From compound word to metropolitan station: Semantic similarity analysis using smart card data

Dingyi Zhuang[a,b], Siyu Hao[a], Der-Horng Lee[a,*], Jian Gang Jin[c]

[a] *Department of Civil and Environmental Engineering, National University of Singapore, Singapore 117576, Singapore*
[b] *School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai,200240, China*
[c] *School of Naval Architecture, Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China*

Abstract

Rapid urbanization and modern civilization require sound integration with public transportation system. Meanwhile, understanding the similarity among subway stations is imperative to construct an intelligent transportation system. In this paper, contrary to previous work that utilized aggregated features to find stations' similarity, we proposed a semantic aspect from natural language processing (NLP) to interpret subway stations as compound words. Specifically, we transplanted context and literal meaning of compound words into mobility and location attributes of stations. Using smart card data, we trained stacked autoencoders (SAE) as an embedding method to learn the mobility semantics, a more abstract representation. Subsequently, we have applied affinity propagation clustering to classify 9 point of interest (POI) categories as urban land use division, or location semantics. Combined with urban planning knowledge, we manage to comprehend the land use meanings of 9 POI clusters. Furthermore, we choose meaningful combination of mobility and location semantics for stations' similarity case studies, inspired by which, conclusive discoveries and planning suggestions are summarized.

Keywords: Urban Computing, Smart Card Data, Data Mining, Human Mobility, Urban Planning

1. Introduction

1.1. Background

Rapid urbanization procedures require a sound integration between urban land use and public transportation systems. Meanwhile, subway system is becoming the essential part of public transportation worldwide because of its prominence in speed, reliability and capacity comparing with other transportation modes (Sun et al., 2012). For public transport agencies and operators, it is necessary to understand their similarity in aspects of passengers' traveling demand and nearby land utilization, i.e. mobility and location features, to better construct public intelligent transportation system (Pelletier et al., 2011). Nowadays, the implementation of automated fare collection (AFC) systems and open-source land use data allow public transport agencies and researchers to fully understand the hidden mechanism of urban mobility. As a result, using such data to analyze public transportation system and passenger behaviors is becoming an important part in traffic analysis. Long and Thill (2015) combined smart card data with a household travel survey to recognize the jobhousing locations and commuting trip routes in Beijing. Lee et al. (2012) established two models to analyze the bus service reliability, where different operating strategies were applied and tested in a simulation environment with passenger demand extracted from smart card data in Singapore. Sun et al. (2013) proposed an encounter network based on smart card data to depict the "familiar stranger" social phenomenon. Such sources of data have broadened an ever-expanding aspect on city dynamics.

In the field of NLP, words similarity can be understood through embedding one-hot encoded

*Corresponding author
Email addresses: zdysdsd@sjtu.edu.cn (Dingyi Zhuang), siyuhao@u.nus.edu (Siyu Hao), dhl@nus.edu.sg (Der-Horng Lee), jiangang.jin@sjtu.edu.cn (Jian Gang Jin)

vectors into semantic word vectors according to their context, which is known as distributed word representation or word embedding (Hinton et al., 1986; Bengio et al., 2003). Traditionally, words are recorded as one-hot encoded vectors which have the same dimension with the whole dictionary. For example, if a document contains 1000 different words (i.e. having a dictionary with 1000 dimensions) while word "professor" is the first word in the dictionary (Brownlee, 2017). Then the word can be represented by a 1×1000 vector with the first element as 1 and the others as 0. Embedding is the operation that projects high-dimensional one-hot encoded vectors into a low-dimensional space to obtain embedding vectors. The most popular embedding methods should be CBOW and the skip-gram model in the famous *word2vec* toolbox to learn the meanings of words from their contexts (Mikolov et al., 2013b,a; Li et al., 2015). Notably, embedding vectors, or semantic vectors, generated by those models could preserve the syntactic and semantic correlations between words in a linear perspective. Take an instance, relations in semantic vectors can be shown as $China - France \approx Beijing - Paris$, $good - best \approx high - highest$. Therefore, by clustering semantic word vectors, synonyms can easily be found (Wang et al., 2015).

1.2. Motivation

It is reasonable to correlate these two fields according to similar sequential features of their research objects. The journeys of passengers recorded in smart card data consist of multiple bus stops or subway stations in a sequence. If we regard each stop or station as a word, the whole journey can be interpreted as a sentence. Naturally, we try to transplant the concept of word embedding to stations mobility discovery. Hence, the semantic vectors generated after embedding could reflect the mobility semantics in a sense (Yuan et al., 2015). However, popular methods like *word2vec* might not be applicable, where the contexts of words are learned to infer their meanings (Mikolov et al., 2013b,a). The previous and the consequent stops for each station are majorly fixed, leading to redundant contexts as training data. Therefore, embedding methods require sophisticatedly weighed.

In addition, some words have their literal meanings whose semantic meanings can be easily inferred, typically, like compound words. For example, "superman" consists of "super" and "man", whose meaning could be speculated as a "man" with "super" power. In our context, where we regard stations as words, we would like to further consider the inherent location features of stations, like POI, to understand the semantics of stations as compound words. By combining the mobility and location features of stations, we believe we can obtain deeper understanding of stations analysis.

1.3. Related work

Traditionally, the research focus of smart card data usually lies in the individual or aggregated mobility features, without incorporation with location features (Sun et al., 2013; Liu et al., 2009; Ghaemi et al.; Yang and Ma, 2015; Ma et al., 2017). Liu et al. (2009) have studied both regularity of human mobility patterns in a large scale network and the transportation behaviors in an individual perspective. Furthermore, they take Shenzhen city in China as a case study to understand the mobility functions of metropolitan transportation system. Although it is a great indication to combine mobility analysis and urban planning research, integration with contextual features, like weather, season and so on, can be complementary with their work. Ghaemi et al. propose to project the public transport data into a 2-D semi-circle space with progressive analysis on the aggregated spatial and temporal data. However, they simply try to comprehend the user behaviors with their spatial or temporal similarity, without consideration on the land use or functional features that each station preserves. Therefore, in smart card data research, mobility analysis can answer how passengers travel while location's functional analysis can response why.

With the aid of deep learning, some researchers attempt to grasp deeper, or more abstract, comprehension of human mobility patterns with variant data sources, including smart card data (Song et al., 2016; Zhang et al., 2017; Polson and Sokolov, 2017; Han and Sohn, 2016; Zhengfeng et al., 2017). Specifically, Han and Sohn (2016) have employed deep belief network (DBN) to reduce the dimensions of O-D flow matrix, which could assist in further clustering operations and similarity analysis. With certain ratios, they have proved that DBN-based results outperform traditional zoning method and PCA-based zoning approach. This study has enlightened us that the implementation of deep learning frameworks to extract the features of stations might provide us with deeper insight of their similarity. Moreover, Vincent et al. (2010) have proposed to use autoencoders to learn useful

representation which can bridge the performance gap with DBN. Therefore, Zhengfeng et al. (2017) have implemented SAE network on smart card data for a more sophisticated data representation, then obtain a more accurate traffic prediction. In conclusion, SAE might be an ideal tool to better comprehend human mobility from smart card data.

Besides, semantic models are now widely applied in fields outside NLP, including being transplanted into urban computing and intelligent transportation system. Yuan et al. (2015) have firstly proposed to understand the latent activity behind human behaviors with semantic models to divide the functional zones of Beijing. They have introduced the concept of extracting mobility semantics and location semantics, and then propose to regard them as metadata of a document. With further topic modeling algorithms implemented, like Latent Dirichlet Distribution (LDA) (Blei et al., 2012), abstract semantic meanings of different functional zones are interpretable. Hinted by Yuan et al. (2015), Wang et al. (2017) also propose to understand subway stations as documents. They have applied similar framework to obtain the latent functions (i.e. semantics) of stations with LDA. However, restricted by the number of stations, topic modeling methods seem not to be efficient in understanding semantics of stations, which will be detailed illustrated in section 4.2.2. Therefore, we believe the incorporation with land use knowledge is crucial to interpret the semantics, and we present our own method to understand the mobility and location semantics in the following sections. What's more, the transplantation of semantic models to subway stations lacks in fully integration with urban planning analysis, which would also be our focus.

1.4. Contribution

From above, it is obvious that smart card data have been an ideal target for previous research and the methodology transplantation from fields outside transportation, like NLP, has gradually emerged. In this paper, our contributions are threefold:

1) We transplant the contextual and literal meaning of compound words to the mobility patterns and inherent location features of subway stations, as shown in Table 1
2) We propose to analyze the location semantics of MRT stations in a planning perspective, with 9 functional categories recognized

Table 1: Mapping between document semantics extraction and station features

Compound word	$\longrightarrow$	MRT station
Context	$\longrightarrow$	Previous & next stations
Document	$\longrightarrow$	Journey records
Metadata	$\longrightarrow$	POI and flow matrices
Topics	$\longrightarrow$	Mobility & location semantics

3) We conduct 4 similarity cases studies based on semantic models, and then suggest several commercial and planning recommendations

The rest of the article is organized as follows. In section 2, we describe the data and how we preprocess them. We introduce the methodology to extract semantics in section 3. The detailed transplantation and comprehension of semantic models are presented in section 4. Case studies and discussion about commercial and planning recommendations are conducted in section 5. Finally, we conclude in section 6 and express our acknowledgement.

2. Data and pre-processing

2.1. Smart card data

2.1.1. Data description

This research uses anonymous smart card data from a week in 2012 that cover 96% of public transit trips in Singapore as the mobility analysis example. Data record trips from Singapore's fare collection system, kindly provided by the Singapore Land Transport Authority (LTA). The smart card has been used to pay for public bus and subway trips since April 2002. Different from other smart card data sources as Munizaga and Palma (2012) and Trpanier et al. (2007) use, the most important advantage of this dataset is that it provides the precise location and time information for both boarding and alighting (Sun et al., 2012). This advantage owes to the tapping-in and tapping-out regulation that could ultimately depict the whole journey of each passenger. Whereas, situations that passengers forget to tap out happen frequently. We have detected and cleaned out these items.

Subways, including massive rapid transit (MRT) and light rail transit (LRT), with nearly 2,000,000 transaction counts each day, have taken up nearly 40% of public transportation usage as the dataset has reflected. However, there are 4560 bus stops in Singapore and only 122 subway stations. We can

Table 2: Subway lines with corresponding codes and types

Name	Code	Type
North South line	NS	MRT
East West line	EW	MRT
North East line	NE	MRT
Circle line	CC	MRT
Circle line Extension	CE	MRT
Bukit Panjang LRT line	BP	LRT
Changi Airport branch line	CG	MRT
Punggol line (East)	PE	LRT
Sengkang line (West)	SW	LRT
Sengkang line (East)	SE	LRT

only obtain the similarity among stops/stations by taking all of them into consideration. Large number of bus stops would lead to curse of dimensionality if we construct the origin-destination (O-D) matrices. Hence, for mobility analysis of this study, only records on boarding and alighting stops, along with boarding and alighting time, of subway passengers are selected. Specifically, information of subway lines is listed in Table 2.

2.1.2. Temporal segmentation based on variability

Given that the smart card data are recorded in the sequence of local time, it is necessary to aggregate them into time intervals for a more convenient processing. Here we introduce the concept of variability combined with our comprehension of temporal distribution of public transportation demand to divide daily data into seven segmentations (Zhong et al., 2016). Different from Zhong et al. (2016), we consider the station as the object and the boarding passengers number as the feature, and determine the station k as a vector $X_k(i)$

$$X_k(i) = [x_1, x_2, ...x_t, ...x_n]$$

Where t is the time slot with $t \in [1, ..., n]$ and n is the number of time intervals. Since the boarding passengers number is the feature x_t, $[x_1, x_2, ...x_t, ...x_n]$ can reflect the temporal pattern of the stations, which relies on the number of time slots n. In addition, i is the denotation of the day. Based on that, for day i and time slot n, we describe the station matrix $\mathbf{X}_n(i)$ as

$$\mathbf{X}_n(i) = \begin{bmatrix} X_1(i) \\ X_k(i) \\ ... \\ X_m(i) \end{bmatrix}$$

m is the total number of subway stations, which equals to 122. Moreover, we define the variability formula between two station matrices $\mathbf{X}_n(i)$ and $\mathbf{X}_n(j)$ with the same time slots n as

$$Var = 1 - \frac{\sum ColCosSim(\mathbf{X}_n(i), \mathbf{X}_n(j))}{n}$$

Where $ColCosSim$ is the operation to calculate the cosine similarity of the corresponding columns of $\mathbf{X}_n(i)$ and $\mathbf{X}_n(j)$.

Variability is to reflect the internal correlation under the influence a given variance, which indicate that patterns with low variability are more predictable and apparent (Zhong et al., 2016). Variability is clearly within [0,1], and lower values indicate better internal correlation. Here we introduce variability as an index to determine how different time-interval division can influence the internal correlation of total number of boarding passengers. Time intervals vary from 1min to 24hours and we average variability between different station matrices. The result is shown in Figure 1, the variability decreases as the size of time intervals increases.

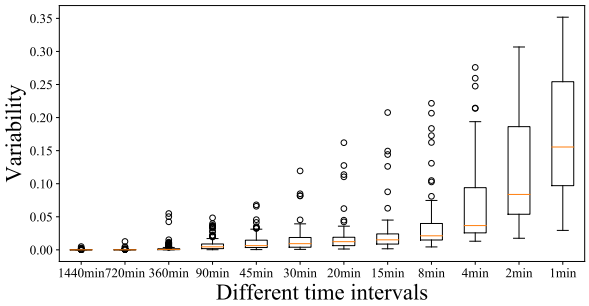

Figure 1: Average variability boxplots across different time intervals

In the aspect of suitable variability, 1 hour or larger might be an ideal time interval as Mahrsi et al. (2017) have inferred. Here we choose majorly 2-3 hours combined with our empirical knowledge on daily transport flow to aggregate smart card data into seven time intervals, which is shown in Table 3.

Table 3: Segmentation of time intervals

Time	Time interval name
5:00-7:00	pre-morning peak
7:00-10:00	morning peak
10:00-16:00	morning off-peak
16:00-17:00	pre-evening peak
17:00-19:00	evening peak
19:00-22:00	late evening peak
22:00-24:00	evening off-peak

2.1.3. Flow matrices construction

As introduced in section 1.2 that, under certain public transportation system, the previous and the next stops for each station are usually fixed, which means that straightly applying current context-based word embedding methods might not be practical. Sequential journey records are actually a continuous series of O-D pairs that occur within differnt time intervals. By preserving the sequential features and temporal patterns of journeys and referring to O-D matrices, we propose to construct flow matrices based on our temporal segmentation.

For each station, we consider the inflow from all the stations to this one and the outflow from this station to all the stations within certain time interval in seven different days. Specifically, the inflow and outflow from the station to itself are always 0, which can serve as the identification information for the station if data are trained in the neural network. Moreover, we use one-hot coded vectors with 7 dimensions for the representation of days. In this way, we believe the flow matrix is the integrated representation of O-D information and temporal features.

Therefore, for the flow matrix in a given time interval, it has $7 \times 122 = 854$ rows (122 stations in 7 days) and $122+122+7 = 251$ (inflow, outflow and index for the day) columns. Totally, there are seven flow matrices that can be processed to extract the mobility semantics.

2.2. POI data

2.2.1. Data collection

To better understand the location features, or land use patterns, of each subway station, we collect POI data from Google Maps. Given that Google Maps has detailed categories for various POI, we pick up 22 of them. Hinted by Wang et al. (2017), we search all the POIs within 500 meters from the target station and count the term frequency of different POI categories for it. Detailed information of POI categories is listed in Table 4.

Table 4: POI categories and codes

POI category	ID	POI category	ID
Atm	1	Storage	12
Bank	2	Intersection	13
Bus station	3	Lodging	14
Transit station	4	Hospital	15
Place of worship	5	Car rental	16
Supermarket	6	Car dealer	17
Shopping mall	7	Car repair	18
Education	8	Bar	19
Parking	9	Cafe	20
Park	10	Government	21
Political	11	Bicycle store	22

2.2.2. Inverse document frequency (IDF)

However, the POI data collected have the upbound of 200 which is the maximum returns restricted by Google Maps. Therefore, the term frequency of specific category, like food, will reach 200 nearly for all the stations. To solve this problem, we introduce inverse document frequency in the natural language processing (NLP) to select important POI categories that better distinguish each station (Jones, 2004). In our context, for each POI category, the formula of IDF can be presented as (Leskovec et al., 2014):

$$idf(t, D) = log\frac{N}{|d \in D : t \in d|}$$

Where

- N: total number of stations
- t: certain POI category
- D: set of stations
- d: certain station
- $|d \in D : t \in d|$: number of the stations whose POI category t is not zero

By introducing the IDF factor, POI that do not frequently appear will achieve higher weight, which is important in POI clustering and location semantics identification . We determine the data after multiplying the POI term frequency data and the corresponding IDF factors as POI-IDF data.

3. Methodology

3.1. SAE network

By representing its input, an autoencoder (AE) is a neural network can extract deeper representation of the data (Lv et al., 2015; Baldi, 2012). An AE consists of two parts , encoder and decoder. Encoder part is a fully connected network that compress the input vectors into semantic vectors. Encoder and decoder parts share the same semantic vectors. Given that AE is used for compression of the data, overfitting might be acceptable. Semantic vectors are decoded The Loss function of AE is defined as:

$$Loss = \frac{(Output - Input)^2}{SampleSize}$$

SAE network is constructed by stacking AEs to create a deep network. It takes the output of previous layer as the input of current layer (Bengio et al., 2007). More specifically, considering SAE network with l layers. With the training data as input, the first layer is trained as an AE. After training the first hidden layer, the output of $m-th$ hidden layer is used as the input of the $(m+1)-th$ hidden layer. Through this way, multiple AEs can be stacked sequentially and hierarchically.

In the context of our research, the compressed vectors are the targets we need. Since traveling demand varies greatly in different time intervals, we train seven SAEs accordingly to obtain the more accurate mobility semantics. Network structure is illustrated in Figure 2

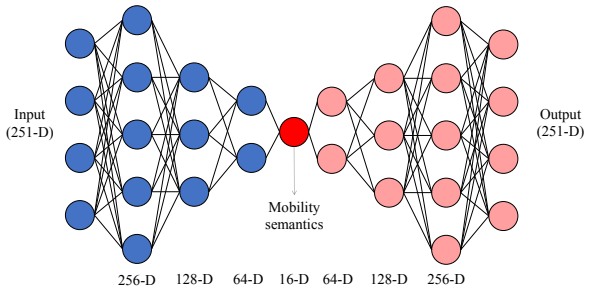

Figure 2: Structure of SAE network

3.2. Singular Value Decomposition (SVD)

We choose SVD to decompose data for better representation. Consider that $\mathbf{M}$ is a matrix with dimension $m \times n$, whose elements could be either the real numbers or complex numbers. Then there exists a factorization with the form (Golub and Kahan, 1965):

$$\mathbf{M} = \mathbf{U} \sum \mathbf{V}^*$$

Where

- $\mathbf{U}$ is an $m \times m$ unitary matrix
- $\sum$ is a diagonal $m \times n$ matrix with non-negative real numbers on the diagonal
- $\mathbf{V}$ is an $n \times n$ unitary matrix while $\mathbf{V}$ is the conjugate transpose of $\mathbf{V}$

The diagonal elements of $\sum$ are named as the singular values of $\mathbf{M}$. Specifically, the diagonal matrix $\sum$ is exclusively determined by matrix $\mathbf{M}$.

3.3. Affinity propagation (AP)

Depending on the concept of "message passing" between data points, AP is the clustering algorithm that does not require to determine the expected number of clustering centers (Frey and Dueck, 2007). The details of the algorithm are illustrated as follow:

Let x_1 to x_n be a a group of data points, and f be the function that measures the similarity between any two points. Here we use Euclidean distance as the measurement function f , such that $f(x_i, x_j) > f(x_i, x_k)$ if x_j is closer to x_i than x_k. Based on the measurement function, the algorithm aims to update two matrices:

- The "responsibility" matrix $\mathbf{R}$, which has value $r(i,k)$ that assess the how suitable x_k is to be regarded as the clustering center for x_i comparing with other candidates points for x_i.
- The "availability" matrix $\mathbf{A}$, which has value $a(i,k)$ that indicate how appropriate for x_i to pick x_k as its clustering center, considering other points' preference for x_k as clustering center.

Both matrices are initialized to all zero then repeat two updating steps:

- Firstly, update all the values in matrix $\mathbf{R}$ by $r(i,k) \leftarrow f(i,k) - \max_{k' \neq k} a(i,k') + f(i,k')$
- Then, update all the values in matrix $\mathbf{A}$ by

$$a(i,k) \leftarrow min(0, r(k,k) + \sum_{i' \notin i,k} \max(0, r(i',k)))$$

$$a(k,k) \leftarrow \sum_{i' \neq k} \max(0, r(i',k))$$

The iterations are repeated until either no further cluster boundaries changes over several iterations, or after the predetermined number of iterations. The clustering centers are extracted from the final matrices with 'responsibility + availability' for themselves being positive (i.e.$(r(i,i)+a(i,i))>0$).

4. Results and discovery

We train seven SAE networks to extract the semantic representation of mobility patterns using seven flow matrices and apply AP on the POI-IDF data to obtain the location semantics.

4.1. Mobility semantics

4.1.1. Training results

- The structure of our SAE network is illustrated in section 3.1. With flow matrices as inputs, we implement our SAE model on a 64-bit server with an Octa-Core 2.20G CPU and 16GB RAM. Each SAE network is trained with 200000 epochs, 128 batch size, GeForce GTX 1060 3GB and an Adaptive Gradient Optimizer with learning rate 0.01. It takes 7 hours to train each model. The convergence of loss function is presented in Figure 3.

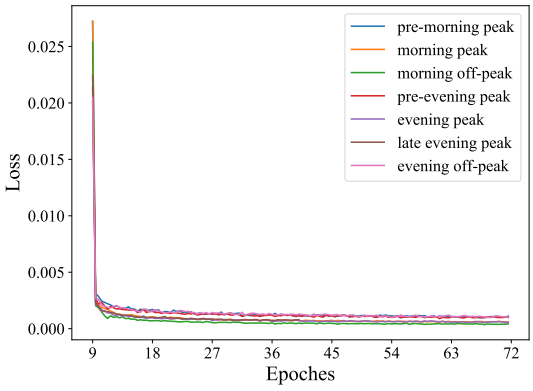

Figure 3: Trainning records of SAE network

Given that the target of SAE network is to reproduce the input data, we could depict the training performance by calculating the R-squared values of each model. Closer to 1, better the training result. Detailed information can be found in Table 5.

Table 5: R-squared examination of SAE performance

Time interval	R-squared value
pre-morning peak	0.881
morning peak	0.951
morning off-peak	0.959
pre-evening peak	0.882
evening peak	0.948
late evening peak	0.947
evening off-peak	0.865
Mean	0.919

4.1.2. Validation of transplantation

In the context of word embedding, words with similar semantic meanings can be clustered into the same group in the embedded space (Mikolov et al., 2013b). Also, semantic vectors preserve the syntactic and semantic correlations between words in a linear perspective, like $China - France \approx Bejing - Paris$ and $good - best \approx high - highest$. Thus, it is logical to assume that mobility semantics can also be separated into different groups. Besides, groups should likewise contain specific meanings that verify from each other. Therefore, we propose to validate our transplantation of word embedding in two aspects:

1) First, whether mobility semantics vectors of seven SAE models can be separated into different groups.
2) Second, whether corresponding elements between two groups would all share similar distance.

Firstly, we implement TSNE, a widely-used decomposition method for high dimensional data visualization, on the mobility semantic vectors we train. The visualization of mobility semantics is given in Figure 4. It is obvious that mobility semantic vectors are formed into seven groups in accordance with seven time-intervals. The elements in each group are thus mobility semantics for different stations in different days.

This result can actually be explained through variability analysis. On account of the 854 elements with 16 dimensions trained by each SAE network, only time interval and the day can influence the formation of groups. Therefore, we depict the variability of mobility semantics under different time intervals and days, which is illustrated in Figure 5.

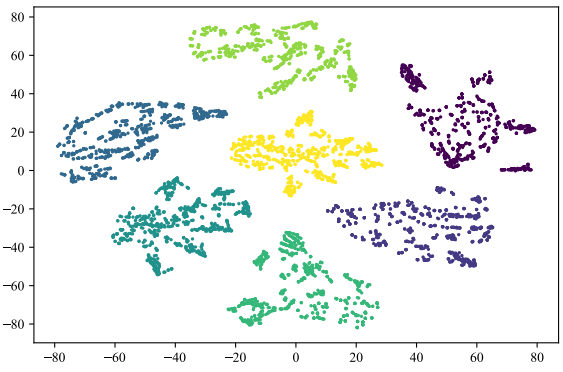

Figure 4: 2-D projection of mobility semantic vectors with TSNE

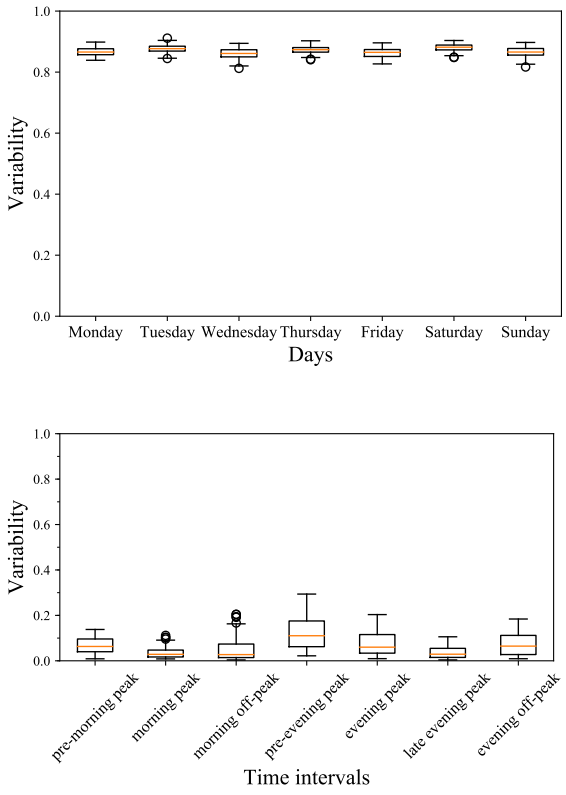

Figure 5: Variability comparison for mobility semantic vectors between within the same day and during the same time interval

Secondly, we would like to further analyze the similarity of distance of corresponding elements in every two groups. In our context where time-interval is the meaning of each group, we need to validate whether distance between "Station 1 in Monday's morning peak" and "Station 1 in Monday's evening peak" is similar to distance between "Station 2 in Friday's morning peak" and "Station 2 in Friday's evening peak". We implement the idea on each pair of groups based on cosine similarity and then calculate their mean cosine similarity to draw a heatmap, which is presented in Figure 6. Concerning that no baseline can be compared, we just roughly judge that an average 0.8 cosine similarity is acceptable.

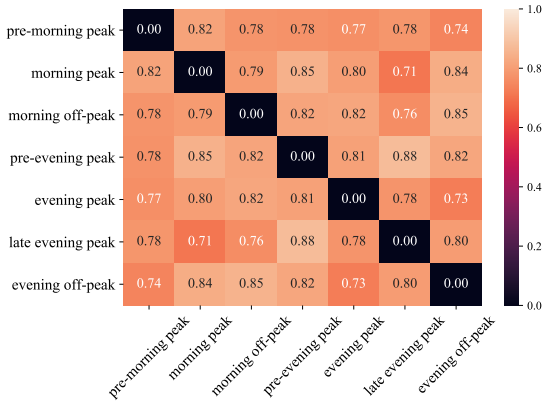

Figure 6: Average cosine similarity comparison between each pair of mobility semantics groups

In conclusion, our transplantation still preserves the distinct attributes of word embedding.

4.2. Location semantics

As introduced above, compound words can also be comprehended through their literal meaning, which can be considered as the inherent location features of stations. Given the fact that data might lose the interpretability when conducting decomposing operations or multiplying IDF factors, we propose to cluster the POI-IDF features of stations to understand their location semantics using our knowledge of urban planning. By studying the stations clustered in the same group, we could somehow infer the meaning of their location semantics. The method we apply is affinity propagation, since it does not require the predefinition of expected number of clusters.

4.2.1. POI-IDF data decomposition

However, before clustering, as demonstrated by (Yuan et al., 2015), POI-IDF vectors are still not a

good representation of the location semantics considering two limitations: 1) *Missing values*. Some POI that featuring the location semantics of the station might not be recorded in the POI database of Google Maps. 2) *Latent semantics*. The frequency of POI, together with IDF factors, are not the intrinsic depiction of the latent semantics of POI configuration for each station.

Therefore, further modification on the POI data is required. We apply the SVD method to find the latent location semantics of each stations, which inherently solve the above limitations(Yuan et al., 2015; Su and Khoshgoftaar, 2009). As a result, the decomposed result could be regarded as the latent semantics of POI-IDF vectors. The benefit of implementing SVD before AP clustering could be reflected in Table 6, where both silhouette score (Rousseeuw, 1987) and Calinski-Harabaz score (Caliński and Harabasz, 1974) increase, indicating a better clustering performance.

Table 6: Performance of SVD on clustering results

	Before SVD	After SVD
Silhouette score	0.12197	0.29851
Calinski-Harabaz score	42.462	156.810

4.2.2. Land use meanings of clusters

In the context of NLP, employing SVD on TF-IDF vectors to discover their latent semantics is known as latent semantic analysis (LSA). In our context, the transplantation of LSA can unveil the location semantics of stations. As mentioned above, in order to understand the specific meaning of location semantics, we need to further comprehend the result of LSA using our urban planning knowledge on their clusters. Therefore, we apply AP on the POI-IDF after SVD. 13 clusters are generated, with 4 clusters only consist of 1 station. Since we focus on the similarity analysis among stations, we only consider the 9 clusters containing multiple stations. We then calculate the average POI term frequency of each cluster, which is illustrated in Figure 7.

Previous work has introduced topic modeling, like LDA (Blei et al., 2012), to interpret the meaning of POI clusters as document (Wang et al., 2017; Yuan et al., 2015). LDA could decompose the documents-words matrix into documents-topics and topics-words matrices to describe the distribution of topics for each document and the meaning of each document. However, if we implemted LDA, with only 122 stations as documents and 22 POI

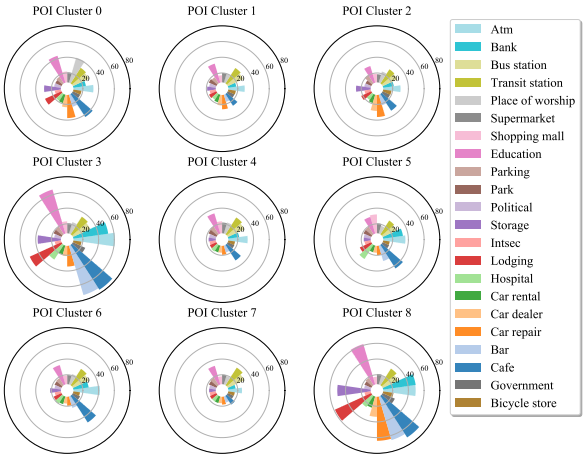

Figure 7: POI configurations of location semantics clusters

categories as words, small volume made it difficult for LDA to learn the topics (i.e. meanings) of each clustering groups. Consequently, we manage to understand the meaning of location semantics in the aspects of land use patterns as well as POI category distribution in different regions of Singapore. The distribution of location semanticscluster for differnt stations is presented in Figure 8, where four stations marked white belong to the clusters with only one element.

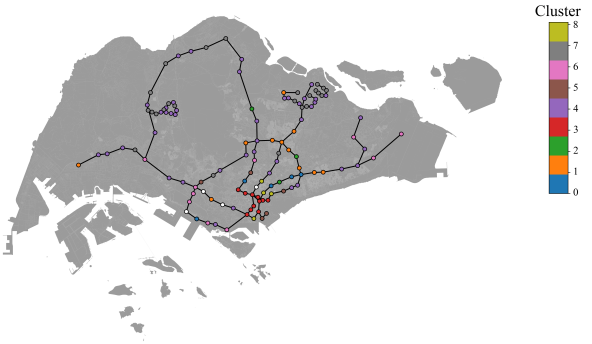

Figure 8: Distribution of location semantic clusters

The detailed comprehension of meanings of different groups can be annotated as follows:

Developed residential area [Cluster 0]. The average POI category distribution of this cluster is quite even, demonstrating that basic infrastructures are quite complete. Besides, the number of each POI category is also above average. Therefore, it is clear that stations in Cluster 0 lie in mature residential areas with an adequate number of service

support like bus stations, cafs, hospitals, parking lots and shopping malls. Lavender MRT station, as an example, is in a largely commercial region, with several blocks of public housing.

Under-development areas [Cluster 1]. With stations scattered in many planning areas provided by the Singapore Urban Redevelopment Authority, many stations and their surroundings in this cluster are now under development. Joo Koon MRT Station, the terminus of East-West Line in 2012, is turned into an industrial estate and is under development of relevant fields like food centers and so on.

Old residential areas [Cluster 2]. This cluster, with evenly distributed POI categories, has more lodging and financial service. However, comparing with Cluster 0, it lacks in education and car service. Kallang, a typical residential town among the first new town of Singapore (Wikipedia, a), is known for its public housing scale along with other landmarks including the old National Stadium and the country's first planning civil airport, the Kallang Airport.

Developed commercial areas [Cluster 3]. With the highest number in many POI categories including financial, education and car service, this cluster could reflect the mature commercial development of areas around the stations. Raffles Place MRT Station and Somerset MRT Station in this cluster are well known for their location in the prosperous east coast of Singapore.

Potential planning areas [Cluster 4]. Areas might be potential for future planning, which can be inferred from the deficient amount of POI configurations of these stations. Also, many areas around the stations, including Pioneer, Sengkang and Clementi, are listed in the planning area of Singapore (Wikipedia, b).

Entertainment areas [Cluster 5]. Those stations are situated in the mature business circles, serving the typical entertainment and commercial zones in Singapore, such as Bayfront, Holland Village and Marina Bay.

Scientific/Educational areas [Cluster 6]. Many educational and scientific institutes are located near those stations. Take Kent Ridge MRT Station as an example; it is surrounded by Science Park and National University of Singapore, which are famous research center and college. It is the same with Jurong East, surrounded by the Science Centre Singapore, famous for its promotion of scientific education for the general public.

Emerging residential area [Cluster 7] Stations in this cluster usually locate in regions in the north or belong to the LRT. These regions are becoming public housing estates.

Emerging commercial areas [Cluster 8] With POI category distribution similar to but less than Cluster 3, stations could be considered as locating in emerging commercial areas. Tanjong Pagar, take an instance, might be attractive to many customer groups, but it is still lack of scale when comparing with Orchard or other large shopping centers.

5. Case studies and discussion

5.1. Semantics case studies

Our ultimate purpose is to discover the similarity among stations with our discovered mobility and location semantics. We choose four combinations in the aspects of mobility semantics, location semantics and whether stations are in the same subway line. We take subway lines into consideration because we aim to increase the interpretability of the results. Otherwise, whimsical correlations among stations might confuse us. To specify, we regard stations in the same POI clusters as same location semantics. Besides, we introduce cosine similarity to assist in the determination of mobility semantics similarity. We first extract the mobility semantics matrix of each station. Considering seven time intervals in seven days, the mobility semantics matrix should have the 49×16 dimensions. Then we calculate the cosine similarity between each two mobility semantics matrices to obtain a correlation matrix with 49×49 dimensions. We consider the number of elements larger than 0.99 or less than 0 in the correlation matrix as the similarity or dissimilarity degree. Detailed case studies are illustrated as follows:

5.1.1. Different lines, same location semantics, same mobility semantics

Will trains running in different lines serve similar groups of passengers? This combination targets at this question. As shown in Figure 9, remote stations or LRT stations are discovered. These stations usually do not possess abundant POI configurations nor passenger flow, which leads to their resemblance in both service and mobility semantics. In addition, discovered stations might usually share the same interchange station, like Farmway

Station and Woodleigh Station, marked in the figure, both share similar passenger flow patterns from Sengkang. Therefore, although land use and passenger flow vary greatly along different lines, trains running in different lines still might serve the same group of passengers.

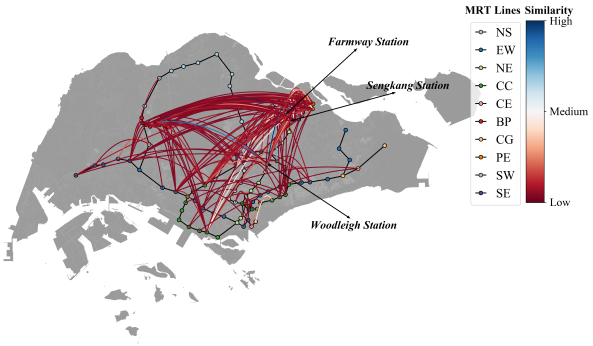

Figure 9: Mobility simialrity relationship for stations in the same POI cluster but differnt subway lines

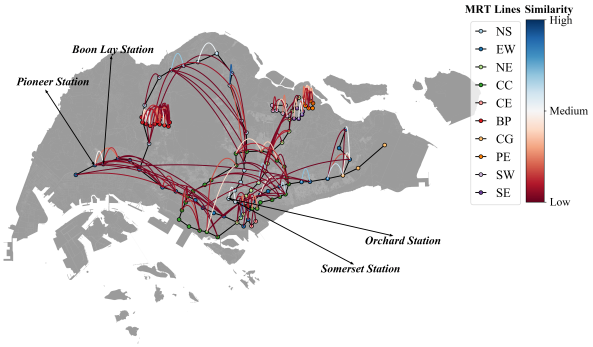

Figure 10: Mobility simialrity relationship for stations in the same POI cluster and subway line

5.1.2. *Same line, same location semantics, same mobility semantics*

As presented in Figure 10, even though many discovered stations pairs share their similarity in a long distance, it is explicit that adjacent stations also take a major part. Typical station pairs with similar land use and passenger flow are included, such as Somerset and Orchard and Pioneer and Boon Lay. Somerset and Orchard, as characteristic developed commercial areas, serve passengers with high purchasing power. On the other hand, Pioneer and Boon Lay, located in the potential planning area, mainly serve the salariats. Hence, further commercial discussion may be proposed.

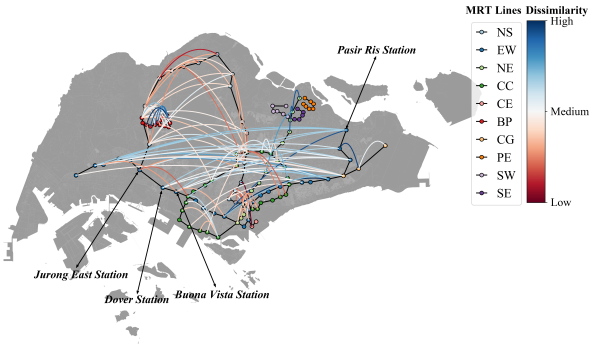

Figure 11: Mobility dissimialrity relationship for stations in the same POI cluster and subway line

5.1.3. *Same line, same location semantics, different mobility semantics*

Analogous to the first case, remote station pairs, such as Pasir Ris and Dover in the Figure 11, are found. This could be interpreted in the same way as the first case does. Moreover, stations in residential region like Jurong East and Buona Vista, are the intersections of different lines to connect flow demand from different places. Although these two stations are three stations away from each other in the East-West Line, their strong mobility semantics dissimilarity could reflect the diversity among them.

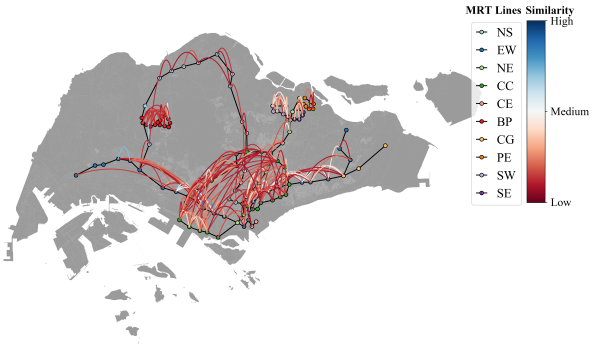

Figure 12: Mobility simialrity relationship for stations in the diffent POI clusters but the same subway line

5.1.4. *Same line, different location semantics, same mobility semantics*

As is shown in Figure 12, stations in the Circle line (CC) and LRT lines take up the majority

as they serve only particular regions. Therefore, POI configurations might vary greatly, especially for opposite stations in the Circle line. While the mobility semantics remain similar since the passenger flow in those circle lines is steady. Considering such an integrated closed-loop system, maintaining its circulation and prevent disasters are essential for those stations in the Circle Line and LRT lines.

5.2. Commercial and planning discussion

We then further propose three recommendations in the aspects of commercial interest and urban planning based on the discussion of previous four case studies.

Advertisement. In general, advertise among adjacent stations in the same line. If we assume that stations with the same location semantics serve similar regular passenger groups, advertisers can focus on stations with same service mobility semantics. What's more, as demonstrated in the first case study, stations with the same service and mobility semantics but in different lines are usually remote and unpopular ones. Consequently, it would be wise to consider the second case study, where adjacent stations in the same line take up the majority.

Infrastructure construction. Infrastructures, including lanes, bus stops and so on, can be constructed according to the same service and mobility semantics. Take Tampines and Jurong East as an example, they are both determined as planning areas and residential towns and they share the same location semantics, mobility semantics and subway line. Constructing similar infrastructures to fulfill the similar traveling demand around those two stations is worth consideration.

Land reuse. Stations with low ridership and POI configurations are recommended to be abolished for better land use. Since Ten Mile Junction LRT station is announced for permanent closure, Woodleigh MRT station and Farmway LRT station, with the same service and mobility semantics but in different lines, could also be abolished to reuse the land for the better purpose.

6. Conclusion

In this article, we propose to study the similarity among subway stations by regarding them as compound words where contextual and literal meanings of words correspond to the mobility and location features of stations. Using smart card data in Singapore, we construct seven flow matrices with our variability-based temporal segmentation, and train SAE networks to extract the mobility semantics. We then validate reasonability of our transplantation from word embedding in two apsects with mobility semantics. Location semantics is discovered through studying 9 POI clutering groups in virtue of our knowledge on land use patterns. We further conduct four case stuides to understand the insight of station similarity with different combinations of mobility and location semantics, based on which commercial and urban planning implications are summarized.

In our future work, we would like to learn the meaning of location semantics in a more quantitative way and conduct more detailed surveys on case studies. Besides, some other information in smart card data, including passengers ages, public bus usage and so on, could be better utilized for a more accurate mobility semantics extraction.

Acknowledgement

The authors would like to express their sincere gratitude to the Singapore Land Transport Authority for supporting this research and providing all the necessary data.

References

Baldi, P., 2012. Autoencoders, unsupervised learning, and deep architectures, in: Proceedings of ICML workshop on unsupervised and transfer learning, pp. 37–49.

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., 2003. A neural probabilistic language model. Journal of machine learning research 3, 1137–1155.

Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., 2007. Greedy layer-wise training of deep networks, in: Advances in neural information processing systems, pp. 153–160.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2012. Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.

Brownlee, J., 2017. Why one-hot encode data in machine learning? https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/.

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods 3, 1–27.

Frey, B.J., Dueck, D., 2007. Clustering by passing messages between data points. science 315, 972–976.

Ghaemi, M.S., Agard, B., Nia, V.P., Trépanier, M., . Public transport identifying temporal user behavior through smart card .

Golub, G., Kahan, W., 1965. Calculating the singular values and pseudo-inverse of a matrix. Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis 2, 205–224.

Han, G., Sohn, K., 2016. Clustering the seoul metropolitan area by travel patterns based on a deep belief network, in: Proc. 3rd MEC Int. Conf. Big Data Smart City (ICBDSC), pp. 1–6.

Hinton, G.E., et al., 1986. Learning distributed representations of concepts, in: Proceedings of the eighth annual conference of the cognitive science society, Amherst, MA. p. 12.

Jones, K.S., 2004. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28, 493–502.

Lee, D.H., Sun, L., Erath, A., 2012. Study of bus service reliability in singapore using fare card data, in: 12th Asia-Pacific Intelligent Transpotation Forum.

Leskovec, J., Rajaraman, A., Ullman, J.D., 2014. Mining of massive datasets. Cambridge university press.

Li, Y., Xu, L., Tian, F., Jiang, L., Zhong, X., Chen, E., 2015. Word embedding revisited: A new representation learning and explicit matrix factorization perspective., in: IJCAI, pp. 3650–3656.

Liu, L., Hou, A., Biderman, A., Ratti, C., Chen, J., 2009. Understanding individual and collective mobility patterns from smart card records: A case study in shenzhen, in: Intelligent Transportation Systems, 2009. ITSC'09. 12th International IEEE Conference On, IEEE. pp. 1–6.

Long, Y., Thill, J.C., 2015. Combining smart card data and household travel survey to analyze jobs–housing relationships in beijing. Computers, Environment and Urban Systems 53, 19–35.

Lv, Y., Duan, Y., Kang, W., Li, Z., Wang, F.Y., et al., 2015. Traffic flow prediction with big data: A deep learning approach. IEEE Trans. Intelligent Transportation Systems 16, 865–873.

Ma, X., Liu, C., Wen, H., Wang, Y., Wu, Y.J., 2017. Understanding commuting patterns using transit smart card data. Journal of Transport Geography 58, 135–145.

Mahrsi, M.K.E., Cme, E., Oukhellou, L., Verleysen, M., 2017. Clustering smart card data for urban mobility analysis. IEEE Transactions on Intelligent Transportation Systems 18, 712–728.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013a. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 .

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013b. Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, pp. 3111–3119.

Munizaga, M.A., Palma, C., 2012. Estimation of a disaggregate multimodal public transport origindestination matrix from passive smartcard data from santiago, chile. Transportation Research Part C 24, 9–18.

Pelletier, M.P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies 19, 557–568.

Polson, N.G., Sokolov, V.O., 2017. Deep learning for short-term traffic flow prediction. Transportation Research Part C: Emerging Technologies 79, 1–17.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics 20, 53–65.

Song, X., Kanasugi, H., Shibasaki, R., 2016. Deeptransport: Prediction and simulation of human mobility and transportation mode at a citywide level., in: IJCAI, pp. 2618–2624.

Su, X., Khoshgoftaar, T.M., 2009. A survey of collaborative filtering techniques. Advances in artificial intelligence 2009.

Sun, L., Axhausen, K.W., Lee, D.H., Huang, X., 2013. Understanding metropolitan patterns of daily encounters. Proceedings of the National Academy of Sciences 110, 13774–13779.

Sun, L., Lee, D.H., Erath, A., Huang, X., 2012. Using smart card data to extract passenger's spatio-temporal density and train's trajectory of mrt system, in: Proceedings of the ACM SIGKDD international workshop on urban computing, ACM. pp. 142–148.

Trpanier, M., Tranchant, N., Chapleau, R., 2007. Individual trip destination estimation in a transit smart card automated fare collection system. I V H S Journal 11, 1–14.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A., 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research 11, 3371–3408.

Wang, J., Kong, X., Rahim, A., Xia, F., Tolba, A., Al-Makhadmeh, Z., 2017. Is2fun: Identification of subway station functions using massive urban data. IEEE Access 5, 27103–27113.

Wang, P., Xu, J., Xu, B., Liu, C., Zhang, H., Wang, F., Hao, H., 2015. Semantic clustering and convolutional neural network for short text categorization, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 352–357.

Wikipedia, a. New towns of singapore. https://en.wikipedia.org/wiki/New_towns_of_Singapore.

Wikipedia, b. Planning areas of singapore. https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore.

Yang, J., Ma, J., 2015. A big-data processing framework for uncertainties in transportation data, in: Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on, IEEE. pp. 1–6.

Yuan, N.J., Zheng, Y., Xie, X., Wang, Y., Zheng, K., Xiong, H., 2015. Discovering urban functional zones using latent activity trajectories. IEEE Transactions on Knowledge and Data Engineering 27, 712–725.

Zhang, J., Zheng, Y., Qi, D., 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction., in: AAAI, pp. 1655–1661.

Zhengfeng, H., Pengjun, Z., Wenjun, X., Gang, R., 2017. Sae for the prediction of road traffic status from taxicab operating data and bus smart card data. International Journal of Modern Physics C 28, 1750121.

Zhong, C., Batty, M., Manley, E., Wang, J., Wang, Z., Chen, F., Schmitt, G., 2016. Variability in regularity: Mining temporal mobility patterns in london, singapore and beijing using smart-card data. Plos One 11, e0149222.