A Multi-dimensional City Data Embedding Model for Improving Predictive Analytics and Urban Operations

Purpose- A smart city is a potential solution to the problems caused by the unprecedented speed of urbanization. However, the increasing availability of big data is a challenge for transforming a city into a smart one. Conventional statistics and econometric methods may not work well with big data. One promising direction is to leverage advanced machine learning tools in analyzing big data about cities. In this paper, we propose a model to learn region embedding. The learned embedding can be used for more accurate prediction by representing discrete variables as continuous vectors that encode the meaning of a region.

Design/methodology/approach- We use the random walk and skip-gram methods to learn embedding and update the preliminary embedding generated by Graph Convolutional Network (GCN). We apply our model to a real-world dataset from Manhattan, New York, and use the learned embedding for crime event prediction.

Findings- Our results show that the proposed model can learn multi-dimensional city data more accurately. Thus, it facilitates cities to transform themselves into smarter ones that are more sustainable and efficient.

Originality- We propose an embedding model that can learn multi-dimensional city data for improving predictive analytics and urban operations. This model can learn more dimensions of city data, reduce the amount of computation, and leverage distributed computing for smart city development and transformation.

Keywords: Smart city; Big data; Machine learning; Region embedding; Graph Convolutional Network (GCN)

1. Introduction

With the continuous growth of the global population and the fast development of urbanization, the urban population is increasing rapidly (World Bank, 2017). The ever-increasing urban population and rapidly changing demographics complicate the urban structure (Boeing, 2018). At the same time, the natural environment is susceptible to various threats, such as energy shortages, air pollution, and global warming (Dong et al., 2018; Ogura & Jakovljevic, 2018). Nowadays, many people living in cities face various risks and problems, such as the shortage of water resources and the unbalanced distribution of medical resources (Hadadin et al., 2010). Therefore, better developing and managing urban areas has become increasingly important in addressing these problems.

A smart city is a potential solution to the problems caused by the unprecedented speed of city development and urbanization (Hall et al., 2000). The concept of smart cities can be traced back to 1974, when the first big data project for cities was created in Los Angeles (Los Angeles Community Analysis Bureau, 1974). Since then, academia and industry have invested time and effort in advancing smart city research. IBM proposed that policymakers treat a city as a complex interconnected network that can proactively predict and solve problems, maximize resources, and use the information to make better decisions (Wiig, 2015). Many academic studies in this domain have explored the constituent elements of smart cities and the interrelationships among them (Hollands, 2008; Allwinkle & Cruickshank, 2011; Lombardi et al., 2012; Chourabi et al., 2012). Based on the definitions and concepts of smart cities put forward by different scholars, a smart city should be able to make conscious efforts to use information systems strategically, seeking to achieve prosperity, effectiveness, and competitiveness at multiple levels of the urban society (Angelidou, 2014). While the goals of a smart city are relatively straightforward, the approaches to transforming a city into a smart one remain unclear.

In recent years, the development of the digital infrastructure, such as the Internet of Things (IoT) and information and communication technologies (ICT), has enabled the

rapid growth of big data at the city level (Batty, 2013; Hashem et al., 2016; Chen et al., 2017). The big data of a city can be considered a spontaneous, objective, and accurate recording of the multi-dimensional characteristics of the city. Big data is usually generated by the passive recording of various people's activities (Rathore et al., 2016). As a result, big data can more comprehensively, objectively, and accurately capture the information of city residents and other physical objects (George et al., 2014; Chen et al., 2015). Therefore, the availability of multi-dimensional city data provides a valuable opportunity to develop smart cities.

However, the increasing availability of big data is also a challenge for transforming a city into a smart one (Li et al., 2019). One of the biggest challenges scholars and governments face is the diversity and hierarchy of data sources—sensors, mobile phone apps, social media, web activities history, and tracking devices, all of which can generate enormous amounts of data (Ghosh et al., 2016). Thus, leveraging big data to achieve smart city transformation has become an influential research topic. In the past, many scholars focused on studying how to develop sustainable and smart cities by analyzing data with statistical and econometric tools. For example, Neirotti et al. (2014) performs a regression analysis of a sample of 70 international cities to identify the crucial factors that influence the coverage index that measures the impacts on the development of smart-city initiatives. Their study helps policymakers under budget constraints prioritize smart-city initiatives, thereby maximizing the return of smart city investments. Liu et al. (2021) use a spatial econometric model to identify key factors influencing smart city development with a sample of 83 Chinese cities. Their results show that governmental support, innovativeness, economic development, and human capital are the four key factors that help policymakers make decisions to develop smart cities.

However, conventional statistics and econometric methods may not work well with big data (Varian, 2014). First, the massive dynamic data renders data manipulation tools in econometrics useless. Second, in many cases, people have to select appropriate predictors from a large number of available variables to improve predictive accuracy.

However, this task cannot be efficiently achieved with conventional econometric models. Third, linear models often do not accurately reflect the relationships among variables in big data. Thus, we need to introduce more flexible models to examine the complex relationships among many variables.

Machine learning techniques such as decision trees, neural networks, and deep learning support the analysis of multi-dimensional city data. The predictions based on machine learning enable cities to achieve the goal of high efficiency and sustainable development (e.g., Din et al., 2019; Shafiq et al., 2020; Zekić-Sušac et al., 2021). Recent studies attempt to predict regional characteristics such as crime rate or traffic flow by learning region embedding from big data to support the development of smart cities (Zhang et al., 2018; Liu et al., 2020). This approach represents discrete variables as continuous vectors that encode the meaning of regions. The learned embedding can be used for identification or prediction, as the regions that are closer in the vector space are expected to have similar regional characteristics (Jurafsky, 2000). The advantages of this method are mainly in two aspects. First, region embedding can be learned from data of different types and from different sources. For example, the data of location, people mobility, and building type are difficult to be handled in conventional econometrics and statistics. However, machine learning can use region embedding to represent such data more effectively. Second, identification or prediction through the learned region embedding is less limited by the scene. The principle of prediction is to leverage the similarity of region embedding. With the appropriate data and a reliable method of learning embedding, people can accurately predict many things in urban areas, such as identifying urban functional areas and predicting local crime rates. These two advantages make this approach particularly suitable for supporting the better development of smart cities.

Accurately learning embedding from big data has become a very important research field. Previous studies have used city data to learn embedding. Some scholars use human mobility flow data to learn embedding. For example, Pan et al. (2012) explore the relationship between taxi trajectory and urban land use. Zheng et al. (2014) use

human as a sensor and model the New York city noise situation with embeddings that include regions, noise categories, and time slots information. Yao et al. (2018) represent urban function through learned zone embeddings by exploiting large-scale taxi traces. However, these methods only consider region correlations hidden in a single dimension of data, such as people's mobility. A few other studies have combined region attributes with human mobility data (Zhang et al., 2019; Fu et al., 2019). While multi-dimensional data are used in these studies, there are at least two directions from which we can improve the learning of region embedding. First, the weights of different attributes of the same region should be considered during the learning process. For example, Zhang et al. (2019) assign the same weight to different attributes in the point of interest (POI) in learning, which reduces the accuracy of learned embeddings. It is challenging to assign weights for different attributes because the weights change dynamically with the city and selected regional attributes. Second, more dimensions of city data need to be analyzed to learn embedding more accurately. In the existing research, embedding has been learned mainly from mobility and POI data. However, there are other types of data, such as labels of regional functions, which represent such main functional areas of the region as the business area. This regional label data is different from mobility and POI data, so it is not easy to learn using past methods. Therefore, we propose a new model to learn region embedding. The proposed method can learn from high-dimensional city data so that the learned embedding can more accurately capture the characteristics of a region. This method can significantly improve the accuracy of our forecasts, thereby facilitating cities to transform themselves into smarter ones that are more sustainable and efficient.

The proposed method in this study is different from those in the previous studies in three aspects. First, we propose a comprehensive method to learn the embedding of different regions from multi-dimensional city data. In this research, we first use self-supervised deep learning methods to learn embedding for point of interest (POI), the type of data describing the attributes of a point in the city. Then we employ Graph Convolutional Network (GCN), an approach for semi-supervised learning to update the

learned embeddings. Our method can integrate more data types into embedding and learn embedding more accurately compared with existing methods. Second, our model is more suitable for practical use in the smart city setting because of its nature of unsupervised learning (our model uses a nearly unsupervised learning method, i.e., only a tiny amount and readily available samples are needed for training). In the preliminary embedding learning process, we use deep walk (an unsupervised deep learning method) and conduct adaptive learning to get the weights of different dimensions in POI data. Therefore, with our method, the only manual intervention for supervised learning is to label each region with its urban function. There is no need to manually label the regions for supervised learning because the type of functional regions in a city can be easily obtained from the government's urban development plan. Consequently, this model can learn embedding from a large number of multi-dimensional unstructured and structured city data with little manual intervention. Third, we demonstrate the proposed method using a real-world dataset from Manhattan, New York from NYC open data website[1]. The dataset includes 5854 POIs, the people mobility data, 12 labels of regional function, and the number of crime events in different regions. We compare the predictive performance of our method against three others to demonstrate the effectiveness and superiority of our proposed method. To sum up, our method can learn the embedding more accurately, and the learned embedding can better present regional characteristics, thereby improving predictive accuracy which is usually imperative for developing smarter cities.

The remainder of our paper is organized as follows. The following section presents the deep learning method and the neural network used in our research. It also includes the method of learning embedding and the description of the datasets. Section 3 shows the images of dimensionality reduction of embedding by deep walk and GCN. It also presents the results that demonstrate the predictive performance of four different methods, including ours. We discuss the limitations and implications of this study in Section 4.

[1] opendata.cityofnewyork.us

2. Method

We propose a framework that utilizes the big data generated by all kinds of sensors, IoT, and manual statistics in a smart city. As shown in Fig. 1, this framework can be divided into three phases. The first phase is to collect city big data, which is in the form of high-dimensional vectors. The second phase is to learn preliminary region embeddings from the original high-dimensional data in the city using random walk and skip-gram, then update it using a graph convolutional network (GCN). The last phase shall be predicting urban characteristics (such as function and crime rate) leveraging the learned region embeddings.

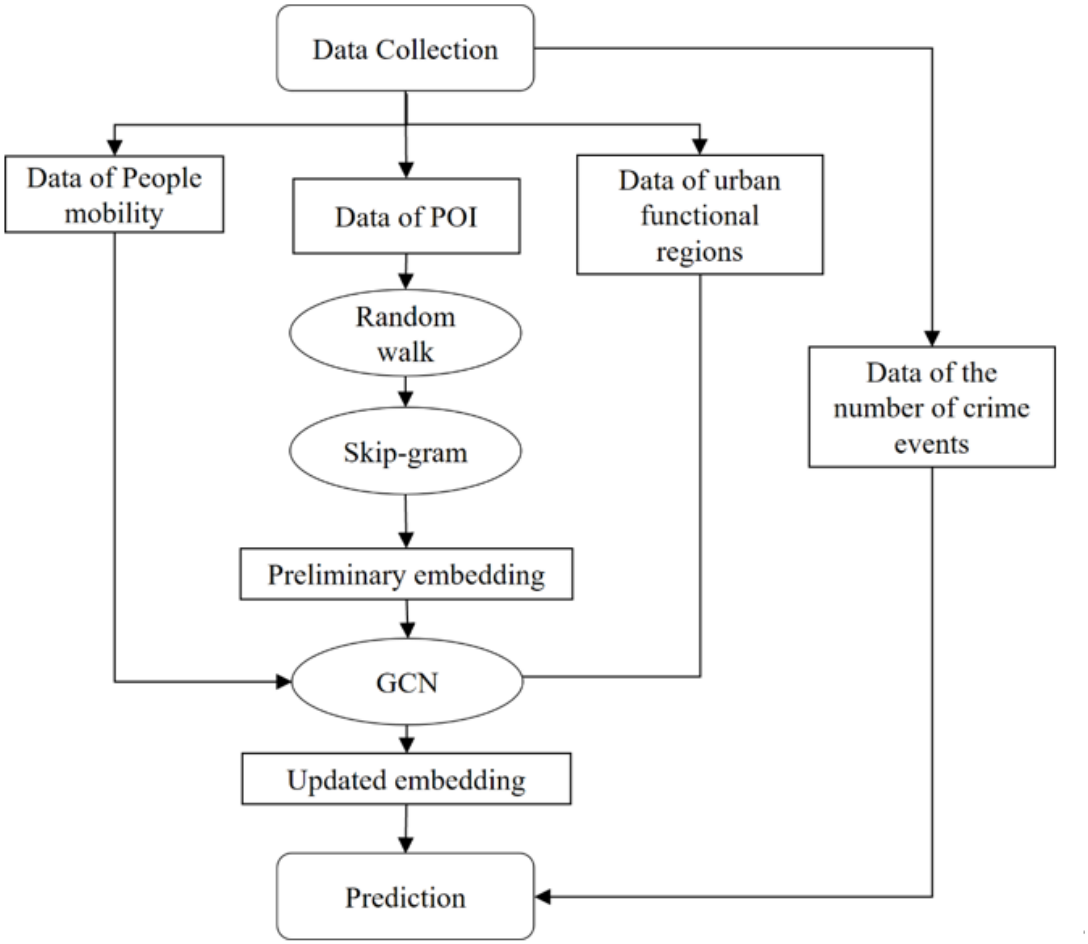

Fig. 1.Framework for learning embedding and prediction based on multi-dimensional city big data

2.1 Preliminary Embeddings

An important task is to learn preliminary region embeddings based on the POI data in the city. Embeddings preserve region characteristics in the form of high-dimensional

vectors. Through this representation, regions with similar characteristics (i.e., have the same function or geographically adjacent) will be close to each other in the embedding space. Based on this key peculiarity, we can further use embeddings to identify or predict a region's characteristics (e.g., number of crimes).

There are four steps to learn the preliminary embeddings (see Fig. 2 and Fig. 3) including pairing connection of POI in a limited distance, constructing a graph, forming paths through the random walk, embedding with skip-gram. These steps are not limited to a specific type of sample and thus have strong applicability to multiple dimensions of information in smart cities.

Types of facilities	**count**	**%**
Commercial	432	7.40%
Cultural Facility	304	5.20%
Education Facility	970	16.60%
Government Facility (non-public safety)	422	7.20%
Health Services	112	1.90%
Miscellaneous	273	4.70%
Public Safety	150	2.60%
Recreational Facility	980	16.70%
Religious Institution	506	8.60%
Residential	850	14.50%
Social Services	375	6.40%
Transportation Facility	447	7.60%
Water	33	0.60%
Total	5854	100%

Table 1 the main types of facilities

Our empirical studies are conducted in Manhattan, New York. which is one of the

most developed areas in the world and is in the process of becoming a smart city. There are two main benefits to using Manhattan as a research environment. First, Manhattan has not only rich and accurate building data, but also taxi mobility data. Second, the blocks of Manhattan use streets as the dividing line, so the boundaries of different blocks are clear, which is convenient for us to learn the region embedding next. We use the data of Point of Interests (POIs), which is a point together with its attributes including its location and facilities such as school, garden, and bus stop. In Manhattan, New York, there is a total of 5854 points (the POI data is publicly available from the Department of Information Technology and Telecommunications of New York). This dataset includes 13 types of facilities (see Table 1). The top three categories are Education Facility (16.6%), Recreational Facility (16.7%), and Residential (14.5%). In addition to the information on facility types, each point has other attributes (there is a total of 11 attributes for a point of interest as shown in Table 2).

Attribute	**Description**	**Field Type**
SEGMENTID	Point is assigned the closest roadbed SegmentID.	double
COMPLEXID	Point is assigned a ComplexID if it is a part of a Complex.	double
SOS	Indicates which side of the street the CommonPlace is on.	text
FACI_DOM	Facility Domains are valid values for each FACILITY_TYPE:	text
BIN	BIN is an abbreviation of Building Identification Number. Point is assigned a BIN if it falls within a building.	double
FACILITY_T	This is a SubType field organizing the CommonPlace points into categories.	integer
SOURCE	The agency that defined the CommonPlace location.	text
B7SC	The Street Code assigned to a CommonPlace.	text
PRI_ADD	The Addresspoint ID if the CommonPlace is related to any Addresspoint	double
NAME	The name of the CommonPlace. Most name come from Feature name table.	text
SAFTYPE	Point is assigned a SAFTYPE if it is a part of a Complex	text

Table 2 the attribute information for point of interest (POI)

2.1.1 Construction of the Urban Functional Corpus

In natural language processing (NLP), corpus is a collection of a large number of processed texts in a predetermined format (Ng & Zelle, 1997), including documents that can be used to simulate natural language on a large scale. In text analysis, the connection between different words in human language can be simulated by the context ranking relationships of different words in the document. Taking advantage of these relationships, we can present the meaning of the word through the context in an unsupervised way. While a POI has accurate latitude and longitude to describe the geographic relationship, the contextual relationships between different POIs are unclear in a city. Therefore, in order to use this method to learn embeddings, we first need to construct the relationships between different POIs. In this paper, we analogize the area to a natural language corpus, since the distribution of POI is similar to the word frequency distribution in a natural language corpus (Yan et al., 2017). We build the corpus through the following steps:

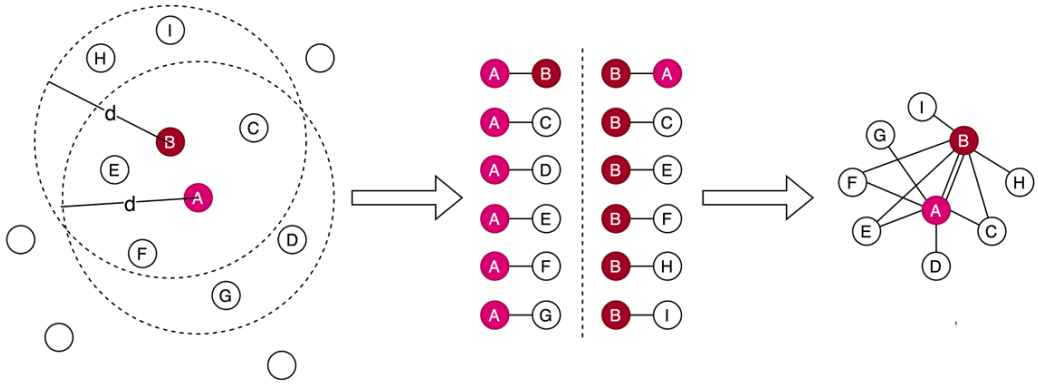

Fig. 2. Construction of the Urban Functional Corpus

1) Pairing

According to the size of the city and the density of the POIs, we choose a distance as the radius. In this study, we choose a radius of 500 meters. We select a POI and form a one-to-many pairing relationship with all other POIs within a chosen radius. After all the POIs are paired, we get 5854 pairing relationships.

2) Composition

According to the obtained pairing relationships, we combined 5854 groups of pairing

relationships to generate a network structure (see Fig. 2). In this way, we transform the geographical coordinates that do not have interrelationships between POIs into a contextual network structure. For example, in Fig. 2, the closest relation between A and B is connected by two lines. From A to C, A can reach C directly, or firstly reach B then to C. However, from A to I, we can only go from A to B and then to I, which means that the distance between A and I is greater than that between A and C.

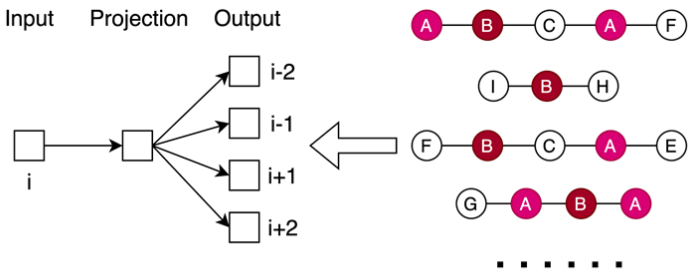

Fig. 3. Random walk and skip-gram

2.1.2 Embedding the Area

In this part, there are two processes including random walk and skip-gram, which are together called the deep walk.

1) Random walk

We have generated a POI network by composition. While this network can capture the distance relationships between different POIs, this global network structure makes it difficult to directly learn embedding by skip-gram. Therefore, we choose the method of random walk to extract local structural information from the global network.

$$P(v_j \mid v_i) = \begin{cases} \frac{\mathrm{M}_{ij}}{\sum_{j \in N_+(v_i)} \mathrm{M}_{ij}} & v_j \in N_+(v_i), \\ 0 & e_{ij} \notin \mathcal{E}, \end{cases} \qquad (1)$$

In Equation (1), $P(v_j \mid v_i)$ is the probability of the random walk. M_{ij} denotes the weight of the edge from point i to point j. This random walk process can get sequences as shown in Fig. 3. Using random walk here has two advantages. First, random walk is easy to compute parallelly. Through distributed computing or multi-threaded computing, different parts of the structured network can be explored at the same time. Second, when there are some minor changes in the authority domain, we only need to iteratively update the learning model with the information obtained by the

new random walk, instead of recalculating the entire structure diagram.

2) skip-gram with attributes

We learn embeddings using the skip-gram algorithm, a self-supervised method. This method maximizes the co-occurrence probability of two points in the obtained order from the random walk. Therefore, this method transforms the problem into the following optimization problem:

$$\underset{\Phi}{\text{minimize}} - \log \text{Prob}\left(\{v_{i-w}, \cdots, v_{i+w}\} \setminus v_i \mid \Phi(v_i)\right) \qquad (2)$$

where w is the window size of the context points that we set. Because the probability distribution is independent, we get

$$\text{Prob}\left(\{v_{i-w}, \cdots, v_{i+w}\} \setminus v_i \mid \Phi(v_i)\right) = \prod_{j=i-w, j\neq i}^{i+w} \Pr\left(v_j \mid \Phi(v_i)\right) \qquad (3)$$

Using negative sampling, the prob function can be transformed into the following objective function.

$$\underset{\Phi}{\text{minimize}} \log \sigma(\Phi(v_j)^T \Phi(v_i)) + \sum_{t\in N(v_i)'} \log \sigma(-\Phi(v_t)^T \Phi(v_i)) \qquad (4)$$

where $N(v_i)'$ is the negative samples for v_i and $\sigma(x) = \frac{1}{1+e^{-x}}$.

With this embedding method, we can capture the similarity among POIs in the city. However, it is still impossible to get accurate embedding for POI because the information contained in each point is high-dimensional. As shown in Table 2, each POI contains 11 types of information including main category, subcategory, etc., which are important information that can reflect regional characteristics. To get the required aggregated embedding of POIs, we need to add the information of different dimensions of a POI according to their weights. We can average the sum of each dimension of information to get the aggregated embedding of the POI, which implies that the contributions of different dimensions to embedding are the same. However, for real world data from a city, different dimensions of information usually have different importance. For example, in our POI data, the contribution of the *main type* and *subtype*

to the expression of regional characteristics is greater than that of the *source*. Therefore, we adopt a self-adaptive method to get the weights of different dimensions (see Fig. 4).

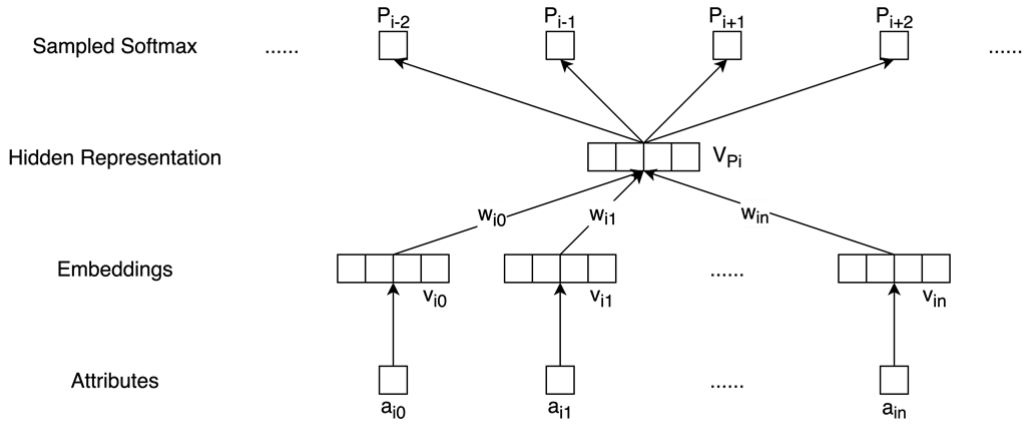

Fig. 4. Skip-gram with attributes

we use a_i^s to denote the weight of the s-th type of attributes of point i, and use a_i^0 denote the weight of the first type of attributes of point i. Then the weighted average layer is defined by the following formula:

$$\mathbf{H}_i = \frac{\sum_{j=0}^{n} e^{a_i^j} \mathbf{W}_v^j}{\sum_{j=0}^{n} e^{a_i^j}} \qquad (5)$$

The core idea of this method is to construct an objective function. For point i and its context point u in the training data, we use y to denote the label. Then, the objective function is

$$\mathcal{L}(i,u,y) = -[y\log(\sigma(\mathbf{H}_i^T\mathbf{Z}_u)) + (1-y)\log(1-\sigma(\mathbf{H}_i^T\mathbf{Z}_u))] \qquad (6)$$

2.2 GCN

Convolutional Neural Network (CNN) (Krizhevsky et al., 2012; He et al., 2016) has been a popular deep learning tool for many years. However, CNN can only be applied in Euclidean space. For non-Euclidean data structure, especially graph data, it has limited potential. Thus, Graph Convolutional Neural Network (GCN) (Kipf & Welling, 2016) is proposed to apply convolution to graph data. As a feature extractor that targets graph data, GCN can update node embeddings by aggregating the

information from neighboring nodes. In this study, we consider a two-layer GCN for a semi-supervised node classification task on a graph (note that too many layers in GCN will lead to the over-smoothing problem, Chen et al., 2020). The graph can be denoted as $\varsigma = (v, \varepsilon)$, with N nodes $v_i \in v$, edges $(v_i, v_j) \in \varepsilon$. The symmetric adjacency matrix is marked as A,

$$A \in \mathbb{R}^{N \times N}$$

and the degree matrix is denoted as D_{ii}.

$$D_{ii} = \sum_j A_{ij}$$

$$\tilde{A} = A + I_N$$

$\tilde{A}$ is denoted as the adjacency matrix of the graph added with self-connections, where I_N is the identity matrix. And $\tilde{D}_{ii}$ is calculated by

$$\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}.$$

Thus, our GCN model can be described as:

$$Z = f(X, A) = \text{Softmax}\,(\hat{A}\text{ReLU}\,(\hat{A}XW^{(0)})W^{(1)}) \qquad (7)$$

$$\hat{A} = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}} \qquad (8)$$

$$\text{Softmax}\,(x_i) = \frac{1}{\mathcal{Z}}\exp\,(x_i), \mathcal{Z} = \sum_i \exp\,(x_i) \qquad (9)$$

where $W^{(0)} \in \mathbb{R}^{C \times H}$ is a trainable weight matrix for the first hidden layer with H dimensions, and $W^{(1)} \in \mathbb{R}^{H \times F}$ is the trainable weight matrix for the second hidden layer with F-dimendional output. For the task of semi-supervised multi-class classification, we optimize the model using cross-entropy loss over all labeled examples:

$$L = \frac{1}{N}\sum_i L_i = -\frac{1}{N}\sum_i \sum_{c=1}^{M} y_{ic}\log\,(p_{ic}) \qquad (10)$$

where M is the number of categories. y_{ic} is a symbolic function. If the ground truth category of sample i is c, then y_{ic} =1. Otherwise, y_{ic} =0, where y_{ic} is the probability of sample i belongs to category c.

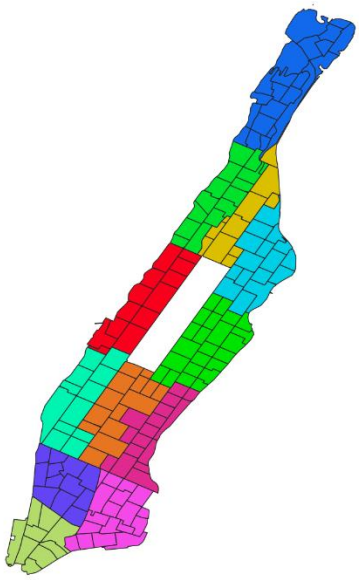

Fig. 5 Boundaries of 180 regions split

by streets and 12 types of functional regions in Manhattan, New York.

In this paper, we construct the graph by regarding the 180 regions as 180 nodes (see Fig. 5). Two nodes are connected (i.e., there is an edge between two nodes) if there are taxi flows between the corresponding two regions. The adjacency matrix A is a 180×180 matrix and its element A_{ij} is weighted by the amount of taxi flow from the node_i to the node_j. The diagonal elements of the matrix A are zero because there is no flow between the same nodes. For the GCN model, the inputs are preliminary node embeddings of the 180 regions obtained from POIs and the adjacency matrix A. The labels are 12 types of functional regions Thus, in Equation 10, we have $N = 180$, $M = 12$.

We divide the dataset into train (40 samples), validation (20 samples), and test splits (120 samples). It is worth noting that we use stratified sampling when splitting the dataset. We train the model for 500 epochs with learning rate 0.001 using Adam optimizer. For all experiments, the Dropout rates are set as 0.5. After finishing training the model on train split, the parameters can be saved to inference F-dimensional representations of all the samples. In such a way, the updated region embeddings can be obtained.

2.3 Predicting the number of crimes

After updating the previously learned embeddings through the classification labels of urban functional areas and people mobility data, we performed the task of crime rate prediction to validate the accuracy of the learned embedding. By predicting using the same set of real crime data, we can compare the goodness of fit of embeddings learned using different methods. In this task, we use the data of different dimensions of the city including POI, urban functional area, people mobility, and learn embeddings through deep learning and GCN. The basic principle of forecasting is that, as the embedding of a high-dimensional vector can accurately express the data, the similarity of embeddings can be used to predict the regional attributes of the city.

We use the Lasso (Least absolute shrinkage and selection operator) regression model for our prediction (Tibshirani, 1996). Lasso estimate compresses some coefficients by constructing a penalty function, and makes some coefficients zero. Thus, Lasso regression has the functions of shrinkage and selection. Compared with OLS, Lasso regression can quickly and effectively extract important variables and simplify the model when there are many variables. Because the regional embedding is a high-dimensional sparsity vector and the information of each dimension is low, we need to exclude some covariates that have little influence on the dependent variable to improve the accuracy of our prediction.

The Lasso regression model consists of two parts. In Equation 11.a, this is an objective function, which is similar to the objective function of OLS, but Lasso regression added a restriction function (see Equation 11.b).

$$\hat{\beta}^{lasso} = \underset{\beta}{argmin} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j \right)^2 \quad (11.\text{a})$$

$$\text{Subject to } \sum_{j=1}^{P} |\beta_j| \leq t \quad (11.\text{b})$$

The smaller the t, the stronger the compression effect on the estimated parameters. When the objective function is minimized, the coefficients of some unimportant

independent variables will be compressed to 0 to achieve the selection of variables. We set the t equal 3 in this research.

The data of the number of crimes comes from the nearly 40,000 criminal records recorded by the New York police department (NYPD) during the year of 2020 in Manhattan, New York. We aggregate these criminal records based on the 180 regions we studied (see Table 3). The highest number of crimes in a region is 946, the lowest is only 17, and the average is 196. The large fluctuations in the number of crimes in different regions make the dataset suitable for evaluating predictive performance. The effectiveness of our method is tested by comparing the size of the goodness of fit obtained using different methods.

describe	**value**
count	180
mean	196.31
std	148.27
min	17
25%	90
50%	155
75%	251
max	946

Table 3 The describe of the number of crime events in 180 regions in a year

3. Results

This section summarizes the initial embedding results learned by the deep walk method. We first illustrate the clustering effect of learned embedding by two dimensionality reduction methods, TSNE and PCA. Then we learn the urban regional function labels with some samples using the GCN method. The accuracy of the

prediction can be used to evaluate the effectiveness of the embeddings. Next, we use Lasso regression to predict the number of crimes based on the learned embeddings. Finally, we compare our method with previous embedding methods for learning urban areas to test the effectiveness of our method.

3.1 Graph Embedding

We learned the embedding of 180 regions in Manhattan, New York from POI data by the deep walk method. There is a total of 5854 points, each contains 11 attributes including location, building types and others (see Table 2). Embedding is a high-dimensional vector, which is 128 dimensions in this study. Every vector of a single dimension has no practical meaning. Therefore, we can show the learned embedding graphically through dimensionality reduction. Dimensionality reduction aims to map the data from original dimension (high dimension) to lower dimension space while minimizing information loss. Because dimensionality reduction will lead to some information loss, we choose two different types of dimensionality reduction methods, Principal Component Analysis (PCA) and T-SNE, to reduce the dimensionality of the embedding.

PCA is a type of linear mapping, projecting high dimensional data into lower dimensional space, while T-SNE is a type of non-linear mapping, modeling high dimensional data into lower dimensional space. PCA is a very well-known dimensionality reduction method because it is simple, fast, and easy to use, but it can only retain overall variance. Compared to PCA, T-SNE is capable of preserving the local and global structure of the data, which is suitable for converting high dimensional data into lower dimensional data for visualization. The core concept of the T-SNE method is to measure the pairwise similarity between high-dimensional and low-dimensional objects. It first converts the high-dimensional Euclidean distance between data points into a conditional probability that represents similarity. We use Kullback-Leibler Divergence as the objective function to measure the similarity between two probability distributions.

$$p_{j|i} = \frac{\exp(-\| x_i - x_j \|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\| x_i - x_k \|^2 / 2\sigma_i^2)} \quad (12)$$

$$q_{j|i} = \frac{\exp(-\| y_i - y_j \|^2)}{\sum_{k \neq i} \exp(-\| y_i - y_k \|^2)} \quad (13)$$

$$C = KL(P \| Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (14)$$

In the cost function of Equation 14, the p_{ij} is similarity of data points in high dimension, while $q_{j|i}$ is similarity of data points in low dimension. The smaller C is, the closer the distribution probability of high-dimensional and low-dimensional is. By solving the smallest C, we can preserve the distribution information of the high-dimensional vector as much as possible.

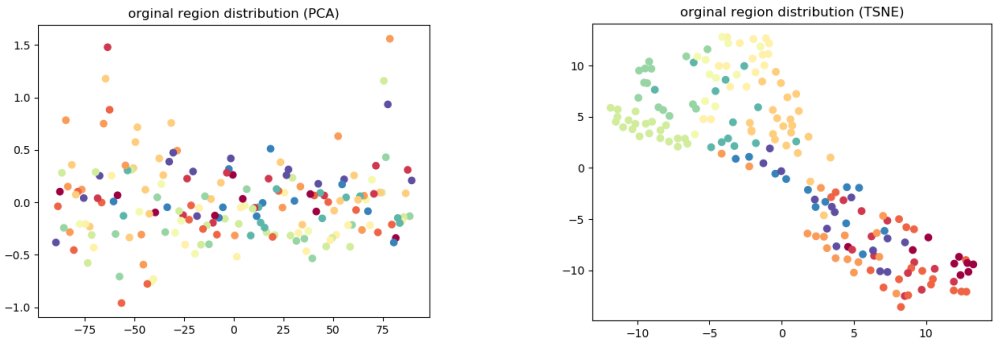

Fig. 6 The original region distribution

From the graphical presentation of the results, we can observe the results of preliminary embedding. In addition, we can use this graph as a benchmark embedding to observe the effect of GCN's update in the next step. Fig. 6 shows the embedding learned from the 11-dimension data of 5854 POIs in Manhattan, New York (reduced to two dimensions through dimensionality reduction). Each point represents an area of Manhattan, which is divided into 180 districts. Thus, there are 180 points in total. There are a total of 12 colors, representing the 12 types of functional areas as mentioned in Section 2 (see Fig. 4). In the dimensionality reduction graph by the T-SNE method, the regions belonging to the same functional area are clustered closer together, suggesting

that our preliminary embedding results are relatively good. The result of dimensionality reduction by PCA is not very good, which may be due to the loss of some information during PCA linear dimensionality reduction. Nevertheless, we can still show it as a benchmark embedding to observe the effect of updating the learned embedding by GCN in the next step.

3.2 GCN

In this part, we have completed two tasks using the GCN method. The first one is to learn the classification labels of 12 types of urban functional areas (see Fig. 5) from 40 random samples and use the remaining 140 samples for prediction. The overall accuracy rate of the prediction is 0.85. We have a total of 12 types of labels, and 40 samples are drawn for learning. Although each label is learned only 3.33 times on average, the accuracy rate is as high as 0.85 when we use them to predict the remaining 140 regional samples. This result shows that the embedding initially learned through the deep walk method can accurately learns the 11 dimensions of POI. In deep learning and neural networks, the unique attributes of a region are measured by high-dimensional vectors. Only when the result of embedding is accurate can the unique attributes of each region be accurately expressed through high-dimensional vectors. Therefore, our method can achieve a higher recognition success rate by learning from a relatively small sample.

Next, we will use GCN to learn the classification labels of the 12 types of urban functional areas in all 180 areas. During the learning process, the population flow data is also used to update the embedding. We take the preliminary embedding as input and get the updated embedding. We also illustrate our result with PCA and T-SNE two dimensionality reduction methods (shown in Fig. 7), so that we can compare the embedding before and after the update.

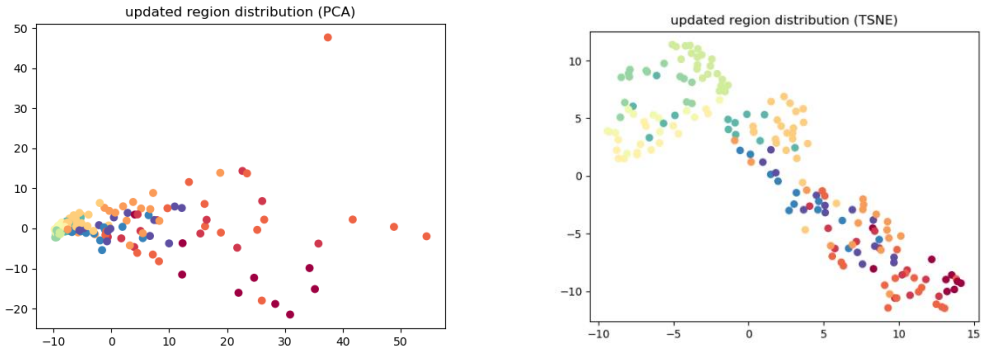

Fig. 7 The updated region distribution

We can see that the clustering situation of Fig. 7 through PCA method is greatly improved, although the effect is still slightly inferior compared to the graph of T-SNE method. Through T-SNE dimensionality reduction, the previous good clustering effect is still maintained (see Fig. 7). However, classifying regions through embedding is not the ultimate goal of our embedding. The existing embedding incorporates multiple dimensional data, including 11 dimensions of POI points, urban functional areas data and population flow data.

3.3 Predicting the Number of Crimes

Our goal is to integrate multi-dimensional data to represent each area through embedding, so that embedding can reflect the unique attributes of each area. In this way, we can perform more downstream tasks through embedding. Through the crime rate prediction task, we demonstrated our method of predicting some characteristics of the city through embedding. We use the number of crimes as the dependent variable and the learned embedding as the independent variable in the Lasso regression. The dataset of the number of crimes comes from the nearly 40,000 criminal records recorded by the New York police department (NYPD) during the year of 2020 in Manhattan, New York. We aggregate these criminal records based on the 180 regions to calculate the number of crimes in each region (see Table 3).

Method	MSE:	RMSE:	R^2:
Skip-gram POI embedding	18710.98	119.64	0.12
TF-IDF POI embedding	15480.27	103.35	0.16
EGES	16266.65	97.53	0.21
EGES+GCN(Ours)	6098.38	65.54	0.68

Table 4 the prediction error and goodness of fit by different method

We use three previous research methods and the methods proposed in this study to make predictions based on the same dataset. These methods include Skip-gram POI embedding, TF-IDF POI embedding (Yao et al., 2018), and EGES (Wang et al, 2018). The Skip-gram POI embedding method only learns embedding from POI data and assigns the same wight on different POI attributes. The TF-IDF POI embedding method uses the number of unique POI categories as the number of the vector dimension. The vector measures the importance of different POI categories to a node. The EGES method adopts a self-adaptive method to get the weights of different dimensions but lacks a GCN compared to our method. We use Mean Square Error (MSE) and Root Mean Square Error (RMSE) to measure the prediction error and use coefficient of determination (R^2) to measure the goodness of fit of models. Our results are shown in Table 4. From these results, we can observe that, compared with these methods from previous studies, our method not only reduces the prediction error to a large extent, but also greatly improves the goodness of fit (reaching 0.68). These results demonstrate that the embedding learned using our method can more accurately reflects the regional characteristics.

4. Discussions and Conclusion

This study proposes a regional embedding model for learning multi-dimensional city data. The learned embedding is a high-dimensional vector that can accurately reflect the attribute characteristics of the region. As a result, we can perform tasks such as prediction or recommendation through embedding. We use machine learning and neural network methods to learn embedding. First, we learn preliminary region

embedding from POI data through random walk and skip-gram methods. In this process, we transform the global network into a local network through random walk. This change not only greatly reduces the computational workload, but also facilitates distributed computing. In skip-gram process, we use an adaptive weight calculation method to learn the multiple attributes of POI data. Using two different dimensionality reduction methods, we show the effect of embedding learned. In the next step, we update the learned embedding using the GCN method. We use the initially learned embedding as input, the regional population mobility data as the adjacency matrix, and the type of urban functional regions as the label for supervised learning to generate more accurate embedding. Although supervised learning is used in this step, there is no need for excessive manual intervention. Finally, to verify the effect of learned embedding, we used different embedding methods to perform the task of predicting the number of crimes in different regions. The results show that our method has better predictive performance over the other methods. We also use Lasso regression to evaluate the prediction effect. Compared with other methods, the coefficient of determination is as high as 0.68 and the smaller MSE and RMSE show that the embedding we have learned can more accurately represent regional attributes.

This research makes two contributions to the literature on smart city. First, it proposes a method to use multiple dimensions of data in the city to learn region embedding. In a smart city, the large amount of data collected through various sensors usually have different dimensions. Thus, it is challenging to use the big data to help cities improve their operations and decision-making. Our study shows that, through our machine learning and neural network methods, we no longer need to pre-suppose predictive equations, and we can rigorously analyze multiple types of data including intra-regional data and inter-regional population mobility data. Second, because the learned embedding can accurately reflects the data, it can be used as an independent variable of the attributes of urban areas for various prediction and recommendation tasks for managing smart cities.

Our study has several limitations. First, the data of our empirical research is limited to Manhattan, New York. Future studies need to analyze data from other cities to assess the general applicability of the model. Second, this research used three types of city data to learn embedding. Future studies can improve the method by incorporating more types of data to increase the accuracy of learned embedding. Third, this research uses crime number prediction to evaluate the learned embedding. Other prediction or recommendation tasks related to smart city management can be used to assess this method, suggesting a rich avenue for future research.

References:

Allwinkle, S. and Cruickshank, P. (2011), “Creating smart-er cities: An overview”, *Journal of urban technology*, 18(2), pp.1-16. https://doi.org/10.1080/10630732.2011.601103

Angelidou, M., (2014), “Smart city policies: A spatial approach”, *Cities*, 41, pp.S3-S11. https://doi.org/10.1016/j.cities.2014.06.007

Batty, M. (2013), “Big data, smart cities and city planning”, *Dialogues in human geography*, 3(3), pp.274-279. https://doi.org/10.1177/2043820613513390

Boeing, G. (2018), “Measuring the complexity of urban form and design”, *Urban Design International*, 23(4), pp.281-292. https://doi.org/10.1057/s41289-018-0072-1

Chen, D., Lin, Y., Li, W., Li, P., Zhou, J., and Sun, X. (2020, April), “Measuring and relieving the over-smoothing problem for graph neural networks from the topological view”, In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 04, pp. 3438-3445). https://doi.org/10.1609/aaai.v34i04.5747

Chen, K., Li, X., & Wang, H. (2015). On the model design of integrated intelligent big data analytics systems. *Industrial Management & Data Systems*. https://doi.org/10.1108/IMDS-03-2015-0086

Chen, Q., Zhang, M., & Zhao, X. (2017). Analysing customer behaviour in mobile app usage. *Industrial Management & Data Systems*. https://doi.org/10.1108/IMDS-04-2016-0141

Chourabi, H., Nam, T., Walker, S., Gil-Garcia, J.R., Mellouli, S., Nahon, K., Pardo, T.A. and Scholl, H.J. (2012), “Understanding Smart Cities: An Integrative Framework”, In *2012 45th Hawaii international conference on system sciences* (pp. 2289-2297). https://doi.org/10.1109/HICSS.2012.615.

Din, I. U., Guizani, M., Rodrigues, J. J., Hassan, S., & Korotaev, V. V. (2019). Machine learning in the Internet of Things: Designed techniques for smart cities. *Future Generation Computer Systems*, 100, 826-843. https://doi.org/10.1016/j.future.2019.04.017

Dong, K., Hochman, G., Zhang, Y., Sun, R., Li, H. and Liao, H. (2018), “CO2 emissions, economic and population growth, and renewable energy: Empirical evidence across regions”, *Energy Economics*, 75, pp.180-192. https://doi.org/10.1016/j.eneco.2018.08.017

Fu, Y., Wang, P., Du, J., Wu, L., and Li, X. (2019, July), “Efficient region embedding with multi-view spatial networks: A perspective of locality-constrained spatial autocorrelations”, In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 906-913). https://doi.org/10.1609/aaai.v33i01.3301906

George, G., Haas, M.R. and Pentland, A. (2014), “Big data and management”, *Academy of management Journal*, 57(2), pp.321-326. https://doi.org/10.5465/amj.2014.4002

Ghosh, D., Chun, S. A., Shafiq, B., & Adam, N. R. (2016, June). Big data-based smart city platform: Real-time crime analysis. In *Proceedings of the 17th International Digital Government Research Conference on digital government research* (pp. 58-66). https://doi.org/10.1145/2912160.2912205

Hadadin, N., Qaqish, M., Akawwi, E. and Bdour, A. (2010), "Water shortage in Jordan—Sustainable solutions", *Desalination*, 250(1), pp.197-202. https://doi.org/10.1016/j.desal.2009.01.026

Hall, R. E., Bowerman, B., Braverman, J., Taylor, J., Todosow, H., and Von Wimmersperg, U. (2000), "The vision of a smart city (No. BNL-67902; 04042)", *Brookhaven National Lab.*, *Upton, NY (US).*

Hashem, I.A.T., Chang, V., Anuar, N.B., Adewole, K., Yaqoob, I., Gani, A., Ahmed, E. and Chiroma, H. (2016), "The role of big data in smart city", *International Journal of information management*, 36(5), pp.748-758. https://doi.org/10.1016/j.ijinfomgt.2016.05.002

He, K., Zhang, X., Ren, S., and Sun, J. (2016), "Deep residual learning for image recognition", In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90

Hollands, R.G., (2008), "Will the real smart city please stand up? Intelligent, progressive or entrepreneurial?", *City*, 12(3), pp.303-320. https://doi.org/10.1080/13604810802479126

Jurafsky, D., (2000), "*Speech and language processing*. *Pearson Education India*.

Kipf, T. N., and Welling, M. (2016), "Semi-supervised classification with graph convolutional networks"., *arXiv preprint arXiv:1609.02907*.

Krizhevsky, A., Sutskever, I. and Hinton, G.E., (2012), "Imagenet classification with deep convolutional neural networks", *Advances in neural information processing systems*, 25, pp.1097-1105.

Li, S., Peng, G. C., & Xing, F. (2019). Barriers of embedding big data solutions in smart factories: insights from SAP consultants. *Industrial Management & Data Systems*. https://doi.org/10.1108/IMDS-11-2018-0532

Liu, K., Wang, M., Li, J., Huang, J., Huang, X., Chen, S., & Cheng, B. (2021). Developing a Framework for Spatial Effects of Smart Cities Based on Spatial Econometrics. *Complexity*, *2021*. https://doi.org/10.1155/2021/9322112

Liu, K., Yin, L., Lu, F. and Mou, N. (2020), "Visualizing and exploring POI configurations of urban regions on POI-type semantic space", *Cities*, 99, p.102610. https://doi.org/10.1016/j.cities.2020.102610

Lombardi, P., Giordano, S., Farouh, H. and Yousef, W. (2012), "Modelling the smart city performance", *Innovation: The European Journal of Social Science Research*, 25(2), pp.137-149. https://doi.org/10.1080/13511610.2012.660325

Los Angeles Community Analysis Bureau. (1974), "The State of the City: A Cluster Analysis of

Los Angeles", *City of Los Angeles*

Neirotti, P., De Marco, A., Cagliano, A. C., Mangano, G., & Scorrano, F. (2014). Current trends in Smart City initiatives: Some stylised facts. *Cities*, 38, 25-36. https://doi.org/10.1016/j.cities.2013.12.010

Ng, H.T. and Zelle, J. (1997), "Corpus-based approaches to semantic interpretation in NLP", AI magazine, 18(4), pp.45-45. https://doi.org/10.1609/aimag.v18i4.1321

Ogura, S. and Jakovljevic, M.M. (2018), "Global population aging-health care, social and economic consequences", *Frontiers in public health*, 6, p.335. https://doi.org/10.3389/fpubh.2018.00335

Pan, G., Qi, G., Wu, Z., Zhang, D., & Li, S. (2012). Land-use classification using taxi GPS traces. *IEEE Transactions on Intelligent Transportation Systems*, 14(1), 113-123. 10.1109/TITS.2012.2209201

Rathore, M.M., Ahmad, A., Paul, A. and Rho, S. (2016), "Urban planning and building smart cities based on the internet of things using big data analytics", *Computer networks*, 101, pp.63-80. https://doi.org/10.1016/j.comnet.2015.12.023

Shafiq, M., Tian, Z., Sun, Y., Du, X., & Guizani, M. (2020). Selection of effective machine learning algorithm and Bot-IoT attacks traffic identification for internet of things in smart city. *Future Generation Computer Systems*, 107, 433-442. https://doi.org/10.1016/j.future.2020.02.017

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), pp.267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Varian, H.R. (2014), "Big data: New tricks for econometrics", *Journal of Economic Perspectives*, 28(2), pp.3-28. DOI: 10.1257/jep.28.2.3

Wang, J., Huang, P., Zhao, H., Zhang, Z., Zhao, B., and Lee, D. L. (2018, July), "Billion-scale commodity embedding for e-commerce recommendation in alibaba", In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 839-848). https://doi.org/10.1145/3219819.3219869

Wiig, A. (2015), "IBM's smart city as techno-utopian policy mobility", *City*, 19(2-3), pp.258-273. https://doi.org/10.1080/13604813.2015.1016275

World Bank (2014), "World development indicators 2014", *The World Bank.*

Yan, B., Janowicz, K., Mai, G., and Gao, S. (2017, November), "From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts", In *Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems* (pp. 1-10). https://doi.org/10.1145/3139958.3140054

Yao, Z., Fu, Y., Liu, B., Hu, W., and Xiong, H. (2018, July), "Representing urban functions through zone embedding with human mobility patterns", In *Proceedings of the Twenty-*

Seventh International Joint Conference on Artificial Intelligence (pp. 3919-3925). https://doi.org/10.24963/ijcai.2018/545

Zekić-Sušac, M., Mitrović, S., & Has, A. (2021). Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities. *International journal of information management*, 58, 102074. https://doi.org/10.1016/j.ijinfomgt.2020.102074

Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H.H., Lin, H. and Ratti, C. (2018), "Measuring human perceptions of a large-scale urban region using machine learning", *Landscape and Urban Planning*, 180, pp.148-160. https://doi.org/10.1016/j.landurbplan.2018.08.020

Zhang, Y., Fu, Y., Wang, P., Li, X., and Zheng, Y. (2019, July), "Unifying inter-region autocorrelation and intra-region structures for spatial embedding via collective adversarial learning", In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1700-1708). https://doi.org/10.1145/3292500.3330972

Zheng, Y., Liu, T., Wang, Y., Zhu, Y., Liu, Y., and Chang, E. (2014, September), "Diagnosing New York city's noises with ubiquitous data", In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 715-725). https://doi.org/10.1145/2632048.2632102