Urban region representation learning with human trajectories: a multi-view approach incorporating transition, spatial, and temporal perspectives

Yu Zhang, Weiming Huang, Yao Yao, Song Gao, Lizhen Cui & Zhongmin Yan

Published online: 04 Sep 2024.

Submit your article to this journal

Article views: 1660

View related articles

View Crossmark data

Citing articles: 1 View citing articles

Taylor & Francis
Taylor & Francis Group

Urban region representation learning with human trajectories: a multi-view approach incorporating transition, spatial, and temporal perspectives

Yu Zhang [a,b]*, Weiming Huang [c]*, Yao Yao [d,e], Song Gao [f], Lizhen Cui [a,b] and Zhongmin Yan [a,b]

aSchool of Software, Shandong University, Jinan, China; bC-FAIR, Shandong University, Jinan, China; cSchool of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore; dSchool of Geography and Information Engineering, China University of Geosciences, Wuhan, China; eCenter for Spatial Information Science, The University of Tokyo, Chiba, Japan; fGeospatial Data Science Lab, Department of Geography, University of Wisconsin-Madison, Madison, WI, USA

ABSTRACT

Mining latent information from human trajectories for understanding our cities has been persistent endeavors in urban studies and spatial information science. Many previous studies relied on manually crafted features and followed a supervised learning pipeline for a particular task, e.g. land use classification. However, such methods often overlook some types of latent information and the commonalities between varying urban sensing tasks, making the features engineered for one specific task sometimes not useful in other tasks. To tackle the limitations, we propose a multi-view trajectory embedding (MTE) approach to learn the features of urban regions (region representations) in an unsupervised manner, which does not rely on a specific task and thus can be generalized to varying urban sensing tasks. Specifically, MTE incorporates three salient information views carried by human trajectories, i.e. transition, spatial, and temporal views. We utilize skip-gram to model human transition patterns exhibited from massive amounts of human trajectories, where long-range dependency is meaningful. Subsequently, we leverage unsupervised graph representation learning to model spatial adjacency and temporal pattern similarities, where short-range dependency is favorable. We perform extensive experiments on three downstream tasks, i.e. land use classification, population density estimation, and house price prediction. The results indicate that MTE considerably outperforms a series of competitive baselines in all three tasks, and different information views have varying levels of effectiveness in particular downstream tasks, e.g. the temporal view is more effective than the spatial view in land use classification, while it is the opposite in house price prediction.

1. Introduction

Mining latent information from human trajectories for understanding our cities has been persistent endeavors in the communities of urban studies and spatial information science (Mazimpaka and Timpf 2016). The spatiotemporal structures and regularities exhibited in human trajectories have tight linkages to many socioeconomic aspects of urban spaces (Barbosa et al. 2018). For example, the collective mobility patterns and interactions between different urban regions are indicative of the spatial distribution of urban functions (Wang et al. 2023; Yuan, Zheng, and Xie 2012). Therefore, various data mining techniques have been developed or tailored to mining trajectory data for a variety of applications in urban planning and management, e.g. land use classification or clustering (e.g. Liu et al. 2012; Zhang et al. 2021; Li, Huang, et al. 2024), population estimation (e.g. Chen et al. 2018; Douglass et al. 2015), house price prediction (e.g. Kang et al. 2021; Wang and Li 2017), the detection of social events (e.g. Zheng et al. 2013), traffic signal control (e.g. Lin et al. 2023), and traffic flow prediction (e.g. Qu et al. 2022; Zhang, Gong, Zhang, et al. 2023).

Many previous studies in mining human trajectories for urban applications have two prominent traits. First, feature engineering methods are prevalent (e.g. Ji et al. 2023), in which varying manually crafted features are constructed from the human trajectories. Commonly used features include pick-up and drop-off numbers of each urban area from vehicle traces (e.g. Liu et al. 2012; Pan et al. 2012), hourly proportions of callings and total calling numbers derived from mobile phone data (Pei et al. 2014).

CONTACT Zhongmin Yan ✉ yzm@sdu.edu.cn

*Yu Zhang and Weiming Huang contributed equally to this study and share first authorship.

Such features have been proven effective, while they heavily rely on human ingenuity and laborious efforts, and often neglect some types of latent information, e.g. the long-range (multi-hop) dependency between urban regions reflected from human travel patterns, and the inherent region similarity entailed from spatial proximity. Second, many previous studies rely on supervised learning for specific tasks (e.g. Hu et al. 2021), which is sometimes impractical to deal with unavailable or sparse ground truth data. In addition, models and features learned for one specific task are not necessarily useful in other tasks, in spite of the commonalities shared among different socioeconomic aspects of our cities, e.g. land use is often correlated with population distribution (Wang, Li, and Rajagopal 2020).

In order to tackle the limitations, representation learning approaches have been introduced into trajectory mining. Representation learning aims at transforming raw data to meaningful representations (vector embeddings) that can support effective machine learning for downstream tasks (Bengio, Courville, and Vincent 2013). The idea of learning representations has led to a wide array of successful stories in several domains in machine learning, e.g. in natural language processing (e.g. Devlin et al. 2019; Mikolov et al. 2013), time series data mining (e.g. Zhu et al. 2022), video analysis (e.g. Liu et al. 2022), and skeleton action recognition (e.g. Yan et al. 2023). These works have fueled the idea of learning effective region representations (embeddings) utilizing trajectory data in an unsupervised manner (e.g. Wang and Li 2017; Yao et al. 2018; Zhang et al. 2020). The learned region embeddings should ideally carry the latent information from human trajectories for sensing region relevance, similarity, and discrimination, and thus a wide range of downstream tasks would be benefitted.

One predominating question of learning effective region embeddings from human trajectories is that what types of latent information are desirable to be incorporated to benefit downstream applications. Intuitively we have three major perspectives of information that can reflect urban region semantics:

(1) Transition view: The transitions between different urban locations (e.g. cell towers in trajectories gathered from mobile phone usage) or regions are a primary source of information carried by human trajectories (e.g. *location a* -> *location b* -> *location a* -> *location c*), and they reflect the connectivity and relevance between urban regions from a human movement and behavior perspective (Barbosa et al. 2018; Chen et al. 2021; Cheng et al. 2021). Transitions in human trajectories often entail strong ties between varying locations, where long-range dependency can be revealed, such as remote commuting between homes and workplaces for some people.

(2) Spatial view: Adjacent urban regions or locations usually carry substantial similarities and relevance in many socioeconomic factors, e.g. population density, and house price. In fact, spatial proximity can be partially captured in the transition view, whereas explicitly modeling the spatial view can strengthen the relevance between adjacent urban areas, and largely fill the uncaptured transitions (as people are more likely to travel to contiguous places when their locations are not captured in trajectories) (Wang and Li 2017).

(3) Temporal view: Human trajectories are sequences of visited locations with time stamp information, therefore they are a natural instrument to reveal the temporal regularities and patterns of urban locations. Temporal regularities are believed to be indicative to urban functions, e.g. people's visiting time to residential and industrial regions largely differ (Wan et al. 2021). In fact, many previous feature engineering methods for trajectory mining focus on constructing features to delineate the temporal perspective of human trajectories (e.g. Kang et al. 2012; Pei et al. 2014).

The three views of information in human trajectories carry their respective traits, which are partially illustrated in Figure 1. For the transition view, long-range (multi-hop) transitions between different locations or regions can be meaningful, in light of the explicit human movements carried between them. For example, the blue trajectory in Figure 1 reflects an individual's movement in a particular day, traversing between accommodation, work, catering, and transportation places. These places are strongly tied by the intended movements between them, and such ties stand in a long-range (multi-hop) fashion, e.g. residence 1 and company 2 are several hops away in
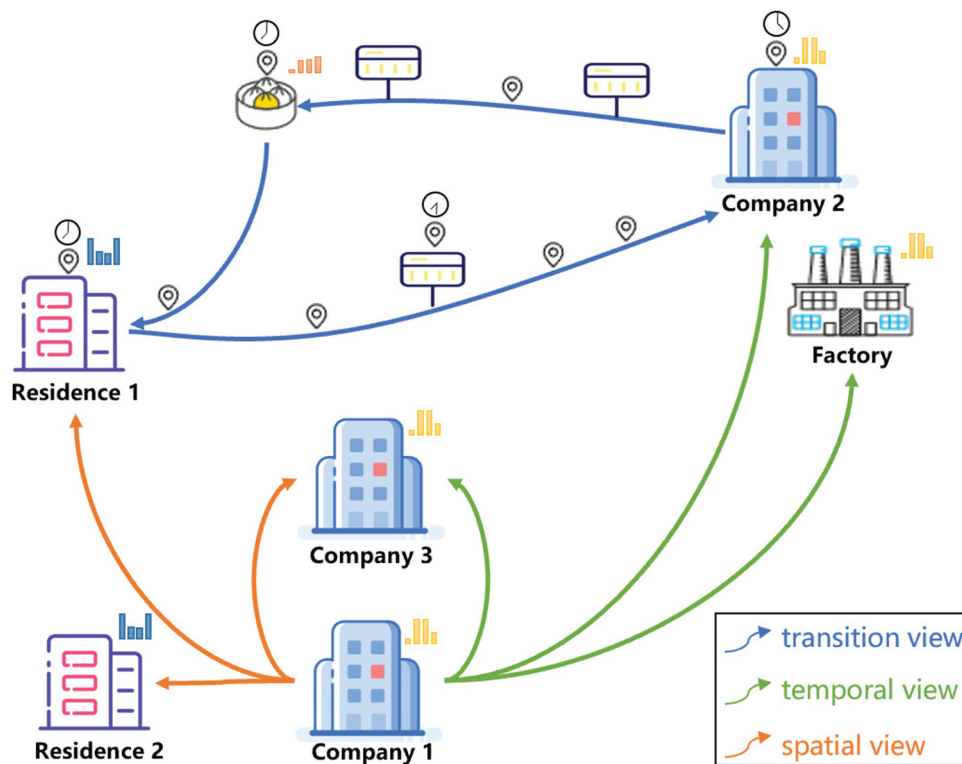
Figure 1. An illustration of multi-view modeling of locations (regions) utilizing human trajectory data. The transition view (blue) shows the sampled locations of an individual, highlighting long-range correlations between locations. The temporal view (green) identifies locations that exhibit similar temporal patterns to company 1. Meanwhile, the spatial view (orange) connects three locations that are spatially proximate to company 1.

the trajectory sequence (by commuting), while they are tightly connected. However, for the spatial view, long-range dependency can hardly hold at the granularity of urban regions or cell towers without explicit connectivity expressed by human movements. In this case, we assume that modeling short-range (one-hop) dependency would yield more meaningful region embeddings. For instance, when finding the correlated neighbors of company 1 in the spatial view, only the one-hop-away neighbors are considered (orange links), since further away places could have weak correlations if not explicitly tied by human movements. For the temporal view, we draw an analogy to the spatial view, for which we can generate temporal features (e.g. hourly temporal distributions) as the "temporal coordinates" of urban locations or regions, and use them to capture temporal proximity. In this case, we still assume that short-range temporal dependency is preferable, and long-range temporal correlations usually does not hold. For example, we build the green links from company 1 to company 2 and 3 as well as the factory, as their temporal visiting patterns are similar.

With the above observations and assumptions, we propose an unsupervised multi-view trajectory embedding (MTE) model for learning effective region representations, which can subsequently be used in various urban analytical tasks. MTE combines the neural language model skip-gram (Mikolov et al. 2013; as used in Word2Vec) and unsupervised graph neural networks (GNNs) based on graph diffusion and the infomax training objective (Hassani and Khasahmadi 2020; Veličković et al. 2019). Specifically, we utilize skip-gram to model the transition view, in which multi-hop and long-range dependencies are meaningful, as this is where the neural language models in the family of Word2Vec shine with strong expressiveness. We then leverage unsupervised GNNs to model the spatial and temporal views, as GNNs often prevail in modeling the correlation in close proximity (Oono and Suzuki 2019). MTE incorporates three views of information from human trajectories into region embeddings, and leverages expressive representation learning techniques based on the traits of different views.

We apply MTE for learning region embeddings in the study area of Shenzhen, China, and utilize the generated region embeddings in three downstream tasks: land use classification, population density estimation, and house price prediction. We compare MTE with several competitive baseline models, which reveals that our assumptions and observations for the three views stand, and our approach consistently prevails in all the downstream tasks. In addition, we discover that the three views have varying levels of effectiveness for different tasks.

Following this introduction, we review related works in Section 2 and introduce the main datasets in Section 3. In Section 4, we provide the details and intuitions of the proposed MTE approach. In Section 5, we demonstrate the experiments and results, including an exploratory similar region search and three quantitative downstream tasks, i.e. land use classification, population density estimation, and house price prediction. The paper ends with a discussion in Section 6 and conclusions in Section 7.

2. Related works

2.1. Trajectory mining for characterizing urban spaces

The employment of human mobility data for understanding our cities and facilitating urban planning and management has boomed in the past decade, thanks to the proliferation of various human mobility data, e.g. taxi traces, and mobile phone location data. Early studies in this direction mainly utilized feature engineering methods, in which hand-crafted feature extraction from trajectories played a pivotal role. For example, Liu et al. (2012) used taxi traces to analyze urban land use distributions, in which they created features from pick-up and drop-off numbers. Pan et al. (2012) crafted a set of pick-up and drop-off features at varying temporal granularities, and utilized an improved DBSCAN clustering technique for land use classification. Pei et al. (2014) created the features of hourly patterns of mobile phone data as well as the total calling volume for land use classification. J. Chen et al. (2018) utilized mobile phone location data for spatially and temporally fine-grained population prediction, in which several features are developed to delineate the stationery and inflow populations for different locations. In addition, topic models once

gained momentum for trajectory data mining. A seminal study by Yuan, Zheng, and Xie (2012) proposed a topic model-based framework for urban functional region discovery, which regards a region as a document, a function as a topic, the categories of points-of-interest (POIs) as metadata, and human mobility patterns as words.

The past years have witnessed a shift of trajectory mining methods toward various sophisticated machine (deep) learning techniques. To this end, Hu et al. (2021) proposed a framework combining skip-gram and a graph convolutional network (GCN) for urban function classification at the level of road segments. Specifically, they first utilized taxi traces to obtain the embeddings of road segments using skip-gram, and performed graph convolution in a road segment adjacency graph for classifying urban functions. This study accomplished superior results, while it generally omits the temporal information in trajectories. Sun et al. (2022) utilized a deep convolutional autoencoder to reconstruct the aggregated temporal features from mobile phone usages, and the learned embeddings are used for clustering analysis and land use classification. This work deeply mined the temporal patterns of each cell tower and urban region, while the correlations embedded in adjacent regions and human transitions are yet to be explored. In addition, such deep learning-based studies generally rely on task-specific supervised learning frameworks, and thus can hardly be generalized to other tasks.

Concurrently, the idea of learning region representations with human mobility data has become increasingly prosperous. Such studies have two key features that make them distinctive: (1) they are unsupervised (or self-supervised) with no prior knowledge from the ground truth data of downstream tasks, and (2) they learn general and multi-task representations. A pioneering work in region representation learning is Wang and Li (2017), who considered temporal dynamics, multi-hop transitions, and spatial adjacency between regions with taxi flow data. Specifically, they constructed a flow graph and a spatial graph in different time slots and conducted random walks in the two graphs; finally, they learned region embeddings using a skip-gram-based reconstruction objective. This study is particularly inspirational, as it considers all the three types of relations between regions, i.e. transition, spatial, and temporal. However, this work has several limitations: (1) long-range

spatial dependency can be easily captured, which over-smooths the embeddings of distant regions; (2) each region is duplicated for every time slot (e.g. each hour), thereby the constructed graph is hardly tractable for a large city.

Later, Yao et al. (2018) developed a skip-gram-like model for learning region embeddings with human mobility data, in which they leveraged the co-occurrence between regions and mobility events (transitions between regions) to learn region embeddings. Fu et al. (2019) used both POIs and human transitions between the POIs for learning region embeddings, in which they constructed two complete graphs (distance graph and mobility graph) and utilized a graph reconstruction objective to optimize the model. Zhang et al. (2020) used taxi traces, POIs, and check-in data to compose a mobility graph and a spatial graph of regions; they finally combined skip-gram-like reconstruction loss functions to learn region embeddings. Shimizu, Yabe, and Tsubouchi (2020) proposed to learn place embeddings in multiple spatial scales with human trajectories using an autoregressive model, namely a long short-term memory (LSTM); this model mainly relied on human transitions between different locations and also incorporated the transitions' time stamps. Wu et al. (2022) proposed multi-graph fusion networks for learning region embeddings with human trajectories, in which they first used several custom similarity metrics to cluster mobility graphs, and then applied a multi-level cross-attention mechanism to enable intra- and inter-cluster message passing; the model is trained also with a reconstruction loss. In the geospatial domain, Zhang et al. (2021) used Word2Vec to learn region embeddings from mobile phone location data, in which they regarded each cell tower as a word, and each trajectory as a sentence. Although this study was dedicated to sensing urban land use, its method does not rely on supervisory signals from a specific task, and thus can be potentially generalized to other tasks.

The above studies in region representation learning using trajectory data, besides Wang and Li (2017), omit certain aspects in trajectory data among the three information views discussed in the article. In addition, the expressiveness of their methodological frameworks is generally limited, as we believe the commonly used skip-gram-based and graph-based reconstruction objectives are limited in modeling the interactions between spatially adjacent and temporally relevant regions. In this study, we propose the MTE model to unleash the rich information carried by the transition, spatial, and temporal views of human trajectories.

2.2. Region representation learning: skip-gram and GNNs

From a methodological viewpoint, many unsupervised region representation learning studies rely on two types of techniques, i.e. skip-gram (e.g. Huang et al. 2022; Wang and Li 2017; Zhang et al. 2021; here we do not differentiate the two variants of Word2Vec, skip-gram and CBOW) and graph representation learning (e.g. Wu et al. 2022; Zhang et al. 2020). In fact, these two strands have overlaps, as graph representation learning can also use skip-gram-like objectives through random walks (e.g. Huang et al. 2022). We believe that skip-gram is particularly useful to model long-range and multi-hop dependency, which is prevalent in the transition view of human trajectories. However, for the spatial and temporal views that are not bonded by human movement, GNNs have better expressiveness (Oono and Suzuki 2019) with the message passing mechanism to naturally capture interactions between adjacent locations and regions. The question then boils down to how to train GNNs in an unsupervised manner for the spatial and temporal views.

There are some unsupervised graph representation learning techniques relying on random walks or the reconstructions of adjacency information (e.g. Grover and Leskovec 2016; Hamilton, Ying, and Leskovec 2017; Kipf and Welling 2017), whereas it is unclear whether such objectives are useful, as graph convolutional encoders already enforce smoothing over adjacent nodes (Veličković et al. 2019). Some state-of-the-art graph representation learning techniques rely on the *infomax* principle (Linsker 1988) that encourages the graph encoder to learn representations through maximizing mutual information of the representations from different scales. A seminal work is *deep graph infomax* (DGI; Veličković et al. 2019), which relies on mutual information maximization between node embeddings and a graph-level embedding, thus making the node embeddings globally relevant. Hassani and Khasahmadi (2020) developed further along this line, and proposed a contrastive multi-view

graph representation learning approach (MVGRL) for unsupervised graph learning, and their fundamental idea is to maximize the mutual information between different structural forms of graphs. Specifically, MVGRL applies diffusion to the structural information of a graph (i.e. the adjacency matrix) to obtain a more global view of the original graph. MVGRL then maximizes the mutual information between the node embeddings in one form and the graph-level embedding in the other form to learn node and graph representations. MVGRL is an expressive graph representation learning methods and prevails in a series of node and graph level tasks. And it deeply explores the interactions between local and global scales, which particularly fits the task of learning region representations, in which both local and global relevance is profitable. Therefore, in this study, we combine skip-gram and MVGRL to learn region embeddings, and the two techniques are used to model different information views based on their traits.

3. Study area and data

As there are different types of trajectory data, we introduce the study area and datasets before diving into the methodology. The study area is one of the most developed cities in China, Shenzhen, which has about 18 million residents. The study area is partitioned into 6,890 regions, which are the urban planning land parcels from the Municipal Planning and Natural Resources Bureau of Shenzhen. Essentially, these regions are divided by the major roads, and serve as the basic unit for urban planning and management. The trajectories used are anonymous mobile phone location data in two days, i.e. March 22 and 23 in 2012. The data records the transitions of more than 16 million mobile phone users among 5,818 cell towers (each user produces one trajectory), providing a reliable representation of the city's population (given that the year-end permanent population figures recorded in the 2012 Shenzhen Statistical Yearbook, totaling 10.46 million individuals). Specifically, each trajectory $s \in S$ is a sequence of transitions between cell towers C, and formally can be represented as $\{r_1, r_2, \ldots, r_n\}$, in which $r_i = \{c_j, t_i\}$ consisting of a cell tower identifier $c_j \in C$ (it is possible that the records of different time

stamps direct to a same cell tower) and a timestamp t_i associated r_i. In addition, each cell tower c_j is associated with a 2D geographic location $\{x_{c_j}, y_{c_j}\}$. The average number of visited cell towers of the trajectories is 20.67. The dataset has a comprehensive coverage of Shenzhen's population and contains rich spatial and temporal information; thus, we utilize this dataset to learn region embeddings for different downstream tasks. See Figure 2 for the geographic locations of cell towers and urban regions.

4. Methodology

The overarching architecture of the proposed MTE model is demonstrated in Figure 3. MTE is predominately composed of four components. First, we feed the trajectory sequences of human transitions between cell towers into a skip-gram model to learn the transition embeddings of cell towers, which captures their long-range (multi-hop) dependencies entailed from the massive human trajectories. Second, we interconnect the cell towers into two graphs based on their spatial adjacency and their similarity in temporal patterns, i.e. a spatial graph and a temporal graph; the embeddings from the transition view are used as the initial node (cell tower) features. Third, the two graphs are respectively fed into a graph encoder, i.e. the unsupervised graph representation learning model MVGRL, to learn cell tower embeddings of the spatial and temporal views. In this process, the cell tower embeddings are learned through graph diffusion and mutual information maximization between local and global representations. Finally, we utilize a Voronoi diagram to map multiview cell tower embeddings to region embeddings that are used in several downstream tasks.

4.1. Learning cell tower embeddings of the transition view

It has been widely acknowledged that the collective travel patterns of urban residents exhibit locational semantics through the explicit connections carried by human transitions (Barbosa et al. 2018; Hu et al. 2021). Intuitively, people often travel with certain purposes, such as the sequence {*home*, *workplace_1*, *dining_1*, *workplace_2*, *dining_2*, *entertainment*, *home*} indicates strong relevance between these locations. Although
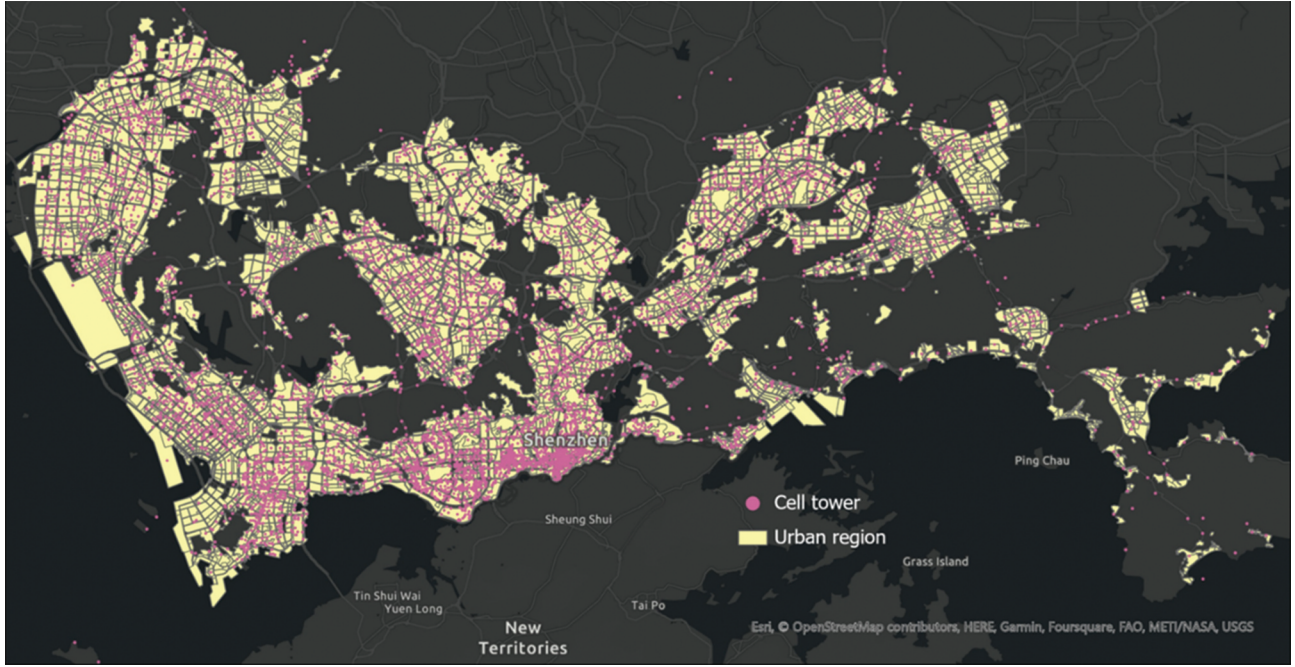
Figure 2. The study area, Shenzhen, encompassing 5,818 cell towers and 6,890 distinct regions.
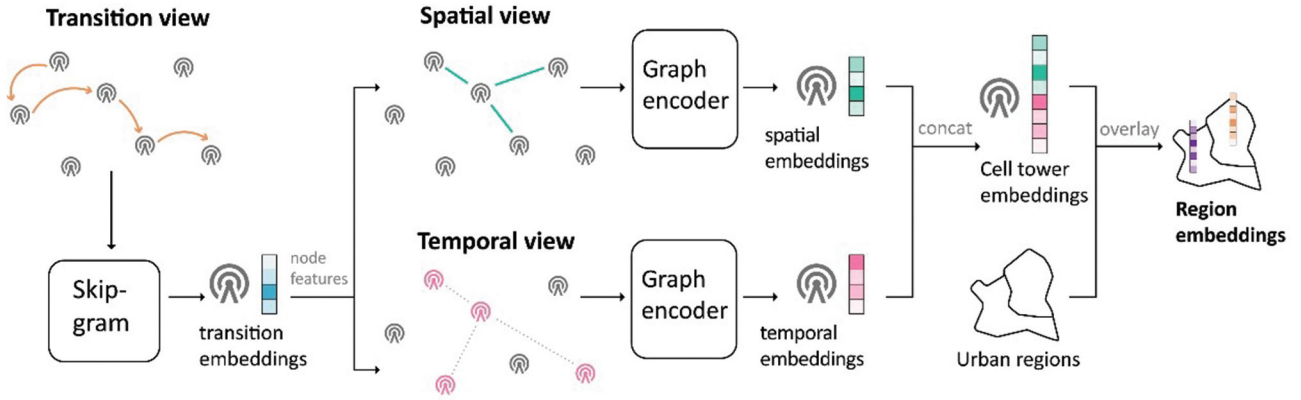

Figure 3. The overarching architecture of MTE.

individual travels often have certain randomness, the collective patterns entailed from the massive human trajectories in the mobile phone location dataset provisions an ideal proxy to understand and capture location and region semantics. We believe that with explicit human transitions, long-range and multi-hop dependency between cell towers holds, where skip-gram is effective.

Feeding the trajectories into skip-gram is straightforward, as the dataset is highly tailored to this model if we omit the timestamps. In this case, each trajectory is simplified to a sequence of cell tower identifiers, which can be regarded as a sentence of several words.

With a window size w, for each location (cell tower) in a trajectory, we retrieve all its context cell towers in the window (w steps both to the left and to the right). We can then form many co-occurrence pairs between cell towers, which captures multi-hop dependencies between them. With these co-occurrence pairs and negative sampling (Mikolov et al. 2013), cell tower embeddings in the transition view can be obtained through minimizing the objective function according to Equation 1:

$$\mathcal{L}_{transition} = \sum_{c \in \mathcal{C}} \sum_{c_q \in N_{R(c)}} - \left(\log \left(\sigma \left(\mathbf{c}_i^{r\mathsf{T}} \mathbf{c}_q' \right) \right) - \sum_{i=1}^{k} \log \left(\sigma \left(\mathbf{c}_i^{r\mathsf{T}} \mathbf{c}_{n_i}' \right) \right) \right)$$

(1)

where $\mathbf{c}_i^r$ denotes the *target embedding* of the cell tower c_i in the transition view, and $N_{R(c)}$ represents the set of context cell towers of c_i within the given window. $\mathbf{c}_q'$ denotes the *context embedding* of the cell tower c_q co-occurred with c_i. σ denotes the *sigmoid* function. c_{n_i} means the cell towers obtained by the negative sampling process.

4.2. Constructing spatial and temporal graphs

To capture spatial and temporal correlations, we construct two undirected graphs to explicitly connect spatially and temporally relevant locations (cell towers), which is a prerequisite for the following graph representation learning process. We regard each cell tower as a node in a graph, and utilize the trained cell tower embeddings in the transition view as initial node features. Then we build spatial edges and temporal edges to form the two graphs.

For the spatial view, the edges are built using k-nearest-neighbors (KNN) method, meaning that each node is connected to all other nodes within its KNN with undirected edges. Each edge is then assigned an unnormalized weight w_s by Equation 2:

$$w_s(c_i, c_j) = \log\left((1 + D^{1.5})/\left(1 + sd_{c_ic_j}^{1.5}\right)\right) \qquad (2)$$

in which D denotes the diagonal length of the minimum bounding rectangle of all the cell towers, and $sd_{c_ic_j}$ represents the spatial distance between two cell towers c_i and c_j. All the spatial edge weights are finally linearly rescaled to [0, 1]. The reason behind the distance decay of $w_s \sim sd^{-1.5}$ is in view of previous practices, e.g. in Huang et al. (2023). In this way, a spatial graph $G_s = (\mathbf{C}_r, \mathbf{A}_s)$ is constructed.

For the temporal view, we first construct the "temporal coordinates" in a latent temporal space. Specifically, we summarize a proportional hourly visit distribution for each cell tower in the dataset, then the "temporal coordinates" tc_i of a cell tower is a 24-dimensional vector, formally $tc_i = \{tp_1, tp_2, \ldots, tp_{24}\}$, in which tp_j is the hourly proportion of visits recorded with the time slot ts_j, which satisfies the constraints $tp_j \in [0, 1]$ and $\sum_k tp_j = 1$. An example is that a cell tower has 6% of its visits recorded in the time slot 11:00 am to 11:59 am ($tp_{12} = 0.06$). Such simple temporal features capture the temporal regularities of locations, e.g. intuitively the proportional hourly visit

distributions largely differ between residential and industrial areas. With the "temporal coordinates," we first apply a global minmax scalar for normalization, and then use ℓ_2 distance to capture the temporal similarity between cell towers. We, once again, use the KNN strategy to construct a temporal graph with unnormalized edge weights $td_{c_ic_j}^{-1}$, in which $td_{c_ic_j}$ denotes the ℓ_2 distance between c_i and c_j using their temporal coordinates, and all the temporal edge weights are also finally linearly rescaled to [0, 1]. After this process, the temporal graph $G_t = (\mathbf{C}_r, \mathbf{A}_t)$ is constructed.

The rationale of using simple KNN graphs over other methods, e.g. Delaunay triangulation, is that we empirically find the performance is benefited from moderately lifting edge densities (cf. Section 5.7). Furthermore, in principle the two graphs can be also viewed as one multiplex graph with two types of edges (spatial edges and temporal edges). However, we keep them as two graphs for separately feeding them to a graph representation learning model because we empirically discover that the spatial view and the temporal view have different levels of effectiveness in varying downstream tasks, and completely mixing all the information views sometimes does not guarantee the best performance (cf. Section 5).

4.3. Learning spatial and temporal embeddings with unsupervised graph representation learning

For the spatial and temporal views, we attempt to grasp the semantic similarities between different cell tower locations (and thus regions) based on their adjacency in the geographic space and the latent temporal space. We argue that graph representation learning is particularly useful for these two views, where modeling the relevance between direct and one-hop neighbors is more meaningful than multi-hop and long-range dependency. As we aim to learn multi-task region representations in a fully unsupervised manner, the essential question then boils down to the design of self-supervisory signals for training GNNs.

In this study, we utilize the *infomax* principle to accomplish unsupervised representation learning for both the spatial graph G_s and the temporal graph G_t, specifically using the MVGRL model. Such processes
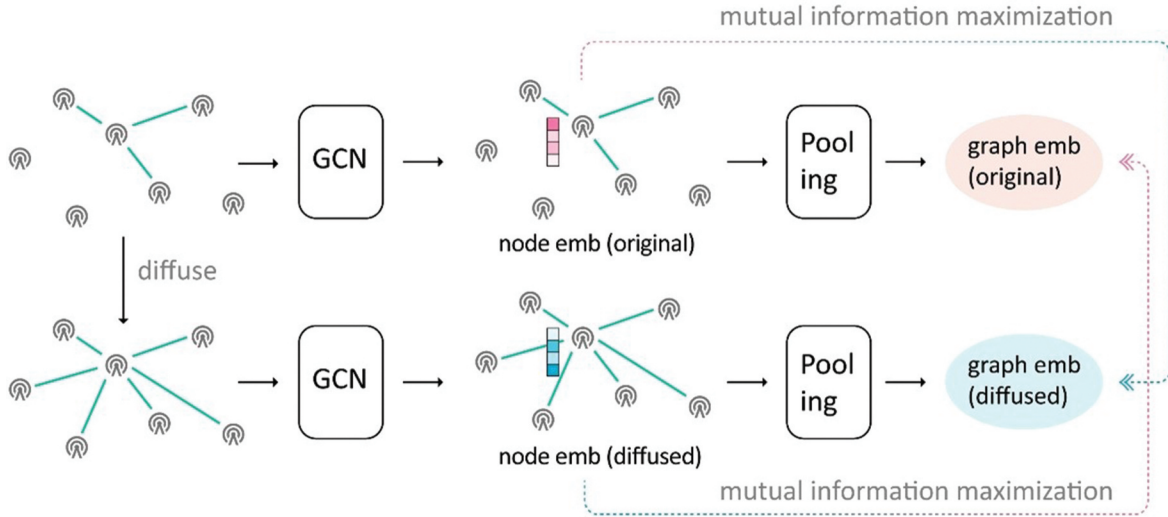
Figure 4. The graph representation learning model for learning cell tower embeddings in the spatial and temporal views.

are identical for the two graphs, so we illustrate only the case for the spatial graph in Figure 4. Overall, for each graph, MVGRL creates another version (form) of the original graph through a graph diffusion technique (Klicpera, Weißenberger, and Günnemann 2019), and both the original and diffused versions undergo graph convolution to obtain node embeddings; finally, the model is trained by maximizing the mutual information between node embeddings in one version and the graph embedding (obtained through pooling) in the other version. In this way, node embeddings become both locally and globally relevant, and carry rich connectivity information entailed from different structural forms of the graph.

Specifically, for the spatial graph G_s, we first apply a graph diffusion technique, which discovers a more global structural form of the graph to capture nuance linkages between unconnected cell towers. Graph diffusion is realized by the Personalized PageRank algorithm. Given the graph G_s, its adjacency matrix $\mathbf{A}_s$ is transformed to $\mathbf{A}_s^d$ with the following rule:

$$\mathbf{A}_s^d = \beta\left(\mathbf{I} - (1-\beta)\mathbf{D}_s^{-\frac{1}{2}}\mathbf{A}_s\mathbf{D}_s^{-\frac{1}{2}}\right) \quad (3)$$

where β is a tunable teleport probability, $\mathbf{I}$ denotes an identity matrix, and $\mathbf{D}_s$ is the degree matrix of the adjacency matrix $\mathbf{A}_s$.

We then have a diffused version of the spatial graph G_s, i.e. $G_s^d = \left(\mathbf{C}_r, \mathbf{A}_s^d\right)$. Subsequently, graph convolutions (two dedicated one-layer GCNs; Kipf and Welling 2017) are applied to both the original spatial

graph and the diffused spatial graph. Formally, the node embeddings are generated with the information propagation rule:

$$\mathbf{C}_s^o = \tau\left(\widehat{\mathbf{D}}_s^{-\frac{1}{2}}\widehat{\mathbf{A}}_s\widehat{\mathbf{D}}_s^{-\frac{1}{2}}\mathbf{C}_r\mathbf{\Theta}_s^o\right) \quad (4)$$

$$\mathbf{C}_s^d = \tau\left(\widehat{\mathbf{D}}_s^{d-\frac{1}{2}}\widehat{\mathbf{A}}_s^d\widehat{\mathbf{D}}_s^{d-\frac{1}{2}}\mathbf{C}_r\mathbf{\Theta}_s^d\right) \quad (5)$$

in which τ is a *parametric ReLU* (*PReLU*) function, $\widehat{\mathbf{A}}$ is an adjacency matrix with self-loop, $\widehat{\mathbf{D}}$ is the degree matrix of $\widehat{\mathbf{A}}$, and $\mathbf{\Theta} \in \mathbb{R}^{F_r \times F_s}$ is a learnable linear transformation applied to every node ($\mathbf{\Theta}$ is different for G_s^o and G_s^d), and $\mathbf{C}_s^o$ and $\mathbf{C}_s^d$ are the node embeddings of the original spatial graph G_s and the diffused spatial graph G_s^d after graph convolution.

We further carry out pooling operations to generate graph embeddings, in order to provide global-level supervisory signals for training the graph encoders. In addition, we introduce an additional activation function to provide nonlinearity for the network and enhance the expressive power of the graph representation. For each of the two graph versions, all the node embeddings are averaged respectively, before going through *sigmoid* function, which can be formally represented as:

$$\mathbf{g}_s^o = \sigma\left(\frac{1}{n_c}\sum_{i=1}^{n_c}\mathbf{c}_{s,i}^o\right) \quad (6)$$

$$\mathbf{g}_s^d = \sigma\left(\frac{1}{n_c}\sum_{i=1}^{n_c}\mathbf{c}_{s,i}^d\right) \tag{7}$$

where σ is the *sigmoid* function, n_c is the number of nodes (cell towers), $\mathbf{c}_{s,i}^o$ and $\mathbf{c}_{s,i}^d$ are the node embeddings in the original spatial graph and the diffused spatial graph respectively (after graph convolution). After this step, we obtain an embedding for both the original spatial graph and the diffused spatial graph, i.e. $\mathbf{g}_s^o$ and $\mathbf{g}_s^d$.

We can now learn rich node (cell tower) embeddings using the *infomax* principle, which is agnostic to downstream tasks. The model is optimized by maximizing the mutual information between the node embeddings in one version of the graph (e.g. the original spatial graph) and the graph embedding in the other version (e.g. the diffused spatial graph), and vice versa. The objective is formally defined as:

$$\mathcal{L}_{spatial} = -\frac{1}{n_c}\sum_{i=1}^{n_c}\left(\mathcal{D}\left(\mathbf{c}_{s,i}^o, \mathbf{g}_s^d\right) + \mathcal{D}\left(\mathbf{c}_{s,i}^d, \mathbf{g}_s^o\right)\right) \tag{8}$$

in which $\mathbf{c}_{s,i}^o$ is a node embedding in the original spatial graph, and $\mathbf{g}_s^d$ is the graph embedding of the diffused spatial graph. $\mathcal{D}$ is a discrimination modeled using a noise-contrastive type objective based on Jensen-Shannon divergence:

$$\mathcal{D}\left(\mathbf{c}_{s,i}^o, \mathbf{g}_s^d\right) = \log\left(f_\varphi\left(\mathbf{c}_{s,i}^o, \mathbf{g}_s^d\right)\right) \\ - \log\left(1 - f_\varphi\left(\widetilde{\mathbf{c}}_{s,i}^o, \mathbf{g}_s^d\right)\right) \tag{9}$$

where f_φ is a bilinear transformation, and $\widetilde{\mathbf{c}}_{s,i}^o$ is the embedding of the node c_i in the corrupted original spatial graph through row-wise shuffling of node initial features (transition embeddings), i.e. replacing the initial features of a node (cell tower) with the initial features of another randomly picked node. The definition of $\mathcal{D}\left(\mathbf{c}_{s,i}^d, \mathbf{g}_s^o\right)$ is a mirror replication of $\mathcal{D}\left(\mathbf{c}_{s,i}^o, \mathbf{g}_s^d\right)$.

Through minimizing the objective function in formula (8), we can then obtain the spatial embedding of a cell tower through summarizing its corresponding node embeddings in both the original spatial graph and the diffused spatial graph, i.e. $\mathbf{c}_{s,i} = \mathbf{c}_{s,i}^o + \mathbf{c}_{s,i}^d$ where $\mathbf{c}_{s,i}$ is the spatial embedding of cell tower c_i (note that spatial embeddings also carry transition information, which is the initial features). Following an identical process, we can also obtain a temporal embedding $\mathbf{c}_{t,i}$ for each cell tower. Finally, the multi-view

embedding of a cell tower is obtained by simply concatenating its spatial embedding and temporal embedding (transition embedding is also carried by them), i.e. $\mathbf{c}_i = \mathbf{c}_{s,i}\mathbf{c}_{t,i}$.

4.4. Mapping cell tower embeddings to region embeddings

As we intend to perform downstream analytical tasks in the scale of urban regions, we map the learned cell tower embeddings to region embeddings. The mapping process follows the previous practice in Zhang et al. (2021), in which Voronoi diagram was utilized to delineate the service area of each cell tower, and a region embedding is defined as an area-weighted summation of the cell tower embeddings whose Voronoi polygons spatially intersect with the region. Formally, this process is defined as Equation 10:

$$\mathbf{r}_i = \frac{1}{a_{r_i}}\sum_j a_{r_i,c_j}\mathbf{c}_j \tag{10}$$

in which a_{r_i} is the area of the region r_i, a_{r_i,c_j} is the interaction area between region r_i and the Voronoi polygon of the cell tower c_j, and $\mathbf{c}_j$ is a cell tower embedding.

5. Experiments and results

5.1. Implementation details

We first utilize skip-gram to generate the transition embeddings of cell towers (64-dimensional), and in this process we tune the window size w in {3, 5, 7, 9, 11} to find out the best range to model the dependency between cell towers in the transition view. We train the skip-gram for 10 epochs in a minibatch mode, and each minibatch trains 10,000 target cell towers. The dimension of the transition view is 64. Additionally, the temporal and spatial views are concatenated through graph contrastive learning, resulting in a combined dimension of 128, representing the concatenation of the original and diffused versions. After obtaining the transition embeddings, they are used as initial features in both the spatial and temporal graphs. For each graph, we train the graph representation learning model MVGRL for 2,000 epochs without a minibatch mode, and with the learning rate of 0.001. And we tune the number of nearest neighbors in the graph construction

processes, i.e. we tune k_s (for the spatial graph) and k_t (for the temporal graph) both in {4, 16, 64, 128, 256} to find the best graph forms to capture short-range dependencies in the geographic and temporal spaces. The hyperparameter tuning of w, k_s, and k_t is carried out and weighed in all three downstream tasks, and we find that the best parameters that produce generally favorable results in all the downstream tasks are $w = 9$, $k_s = 64$, and $k_t = 128$ (see Section 5.7).

5.2. Baseline models

We compare the proposed MTE model for region representation learning with several baseline models:

(1) Traj2Vec (Zhang et al. 2021): This model is essentially the skip-gram model used for learning region embedding also with mobile phone mobility data. It is equivalent to only using the transition view of MTE to generate region embeddings.

(2) MNE (Zhang et al. 2018): This method treats the three views as a multiplex graph, i.e. cell towers are nodes which are connected by three types of edges. Transition edges are built based the transition frequencies between cell towers, and spatial and temporal edges remain the same as in our model. For each node, MNE learns specific embeddings of each edge types and a common embedding to bridge them. The final node embeddings also incorporate the information from all three views.

(3) HDGE (Wang and Li 2017): This model constructs a flow graph and a spatial graph for cell towers to learn a specific cell tower embedding for each time slot. We concatenate the embeddings of each cell tower in all time slots to generate region embeddings.

(4) HIER (Shimizu, Yabe, and Tsubouchi 2020): This model uses LSTM and a next location prediction training objective to learn cell tower embeddings (and thus region embeddings). In this process, the visit times are directly used as input features to have a sense of the temporal patterns.

(5) Sk-3views: In this variant, we learn the embeddings of all three views separately using skip-gram. For the spatial and temporal views, we

conduct biased random walks in the spatial and temporal graphs (the biases are equal to unnormalized edge weights) to generate cell tower sequences, which are then fed into skip-gram. The embeddings from the three views are finally concatenated to form region embeddings.

(6) Graph-3views: In this variant, we replace skip-gram in MTE with the graph representation learning model MVGRL. Like the baseline MNE, the transition graph is constructed based on transition frequencies between cell towers.

In addition, we also compare MTE with two ablations, MTE-spatial and MTE-temporal, in order to verify the necessity of the spatial view and the temporal view.

5.3. Similar location search

Before diving into the quantitative evaluations of our proposed model in downstream tasks, we carry out an exploratory search of similar locations (cell towers) through measuring the cosine similarities of their embeddings (Gao, Janowicz, and Couclelis 2017; Liang et al. 2022). The similarity search results can help us gain intuitions of the information content and effectiveness of different views.

We demonstrate a case of similar location search in Figure 5. In this case, we use an anchor location, i.e. the pink point located in an industrial park, to search its similar locations. If we only use the embeddings from the spatial graph (which are initialized with transition embeddings), we can find the two yellow points as the most similar locations, which are very spatially close while both located in a residential area. This indicates that spatial and transition views combined focuses mainly on spatial proximity, while not on functional similarity (although often spatially close regions are also functionally similar). If we only use the embeddings from the temporal graph, we then find the two orange locations as the most similar ones, which are located far apart while both are located in industrial parks. This means that the temporal and transition views combined emphasizes functional similarity while generally is irrelevant to spatial proximity. Finally, we combine the embeddings from the spatial and temporal graphs, we then find the two blue points located in industrial parks as

Figure 5. The results of similar location (cell tower) search. A selected anchor point (pink) is presented along with the two most similar locations using the embedding fusing three views (blue), fusing transition and spatial views (yellow), and fusing transition and temporal views (orange).

the most similar locations, which are both spatially adjacent (further than the yellow points though) and functionally similar. This illustrates the effectiveness of our model, which balances spatial proximity and functional resemblance.

5.4. Land use classification

The spatial distribution of urban land use (urban function) is of paramount relevance for urban planning and management. Over the last decade, numerous studies have utilized human trajectories for sensing urban land use, in view of the explicit linkage between human mobility patterns and the spatial distribution of land use in our cities (e.g. Liu et al. 2012; Pei et al. 2014; Shimizu, Yabe, and Tsubouchi 2020; Zhang et al. 2021). In principle, more meaningful region embeddings learned from human trajectories should better capture the linkage between city structures and human movement, and thus can yield better performance in land use classification.

In this task, we utilize the MTE region embeddings for land use classification against the baseline models. The ground truth data is the authoritative land use data of Shenzhen in 2014 and is available in 5,487 regions. We merge the fine-grained land use types in the authoritative dataset into six: (1) *natural and open space* (*nat.*), (2) *commercial* (*com.*), (3) *residential* (*res.*), (4) *industrial* (*ind.*), (5) *public service* (*pub.*), and (6) *transportation and logistics* (*trans.*). Specifically, we use a random forest (RF) classifier (with 100 decision trees) and randomly choose 2/3 of the regions as training data, and the remaining 1/3 of the regions as test data (from the 5,487 regions with ground truth information). We repeat the experiments for 100 times with random training test set splits, and finally report the average performance metrics. In view of the unbalanced nature of the land use types in the study area (*industrial* and *residential* dominant), we use the metrics of accuracy (ACC; it is equivalent to weighted recall), weighted precision (WP), and weighted F1 score (WF1).

5.4.1. Performance

The results are presented in Table 1, and we observe that MTE ($w = 9$, $k_s = 64$, and $k_t = 128$) that incorporates the information from all three views outperforms all baselines in all evaluation metrics, suggesting that it can lead to not only more accurate predictions, but also more balanced results among all land use types (in view of the weighted F1). We also report the performance of using only the region embeddings generated from the spatial graph (MTE-spatial) and the temporal graph (MTE-temporal) respectively. We find that the MTE-temporal leads to better performance than MTE-spatial (note that both spatial embeddings and temporal embedding also carry the information from the transition view), meaning that the similarities exhibited from the temporal regularities are more indicative than spatial adjacency for this task. However, the information from all three views is useful, which is suggested by the superior performance of MTE.

For the baselines, Traj2Vec is essentially the transaction view of MTE, and it produces similar results as Sk-3views, suggesting that incorporating the spatial and temporal similarities through modeling the long-range dependencies in the two domains generally does not help for this task. Likewise, the unsatisfactory performance from Graph-3views indicates that using a (one-layer) GNN to capture the dependencies carried by the human transaction sequences is suboptimal. HDGE is also less effective than Traj2Vec, and we speculate that its approach to capture long-range dependency based on spatial adjacency undermines its performance for this task, as such long-range relations usually do not hold, especially with our dataset (trajectories from mobile phone data). HIER generates slightly better performance than other baselines except for MNE, as it considers all three perspectives. The performance of MNE is the best among all baseline methods, indicating that explicitly capturing the commonalities among the three views helps mitigate the (sometimes backward) effects of the long-range spatial and temporal dependencies in a skip-gram like architecture. However, it still does not reach the effectiveness of MTE.

5.4.2. Classification confusion analysis

We further analyze the performance of MTE, MTE-spatial, and MTE-temporal through visualizing their confusion matrixes of land use classification, which are shown in Figure 6. We observe that the accuracy scores for *residential* and *industrial* are generally higher than other land use types, as they are dominating in the study as the predictions from RF are mostly inclined to them. We notice that MTE-spatial and MTE-temporal have similar performance for *industrial*, which is likely because that industrial areas both exhibit spatial clustering and distinct temporal patterns; for residential areas, MTE-temporal leads to a large performance gain compared to MTE-spatial, which suggests that the temporal regularity is more indicative than spatial adjacency for residential areas. For the land use types of *natural and open space*, *transportation and logistics*, MTE-spatial produces better accuracy scores than MTE-temporal, suggesting that the spatial clustering effects of these land use types are more salient than their temporal regularities. We believe that such an analysis at the individual land use type level is useful in a binary classification scenario, e.g. discerning residential and nonresidential areas. At last, MTE takes the advantages of MTE-spatial and MTE-temporal, and produces the best performance for most of the land use types.

5.5. Population density estimation

The correlation between human mobility patterns (particularly those exhibited from mobile phone data) and the spatial distribution of population has

Table 1. Performance of land use classification using MTE and several baseline methods, with the evaluation measures of accuracy (ACC), weighted precision (WP), and weighted F1 score (WF1).

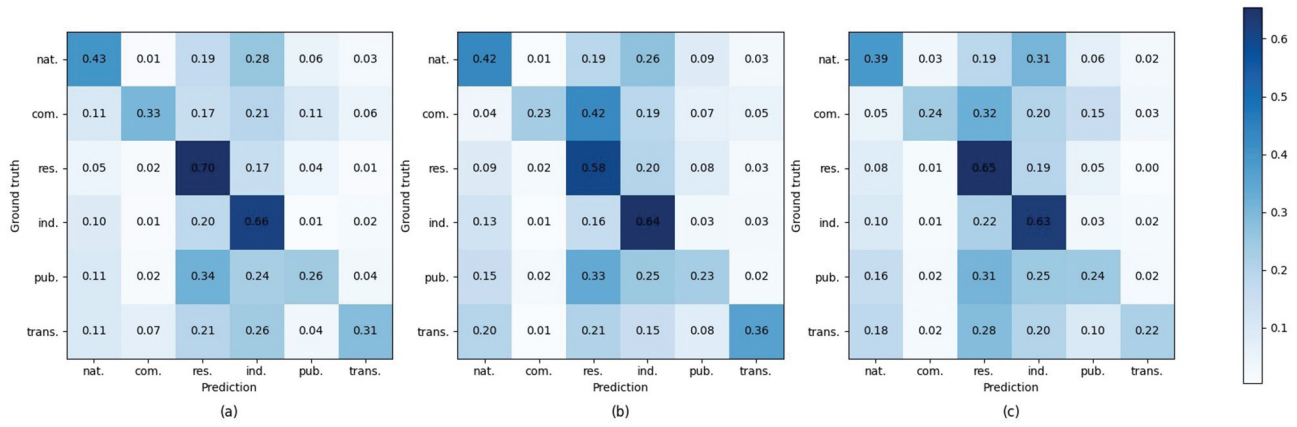| Model	ACC	WP	WF1
Traj2Vec	0.527±0.010	0.519±0.011	0.508±0.011
MNE	0.543±0.010	0.535±0.012	0.525±0.012
HDGE	0.520±0.009	0.511±0.009	0.505±0.009
HIER	0.537±0.011	0.532±0.011	0.513±0.011
Sk-3views	0.528±0.010	0.525±0.012	0.502±0.010
Graph-3views	0.521±0.010	0.516±0.010	0.506±0.010
MTE-temporal	0.530±0.010	0.519±0.011	0.509±0.010
MTE-spatial	0.520±0.010	0.508±0.010	0.508±0.010
MTE	**0.554±0.011**	**0.546±0.012**	**0.538±0.012**

Figure 6. The confusion matrixes for (a) MTE, (b) MTE-spatial, and (c) MTE-temporal.

been verified in many previous studies (e.g. Ahas et al. 2010; Chen et al. 2018; Douglass et al. 2015; Kang et al. 2012; Ratti et al. 2006). Therefore, in the second downstream task, we utilize the MTE region embeddings learned from the massive amount of mobile phone human trajectories for estimating population density for urban regions, which is useful to provide basic evidence for planning regional economic and social development.

In the experiment, we use the population density data in 2013 from WorldPop[1] as the ground truth data (6,789 regions have ground truth data). We input the region embeddings produced from MTE and the baseline models into a RF regression model (with 100 decision trees), and randomly choose 2/3 of the regions as training data, and the remaining 1/3 of the regions as test data (from the 6,789 regions with ground truth information). We repeat the experiments for 100 times with random training test set splits, and finally report the average performance metrics, i.e. R^2, root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

The results are demonstrated in Table 2, and we further spatially visualize the absolute estimation errors in each region in Figure 7. We observe that MTE produces impressive performance for this task, with $R^2 > 0.8$, meaning that the variation of population density across the study area can be very well explained by the variation of MTE embeddings. Interestingly, we can observe that MTE-spatial produces much better performance than MTE-temporal, and is also slightly better than the full version of MTE. This means that spatial adjacency is much more indicative for population density estimation than temporal similarities; adding the information from the temporal view only worsens the performance. This is likely because that densely or sparsely populated areas are generally respectively clustered, and thus spatial adjacency is pivotal for this task. Temporal regularities are useful for discerning residential and industrial areas with other land use types, but they are not indicative to different residential areas with varying levels of population density. For example, the temporal regularities in a region with high-rise apartment buildings (densely populated) and another region with detached houses (sparsely populated) are similar, while their population densities largely differ.

Table 2. Performance of population density estimation. Units for RMSE and MAE are number of people/km[2].

| Model	R^2↑	RMSE↓	MAE↓	MAPE↓
Traj2Vec	0.669±0.023	5355.76±322.30	2996.36±75.13	0.793±0.066
MNE	0.691±0.020	5165.66±259.55	2912.74±77.24	0.746±0.06
HDGE	0.705±0.021	5080.28±292.38	2802.99±78.36	0.711±0.062
HIER	0.506±0.025	6557.95±329.47	3740.22±88.79	1.134±0.083
Sk-3views	0.644±0.025	5513.84±341.16	3118.17±86.53	0.846±0.062
Graph-3views	0.680±0.026	5273.55±334.89	2858.80±78.01	0.673±0.049
MTE-temporal	0.313±0.023	7692.49±292.24	4751.09±94.05	1.574±0.114
MTE-spatial	**0.827±0.016**	**3863.49±244.08**	**2132.52±71.66**	**0.467±0.036**
MTE	0.816±0.015	3987.62±243.51	2226.17±65.64	0.493±0.037
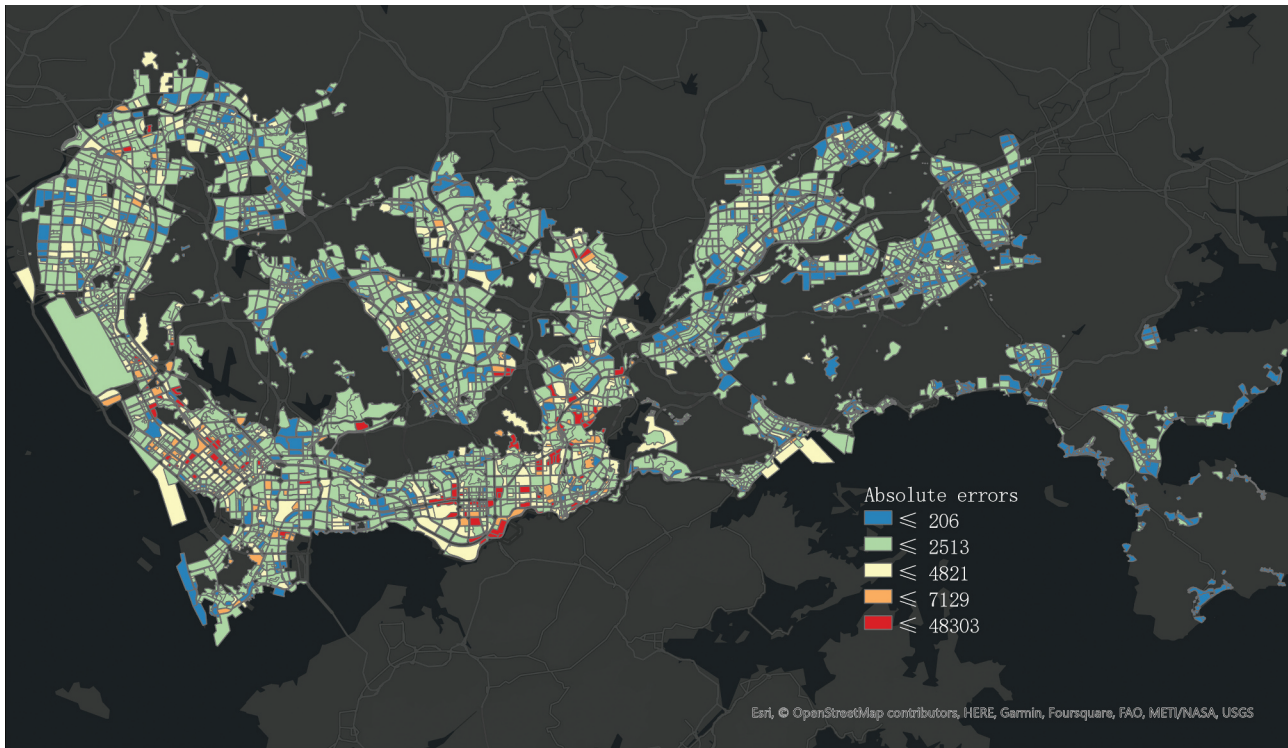
Figure 7. Visualization of absolute errors of population density estimation on a map. Unit is number of population/km^2.

As to the baselines, the ones with explicit modeling of spatial adjacency (e.g. MNE and HDGE) generally outperform others. Although Sk-3views and Graph-3views also encode spatial adjacency, the information from the temporal perspective could have slightly backward effects for this task. Traj2Vec also produces average-level performance, as the human transition patterns can be useful in discriminating residential areas with different levels of population density, and the spatial adjacency information can partially be reflected by massive amounts of human transitions. At last, HIER produces the worst performance among all baseline models, as it uses an autoregressive model to learn location embeddings, which neither enforces the frequently co-occurred (from the transition view) nor the spatially close locations to be similar in the embedding space.

5.6. House price prediction

House price is an important factor for human habitation and economic development in our cities. Modeling and predicting house price using different types of geospatial data have been explored in several disciplines, such as geography, urban studies, and economics. Among other data sources, human mobility data has been proved to be effective for modeling house price (e.g. Kang et al. 2021). In the third downstream task, we use the learned region embeddings for house price prediction.

In the experiment, we use the house price data in 2020 from a Chinese real estate agent Lianjia[2] as the ground truth data (970 regions have ground truth data). The human trajectories and house price data do not perfectly temporally align, while the general patterns of house price should be largely similar. We use the same RF regression model, dataset split means, number of times of experiment repetition, and evaluation metrics and for this task, as in the population density task.

The results are demonstrated in Table 3, and the absolute errors of the regions are spatially visualized in Figure 8. We observe that MTE also prevails in this task, and the information from the transition and spatial views is more indicative than the temporal view. However, the temporal view also has certain effectiveness, as MTE with all three views has the best performance (only slight performance gain compared to MTE-spatial though). This indicates that house with similar price ranges tends to be spatially close, which conforms with our common cognition. However, it seems that house with varying price

Table 3. Performance of house price prediction. Units for RMSE and MAE are CNY/m².

Model	$R^2\uparrow$	RMSE↓	MAE↓	MAPE↓
Traj2Vec	0.532±0.043	16678.13±1310.83	11591.79±528.89	0.203±0.012
MNE	0.561±0.040	16184.94±1533.87	11240.92±572.29	0.195±0.012
HDGE	0.552±0.045	16454.15±1556.21	11090.67±539.56	0.188±0.010
HIER	0.460±0.038	18062.00±1470.42	12735.77±557.24	0.229±0.015
Sk-3views	0.506±0.041	17292.47±1278.74	12022.85±543.42	0.212±0.013
Graph-3views	0.430±0.053	18578.60±1663.10	12969.44±615.66	0.229±0.014
MTE-temporal	0.073±0.041	23704.17±1366.91	17862.36±663.14	0.324±0.018
MTE-spatial	0.576±0.047	15958.75±1498.07	10523.48±510.15	**0.178±0.011**
MTE	**0.579±0.044**	**15734.94±1380.20**	**10502.69±501.10**	0.179±0.011
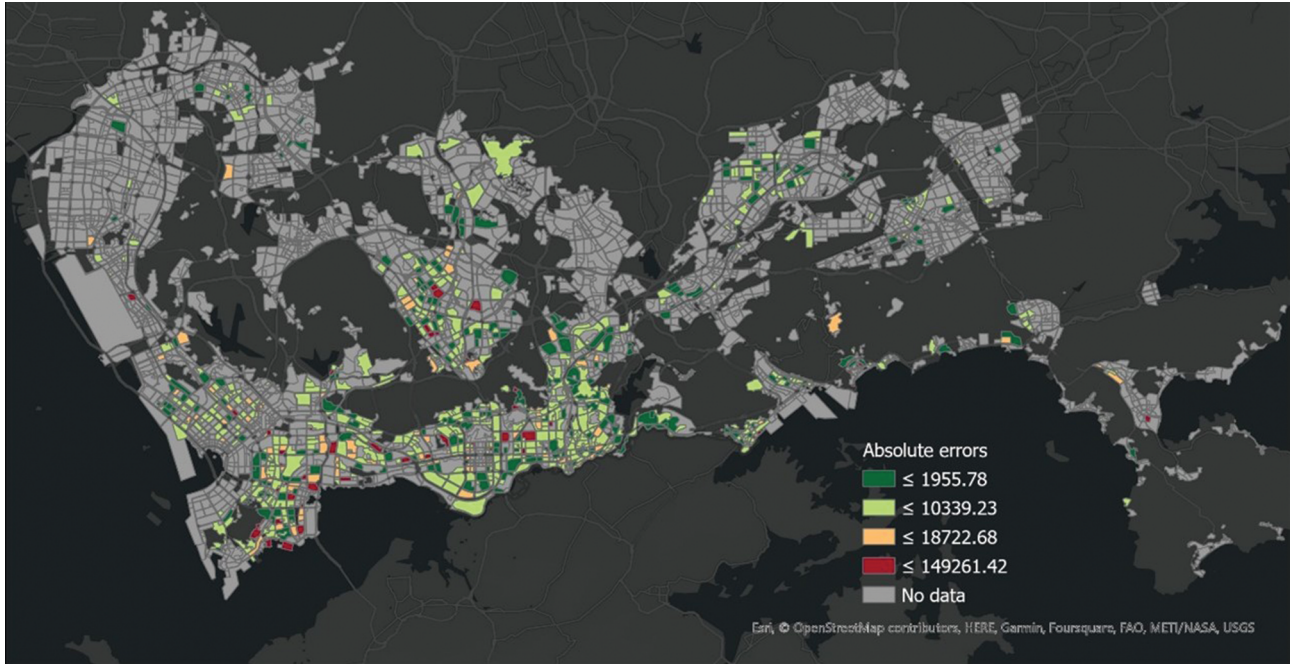
Figure 8. Visualization of absolute errors of house price prediction on a map. Unit is CNY/m².

ranges only has blurry patterns in their temporal regularities, and thus the temporal view merely leads to marginal performance gain for this task.

As to the baselines, we observe that MNE outperforms others, probably as it has a more sophisticated method to fuse the three views of information, which mitigates the negative effects of modeling long-range dependencies in the spatial and temporal domains; this could be the reason of its superior performance compared to Sk-3views. HDGE also has comparable performance as MNE likely due to its emphasis on modeling the spatial adjacency and human transition patterns. The transition view's information is also important for this task, as evidenced by the performance of Traj2Vec, which is better than Sk-3views where the long-range dependencies in the spatial and temporal domains downgrade the effectiveness. HIER and Graph-3views have unsatisfactory

performance, which is likely because that the former does not incorporate spatial information, and the latter only models short-range dependency in the transition view.

5.7. Parameter sensitivity analyses

As described in Section 5.1, we tune three hyperparameters for the three information views: for the transition view, we tune the window size w in {3, 5, 7, 9, 11}; for the spatial and temporal views, we tune k_s (for the spatial graph) and k_t (for the temporal graph) both in {4, 16, 64, 128, 256}. The parameter tuning is weighted in the three downstream tasks to select the best combination. All experiment settings are consistent with the respective downstream tasks.

As the embeddings from the transition view are used as the initial features for the spatial and

Table 4. Parameter analysis in terms of the window size w in the transition view. LU-F1 denotes the weighted F1 score in the land use classification task, PD-MAE denotes MAE in the population density estimation task, and HP-MAE denotes MAE in the house price prediction task. These three columns are in the form of metric (rank).

Window size w	LU-F1↑	PD-MAE↓	HP-MAE↓	Average rank↓
3	0.505(5)	**2951.830(1)**	11885.103(3)	2.33
5	0.506(4)	3018.193(4)	11875.263(2)	3.67
7	0.507(3)	3048.656(5)	11918.788(4)	4.33
9	**0.513(1)**	2982.641(2)	**11591.789(1)**	**1.67**
11	0.509(2)	3001.700(3)	12342.406(5)	3

temporal views, we first solely use the transition embeddings with different window sizes in the three downstream tasks (which is equivalent to hyperparameter tuning for the baseline Traj2Vec). The results of shown in Table 4, from which we observe that the $w = 9$ leads to the generally best performance across the three tasks. It appears that land use and house price tasks tend to benefit from medium or large window size, meaning that long-range dependency is preferable. However, for the population density task, it seems that small window size is favorable. This is interesting as the spatial view weighted the most in this task, which implies that short-range spatial adjacency is preferable for this task, while long-range dependency exhibited from the transition view is not.

The results of the grid search of k_s and k_t in the spatial and temporal graphs are demonstrated in Figure 9. We observe the tendency that generally large k_t leads to better performance in all three tasks, especially in population density and house price tasks. In fact, the temporal view only has marginal and even backward effectiveness for the two tasks, so we believe that capturing temporal similarities from large numbers of "temporal neighbors" could bring slight benefit or mitigate the backward effects in the two tasks. We also observe that medium numbers of neighbors in the spatial graph result in better performance.

Overall, we find that $w = 9$, $k_s = 64$, and $k_t = 128$ result in balanced best performance across the three downstream tasks, implying long-range dependency in the transition view, while medium-sized short-range dependencies in the spatial and temporal views.

6. Discussion

Through extensive experiments, we find that the proposed MTE approach consistently prevails in the downstream tasks compared to several competitive baselines.

It is also encouraging to see that our assumptions stand: (1) The three information views carried by human trajectories indeed have respective advantages in different downstream tasks, and combing them generally produces the most favorable results (although incorporating the temporal view induces slight backward effect in the population density task). (2) Modeling long-range dependency is favorable for modeling the transition view, thereby skip-gram model is a good fit. (3) Modeling the dependencies in close proximities in the spatial and temporal views is preferable, where the powers of GNNs shine.

The different ways of modeling long-range and short-range dependencies could be analogous to the difference between depth-first and breath-first strategies in search algorithms (Cormen et al. 2022). In the transition view, a depth-first-like method is used, where skip-gram could peek at further away locations (9-steps away) that are linked by human trajectories. As to the spatial and temporal views, a breath-first-like approach is used; even though 64 and 128 neighbors are finally used for graph construction, they enclose central locations and compose generally short-radius areas. We believe that this also partially answers another question in the geospatial domain, i.e. skip-gram vs. GNNs, which one should be used in what scenarios? Our answer to this is that skip-gram is favorable when long-range dependency stands, while GNNs are better for modeling short-range correlations.

An interesting observation from the study is that different information views have varying effectiveness in different downstream tasks. The transition view is important for all tasks, and it serves as an indispensable foundation of the learned region embeddings. The spatial view is also important for all tasks, while it is less indicative than the temporal view in the land use task. The temporal view can largely benefit the land use task, and subtly benefit house price prediction, while it is not preferable to be used in estimating population density. This indicates that it is not always
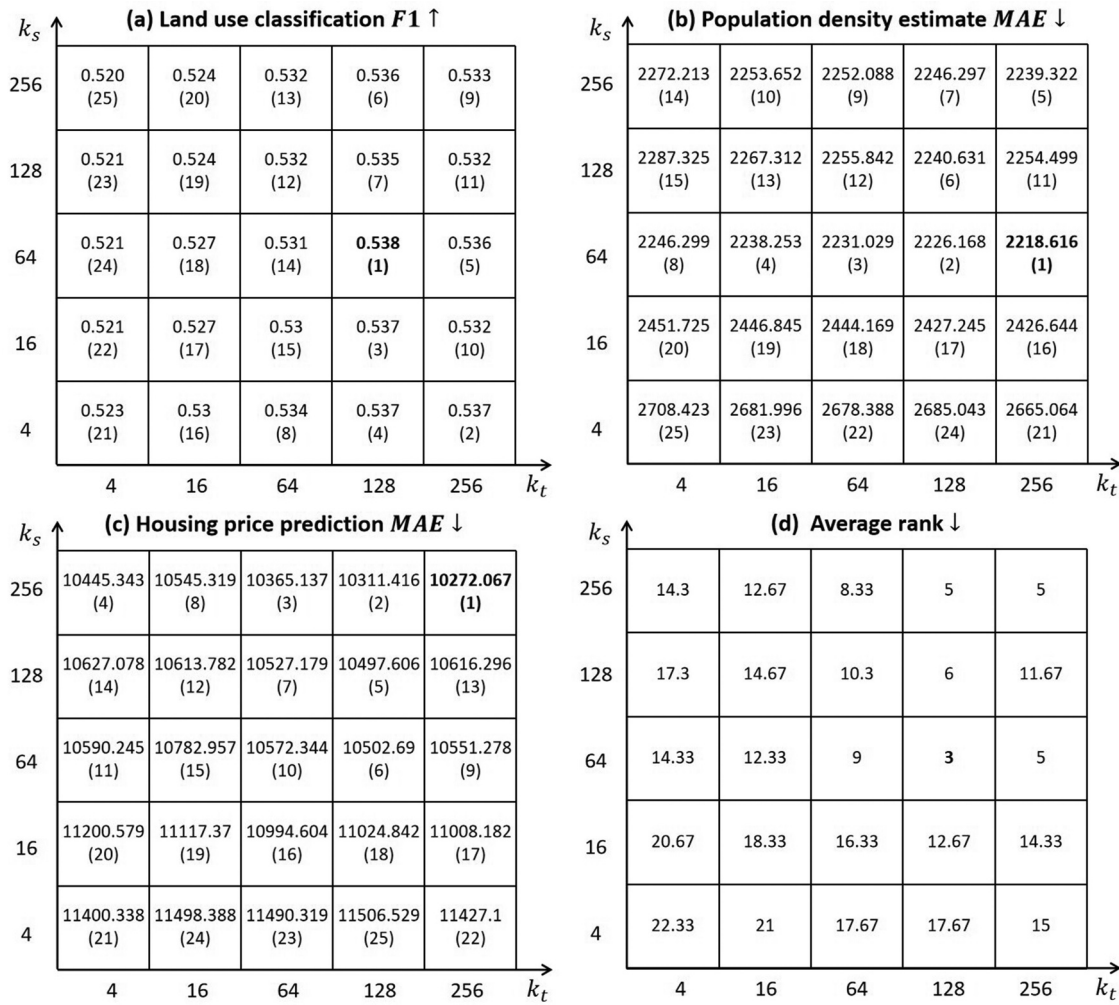
Figure 9. Parameter sensitivity analyses in terms of k_s (for the spatial graph) and k_t (for the temporal graph), i.e. the k number in constructing KNN graphs.

the case that fusing all information views provides the best performance, and it should be weighted on particular downstream tasks. This is also the reason that we use simple concatenation for combining the embeddings from the spatial and temporal graphs, rather than sophisticated means for fusion, e.g. to perform further contrastive learning between them (Park et al. 2020).

Human trajectories are an important component in today's landscape of geospatial big data. In particular, POIs are another pivotal data source that are prevalently used in many urban sensing tasks. We believe that it is also important to compare different types of data in terms of their effectiveness in varying downstream tasks. In this regard, this study utilizes the same ground truth dataset in the population density and house price tasks as a state-of-the-art POI-based

region embedding study (HGI; Huang et al. 2023), so a direct comparison can be drawn between them. We observe that MTE considerably outperforms HGI in the two tasks, which is even more impressive if one considers the temporal misalignment between the used human trajectories and the ground truth data. This finding does not water down the contributions of Huang et al. (2023), while it indicates that human trajectories can be more useful than POIs for the two tasks. This observation implies that we should consider the fitness of different types of geospatial data when using (integrating) them for various urban sensing tasks; simply fusing all data modalities might not be an optimal solution, while developing task-adaptive machine learning solutions might be a promising avenue, i.e. each data modality weighs differently in varying tasks.

At last, we observe that pre-training methods are increasingly popular in spatiotemporal data mining, particularly in urban mobility studies (e.g. Li, Xia, et al. 2024; Zhang, Gong, Zhang, et al. 2023). In such studies, region (location) representations are also learned in varying manners, e.g. using masked auto-encoder, to carry the rich information of different urban areas in terms of their roles (functions) and correlations. These representations are effective in mobility prediction tasks. In this regard, it is a promising research direction to further develop pre-training methods in urban mobility studies, for various (relatively) static urban analyses, e.g. inferring land use and population density.

7. Conclusions

In this paper, we propose a novel approach MTE for learning effective region representations (vector embeddings) with human trajectories in a fully unsupervised manner. MTE models three salient information perspectives of trajectory data, namely the transition, spatial, and temporal views, and utilizes varying machine learning techniques based on the traits of different views. Specifically, long-range dependency is meaningful in the transition view, and thus we utilize skip-gram for this view; short-range dependency is preferable for the spatial and temporal views, so unsupervised graph representation learning is leveraged to learn embeddings for the two views. We use a mobile phone trajectory dataset in Shenzhen for pre-training the proposed MTE model. Through extensive experiments in three downstream tasks, i.e. land use classification, population density estimation, and house price prediction, we observe that MTE region embeddings considerably outperform several competitive baselines in all three tasks. In addition, we observe that the effectiveness of different perspectives varies in particular downstream tasks, e.g. the temporal view is more effective than the spatial view in land use classification, while it is the opposite in house price prediction. This study provides insights into the fitness of distinctive machine learning techniques for modeling different information perspectives carried by geospatial data.

Notes

1. https://www.worldpop.org/
2. https://lianjia.com

ORCID

Yu Zhang http://orcid.org/0009-0004-6404-4958
Weiming Huang http://orcid.org/0000-0002-3208-4208
Yao Yao http://orcid.org/0000-0002-2830-0377
Song Gao http://orcid.org/0000-0003-4359-6302
Lizhen Cui http://orcid.org/0000-0002-8262-8883
Zhongmin Yan http://orcid.org/0000-0002-5271-5417

Data availability statement

The codes and embeddings that support the findings of the paper can be found at https://github.com/ZYuSdu/MTE.

References

Ahas, R., A. Aasa, S. Silm, and M. Tiru. 2010. "Daily Rhythms of Suburban commuters' Movements in the Tallinn Metropolitan Area: Case Study with Mobile Positioning Data." *Transportation Research Part C: Emerging Technologies* 18 (1): 45–54. https://doi.org/10.1016/j.trc.2009.04.011.

Barbosa, H., M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini. 2018. "Human Mobility: Models and Applications." *Physics Reports* 734:1–74. https://doi.org/10.1016/j.physrep.2018.01.001.

Bengio, Y., A. Courville, and P. Vincent. 2013. "Representation Learning: A Review and New Perspectives." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 35 (8): 1798–1828. https://doi.org/10.1109/TPAMI.2013.50.

Chen, J., T. Pei, S.-L. Shaw, F. Lu, M. Li, S. Cheng, X. Liu, and H. Zhang. 2018. "Fine-Grained Prediction of Urban Population Using Mobile Phone Location Data." *International Journal of Geographical Information Science* 32 (9): 1770–1786. https://doi.org/10.1080/13658816.2018.1460753.

Chen, M., Q. Liu, W. Huang, T. Zhang, Y. Zuo, and X. Yu. 2021. "Origin-Aware Location Prediction Based on Historical Vehicle Trajectories." *ACM Transactions on Intelligent Systems and Technology (TIST)* 13 (1): 1–18. https://doi.org/10.1145/3462675.

Cheng, H., W. Liao, M. Y. Yang, B. Rosenhahn, and M. Sester. 2021. "Amenet: Attentive Maps Encoder Network for Trajectory Prediction." *Isprs Journal of Photogrammetry & Remote Sensing* 172:253–266. https://doi.org/10.1016/j.isprsjprs.2020.12.004.

Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein. 2022. *Introduction to Algorithms*. Cambridge, Massachusetts Ave: MIT press.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Douglass, R. W., D. A. Meyer, M. Ram, D. Rideout, and D. Song. 2015. "High Resolution Population Estimates from Telecommunications Data." *EPJ Data Science* 4 (1): 1–13. https://doi.org/10.1140/epjds/s13688-015-0040-6.

Fu, Y., P. Wang, J. Du, L. Wu, and X. Li. 2019. "Efficient Region Embedding with Multi-View Spatial Networks: A Perspective of Locality-Constrained Spatial Autocorrelations." *Proceedings of the AAAI Conference on Artificial Intelligence*, Hawaii, USA, 906–913.

Gao, S., K. Janowicz, and H. Couclelis. 2017. "Extracting Urban Functional Regions from Points of Interest and Human Activities on Location-Based Social Networks." *Transactions in GIS* 21 (3): 446–467. https://doi.org/10.1111/tgis.12289.

Grover, A., and J. Leskovec. 2016. "node2vec: Scalable Feature Learning for Networks." *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, San Francisco, USA, 855–864.

Hamilton, W., Z. Ying, and J. Leskovec. 2017. "Inductive Representation Learning on Large Graphs." *Advances in Neural Information Processing Systems*, Long Beach, California, USA, 30.

Hassani, K., and A. H. Khasahmadi. 2020. "Contrastive Multi-View Representation Learning on Graphs." *International Conference on Machine Learning. PMLR*, Virtual, 4116–4126.

Hu, S., S. Gao, L. Wu, Y. Xu, Z. Zhang, H. Cui, and X. Gong. 2021. "Urban Function Classification at Road Segment Level Using Taxi Trajectory Data: A Graph Convolutional Neural Network Approach." *Computers, Environment and Urban Systems* 87:101619. https://doi.org/10.1016/j.compenvurbsys.2021.101619.

Huang, W., L. Cui, M. Chen, D. Zhang, and Y. Yao. 2022. "Estimating Urban Functional Distributions with Semantics Preserved POI Embedding." *International Journal of Geographical Information Science* 36 (10): 1905–1930. https://doi.org/10.1080/13658816.2022.2040510.

Huang, W., D. Zhang, G. Mai, X. Guo, and L. Cui. 2023. "Learning Urban Region Representations with POIs and Hierarchical Graph Infomax." *Isprs Journal of Photogrammetry & Remote Sensing* 196:134–145. https://doi.org/10.1016/j.isprsjprs.2022.11.021.

Ji, Y., S. Gao, T. Huynh, C. Scheele, J. Triveri, J. Kruse, C. Bennett, and Y. Wen. 2023. "Rethinking the Regularity in Mobility Patterns of Personal Vehicle Drivers: A Multi-City Comparison Using a Feature Engineering Approach." *Transactions in GIS* 27 (3): 663–685. https://doi.org/10.1111/tgis.13043.

Kang, C., Y. Liu, X. Ma, and L. Wu. 2012. "Towards Estimating Urban Population Distributions from Mobile Call Data." *Journal of Urban Technology* 19 (4): 3–21. https://doi.org/10.1080/10630732.2012.715479.

Kang, Y., F. Zhang, W. Peng, S. Gao, J. Rao, F. Duarte, and C. Ratti. 2021. "Understanding House Price Appreciation Using Multi-Source Big Geo-Data and Machine Learning." *Land Use Policy* 111:104919. https://doi.org/10.1016/j.landusepol.2020.104919.

Kipf, T. N., and M. Welling. 2017. "Semi-Supervised Classification with Graph Convolutional Networks." *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France.

Klicpera, J., S. Weißenberger, and S. Günnemann. 2019. "Diffusion Improves Graph Learning." *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vancouver, Canada, 13366–13378.

Li, Z., W. Huang, K. Zhao, M. Yang, Y. Gong, and M. Chen. 2024. "Urban Region Embedding via Multi-View Contrastive Prediction." *Proceedings of the AAAI Conference on Artificial Intelligence* 38 (8): 8724–8732. https://doi.org/10.1609/aaai.v38i8.28718.

Li, Z., L. Xia, Y. Xu, and C. Huang. 2024. "GPT-ST: Generative Pre-Training of Spatio-Temporal Graph Neural Networks." *Advances in Neural Information Processing Systems*, Vancouver, Canada, 36.

Liang, Y., J. Zhu, W. Ye, and S. Gao. 2022. "Region2Vec: Community Detection on Spatial Networks Using Graph Embedding with Node Attributes and Spatial Interactions." In *Association for Computing Machinery, New York, NY, USA, Article*. Proceedings of the 30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '22), 39, 1–4. https://doi.org/10.1145/3557915.3560974.

Lin, J., Y. Zhu, L. Liu, Y. Liu, G. Li, and L. Lin. 2023. "Denselight: Efficient Control for Large-Scale Traffic Signals with Dense Feedback." *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, Macao, China.

Linsker, R. 1988. "Self-Organization in a Perceptual Network." *Computer* 21 (3): 105–117. https://doi.org/10.1109/2.36.

Liu, Y., F. Wang, Y. Xiao, and S. Gao. 2012. "Urban Land Uses and Traffic 'Source-Sink areas': Evidence from GPS-Enabled Taxi Data in Shanghai." *Landscape and Urban Planning* 106 (1): 73–87. https://doi.org/10.1016/j.landurbplan.2012.02.012.

Liu, Y., K. Wang, L. Liu, H. Lan, and L. Lin. 2022. "Tcgl: Temporal Contrastive Graph for Self-Supervised Video Representation Learning." *IEEE Transactions on Image Processing* 31:1978–1993. https://doi.org/10.1109/TIP.2022.3147032.

Mazimpaka, J. D., and S. Timpf. 2016. "Trajectory Data Mining: A Review of Methods and Applications." *Journal of Spatial Information Science* 2016 (13): 61–99. https://doi.org/10.5311/JOSIS.2016.13.263.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." *Proceedings of the Twenty-sixth International Conference on Neural Information Processing Systems*, Nevada, USA.

Oono, K., and T. Suzuki. 2019. "Graph Neural Networks Exponentially Lose Expressive Power for Node Classification."

International Conference on Learning Representations, New Orleans, USA.

Pan, G., G. Qi, Z. Wu, D. Zhang, and S. Li. 2012. "Land-Use Classification Using Taxi GPS Traces." *IEEE Transactions on Intelligent Transportation Systems* 14 (1): 113–123. https://doi.org/10.1109/TITS.2012.2209201.

Park, C., D. Kim, J. Han, and H. Yu. 2020. "Unsupervised Attributed Multiplex Network Embedding." *Proceedings of the AAAI Conference on Artificial Intelligence*, New York, USA, 5371–5378.

Pei, T., S. Sobolevsky, C. Ratti, S.-L. Shaw, T. Li, and C. Zhou. 2014. "A New Insight into Land Use Classification Based on Aggregated Mobile Phone Data." *International Journal of Geographical Information Science* 28 (9): 1988–2007. https://doi.org/10.1080/13658816.2014.913794.

Qu, H., Y. Gong, M. Chen, J. Zhang, Y. Zheng, and Y. Yin. 2022. "Forecasting Fine-Grained Urban Flows via Spatio-Temporal Contrastive Self-Supervision." *IEEE Transactions on Knowledge and Data Engineering* 35 (8): 8008–8023. https://doi.org/10.1109/TKDE.2022.3200734 .

Ratti, C., D. Frenchman, R. M. Pulselli, and S. Williams. 2006. "Mobile Landscapes: Using Location Data from Cell Phones for Urban Analysis." *Environment & Planning. B, Planning & Design* 33 (5): 727–748. https://doi.org/10.1068/b32047.

Shimizu, T., T. Yabe, and K. Tsubouchi. 2020. "Enabling Finer Grained Place Embeddings Using Spatial Hierarchy from Human Mobility Trajectories." *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, Seattle, Washington, USA, 187–190.

Sun, Z., Z. Peng, Y. Yu, and H. Jiao. 2022. "Deep Convolutional Autoencoder for Urban Land Use Classification Using Mobile Device Data." *International Journal of Geographical Information Science* 36 (11): 1–31. https://doi.org/10.1080/13658816.2022.2105848.

Veličković, P., W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm. 2019. "Deep Graph Infomax." *International Conference on Learning Representations*, New Orleans, USA.

Wan, H., Y. Lin, S. Guo, and Y. Lin. 2021. "Pre-Training Time-Aware Location Embeddings from Spatial-Temporal Trajectories." *IEEE Transactions on Knowledge and Data Engineering* 34 (11): 5510–5523. https://doi.org/10.1109/TKDE.2021.3057875.

Wang, H., and Z. Li. 2017. "Region Representation Learning via Mobility Flow." *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, Singapore, Singapore, 237–246.

Wang, K., L. Liu, Y. Liu, G. Li, F. Zhou, and L. Lin. 2023. "Urban Regional Function Guided Traffic Flow Prediction." *Information Sciences* 634:308–320. https://doi.org/10.1016/j.ins.2023.03.109.

Wang, Z., H. Li, and R. Rajagopal. 2020. "Urban2vec: Incorporating Street View Imagery and Pois for Multi-Modal Urban Neighborhood Embedding." *Proceedings of the AAAI Conference on Artificial Intelligence*, New York, USA, 1013–1020.

Wu, S., X. Yan, X. Fan, S. Pan, S. Zhu, C. Zheng, M. Cheng, and C. Wang. 2022. "Multi-Graph Fusion Networks for Urban Region Embedding." *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, Vienna, Austria.

Yan, H., Y. Liu, Y. Wei, Z. Li, G. Li, and L. Lin. 2023. "Skeletonmae: Graph-Based Masked Autoencoder for Skeleton Sequence Pre-Training." *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France, 5606–5618.

Yao, Z., Y. Fu, B. Liu, W. Hu, and H. Xiong. 2018. "Representing Urban Functions Through Zone Embedding with Human Mobility Patterns." *Proceedings of the Twenty-Seventh International Conference on International Joint Conferences on Artificial Intelligence*, Stockholm, Sweden, 3919–3925. IJCAI.

Yuan, J., Y. Zheng, and X. Xie. 2012. "Discovering Regions of Different Functions in a City Using Human Mobility and POIs." *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, Beijing, China, 186–194.

Zhang, H., L. Qiu, L. Yi, and Y. Song. 2018. "Scalable Multiplex Network Embedding." *IJCAI*, 3082–3088.

Zhang, J., X. Li, Y. Yao, Y. Hong, J. He, Z. Jiang, and J. Sun. 2021. "The Traj2Vec Model to Quantify residents' Spatial Trajectories and Estimate the Proportions of Urban Land-Use Types." *International Journal of Geographical Information Science* 35 (1): 193–211. https://doi.org/10.1080/13658816.2020.1726923.

Zhang, M., T. Li, Y. Li, and P. Hui. 2020. "Multi-View Joint Graph Representation Learning for Urban Region Embedding." *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, Yokohama, Japan, 4431–4437.

Zhang, X., Y. Gong, C. Zhang, X. Wu, Y. Guo, W. Lu, and X. Dong. 2023. "Spatio-Temporal Fusion and Contrastive Learning for Urban Flow Prediction." *Knowledge-Based Systems* 282:111104. https://doi.org/10.1016/j.knosys.2023.111104.

Zhang, X., Y. Gong, X. Zhang, X. Wu, C. Zhang, and X. Dong. 2023. "Mask-And Contrast-Enhanced Spatio-Temporal Learning for Urban Flow Prediction." *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, Birmingham, United Kingdom, 3298–3307.

Zheng, K., Y. Zheng, N. J. Yuan, and S. Shang. 2013. "On Discovery of Gathering Patterns from Trajectories." *2013 IEEE 29th international conference on data engineering (ICDE)*, Brisbane, Australia, 242–253. IEEE.

Zhu, Y., Y. Zhang, L. Liu, Y. Liu, G. Li, M. Mao, and L. Lin. 2022. "Hybrid-Order Representation Learning for Electricity Theft Detection." *IEEE Transactions on Industrial Informatics* 19 (2): 1248–1259. https://doi.org/10.1109/TII.2022.3179243.