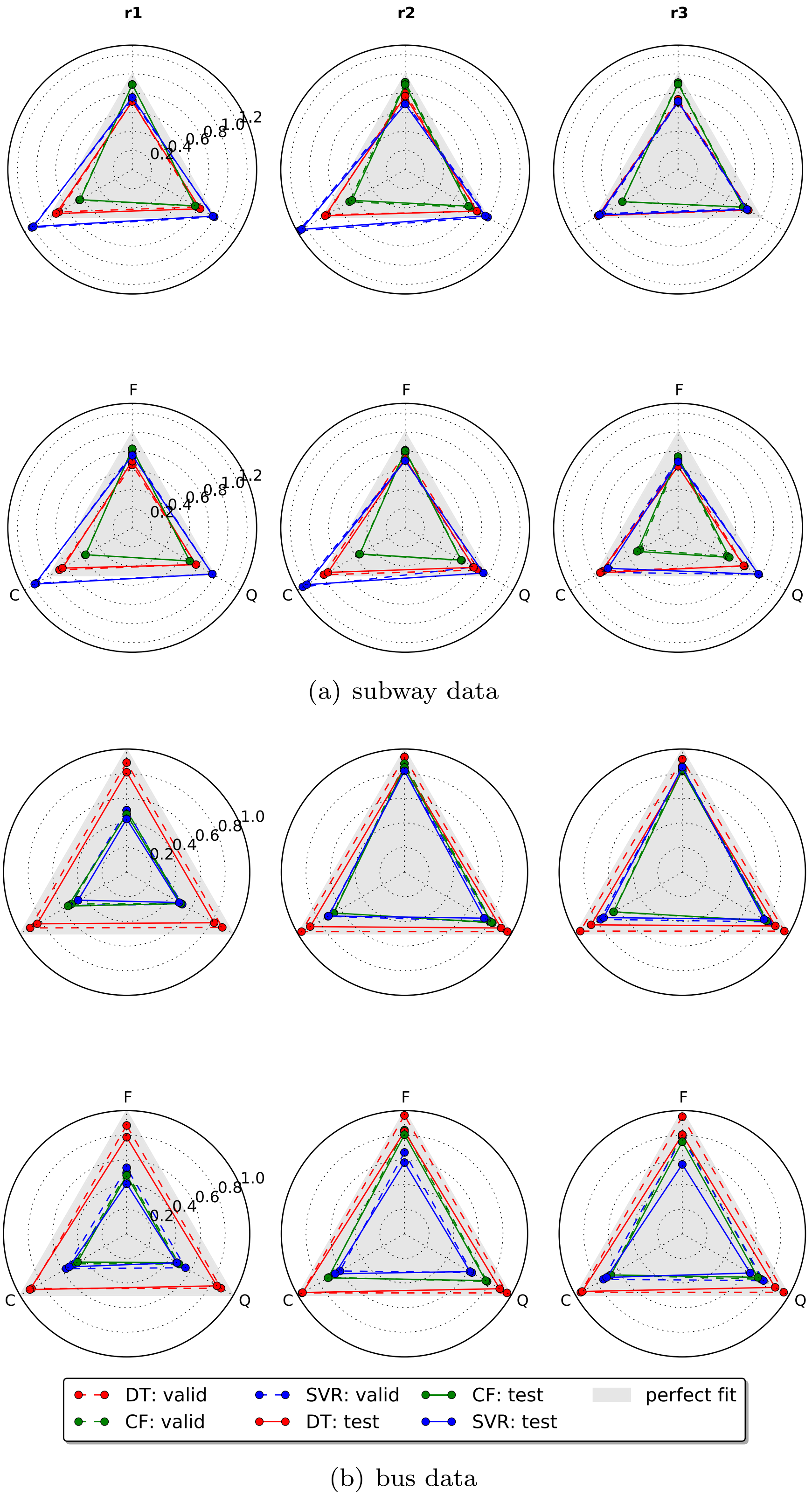
From the resulting *FCQ* measures and residual analyses, we highlight three main findings with regard to the models used: (1) the DT model has the best overall prediction accuracy for both the subway and bus data. However, it is sensitive to the training-test data split. (2) The CF model also achieves good accuracy; in addition, it is more resilient to the random splitting of training-test data. This makes it appealing for case studies involving unknown data. (3) Finally, the SVR model gave the worst *FCQ* measures. Results also suggest a critical range for land-use planning to influence the public transit ridership of a locality. Specifically, a radius of 1 km around the subway stations and 500 m around the bus stations should be taken as the key areas of interest in the planning for transit oriented developments (TOD).

The other criterion that we consider is the generality of the models, which indicates how well the models can be used to predict outcomes with unknown data (test set in this case) based on its fitting on the known data (training set). From the results shown in Figs. 2 and 3, we find that the CF and SVR models have better generality than DT, in which the set of rules are constructed based on the training dataset. Since the threshold value for such variables are determined in the training process, a subtle difference in the test data may result to a different decision path, thus generating a different result. The CF model exhibits good generality because it interpolates the unknown data from the known ones based on the prediction variables; this makes the model more robust to work on unknown (but presumably not fundamentally different) data.

From a more practical perspective, the computational cost is also considered as a factor in contrasting the models. This is especially true if real time feedback of urban design and planning strategies are needed. In this study, SVR and DT exhibited complexity during the model training that are proportional to $O(n^3)$ and $O(n\log(n))$ respectively, where n indicates the number of data samples; but during testing, these were reduced to $O(n)$ and $O(\log(n))$, respectively. However, it is noted that by precomputing the kernels of SVR for both the training and test datasets, the complexity can be significantly reduced. For the CF model, there is no need for training, and the complexity for fitting is $O(n^2)$. In this perspective, both DT and SVR can be efficient in terms of handling large data due to their relatively low computational complexity. The scalability of the CF model is restricted due to its relatively high computational complexity. We note the significant advantage of the DT model in terms of computational cost, which also generate good prediction accuracies for bigger data sizes (i.e., bus data, in this case). This suggests that the DT model can be used in prediction analysis when there is sufficient empirical data, while the other two seems more suitable for scenarios with unknown data due to their better generality.

In terms of the ease of interpretation, the SVR model projects the prediction variables onto higher dimensions, which can include very complex combinations. Thus, it is not intuitive to revert from the support vectors to the prediction variables making it challenging to interpret. The CF model, on the other hand, reveals the similarities of different localities described by their landscape features. Such a similarity matrix can also identify 'outliers' that may have extremely larger (or lower) densities of certain land-use categories and (/or) amenities than others. Correlating such localities with their ridership allows planners to examine the efficiency of the transportation system attributed to the landscape features and thus pinpoint bottlenecks (high ridership and high land-use density) or underutilized (low ridership and high land-use density) areas. Thus, the CF model brings more interpretable findings besides good prediction. In the DT model, feature importances of different prediction variables can be calculated based on the constructed decision tree through either the Gini-impurity or information entropy of a variable. Here, we calculate the feature importances based on the Gini impurity to split data samples. The total value of 1 is split among the prediction variables—higher value translates to higher importance.

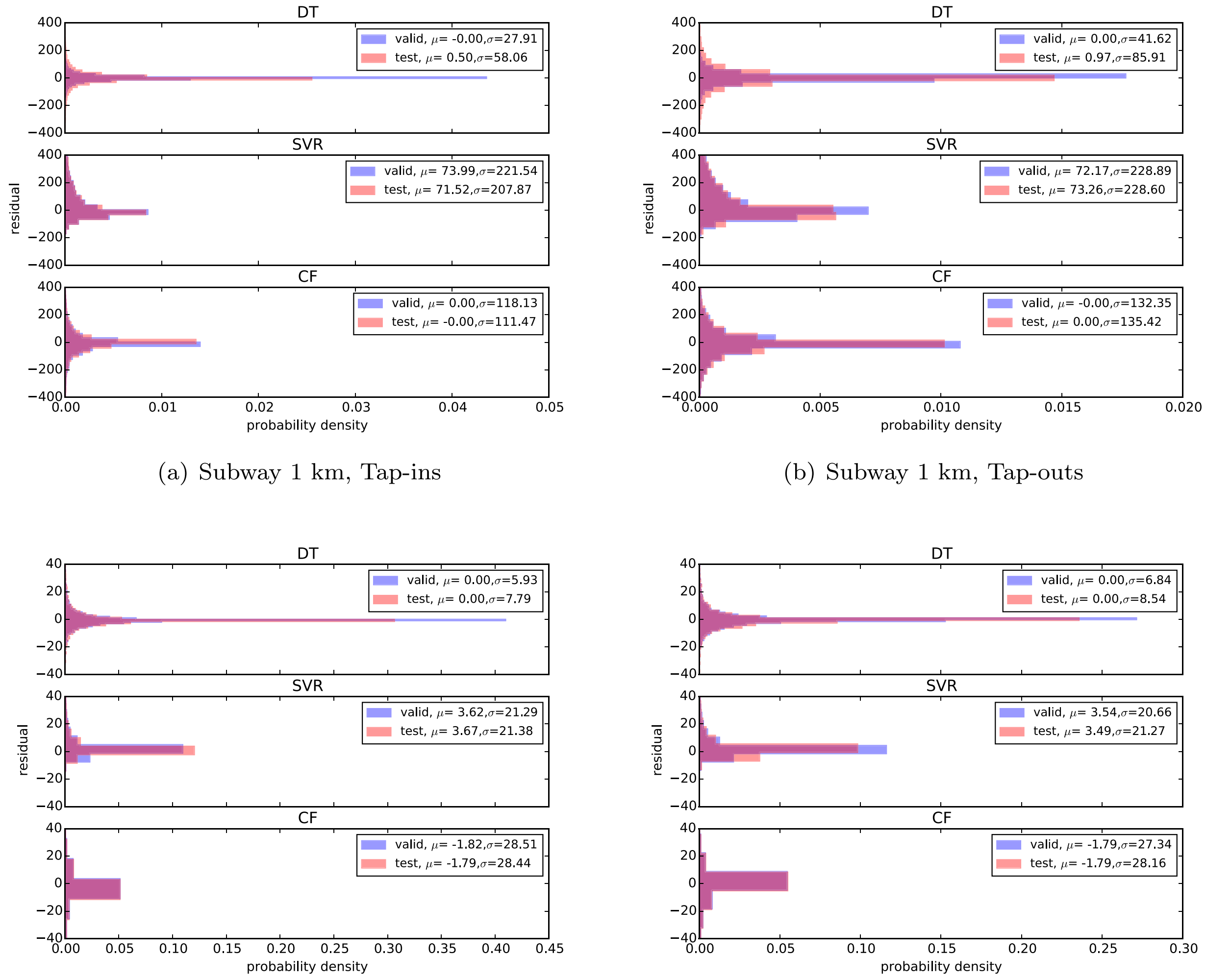
For the subway data, as shown in Fig. 4(a), the most influential variable is the time *T* (around 0.45) as different ridership patterns are displayed along time in the daily ridership, which is related to the travel behavior of individuals (home/work hours and peak vs off-peak hours). Another equally important factor is the commercial amenities A_c (around 0.40). This is quite intuitive as commercial amenities such as shops, malls, and neighborhood grocery stalls are quite closely related to people's daily activities; and, they are mostly deployed around subway stations, at least in the context of Singapore. In addition, some other amenity factors (e.g., A_s, A_{ed}, A_t) have been found to be more important (around 0.1) when compared with other land-use sector variables. Also, the relative importances of features do not vary much with different radii. For the bus data, as shown in Fig. 4(b), the amenity-related prediction variables still add accuracy to the forecast. However, there are some differences between the bus and subway datasets. First, the importance of time and commercial type of amenity do not show the same level of importance in the bus data, while entertainment and other types of amenities become more important. Second, as the radius coverage increases, the relative feature importance changes and



(a) subway data

(b) bus data

Fig. 2. Goodness of fit (FCQ) for the DT, SVR and CF models: the DT model (red lines) achieves the best overall accuracy in both the training and test set over all radii. The CF model generates the least discrepancy between the training and test data set (as reflected by the close solid and dashed green line). The SVR model has the least accuracy as it generally overestimates the ridership (the larger than 1 *C* value). The prediction accuracies become not sensitive only beyond certain threshold (i.e., $r_2 = 500$ m for the subway and 250 m for the bus data). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Subway 1 km, Tap-ins

(b) Subway 1 km, Tap-outs

(c) Bus 500 m, Tap-ins

(d) Bus 500 m, Tap-outs

Fig. 3. Probability density function of residuals for the DT, SVR and CF models: All three models generate small residuals (i.e., mostly within 200 for subway and 20 for bus tap data). The mean (μ) and standard deviation (σ) of residuals can indicate the prediction accuracy of a model. Specifically, a model with μ closest to 0 and the smallest σ achieves the best prediction accuracy (i.e., the DT model in this case). CF model has the least difference between the valid and test datasets, while SVR in general over-predict the ridership in both tap data with a longer tail toward the positive residual values.

their variances also increase. This is likely due to the fact that bus stations are much more dense than subway stations; thus, the locality coverage around the bus stations contains more noise in terms of the land-use related features (e.g., double count for duplicated area with nearby stations, or lack of sufficient landscape features in areas far away from the downtown) as the radius increases.

3.2. Model choice

In Section 3.1, we showed that the SVR exhibited an inferior predictive accuracy compared to both the DT and CF models, which are the foci of the succeeding discussion. It has been demonstrated that DT and CF generate good prediction accuracies and have extra appealing features given their interpretability.

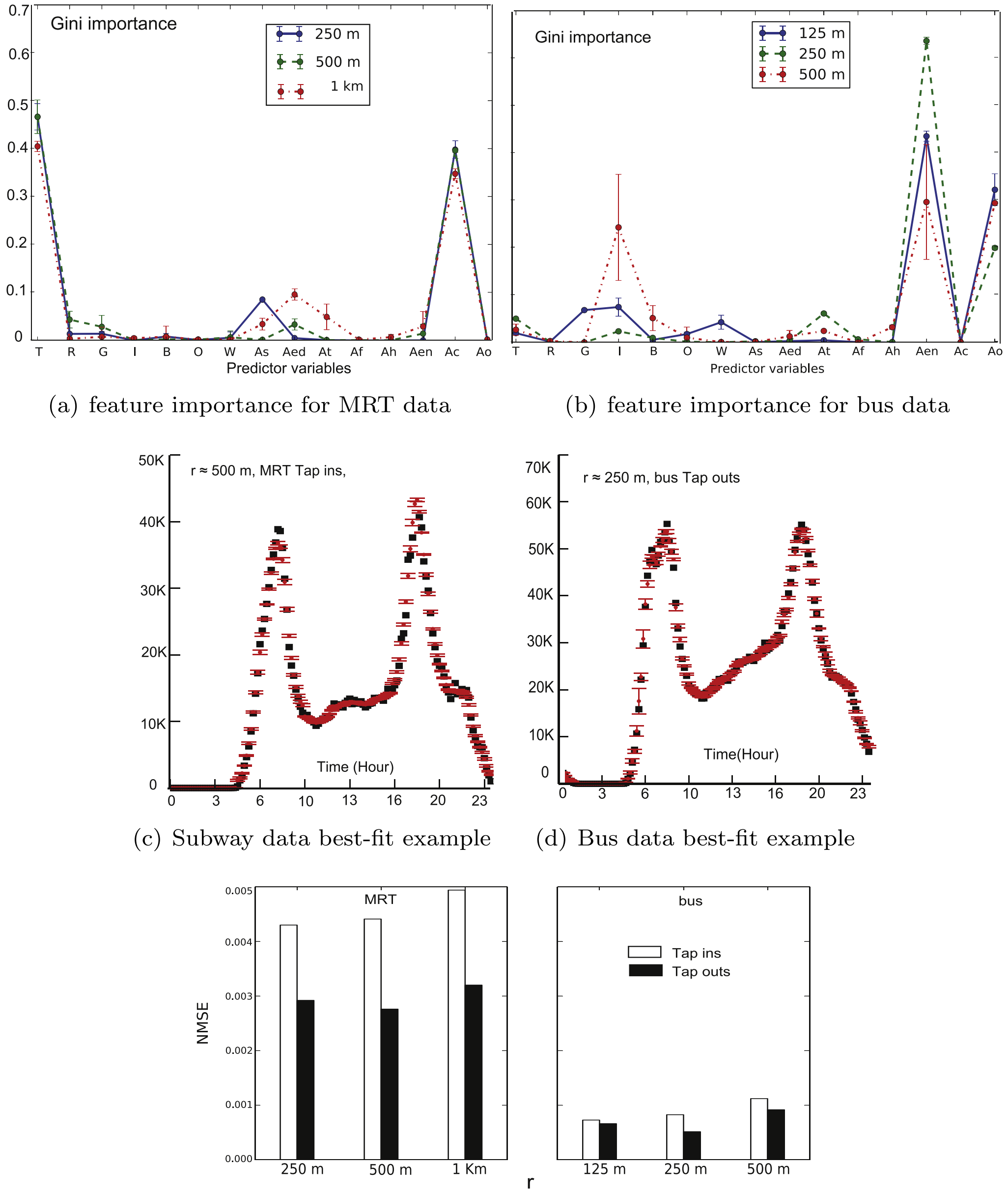
Now, to illustrate the potential of DT and CF, Fig. 4(c) and (d) display the prediction results for the subway and bus data. The Normalized Mean Square Error (NMSE) for all cases in the best-fit models are within 0.5% as shown in Fig. 4(e). Despite the subtle difference in the prediction accuracies of the models, the decision-tree model is favored for two reasons: (1) computational cost for the DT model is much lower than the CF model, which enables DT to be applied for many scenario studies efficiently. (2) With sufficient training data, DT can generate even more accurate predictions than CF. In the CF model, outlier localities in terms of ridership will be likely neutralized since we are averaging across other similar localities; such information loss can cause large errors.

Fundamentally, the methods developed and presented in this manuscript are aimed at enabling urban planners to best estimate future travel demands given a region's land resource allocation. When combined with optimization techniques, one can potentially use this perspective to optimize land resource allocation while considering its impact on the transit system (in terms of commuter experience such as train delays or crowdedness). Finally, since we are utilizing two levels of land-use feature granularity—sector level and amenity level, the practical implications are two-fold: (1) long term strategic planning of how the different types of land-use should be distributed and (2) the more granulated planning and design of amenity locations.

4. Scenarios studies

4.1. Motivation

In Section 3.2, we established that the decision-tree model is capable of accurately capturing the relation between land-use entities and public transit ridership. We also showed that local granular



(a) feature importance for MRT data

(b) feature importance for bus data

(c) Subway data best-fit example

(d) Bus data best-fit example

(e) Error measure

Fig. 4. Best-fit models and results. (a and b) For each locality coverage range, a line with error bar indicates the mean and standard deviation of the Gini impurity factor of the features of 10 distinct experimental runs. It is noted that in addition to time (T), amenity density of commercial type (A_c) and entertainment type (A_{en}) are among the most important features for subway and bus stations respectively. (c and d) Best-fit models prediction results (red error bar with mean as the red solid dot and standard deviation over 10 replications as the errors) shows good agreement with the empirical data (black solid square) as reflected by the small Normalized Mean Square Error (NMSE) values. Here, two examples are demonstrated using the DT and CF models respectively, while the NMSE values for all cases are summarized in (e). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

land-use information (e.g., amenity densities) contributes more significantly to transit ridership prediction in both space and time than the previously reported use of the more general land-use sector information from literature. Building from these, various urban land-use development schemes can now be studied more systematically vis-à-vis their impact on the public transportation system—this is to demonstrate how planners can compare and evaluate different schemes using our approach.

Urban planners at a country level generate concept plans for mid to long term (40–50 years) developments based on comprehensive multidisciplinary studies such as economic growth, population expansion, and the improvement of the quality of life (Tan, 2013). A masterplan is then generated to guide the short to mid term (10–15 years) development plans. For example, in Singapore, masterplans on land-use and transport are diligently revised every several years to cater for the country's rapid development. Such plans outline the general urban development through several "key plans" built mainly upon aggregated statistics. However, these plans cannot be directly implemented without a careful investigation of various what-if scenarios and provision of insights arising from the following objectives: (1) quantify the general impact of the different conceptual plans; and (2) evaluate the impact(s) of one urban factor (e.g., land-use) on another (e.g., transportation).

In the context of Singapore, the total population is projected to increase from 5.3 million to 6.9 million by 2030 to support rapid economic growth. Such a large portion (i.e., 30%) of population growth does not only post challenges to urban planners on land-use planning in the land-scarce city-state, but also on the

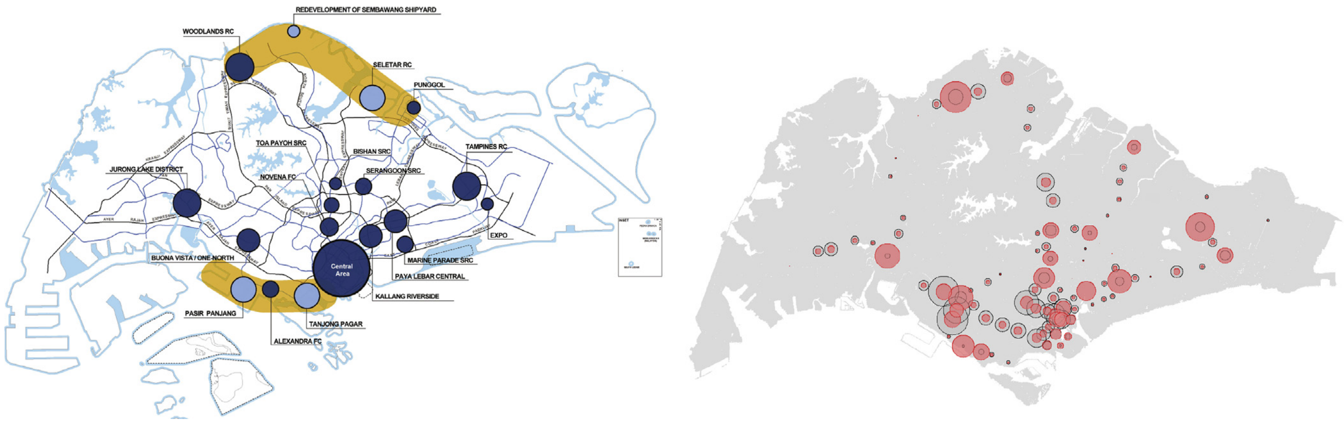


Fig. 5. (a) The "regional center" (*RC*) conceptual plan of Singapore toward 2030. The *RC*s are distributed across the country and we take the size of the circles in the figure to mimic the impact range of the corresponding *RC*s. An intuitive assumption is thus the circle size positively correlates to the amenity resource densities at the corresponding *RC*. (b) The current and synthetic aggregated amenity distribution. The synthetic scenario takes a 100% of total amenity density increment with distribution across different localities corresponding to the *RC* sizes. The hollow black and solid red circles show the normalized aggregated amenity densities in the current and synthetic scenarios respectively. It is noted that the synthetic scenario generates the amenity resource distribution quite consistent with the *RC*s distribution as shown in (a). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

city's public transportation system (i.e., sufficient resources to serve the increased population) (Legara et al., 2014). That is, whatever the masterplan is, it should take into consideration the effects of any land-use feature changes on transit ridership. In this light, we conduct a series of scenario studies based on synthetics urban development schemes at localities within 1 km of each current MRT station and illustrate that our approach can allow stakeholders to achieve the two planning objectives outlined.

4.2. Scenarios

To synthesize the development of urban land-use toward 2030 based on the masterplan generated in 2013, we consider two key planning details. First, the masterplan projects a total of 4% increase in the residential sector, 3% for both the business and industrial land sectors, and with greenery and watery remain unchanged. It is intuitive to project such minor changes in the distribution since land-use sector plan is subject to much longer time frames as compared to local land resource expansions such as amenity developments. In addition, the feature analysis performed using our decision tree model has indicated that the amenity density, as a factor, contributes much more significantly in the prediction of ridership than the more general land-use sector information (*see* Fig. 4(a)). It can then be said that relative changes in the land-use sector densities of each locality are subtle. Consequently, in the synthesized scheme, we assume that the land-use sector configuration remains unchanged.

Next, we consider the plan to develop "regional centers" (*RC*) as shown in Fig. 5(a). These *RC*s are distributed across the country with two hypothetical impacts: (1) to better serve the regional residences through more concentrated amenity resource supply; and (2) to ease the strain on the transportation system by reducing cross-regional transit demand. Many modern urban development studies such as (Makse et al., 1998; Decraene et al., 2013) are based on urban percolation with a single assumed city center (i.e., CBD). Because of its centrality, there are two high peaks in daily public transit ridership (see Fig. 4(c)) driven by the daily activities of commuters traveling from home to work in the morning and back from work to home in the evening. Another series of studies based on the multiple nuclei model (Harris and Ullman, 1945) show that the concept of "many centers" has its advantages in addressing such high-peak transit demand problems. The *RC*s serve as conceptual "many centers" in the context. To help planners quantify the impact of *RC*s on public transit ridership, here we consider a synthetic development scheme through amenity density increases. We take the relative ratio of amenity increments among localities corresponding to the *RC* sizes as shown in Fig. 5(a). For illustration purposes, we assume that 80% of the future amenity resource will be developed around these *RC*s, while the remaining 20% will be randomly distributed to the localities beyond these RCs. The increment of amenity densities at each of the locality is then evenly distributed across each amenity type in the synthetic scenarios. Fig. 5(b) illustrates the normalized amenity density distribution of the current and the synthetic scenario with 100% of amenity density expansion.

It is worth mentioning that we use normalized densities of land-use information (including sector and amenity) according to Eq. (8). This is because the DT model is sensitive to boundary conditions. If the maximum/minimum values are not included in the training set, the predicted results are limited to the existing boundaries of the training set. Normalization in this case addresses this problem by converting the density values to the relative "ranking" among all the localities. As a result, the model used in this study captures the relation between the relative land-use resource among all localities and the public transit ridership.

$$\bar{X} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{8}$$

4.3. Results

With different amounts of total amenity increments, the relative land resource ranking among all localities changes accordingly. We then study the impact of such urban development scheme on the public transit ridership. We first examine the impact on total ridership of tap-ins and tap-outs with varying total amenity resource increments as shown in Fig. 6.

The figure shows that by increasing the amenity resources mainly within the localities of the conceptually planned *RC*s, the total public ridership for MRT eventually decreases after observing an initial ridership increase with different amenity density changes. More specifically, 55% of the total amenity increment becomes a tipping (phase transition) point where further amenity increment will result to a trend reversal in the ridership—from increasing to decreasing. On the other hand, a 110% of amenity increment is another critical point where the ridership goes even lower than the existing ridership level. This result can be very useful

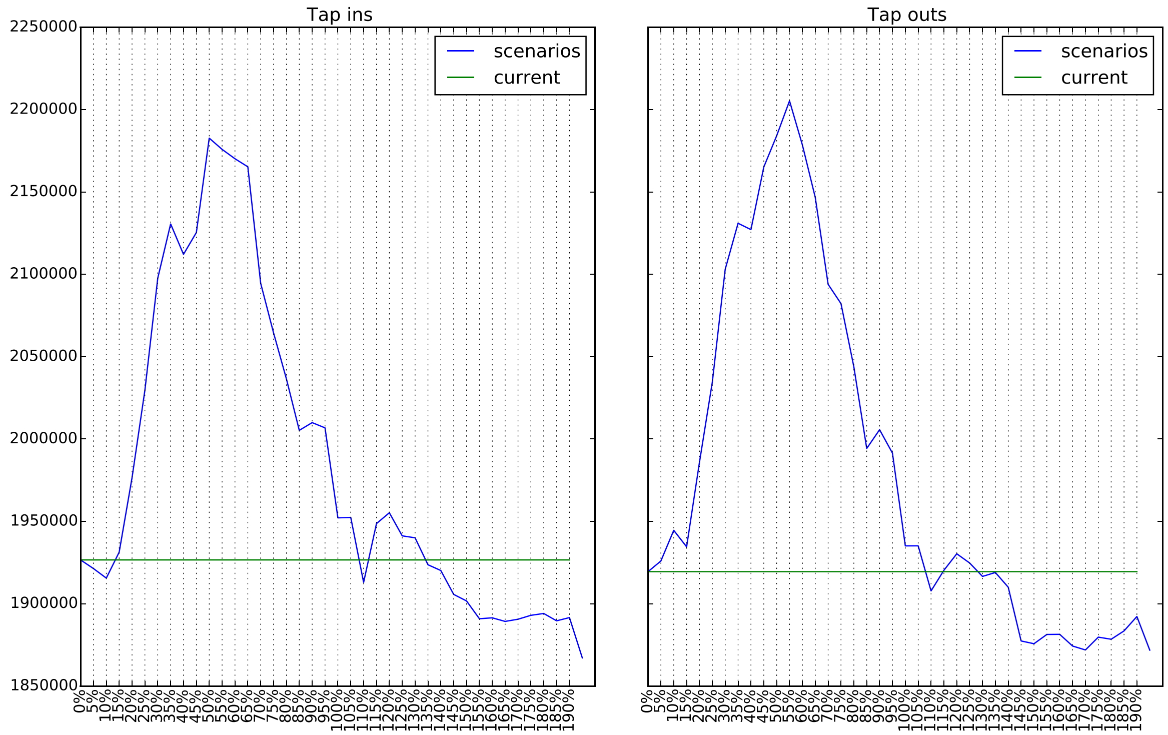


Fig. 6. The total ridership of tap ins and outs eventually decrease with more total amenity sources are increased based on the synthetic scheme, which deploys the amenity resource mainly to the conceptually planned Regional Centers. The blue lines indicate the projected total ridership in the synthetic scenarios of amenity increment while the red flat line indicates the current ridership as a baseline. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to planners vis-à-vis in the evaluation of masterplans. Qualitatively, our result confirms the assumed hypothetical impact of the conceptual development of *RC*s, which aim to bring more workplace and other amenities to places near homes by 2030, which subsequently reduces cross-regional transport movements. Quantitatively, tipping points have been identified and can be used to guide the urban development in achieving the set goals. For instance, an increment of at least 55% of the current amenity resource across the nation is suggested for planners to effectively use land-use resources (i.e., amenities at local scale) as a tool to impact and control the public transportation demand. Note that this type of “tweaking” on the public transportation demand can be carried out without having to modify existing transportation resource supplies. Our study has demonstrated that the strain on transportation infrastructures can be eased effectively by disaggregating land-use resources to intrinsically influence people’s movements, and therefore the demand for transport.

To examine the impact of amenity increment on public transit ridership in more detail, we use total amenity increments of 25%, 55%, 110% and 200% of the current amenity densities in this study. Note that our developed platform can take and evaluate any amenity increment value as input. In any case, the values are monotonically increasing including the critical values (i.e., 55% and 110%) found in the previous set up. Results are shown in Fig. 7.

From the results, we first observe that the overall change ratio of the MRT ridership is mainly bounded within a 20% magnitude even with a total amenity density increase of up to 200% of the current layout. It shows that the urban system exhibits synergistic effects, where components such as land-use and transportation have feedback effect on each other. The temporal distribution of the change ratio is nonlinear, but two patterns can be observed: (1) the variations along time follow similar patterns for synthetic scenarios with total amenity density increment all below or above 55%. As shown in Fig. 7, the curves with 25% and 55% amenity increments (red and green curves) are highly correlated (with 0.924 and 0.813 of R^2 values for tap ins and outs respectively), so are the curves with 110% and 200% increment (blue and purple curves with 0.935 and 0.895 of R^2 values for tap ins and outs). (2) With increment of amenities less than 55%, ridership is almost increased for all hours along, while increment of amenities above 55% results to lower ridership during peak hours but higher ridership during non-peak hours, immediately before or after peak hours.

4.4. Discussion

The results indicate a critical phase transition at 55% amenity density increment. In the first phase of the amenity increase, the urban development plan has not change the current distribution of amenity resources significantly so as to influence the behavioral mobility patterns of the whole population. For example, Central Business Districts (CBD) are still more aggregated amenity resource centers compared to other newly planned *RC*s; thus, the ridership is still concentrated in these areas. In the second phase of amenity increment, with more than 55%, the ridership across time becomes more fluctuating and the peaks become more diverse. In this study, we show that the current peak can be reduced while shifting the ridership to other off-peak hours.

By checking the ridership distribution at each locality, we find two possible reasons for such results. First, the extreme hotspot localities disappear in the synthetic scenario. The ranking of localities in terms of amenity density becomes significantly different as more than 55% of amenity density is increased mainly at all the *RC*s across the administrative region. Since normalized densities are used as predictors in the decision-tree model, such change in ranking will significantly change the ridership of a locality. In fact, there are only a few localities with very concentrated amenity resources that generate very high ridership peaks in the current situation. For example, Raffles Place at the CBD area is the current financial

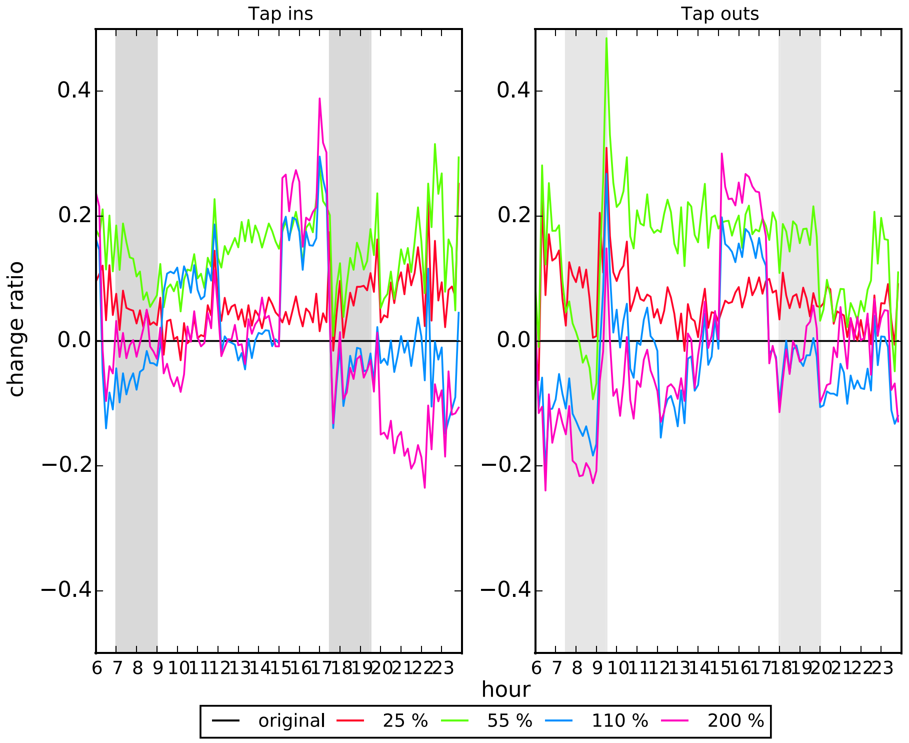


Fig. 7. Change of total daily MRT ridership with different amenity density increments. The change of ridership is mainly bounded within 20% with up to 200% of amenity increment.

and commercial center, which accommodates up to 400 tap-ins per minute during the evening peak hour. As more amenity resources are developed, local hotspots such as Raffles Place will become less congested with amenity resources; thus, the ridership is predicted to become more evenly distributed across other localities with similar amenity resources. Such reduction of ridership during peak hours will significantly reduce the total ridership as shown in Fig. 7. Second, more localities in the synthetic scenario generate small peaks in ridership during off-peak hours—the time window immediately before or after the peak hour. This reflects the comparative importance of amenity densities compared to time as predictors in the decision tree model as shown in Fig. 4(a). In the synthetic scenario, the relative land sector densities (such as residential, business and industrial) of each locality are assumed constant. Thus, the tap-in ridership contributed by the regular work-hour routine in the morning does not change much (because people still travel from home to work in the morning following the "regular" work hours). However, other activities such as shopping, watching movies, visiting parks, and other ad-hoc activities can be more dispersed across the different *RC*s. The increased impact of *RC*s will likely change the ridership along time in terms of many random low peaks. This is because when the ridership is distributed to different *RC*s, the travel duration of each trip becomes more diverse, thus the currently concentrated ridership peak becomes shifted.

Our findings can shed light on urban planning policies with respect to land-use and transport. On one hand, by increasing more than 55% of the current amenity resources to different *RC*s across the country, the ridership of public transit, in particular during the peak hours, can be regulated and even reduced. Such amenity resource deployment can thus improve the transport system by easing the peak hour bottlenecks at the CBD area and improving the utilities of other localities. On the other hand, ridership within a locality can serve as an important indicator for people's activities. Through land-use and amenity deployment planning, it is possible to influence the activity patterns of citizens at the localities so as to boost the utilities of the *RC*s for the purpose of serving regional population. More importantly, the planners can be guided with quantitative results of different synthetic schemes to implement conceptual plans of urban development.

5. Conclusion

In this work, we investigated in great detail the complex relationship between public transit ridership and land-use information at two levels of granularity of land-use related features. The methods and analyses conducted were employed using data on Singapore; the choice of administrative region is mainly due to data accessibility and availability.

We first tested three different multivariate predictive models to find the interdependency of various land-use related features such as amenities and land-use types, and public transport ridership. Among the models explored to predict transit ridership, we found that the decision-tree (DT) was the best based on four criteria: prediction accuracy, generality, computational cost, and interpretability. Through feature analysis using DT, we identified that the amenity-related features, particularly the densities (of the commercial and entertainment types), have the most impact in the prediction accuracy. We then used the DT model to quantitatively assess the future of transportation demand given different synthetic urban development schemes based on Singapore's 2013 Land Use Plan toward 2030, which conceptualizes the development of regional centers (*RC*). The hypothetical scenarios were built under the assumption that the land sector category densities will remain relatively constant toward 2030 and that the amenity resources are the ones to be developed and distributed across *RC*'s (as proposed in the concept plan). Our findings highlighted that land-use amenities are indeed the more influential feature variables when it comes to estimating public transit ridership. It needs to be pointed out, however, that for the scope of this work, we are only comparing between two levels of geo-information granularity—*amenities* and *land-use sectors*.

Be that as it may, with the different amounts of total amenity increments, we found that the total ridership, after exhibiting an initial phase of increasing trend, decreases and hovers around stable levels similar to the current situation. We quantitatively measured these impacts with respect to the increment of amenities and found that an increase in the amenities of 55% is a key phase-transition point where stakeholders can effectively control public transit ridership. Moreover, we demonstrated that the peak hour ridership can be significantly reduced by the development of more than 55% of the existing amenity density around the proposed *RC*s. For all that, however, we note that the amenity densities did not manifest the same level of impact on (tap-in) ridership during regular morning work-hours when most employed individuals follow a fixed routine in going to work (as opposed to shopping, watching movies, and/or visiting parks where individuals exhibit more flexibility in changing routines).

The general framework presented can shed light on certain urban planning processes that aim to help ease bottlenecks in public transit systems by improving on the utilization of other localities via the development of amenities. The models and procedures introduced can be instrumental in constructing decision-support tools for urban planners—tools that will allow them to investigate several urban planning scenarios that involve interacting land-use and transport systems. In terms of land-use planning and design, the implications are two-fold: (1) make provisions for more strategic mapping of land-use sectors for a given administrative region and (2) make way for more tactical blueprints of amenity locations; this allows the utility of our work to be extended to both urban development and redevelopment.

Acknowledgements

This research was supported by Singapore A*SERC Complex Systems Programme research grant (1224504056). We also acknowledge the Land Transport Authority of Singapore for the one-week ticketing data used in this work. The authors would like to thank James Decraene for some useful discussions at the beginning of the study and Guoqi Li for suggesting the collaborative filtering method.

References

Bagchi, M., White, P.R., 2005. The potential of public transport smart card data. Transp. Policy 12 (5), 464–474.

Batty, M., 2007. Cities and Complexity: Understanding Cities With Cellular Automata, Agent-based Models, and Fractals. The MIT Press.

Beimborn, E., Horowitz, A., Vijayan, S., Bordewin, M., 1999 May. An Overview: Land Use and Economic Development in Statewide Transportation Planning. Tech. Rep., U.S. Department of Transportation, Federal Highway Administration.

Bettencourt, L.M.A., 2013. The origins of scaling in cities. Science 340 (6139), 1438–1441.

Chakraborty, A., Mishra, S., 2013. Land use and transit ridership connections: implications for state-level planning agencies. Land Use Policy 30 (1), 458–469.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 27:1–27:27, software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.

Chen, C., Varley, D., Chen, J., 2011. What affects transit ridership? A dynamic analysis involving multiple factors, lags and asymmetric behavior. Urban Stud. 48 (9), 1893–1908.

Choi, J., Lee, Y., Kim, T., Sohn, K., 2012. An analysis of metro ridership at the station-to-station level in Seoul. Transportation 39 (3), 705–722.

Crane, R., 2000. The influence of urban form on travel: an interpretive review. J. Plan. Lit. 15 (1), 3–23.

Decraene, J., Monterola, C., Lee, K., Hung, G., 2013. A quantitative procedure for the spatial characterization of urban land use. Int. J. Mod. Phys. C 24 (1250092), 1–15.

Decraene, J., Monterola, C., Lee, K.K., Hung, G.G., Batty, M., 2013. The emergence of urban land use patterns driven by dispersion and aggregation mechanisms. PLOS ONE 8 (12), 1–9.

Engelen, G., 1988. The theory of self-organization and modelling complex urban systems. Eur. J. Oper. Res. 37 (1), 42–57.

Filion, P., McSpurren, K., 2007. Smart growth and development reality: the difficult co-ordination of land use and transport objectives. Urban Stud. 44 (3), 501–523.

Florian, M., Gaudry, M., Lardinois, C., 1988. A two-dimensional framework for the understanding of transportation planning models. Transp. Res. B: Methodol. 22 (6), 411–419.

Gomez-Ibanez, J.A., 1996. Big-city transit, ridership, deficits, and politics. J. Am. Plan. Assoc. 62 (1), 30–50.

Greer, M.R., van Campen, B., 2011. Influences on public transport utilization: the case of Auckland. J. Public Transp. 14 (2), 51–68.

Harris, C.D., Ullman, E.L., 1945. The nature of cities. Ann. Am. Acad. Polit. Soc. Sci. 242, 7–17.

Hasan, S., Schneider, C.M., Ukkusuri, S.V., González, M.C., 2013. Spatiotemporal patterns of urban human mobility. J. Stat. Phys. 151 (1-2), 304–318.

Hair Jr., J.F., Black, W.C., Babin, B.J., Anderson, R.E., 2009. Multivariate Data Analysis. Prentice Hall.

Lee, J., Kurisu, K., An, K., Hanaki, K., 2015. Development of the compact city index and its application to Japanese cities. Urban Stud. 52 (6), 1054–1070.

Legara, E.F., Monterola, C., 2015, August. Inferring Passenger Type from Commuter Eigentravel Matrices, arXiv:1509.01199, [physics. soc-ph].

Legara, E.F., Monterola, C., Lee, K.K., Hung, G.G., 2014. Critical capacity, travel time delays and travel time distribution of rapid mass transit systems. Physica A 406, 100–106.

Litman, T., 2007. Land Use Impacts on Transport: How Land Use Factors Affect Travel Behavior. Tech. Rep., Victoria Transport Policy Institute www.vtpi.org/landtravel.pdf.

Litman, T., 2008. Valuing transit service quality improvements. J. Public Transp. 11 (2), 43–63.

Makse, H.A., Andrade, J.S., Batty, M., Havlin, S., Stanley, H.E., 1998. Modeling urban growth patterns with correlated percolation. Phys. Rev. E 58, 7054–7062.

Manheim, M., 1979. Fundamentals of Transportation Systems Analysis. MIT Press, Cambridge, MA.

Neil, P., Balcombe, R., Mackett, R., Titheridge, H., Preston, J., Wardman, M., Shires, J., White, P., 2006. The demand for public transport: the effects of fares, quality of service, income and car ownership. Transp. Policy 13 (4), 295–306.

OSM, 2015. Openstreetmap webpage (online) (cited 27.08.15).

Pelletier, M.-P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: a literature review. Transp. Res. C: Emerg. Technol. 19 (4), 557–568.

Putman, S.H., 2013. Integrated Urban Models Volume 1: Policy Analysis of Transportation and Land Use (RLE: The City), vol. 1. Routledge.

Sarwar, B., Karypis, G., Konstan, J., Riedl, J., 2001. Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International Conference on World Wide Web, ACM, pp. 285–295.

Still, B.G., May, A.D., Bristow, A.L., 1999. The assessment of transport impacts on land use: practical uses in strategic planning. Transp. Policy 6 (2), 83–98.

Tan, S.B., 2013. Long-term Land Use Planning in Singapore, Case Study. The Lee Kuan Yew School of Public Policy at the National University of Singapore.

Taylor, B.D., Fink, C.N.Y., 2002. The Factors Influencing Transit Ridership: A Review and Analysis of the Ridership Literature. Working Paper, UCLA Department of Urban Planning.

Wang, J., 2011. Appraisal of Factors Influencing Public Transport Patronage. Research Report 434, NZ Transport Agency.

Wetherell, W.B., 2012. The Calculation of Image Quality, Vol. 8 of Applied Optics and Optical Engineering. Elsevier, pp. 179–180 (Chapter 6).