

Spatiotemporal tensor completion for improved urban traffic imputation

Ahmed Ben Said, Abdelkarim Erradi,

Abstract—Effective management of urban traffic is important for any smart city initiative. Therefore, the quality of the sensory traffic data is of paramount importance. However, like any sensory data, urban traffic data are prone to imperfections leading to missing measurements. In this paper, we focus on inter-region traffic data completion. We model the inter-region traffic as a spatiotemporal tensor that suffers from missing measurements. To recover the missing data, we propose an enhanced CANDECOMP/PARAFAC (CP) completion approach that considers the urban and temporal aspects of the traffic. To derive the urban characteristics, we divide the area of study into regions. Then, for each region, we compute urban feature vectors inspired from biodiversity which are used to compute the urban similarity matrix. To mine the temporal aspect, we first conduct an entropy analysis to determine the most regular time-series. Then, we conduct a joint Fourier and correlation analysis to compute its periodicity and construct the temporal matrix. Both urban and temporal matrices are fed into a modified CP-completion objective function. To solve this objective, we propose an alternating least square approach that operates on the vectorized version of the inputs. We conduct comprehensive comparative study with two evaluation scenarios. In the first one, we simulate random missing values. In the second scenario, we simulate missing values at a given area and time duration. Our results demonstrate that our approach provides effective recovering performance reaching 26% improvement compared to state-of-art CP approaches and 35% compared to state-of-art generative model-based approaches.

Index Terms—Traffic tensor, Tensor completion, CANDECOMP/PARAFAC

I. INTRODUCTION

Modern smart cities are increasingly deploying Internet of Things (IoT) sensors to collect and analyze data to efficiently manage urban assets and services such as public transport, utilities, traffic monitoring and public safety. With the widespread usage of sensors, massive urban data are continuously collected. There has been a great interest in using recent advances in data analytics to exploit these data in order to deliver better urban services and solve critical problems associated to the massive urban growth such as traffic congestion and public transport efficiency. For instance, forecasting traffic flow has been one of the most successful applications of state-of-art deep learning approaches. The collected sensory data are inevitably prone to multiple equipment-related issues and data collection imperfections causing data loss. Such data loss can be associated to multiple causes such as GPS calibration,

connectivity problem or weather conditions. Missing values have dramatic consequences as it may lead to drawing misleading conclusions and therefore wrong decisions. In terms of cost, missing values may force the city planning authority to redo the experiment in order to collect the required data, hence extra budgetary cost and time delay. At a city-wide scale, multiple sensors are deployed to continuously collect traffic data. However, it is costly and technically difficult to deploy traffic sensors across every corner of a metropolitan area, not to mention the challenge of management and maintenance. Practically, urban authority relies on few sensors deployed only on key areas. Hence, traffic data imputation is important to obtain a complete overview of the overall city traffic. Completing the missing data is critical for many tasks such as estimating travel time and congestion-aware route planning. In this paper, we address the problem of urban traffic data completion. We propose a modified CP completion approach that takes into account the urban and time context of the traffic to drive the completion algorithm. The paper contribution can be summarized as follows:

- We model the interaction between regions in the area of study as a spatiotemporal tensor. This tensor suffers from missing data which must be recovered to get better insights about the traffic flow.
- The tensor captures the traffic flow and hence the interaction between regions in the time domain. Our choice for tensor design considers the full traffic records, i.e. our tensor is built using all locations visited during the trip from the start to the end regions. This results in a less sparse tensor compared to the scenario where only source and destination are used to build it as more interaction between regions are derived.
- We propose an urban and time aware CP completion approach. The urban characteristics of each region are taken into consideration. They include the region's richness, diversity, concentration of Points of Interest (POIs) and convenience. These characteristics are used to determine the similarity between regions which is then used to augment the CP completion with additional features.
- We conduct a time series analysis to determine the periodicity of the traffic pattern. This periodicity is used to construct the temporal characteristics incorporated in the CP cost function.
- We propose an alternating least square approach to minimize the CP cost. To address the optimization, we propose to conduct the minimization on a reshaped version of the inputs.
- We provide a comprehensive comparative study with mul-

A. Ben Said is with the Department of Computer Science & Engineering, College of Engineering, Qatar University, Doha, 2713, Qatar. Email aben-said@qu.edu.qa

A. Erradi is with the Department of Computer Science & Engineering, College of Engineering, Qatar University, Doha, 2713, Qatar. Email erradi@qu.edu.qa

multiple completion approaches to validate the effectiveness of our proposal. Our experiments are performed on two real world datasets: T-drive taxi data [19], [20] from Beijing and Porto taxi ¹.

The rest of the paper is organized as follows: Section II discusses important related work. Section III presents some tensor calculation basics used in the proposed approach. We detail in section IV the formulation of our proposed urban-aware CP completion problem. Experiments are presented in section V. The last section concludes the paper and presents an agenda for future work.

II. RELATED WORK

Zhang et al. [1] proposed a spatiotemporal learning approach to predict the citywide crowd flows. The authors introduced inflow and outflow matrices for counting the incoming/outgoing moving objects for a given region at a given time slot. The aggregation of the two matrices are used to predict the flow at time t_{K+1} given the historical traffic flow till time t_K . The authors proposed Deep-ST, a deep neural network architecture trained on the flow tensor. The historical flow tensors are grouped into three time horizons: recent, near and distant. A stack of Convolutional Neural Networks (CNN) is trained on each time horizon flow tensor. Deep-ST exploits additional information such as weather, and type of the day (weekend/weekday) for more accurate forecasting. Hoang et al. [2] focused on the factors affecting crowd flows including seasonal (periodic), trend (changes in periodic patterns) and residual (instantaneous) flows. Gaussian Markov random field is used to model the seasonal and trend flows. The instantaneous flow exploits the spatiotemporal dependencies of different flows in addition to the weather data. This is achieved by applying a regression analysis where information about intra-regions and inter-regions dependence and weather condition are incorporated. Huang et al. [3] proposed a Deep Belief Network (DBN) based architecture for traffic flow prediction. DBN is effective in generating features in an unsupervised fashion. On top of the DBN, a multitask layer is incorporated for supervised traffic flow prediction.

Since most of the research works on traffic and crowd flow prediction are data-driven, it is of paramount importance to maintain clean and complete data. This research problem has been extensively studied and multiple successful approaches have been developed. For instance, Sparsity Regularized SVD (SRSVD) [4] approach focuses on Internet traffic matrix completion. SRSVD uses Singular Value Decomposition to find a global low rank approximation of the matrix and exploits its spatiotemporal structure by augmenting the minimization problem with two spatial and temporal matrices. The problem is then solved using Alternating Least Square (ALS) minimization. Compressive Sensing (CS) [5] is closely related to matrix completion problem. CS accurately recovers information of a sparse matrix using small subset of samples. Roughan et al. [6] proposed Sparsity Regularized Matrix Factorization (SRMF). This approach exploits the low rank and spatiotemporal property of the traffic matrix to estimate the missing

values. SRMF seeks the global low rank approximation which is then augmented with an interpolation technique such as k-nearest neighbours to fully recover the traffic matrix. Wen et al. [7] addressed the computation complexity of the completion problem based on the nuclear norm which requires calculating singular value decompositions. The authors proposed the Low-rank Matrix Fitting (LMaFit), a low complexity algorithm that is based on nonlinear successive over-relaxation approach that requires solving a linear least squares problem at each iteration.

However, the aforementioned solutions for traffic matrix completion operate on the two-dimensional traffic matrices whose columns are stacked. The multi-ways nature of such matrices is unfortunately ignored. Consequently, the matrix representation is simply not enough for efficient data recovery solutions.

In presence of more than two dimensional data, tensor representation for data recovery has been recently investigated. Indeed, a tensor can encompass more global information compared to a matrix such as an additional third dimension representing the time. Long et al. [8] reviewed state-of-art techniques of tensor completion for visual data. The authors identified two groups of approaches based on the optimization techniques used. One sets a predefined rank and optimizes the factors of tensor decomposition while the second group minimizes the rank of the estimated tensor iteratively. Acar et al. [10] proposed a CP weighted optimization algorithm (CP-WOPT). A first-order optimization is utilized to solve the weighted least squares problem. CP-WOPT has been successfully used to estimate missing data in spatiotemporal internet traffic tensor. In [11], the authors studied the convergence of the regularized ALS for tensor decomposition. Regularization is applied to avoid overfitting. The authors proved that ALS does not always converge using the Gauss-Siedel method while the regularized ALS provides better convergence and may decrease the required number of iterations.

In context of urban dynamics and mobility pattern, missing data is a common issue. Li et al [12] and Ni et al [13] surveyed state-of-art techniques for traffic data completion. In [14], the authors addressed the problem of missing values in intelligent transportation system using a probabilistic framework that extends the well known bayesian approach of Salakhutdinov and Mnih [15] to the higher order tensor. However, no urban context information is used. In [16], the data tensor represents the interaction between regions of the area of study. The set of regions is obtained based on the traffic zones provided by the transportation authority. Each data point r_{ijk} is the log transform of the number of moving objects whose start point is zone i and final destination is zone j departing at time k . The temporal dimension represents one hour-slice. For better recovery performance, the authors augmented the completion approach with urban contextual factors. These factors reflect the proportion of each type of POI at each region. Although this approach attempted to take into account the urban context, it does not consider important urban factors such as the convenience and the diversity of the region in terms of POIs. In addition, the authors studied only the mobility interaction based on the start and end regions of the

¹www.kaggle.com/c/pkdd-15-predict-taxi-service-trajectory-i/data

urban mobility data which do not capture the instantaneous interaction between regions while travelling from source to destination. Tan et al. [17] proposed an algorithm that uses the multimode transport information to predict the traffic flow with a low-rank constraint. The authors also addressed the forecasting problem in the presence of missing data. However, the proposed method does not scale well with very large traffic tensor. Li et al. [18] proposed a completion approach for tensor built using passenger flow from a metro service. The completion objective is regularized by introducing weakly dependent penalty and graph penalty and solved using Block Coordinate Descent. The main assumption is that two stations are less likely to be highly-dependent in terms of traffic flow profile. However, such assumption may not be valid for general road network traffic where the traffic flow is not restricted for just one mean of transportation.

With recent advances in deep learning, data imputation has been addressed using generative models. Yoon et al. [21] proposed a Generative Adversarial Network (GAN) [22] based model in which the generator observes parts of the real data and completes the missing components. The discriminator is trained to discriminate between the observed and imputed data while being provided with a hint vector. This vector guides the discriminator to improve the quality of imputation while ensuring that the generator completes the missing information according to the data distribution. In [23], the authors proposed another GAN based data completion approach named MISGAN. Two discriminator-generator pairs are used, one dedicated for the mask and the other for the data. The aim is to strengthen the imputation performance by modeling the distribution of the masks responsible for missing data. In their experiments, the authors considered only the scenario of completely random missing data. Boquet et al. [24] used Variational Autoencoder (VAE) [25] to develop an end-to-end solution for traffic forecasting which can handle data imputation. The imputation module consists of a recognition model that, once trained, can map the traffic samples to a latent space. A decoder, trained to reconstruct the traffic samples from the latent space, can then be used to generate the imputed samples. In [26], GP-VAE, a novel VAE-based technique is proposed. Gaussian process prior and Cauchy kernel are used to model the temporal dependencies of the data. Variational parameters are predicted using the inference model which takes the data with missing information. GP-VAE is validated on benchmark tasks and medical data. Mattei et al. [27] proposed MIWAE, an Importance-Weighted Autoencoder approach dedicated for missing at random data imputation. MIWAE maximizes a lower bound of the observed log-likelihood without any additional computational overhead compared to the Importance Weighted Autoencoders (IWAE) [28]. In [29], the authors presented not-MIWAE, the not-missing-at-random IWAE to deal with data missing not at random. A deep neural network is used to model the conditional distribution of the pattern of missing values, hence acquiring the knowledge about the type of missingness. The proposed model maximizes a lower bound of the joint likelihood and a reparameterisation trick allows deriving the stochastic gradients of the bound for the latent and data spaces. Gondara et al. [30] proposed an unsupervised

approach based on overcomplete deep denoising autoencoder. At the encoder level, the number of neurons per layer increases by a factor as the model goes deeper while at the decoder layer, the number of neurons is scaled back to the original data dimensionality. Although generative models, particularly GAN-based, have achieved state-of-art results for imputation tasks, they are difficult to train. Indeed, they generally involve latent variables that fail to represent the data hence making the interpretation and understanding of the imputation difficult. Furthermore, due to the loss formulation, GAN models suffer from mode collapse in addition to convergence issues [31]. In this work, we address the problem of missing values in context of urban mobility data. Our approach relies on the CP completion method. More specifically, we advocate including urban and temporal information to model the spatiotemporal interaction between regions which leads to better performance in terms of traffic data recovery and imputation.

III. PRELIMINARIES

We present in this section some basic preliminaries and definitions related to tensor calculation.

A tensor is a multidimensional array. The order of the tensor is its number of dimensions. The zero order tensor is a scalar. A first order tensor is a vector. A second order tensor is a matrix. For more than two dimension, the general representation is the tensor.

We use Euler script letter \mathcal{T} to denote a tensor of order $n \geq 3$. We use bold capital letter $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ to denote a matrix in $\mathbb{R}^{I_1 \times I_2}$ and lower case (a, b, c) to denote a vector. The entry of a matrix $\mathbf{A} \in \mathbb{R}^{I_1 \times I_2}$ is denoted by $a_{i_1 i_2}$. The entry of a tensor $\mathcal{T} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_n}$ is denoted by $t_{i_1 i_2, \dots, i_n}$. The nuclear norm of a tensor \mathcal{T} is $\|\mathcal{T}\|_1 = \sum_{i_1} \sum_{i_2} \dots \sum_{i_n} |t_{i_1 i_2, \dots, i_n}|$. The Frobenius norm of a tensor \mathcal{T} is $\|\mathcal{T}\|_F = \left(\sum_{i_1} \sum_{i_2} \dots \sum_{i_n} t_{i_1 i_2, \dots, i_n}^2 \right)^{\frac{1}{2}}$.

We present in followings, some definitions related to the matrix tensor calculus.

Definition 1. The Hadamard product $(*)$ of two tensors of the same size is the element wise multiplication of its entries. Let $\mathcal{T}_1 \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_n}$ and $\mathcal{T}_2 \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_n}$ be two tensors, the Hadamard product, denoted $\mathcal{T}_1 * \mathcal{T}_2$ is the tensor whose entries $(\mathcal{T}_1 * \mathcal{T}_2)_{i_1 i_2, \dots, i_n} = t_{i_1 i_2, \dots, i_n}^{(1)} t_{i_1 i_2, \dots, i_n}^{(2)}$.

Definition 2. The Kronecker product (\otimes) of matrices $\mathbf{A} \in \mathbb{R}^{I_1 \times I_2}$ and $\mathbf{B} \in \mathbb{R}^{I_3 \times I_4}$ is a matrix $\mathbf{C} \in \mathbb{R}^{I_1 I_3 \times I_2 I_4}$ defined as:

$$\mathbf{C} = \mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} \quad (1)$$

Definition 3. The Khatri-Rao product (\odot) of matrices $\mathbf{A} \in \mathbb{R}^{I_1 \times I_3}$ and $\mathbf{B} \in \mathbb{R}^{I_2 \times I_3}$ is a matrix $\mathbf{C} \in \mathbb{R}^{I_1 I_2 \times I_3}$ defined as:

$$\mathbf{C} = \mathbf{A} \odot \mathbf{B} = \begin{bmatrix} a_1 \otimes b_1 & a_2 \otimes b_2 & \dots \end{bmatrix} \quad (2)$$

where a_i and b_j are the i^{th} and j^{th} column of \mathbf{A} and \mathbf{B} respectively.

Definition 4. A mode n -matricization of the N^{th} order tensor, known as unfolding, is the process of organizing the tensor into a matrix. We illustrate in Fig. 1 the first, second and third order matricization of a $N \times M \times T$ tensor. The mode n matricization is denoted as $T_{(n)}$.

Definition 5. The n^{th} order tensor $\mathcal{T} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_n}$ is rank one if it can be written as the outer products of N vectors:

$$\mathcal{T} = a^{(1)} \circ a^{(2)} \circ \dots \circ a^{(N)} \quad (3)$$

$a^{(r)}$, $1 \leq r \leq N$, is a vector in \mathbb{R}^{I_r} . \circ is the outer product.

Definition 6. The CANDECOMP/PARAFAC (CP) approach decomposes the tensor as a sum of vectors from rank one components:

$$\mathcal{T} = \sum_{i=1}^R a_r^{(1)} \circ a_r^{(2)} \circ \dots \circ a_r^{(N)} = [\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}] \quad (4)$$

$a_r^{(i)}$ is the r^{th} vector of matrix $\mathbf{A}^{(i)}$ and $[\]$ denotes the CP decomposition. The set of matrices $\mathbf{A}^{(i)}$ are the latent factor matrices. As an example, let $\mathcal{T} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ a third-order tensor. Its CP decomposition is:

$$\mathcal{T} = \sum_{i=1}^R a_r \circ b_r \circ c_r \quad (5)$$

The factor matrices are the vector combinations from the rank-one components: $\mathbf{A} = [a_1, a_2, \dots, a_R] \in \mathbb{R}^{I_1 \times R}$, $\mathbf{B} = [b_1, b_2, \dots, b_R] \in \mathbb{R}^{I_2 \times R}$ and $\mathbf{C} = [c_1, c_2, \dots, c_R] \in \mathbb{R}^{I_3 \times R}$. Fig 2 illustrates the CP decomposition of a third-order tensor.

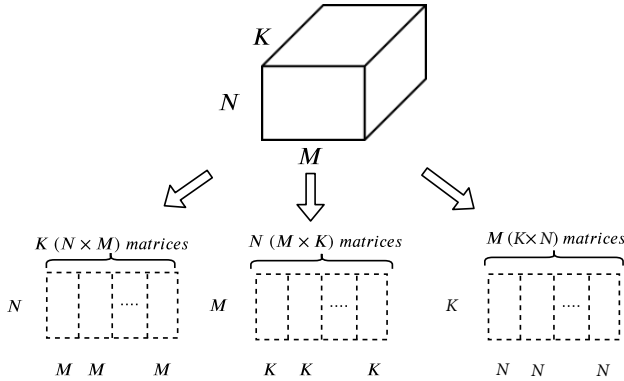


Fig. 1: Matricization (Unfolding) of a 3D tensor of size $(N \times M \times K)$ to three matrices of sizes $(K \times NM)$, $(N \times MK)$ and $(M \times KN)$.

IV. TRAFFIC FLOW DATA COMPLETION USING URBAN AND TIME AWARE CP APPROACH

In this section, we detail the formulation of the traffic tensor completion using an enhanced CP approach. First, we introduce the formulation of the problem. Next, we present a summary of the overall data completion approach and we

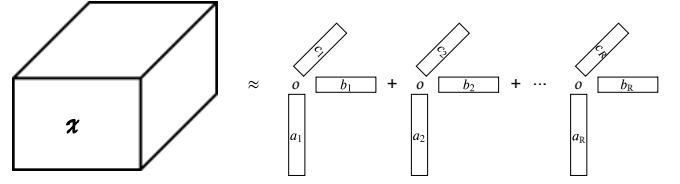


Fig. 2: CP decomposition of third-order tensor.

detail the enhanced CP completion for traffic flow tensor completion. Finally, we show how the spatiotemporal urban features can be integrated with CP to enhance the recovery performance of urban traffic information.

A. Problem formulation

Urban traffic data collected from distributed sensors are prone to multiple imperfections leading to missing measurements. To address the inter-region traffic flow data completion problem, we first segment the area of study into M regions. Then we model the traffic flow from one region to another as a spatiotemporal tensor that may have missing measurements. Multiple approaches can be adopted for the region segmentation including administrative, morphology, grid and road segments based segmentation [32], [33]. In this paper, we simply adopt a grid based approach. Specifically, we set up a boundary box over the area of study and divide it into elementary squares as illustrated in Fig. 3. The size of the elementary squares is adjustable depending on the desired granularity e.g. 1km^2 , 2km^2 , etc. Given M regions and T

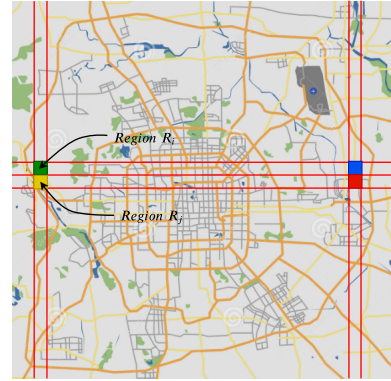


Fig. 3: Grid-based segmentation of Beijing. (For clarity, we illustrate few squares)

time intervals, a traffic flow tensor $\mathcal{X} \in \mathbb{R}^{M \times M \times T}$ is a third order tensor where each entry x_{ijk} represents the number of objects (car, bike, pedestrians ...) located at region R_i at time k and relocated at region R_j at time $k+1$. In previous approaches, this tensor is constructed by considering only the start and end location of the moving objects (e.g. car, bus, etc.). Thus, it does not capture the complete travel patterns of these objects and the intermediate visited segmentation squares. Unlike these approaches, we consider every sampled location from the travel pattern of each moving object to build the traffic flow tensor in order to provide a complete overview

of the traffic during the studied time span.

Let $\mathcal{W} \in \mathbb{R}^{M \times M \times T}$ be a binary tensor such that:

$$w_{ijk} = \begin{cases} 0 & \text{if } x_{ijk} \text{ is missing} \\ 1 & \text{otherwise} \end{cases}$$

\mathcal{W} models the perturbation that leads to the missing information in the data tensor. The observed traffic flow tensor \mathcal{Y} , i.e. the tensor with missing data is the element wise product of the complete tensor \mathcal{X} with the perturbation \mathcal{W} :

$$\mathcal{Y} = \mathcal{W} * \mathcal{X} \quad (6)$$

Our goal is to recover \mathcal{X} given the observed tensor \mathcal{Y} by seeking an approximate tensor $\hat{\mathcal{X}}$ which is as close as possible to the true and complete tensor \mathcal{X} . Our strategy consists of introducing a prior knowledge related to the urban and time context of the traffic flow.

B. Overview of the proposed enhanced CP for traffic flow data completion

Fig. 4 illustrates the proposed traffic data completion approach using an enhanced CP approach taking into account the temporal aspects of the input data and the urban characteristics of the area of study. Given an area of study and its associated database of moving object trajectories and POI, the CP completion approach is a four-stage process: tensor construction, temporal matrix calculation, urban matrix calculation and the urban and time aware CP completion solver. First, the area of study is segmented into a grid of elementary blocks. Each block is identified by its central location. Given a set of trajectories of moving objects e.g. pedestrians, buses, taxis, each trajectory sample is associated to the nearest block with its associated timestamp to construct the traffic tensor. This tensor is corrupted resulting in missing information. Using the POI data, we construct the urban context similarity matrix. For each block, we derive the following metrics: Richness, Diversity, Concentration and Convenience. These metrics are then used to construct the urban similarity matrix using cosine similarity. To derive the temporal matrix, we conduct an entropy analysis to determine the most regular time series in the traffic tensor. Then, we conduct a joint Fourier and correlation analysis to determine the periodicity of this particular time series. The calculated period is used to construct the temporal matrix which is a specific Toeplitz matrix. The two matrices are fed to a modified CP completion objective function which is optimized using an alternating minimization approach. The objective of the minimization is to reduce the error between an approximate traffic tensor and the true one.

We present, in the following, the formulation of CP completion problem and detail the calculation of the urban context and temporal matrices. Then, we present, the modified urban and temporal aware CP completion objective function and how to solve it in order to obtain the approximate complete traffic tensor.

C. CP completion for traffic tensor recovery

CP completion aims at recovering a tensor $\mathcal{X} \in \mathbb{R}^{M \times M \times T}$ given its rank R :

$$\mathcal{X} \approx [\mathbf{A}, \mathbf{B}, \mathbf{C}] = \sum_{r=1}^R a_r \circ b_r \circ c_r \quad (7)$$

The CP optimization problem can be formulated as:

$$\text{minimize } f(A, B, C) = \|\mathcal{W} * (\mathcal{X} - [\mathbf{A}, \mathbf{B}, \mathbf{C}])\|_F^2 \quad (8)$$

with respect to the factor matrices \mathbf{A} , \mathbf{B} and \mathbf{C} . It has been shown [11] that the regularized version of the problem 8 converges faster. It is expressed as follows:

$$\begin{aligned} \text{minimize } f_\lambda(A, B, C) = & \|\mathcal{W} * (\mathcal{X} - [\mathbf{A}, \mathbf{B}, \mathbf{C}])\|_F^2 \\ & + \lambda (\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2) \end{aligned} \quad (9)$$

$\lambda > 0$ is a regularization parameter allowing a tradeoff between the approximation errors and the fitting error. Problem (9) can be solved using the regularized ALS technique. Specifically, three sub-problems are derived:

$$\begin{aligned} \mathbf{A}^{k+1} = & \underset{\mathbf{A} \in \mathbb{R}^{M \times R}}{\text{argmin}} \left\| \mathbf{W}_{(1)} \left(\mathbf{X}_{(1)} - \check{\mathbf{A}}(\mathbf{C}^k \odot \mathbf{B}^k)^T \right) \right\|_F^2 \\ & + \lambda \|\check{\mathbf{A}}\|_F^2 \\ \mathbf{B}^{k+1} = & \underset{\mathbf{B} \in \mathbb{R}^{M \times R}}{\text{argmin}} \left\| \mathbf{W}_{(2)} \left(\mathbf{X}_{(2)} - \check{\mathbf{B}}(\mathbf{C}^k \odot \mathbf{A}^{k+1})^T \right) \right\|_F^2 \\ & + \lambda \|\check{\mathbf{B}}\|_F^2 \\ \mathbf{C}^{k+1} = & \underset{\mathbf{C} \in \mathbb{R}^{T \times R}}{\text{argmin}} \left\| \mathbf{W}_{(3)} \left(\mathbf{X}_{(3)} - \check{\mathbf{C}}(\mathbf{B}^{k+1} \odot \mathbf{A}^{k+1})^T \right) \right\|_F^2 \\ & + \lambda \|\check{\mathbf{C}}\|_F^2 \end{aligned} \quad (10)$$

Where $\mathbf{W}_{(i)}$ and $\mathbf{X}_{(i)}$ are the i^{th} order matricization of tensors \mathcal{W} and \mathcal{X} respectively, T is the transpose operator and k refers to the number of iterations. We can clearly notice that $\lambda \|\check{\mathbf{A}}\|_F^2$, $\lambda \|\check{\mathbf{B}}\|_F^2$ and $\lambda \|\check{\mathbf{C}}\|_F^2$ do not depend on k . It is worth noting that problem 9 always converges toward a global minimum [34]. However, the obtained optimal solution is related to the regularized problem 9, not to problem 8.

D. Urban and time aware CP completion

Moving patterns accross urban area have spatial and temporal dependencies. For example, at 5 PM, the end of working hours and in the city centers, traffic is usually slow with many pedestrians, cars, buses, etc. In addition, traffic flow is also characterized by the so called urban context [16], [35], [36], that is the characteristics of the surroundings such as presence of POIs including transportation facilities (metro, bus and subway stations ...), shopping malls, coffee shops, etc. Wang et al. [16] attempted to incorporate urban context information in the tensor completion problem. Authors defined an urban matrix which captures the similarity between regions in term of POI categories proportion. By category, we refer to the type of POI such as shopping, transportation, restaurant, etc. In addition, temporal information is also incorporated. It reflects the intensity of moving patterns from source to destination. In

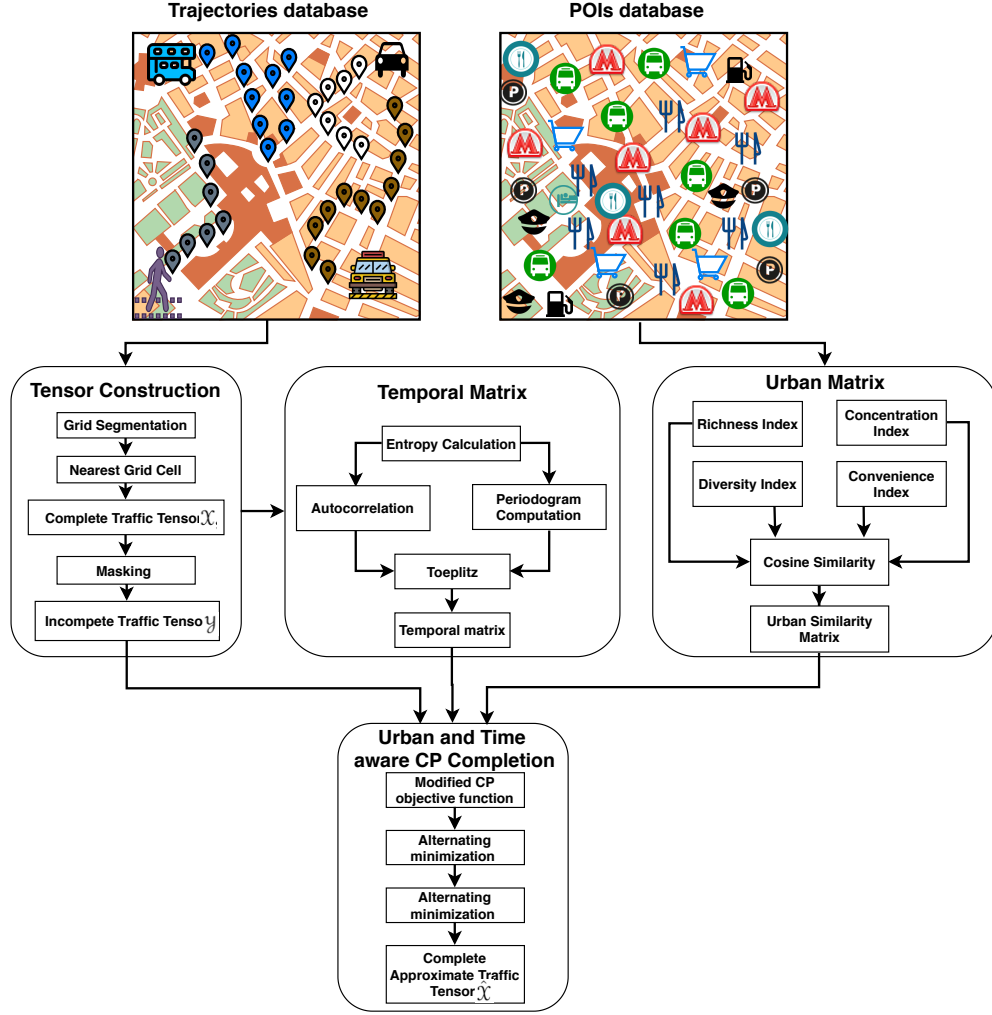


Fig. 4: Traffic tensor data completion using urban and time aware CP approach.

[4], [37], authors modeled this information as a simple Toeplitz matrix \mathbf{T}_0 of the form:

$$\mathbf{T}_0 = \begin{bmatrix} 1 & -1 & 0 & \dots \\ 0 & 1 & -1 & \ddots \\ 0 & 0 & 1 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix} \quad (11)$$

Matrix \mathbf{T}_0 is characterized with the central diagonal of ones, and the first upper diagonal of -1. It simply indicates that traffic flows at adjacent time slots are similar. Authors in [4] recommend incorporating a domain knowledge in the design of \mathbf{T}_0 such as the periodicity of the traffic data rather than assuming similarity with adjacent time slots. The authors proposed another form of temporal matrix in which the temporal similarity is offset by a period of 24h assuming diurnal patterns in the tensor.

We detail in the followings, our proposed urban similarity and temporal matrices which will be incorporated in the CP completion problem.

1) *Urban context similarity matrix*: In ecological and biogeographical studies, statistical measures have been estab-

lished to characterize the diversity of an area in terms of species. This allows to obtain a quantitative estimate of the biological diversity. Inspired by the same concept, an area is characterized by its POIs diversity. For this, we use the study in [42]. Specifically, it considers Hill numbers as a measurement of POIs diversity. They are multifaceted measurements of order q and expressed as:

$${}^q D = \left(\sum_{i=1}^s p_i^q \right)^{\frac{1}{1-q}} \quad (12)$$

where p_i is the i^{th} POI category proportion, s is the number of POIs categories and q is the order. From Hill numbers, we define the following measures for each region R_i :

Definition 7. For $q = 0$, we define the Richness index Rch :

$$Rch = {}^0 D = \left(\sum_{i=1}^s p_i^0 \right)^1 = s \quad (13)$$

In other words, the richness index Rch is the number of POIs categories at region R_i . Therefore, the presence of higher number of POIs categories indicates a richer region. We note that Rch does not depend on the number of POIs of each

category.

Definition 8. For $q = 1$, Eq. 12 is not defined. However, its limit when $q \rightarrow 1$ is the exponential of the Shannon index. We define the Shannon diversity index Sh as:

$$Sh = {}^1D = \lim_{q \rightarrow 1} {}^qD = \exp\left(-\sum_{i=1}^s p_i \log(p_i)\right) \quad (14)$$

Sh expresses the amount of randomness in the POIs categories and the number of POIs. Lower entropy values indicates greater randomness and vice versa.

Definition 9. For $q = 2$, we define the region concentration Ctr :

$$Ctr = {}^2D = 1/\left(\sum_{i=1}^s p_i^2\right) \quad (15)$$

It is the inverse of the Simpson index which reflects the probability that two sampled elements randomly drawn from large community would belong to the same categories.

Definition 10. For a region R_i , we define its traffic convenience Co , that is the proportion of POIs associated to transportation category such as public transport stations, parking lots, etc.

Definition 11. For each region R_i , we define its urban characteristic vector v :

$$v = [Rch, Sh, Ctr, Co] \quad (16)$$

The urban context similarity matrix \mathbf{U} is a matrix whose element u_{ij} reflects the urban similarity between region R_i and R_j that is:

$$u_{ij} = \frac{v_{R_i} \cdot v_{R_j}}{\|v_{R_i}\| \|v_{R_j}\|} \quad (17)$$

Each element is the cosine similarity between each pair of regions. It ranges between -1 and 1. A value closer to 1 indicates high similarity. The urban similarity matrix is driven by the richness, convenience, concentration and diversity indexes. When planning a trip from source to destination, one usually avoids crowded areas with high congestion and concentration of POIs. Such information is captured by the proposed design of the urban similarity matrix.

2) *Temporal similarity matrix*: As highlighted in section IV-D, temporal dependency information is usually manifested across adjacent timestamps or offset by the period of the traffic data. However, simply assuming a 24h period of traffic is inconsistent particularly with presence of missing data as it is problematic to detect the periodicity. Furthermore, traffic patterns are not consistently regular every 24h as the traffic significantly changes from weekdays to weekend and can easily be disrupted by any disturbance on the road network. To overcome this issue, we adopt a time series analysis strategy to construct the temporal similarity matrix. More specifically, we consider the traffic tensor from time series perspective, that is the data across the third dimension of the tensor \mathcal{Y} . Then, we determine the most regular time series as it will provide the most accurate periodicity estimate. A joint

robust Fourier and autocorrelation is conducted to determine the period. This period is then used to construct the temporal matrix using the Toeplitz form.

Definition 12. A traffic time series ts_{ij} represents the traffic information of pair of regions R_i and R_j across the full time horizon T . In the remainder, we refer to ts_{ij} as simply ts .

Definition 13. The most regular time series ts is defined as the one with the least Sample Entropy (SamEn) [40] value. Sample Entropy has been widely used for time series analysis. It calculates the irregularity and reflects the randomness and complexity of the time series. The lower the value of SampEn, the more regular the time series is. Given a time series $ts = ts_1, ts_2, \dots, ts_L$ and a template vector ts^m of length m from ts where $ts_i^m = \{ts_i, ts_{i+1}, \dots, ts_{i+m-1}\}$, the distance function between two template vectors is:

$$d(m, i, j) = d[ts_i^m, ts_j^m] = \max_{k=1 \dots m} \left\{ |ts_{i+k-1} - ts_{j+k-1}| \right\} \quad (18)$$

Let $\Theta_i^m(r)$ be the number of template vectors within distance less or equal a threshold th from ts_i^m is:

$$\Theta_i^m(r) = \sum_{j=1, j \neq i}^{N-m} \Omega(m, i, j, th) \quad (19)$$

where:

$$\Omega(m, i, j, th) = \begin{cases} 1 & \text{if } d(m, i, j) \leq th \\ 0 & \text{otherwise} \end{cases}$$

The probability that two template vectors of length m will match is defined as:

$$\Delta_{th}^m = \frac{1}{N-m} \sum_{i=1}^{N-m} \Theta_i^m(r) \quad (20)$$

Finally, SamEn is expressed as:

$$SampEn(m, th, N) = \ln\left(\frac{\Delta_{th}^m}{\Delta_{th}^{m+1}}\right) \quad (21)$$

SampEn cannot be directly applied as the time series contains missing values. To solve this problem, we adopt the strategy proposed in [41] named KeepSampEn, that is the template vector ts^m must not contain any missing values. Such straightforward approach has shown great stability performance and robustness against missing values.

After identifying the most regular time series, we apply a joint Fourier and autocorrelation analysis to determine its periodicity. The approach consists of transforming the time series into frequency domain, determining the most dominant frequency and then mapping back to the time domain to calculate the period. First, in order to determine the most dominant frequency, we calculate the periodogram of the time series. It is the square of each coefficient of the Fourier Transform of ts . To mitigate the effect of missing data, we use the Lomb-Scargle periodogram [43]. This periodogram is

widely used in astronomy where missing data is a common issue. At a frequency f_k , it is defined as:

$$P(f_k) = \frac{1}{2} \left(\frac{\sum_{i=1}^L \left(ts[i] \cos(2\pi f_k (ts[i] - \tau)) \right)^2}{\sum_{i=1}^L \cos^2(2\pi f_k (ts[i] - \tau))} + \frac{\sum_{i=1}^L \left(ts[i] \sin(2\pi f_k (ts[i] - \tau)) \right)^2}{\sum_{i=1}^L \sin^2(2\pi f_k (ts[i] - \tau))} \right) \quad (22)$$

Where:

$$\tau = \frac{1}{4\pi f_k} \tan^{-1} \left(\frac{\sum_{i=1}^L \sin(4\pi f_k)}{\sum_{i=1}^L \cos(4\pi f_k)} \right) \quad (23)$$

In this particular periodogram, the sine and cosine coefficients are separately normalized by a time constant which depends on the frequency f_k in order to make the transform insensitive to time shift. To identify the most dominant frequency, we use a thresholding approach. Each coefficient is however mapped in the time domain to a period range $[\frac{N}{k}, \frac{N}{k-1}]$. To accurately determine the period, we use the circular autocorrelation. Given a sequence ts , its circular autocorrelation is expressed as:

$$Corr(\theta) = \frac{1}{N} \sum_{i=1}^N ts[i] ts[i + \theta] \quad (24)$$

Therefore, given a time range $[t_1, t_2]$ obtained using Lomb-Scargle periodogram, we look for the presence of peak in $\{Corr(t_1), Corr(t_1 + 1), \dots, Corr(t_2 - 1)\}$ using quadratic fitting. If the obtained fitting is concave, it indicates the presence of a period $t^* = \underset{t_1 \leq t < t_2}{argmax} Corr(t)$. Once t^* is determined, we define the temporal similarity matrix as a Toeplitz matrix in which the difference is offset with t^* :

$$\mathbf{T}_o = \begin{bmatrix} 1 & 0 & \dots & -1 & 0 & \dots \\ 0 & 1 & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 1 & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix} \quad (25)$$

3) *The optimization algorithm:* By introducing the urban and temporal contexts into the CP completion, the modified objective function is expressed as follows:

$$\begin{aligned} \text{minimize } f_{\lambda}^u(\mathbf{A}, \mathbf{B}, \mathbf{C}) &= \|\mathcal{W} * (\mathcal{X} - \llbracket \mathbf{A}, \mathbf{B}, \mathbf{C} \rrbracket)\|_F^2 \\ &+ \lambda (\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2) + \beta (\|\llbracket \mathbf{U}\mathbf{A}, \mathbf{B}, \mathbf{C} \rrbracket\|^2 \\ &+ \|\llbracket \mathbf{A}, \mathbf{U}\mathbf{B}, \mathbf{C} \rrbracket\|^2 + \|\llbracket \mathbf{A}, \mathbf{B}, \mathbf{T}_o\mathbf{C} \rrbracket\|^2) \end{aligned} \quad (26)$$

Where β is a positive regularization parameter. In the above design of f_{λ}^u , two insights are exploited. First, the traffic data with its periodicity and temporal stability are included. Second, the urban similarity matrix reflects how one region is similar to another one in terms of POIs. At a given time, regions with high similarity are highly likely to exhibit the same traffic pattern. This knowledge is incorporated in the modified traffic tensor completion problem. Such design has been used for internet traffic data completion [37] with different space and time context, and shown effective recovery performance.

To solve the objective 26, we adopt an alternating least square procedure. First, we fix \mathbf{B} and \mathbf{C} and we solve for \mathbf{A} . Next, we fix \mathbf{A} and \mathbf{C} and solve for \mathbf{B} . Finally, we fix \mathbf{A} and \mathbf{B} and solve for \mathbf{C} . When fixing two parameters and solving for the third, the problem becomes a simple linear least squares. For example, assuming \mathbf{B} and \mathbf{C} are fixed, the obtained least squares problem can be expressed as follows [37]:

$$\begin{aligned} &\|\mathbf{W}_{(1)} * (\mathbf{X}_{(1)} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T)\|_F^2 + \lambda \|\mathbf{A}\|_F^2 \\ &+ \beta (\|\llbracket \mathbf{U}\mathbf{A}(\mathbf{C} \odot \mathbf{B})^T \rrbracket\|_F^2 + \|\mathbf{A}(\mathbf{C} \odot (\mathbf{U}\mathbf{B}))^T\|_F^2 \\ &+ \|\mathbf{A}((\mathbf{T}_o\mathbf{C}) \odot \mathbf{B})^T\|_F^2) \end{aligned} \quad (27)$$

By writing:

$$\Psi_1 = \mathbf{C} \odot \mathbf{B} \quad \Phi_1 = \mathbf{C} \odot (\mathbf{U}\mathbf{B}) \quad \Gamma_1 = (\mathbf{T}_o\mathbf{C}) \odot \mathbf{B} \quad (28)$$

and taking the derivative of Eq. 27 with respect to \mathbf{A} and setting it equal to zero, we have:

$$\begin{aligned} &(\mathbf{W}_{(1)} * \mathbf{W}_{(1)} * (\mathbf{A} \Psi_1^T)) \Psi_1 + \lambda \mathbf{A} (\mathbf{I}_{[A]} + \Phi_1^T \Phi_1 + \Gamma_1^T \Gamma_1) + \\ &\beta \mathbf{U}^T \mathbf{U} \mathbf{A} \Psi_1^T \Psi_1 = \mathbf{W}_{(1)} * \mathbf{X}_{(1)} \end{aligned} \quad (29)$$

where $\mathbf{I}_{[A]}$ is the identity matrix whose size is the number of rows of \mathbf{A} . Let $vec(\cdot)$ be the operator which creates a column vector from a matrix by stacking its columns one below another:

$$vec(X) = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \quad (30)$$

where x_i is the i^{th} column of matrix X . We use the following formulas:

$$\begin{aligned} vec(\mathbf{A}\mathbf{X}\mathbf{B}) &= (\mathbf{B}^T \otimes \mathbf{A}) vec(\mathbf{X}) \\ vec(\mathbf{A}\mathbf{B}) &= (\mathbf{B}^T \otimes \mathbf{I}_{[A]}) vec(\mathbf{A}) = (\mathbf{I}_{[B^T]} \otimes \mathbf{A}) vec(\mathbf{B}) \\ vec(\mathbf{A}) * vec(\mathbf{B}) &= diag(vec(\mathbf{A})) vec(\mathbf{B}) \\ vec(\mathbf{A} * \mathbf{B}) &= vec(\mathbf{A}) * vec(\mathbf{B}) = vec(\mathbf{B}) * vec(\mathbf{A}) \end{aligned} \quad (31)$$

where $diag(x)$ is a diagonal matrix with the elements of the vector x are in its diagonal.

By applying the vec operator, we have:

$$\begin{aligned} &(\Psi_1^T \otimes \mathbf{I}_{[\mathbf{W}_{(1)}]}) diag(vec(\mathbf{W}_{(1)}) * vec(\mathbf{W}_{(1)})) \cdot \\ &(\Psi_1 \otimes \mathbf{I}_{[A]}) vec(\mathbf{A}) + \beta ((\Psi_1^T \Psi_1) \otimes (\mathbf{U}^T \mathbf{U})) vec(\mathbf{A}) \\ &\lambda ((\mathbf{I}_{[A]} + \Phi_1^T \Phi_1 + \Gamma_1^T \Gamma_1)^T \otimes \mathbf{I}_{[A]}) vec(\mathbf{A}) \\ &= vec(\mathbf{W}_{(1)}) * vec(\mathbf{X}_{(1)}) \\ &\left((\Psi_1^T \otimes \mathbf{I}_{[\mathbf{W}_{(1)}]}) diag(vec(\mathbf{W}_{(1)}) * vec(\mathbf{W}_{(1)})) \right. \\ &(\Psi_1 \otimes \mathbf{I}_{[A]}) + \lambda (\mathbf{I}_{[A]} + \Phi_1^T \Phi_1 + \Gamma_1^T \Gamma_1)^T \otimes \mathbf{I}_{[A]} + \\ &\left. \beta (\Psi_1^T \Psi_1) \otimes (\mathbf{U}^T \mathbf{U}) \right) vec(\mathbf{A}) = \Delta vec(\mathbf{A}) \\ &= vec(\mathbf{W}_{(1)}) * vec(\mathbf{X}_{(1)}) \end{aligned} \quad (32)$$

Finally, we have:

$$\begin{aligned} \Delta \text{vec}(\mathbf{A}) &= \text{vec}(\mathbf{W}_{(1)}) * \text{vec}(\mathbf{X}_{(1)}) \\ \text{vec}(\mathbf{A}) &= \left(\Delta\right)^+ \text{vec}(\mathbf{W}_{(1)}) * \text{vec}(\mathbf{X}_{(1)}) \end{aligned} \quad (33)$$

Where $\left(\cdot\right)^+$ is the Moore-Penrose inverse.

Similarly, to solve for \mathbf{B} , we fix \mathbf{A} and \mathbf{C} , the obtained least square problems is:

$$\begin{aligned} &\|\mathbf{W}_{(2)} * (\mathbf{X}_{(2)} - \mathbf{B}(\mathbf{C} \odot \mathbf{A})^T)\|_F^2 + \lambda \|\mathbf{B}\|_F^2 \\ &+ \beta \left(\|(\mathbf{B}(\mathbf{C} \odot (\mathbf{U}\mathbf{A}))^T)\|_F^2 + \|\mathbf{U}\mathbf{B}(\mathbf{C} \odot \mathbf{A})^T\|_F^2 \right. \\ &\quad \left. + \|\mathbf{B}((\mathbf{T}_o\mathbf{C}) \odot \mathbf{A})^T\|_F^2 \right) \end{aligned} \quad (34)$$

Let:

$$\Psi_2 = \mathbf{C} \odot \mathbf{A} \quad \Phi_2 = \mathbf{C} \odot (\mathbf{U}\mathbf{A}) \quad \Gamma_2 = (\mathbf{T}_o\mathbf{C}) \odot \mathbf{A} \quad (35)$$

We have:

$$\begin{aligned} &\left((\Psi_2^T \otimes \mathbf{I}_{[\mathbf{W}_{(2)}]}) \text{diag}(\text{vec}(\mathbf{W}_{(2)}) * \text{vec}(\mathbf{W}_{(2)})) \right. \\ &(\Psi_2 \otimes \mathbf{I}_{[\mathbf{B}]}) + \lambda \left(\mathbf{I}_{[\mathbf{B}]} + \Phi_2^T \Phi_2 + \Gamma_2^T \Gamma_2 \right)^T \otimes \mathbf{I}_{[\mathbf{B}]} + \\ &\quad \left. \beta (\Psi_2^T \Psi_2) \otimes (\mathbf{U}^T \mathbf{U}) \right) \text{vec}(\mathbf{B}) = \Delta_2 \text{vec}(\mathbf{B}) \\ &= \text{vec}(\mathbf{W}_{(2)}) * \text{vec}(\mathbf{X}_{(2)}) \end{aligned} \quad (36)$$

Therefore:

$$\text{vec}(\mathbf{B}) = \left(\Delta_2\right)^+ \text{vec}(\mathbf{W}_{(2)}) * \text{vec}(\mathbf{X}_{(2)}) \quad (37)$$

Finally, to solve for \mathbf{C} , we fix \mathbf{A} and \mathbf{B} . The obtained least square problems is:

$$\begin{aligned} &\|\mathbf{W}_{(3)} * (\mathbf{X}_{(3)} - \mathbf{C}(\mathbf{B} \odot \mathbf{A})^T)\|_F^2 + \lambda \|\mathbf{C}\|_F^2 \\ &+ \beta \left(\|(\mathbf{C}(\mathbf{B} \odot (\mathbf{U}\mathbf{A}))^T)\|_F^2 + \|\mathbf{C}(\mathbf{U}\mathbf{B} \odot \mathbf{A})^T\|_F^2 \right. \\ &\quad \left. + \|\mathbf{T}_o\mathbf{C}(\mathbf{B} \odot \mathbf{A})^T\|_F^2 \right) \end{aligned} \quad (38)$$

$$\Psi_3 = \mathbf{B} \odot \mathbf{A} \quad \Phi_3 = \mathbf{B} \odot (\mathbf{U}\mathbf{A}) \quad \Gamma_3 = (\mathbf{U}\mathbf{B}) \odot \mathbf{A} \quad (39)$$

We have:

$$\begin{aligned} &\left((\Psi_3^T \otimes \mathbf{I}_{[\mathbf{W}_{(3)}]}) \text{diag}(\text{vec}(\mathbf{W}_{(3)}) * \text{vec}(\mathbf{W}_{(3)})) \right. \\ &(\Psi_3 \otimes \mathbf{I}_{[\mathbf{C}]}) + \lambda \left(\mathbf{I}_{[\mathbf{C}]} + \Phi_3^T \Phi_3 + \Gamma_3^T \Gamma_3 \right)^T \otimes \mathbf{I}_{[\mathbf{C}]} + \\ &\quad \left. \beta (\Psi_3^T \Psi_3) \otimes (\mathbf{U}^T \mathbf{U}) \right) \text{vec}(\mathbf{C}) = \Delta_3 \text{vec}(\mathbf{C}) \\ &= \text{vec}(\mathbf{W}_{(3)}) * \text{vec}(\mathbf{X}_{(3)}) \end{aligned} \quad (40)$$

Therefore:

$$\text{vec}(\mathbf{C}) = \left(\Delta_3\right)^+ \text{vec}(\mathbf{W}_{(3)}) * \text{vec}(\mathbf{X}_{(3)}) \quad (41)$$

By applying **unvec()**, the inverse vec operator, we obtain the solution to the subproblems 27, 34 and 38. We present in Algorithm 1, the pseudocode of the proposed urban and time aware CP tensor completion.

Algorithm 1 Urban and time aware tensor completion

```

1: Input:  $\mathcal{Y}, \mathcal{W}, \mathbf{F}, \mathbf{T}_o, R, \beta, \lambda, tol$ 
2: Output:  $\hat{\mathcal{X}}$ ,
3: Initialize:  $\mathbf{A} \in \mathbb{R}^{M \times R}$ ,  $\mathbf{B} \in \mathbb{R}^{M \times R}$  and  $\mathbf{C} \in \mathbb{R}^{M \times R}$ 
4:  $Eval_0 = f_\lambda^u(\mathbf{A}, \mathbf{B}, \mathbf{C})$  (Eq. 26)
5: Repeat:
6:   Solve for  $\mathbf{A}$  using Eq. 33
7:   Solve for  $\mathbf{B}$  using Eq. 37
8:   Solve for  $\mathbf{C}$  using Eq. 41
9:    $Eval = f_\lambda^u(\mathbf{A}, \mathbf{B}, \mathbf{C})$  (Eq. 26)
10:   $\epsilon = Eval_0 - Eval$ 
11:   $Eval_0 = Eval$ 
12: Until  $\epsilon < tol$ 
13: Output:  $\hat{\mathcal{X}}$  (Eq. 7)
  
```

V. EXPERIMENTAL RESULTS

In this section, we validate the effectiveness of our completion approach. We conduct a set of experiments on two traffic datasets and compare the recovery performance with multiple state-of-art approaches.

A. Data

The data we use in our experiments are road traffic records of taxi from two cities: Porto, Portugal and Beijing, China.

- Porto Taxi: the data contain 442 trajectories of taxi cabs in Porto, Portugal. For each taxi, time stamped geolocations along with metadata are provided. After segmenting the area of study into $1km^2$ cells and aggregating the traffic in each grid cell, we obtain a $(91 \times 91 \times 2880)$ traffic tensor.
- T-drive: The data contains 15 million time stamped GPS records of 10357 taxis from February 2nd to February 8th 2008 from Beijing, China. The average sampling rate is about 177 seconds. The data are preprocessed to eliminate noisy records. By applying grid segmentation using $2 km^2$ and traffic record aggregation, we obtain a tensor of size $(1516 \times 1516 \times 352)$.

B. Methodology

We set up two evaluation protocols. In the first one, given the traffic tensor, we drop measurement at random. This is achieved by randomly generating the binary mask \mathcal{W} and multiply it by the traffic tensor \mathcal{X} to create the observed data \mathcal{Y} . However, in a realistic case, missing data are the results of a failure usually related to sensor or transmission equipment dysfunction and for a some duration. We also simulate this structured missing values scenario by imputing measurements at random cells for a time duration. For each scenario, we vary the rank parameter R and the missing value rate then report the Relative Error (RE):

$$RE = \frac{\|\mathcal{X} - \hat{\mathcal{X}}\|^2}{\|\mathcal{X}\|^2} \quad (42)$$

Where \mathcal{X} and $\hat{\mathcal{X}}$ are the true and recovered traffic tensors. We evaluate the proposed completion approach against:

CP_ALS [44], CP_ARLS [11], CP_OPT [45], CP_WOPT [10], CP_APR [46] with two configurations: row subproblems by projected quasi-Newton CP_APR_PQNR and row subproblems by projected damped Hessian CP_APR_PDNR and GCP_OPT [47]. We set $m = 3$ and $th = 0.3$ for the Sample Entropy.

We also evaluate the proposed approach against GAN-based approaches: GAIN [21], MIDA [30] (for both random and structured missing values), MIWAE (for random missing values) [27] and not-MIWAE [29] (for structured missing values). For each approach, we report the best obtained result after multiple runs.

C. Recovering traffic tensor with random missing values

We illustrate in Fig. 5 the variation of RE with respect to varying missing values rate for Porto Taxi. We run this simulation using the regularization parameters $\lambda = \beta = 0.1$. Results show that the proposed tensor completion approach achieved the best performance for low and high missing value rates and varying R . We notice 23% improvement compared to the closest performance for $R = 4$ with low rate missing values. We notice that GCP_OPT and CP_WOPT completely fail to recover data for high missing value rates. Figure 6 illustrates the recovery performance of Beijing T-drive Taxi data with random missing values. The findings confirm the effectiveness of our approach in completing the traffic tensor with 33% improvement compared to the closest performance for $R = 5$ and low missing values rate. We notice that for high missing rates, GCP_OPT, CP_WOPT could not recover the tensor data.

We illustrate in Fig. 7 the comparison of the proposed completion approach against state-of-art generative models: GAIN, MIDA and MIWAE. We report in this comparison the best performance achieved by the proposed technique. The results show that for low missing rates these models achieved better performance. However, for severe missing rates, the proposed approach achieved significantly better performance with 30% improvement for Porto traffic tensor having 80% missing rate.

D. Recovering tensor with structured missing values

In this simulation scenario, we set $\lambda = 0.1$ and $\beta = 0.01$. Figure 8 depicts the RE results for different settings. The best recovery performance is achieved by our method while GCP_OPT results in high RE. Most of the algorithms showed stable performance except CP_APR_PQNR and GCP_OPT. CP_WOPT again results in similar performance as in the previous scenario. We illustrate in Fig. 9 the recovery performance for Beijing T-drive taxi with structured missing values. Our CP completion approach achieved the lowest RE values for all missing values rate with 26% improvement compared to the closest performance for $R = 3$ and low missing values rate. Fig. 10 depicts the comparison of the proposed completion method against GAIN, MIDA and not-MIWAE. Note that not-MIWAE is designed to recover data with structured missing values. The findings show similar performance to the random missingness experiment. In fact,

although the generative models achieved better performance for low missing rates, a performance degradation is witnessed for higher missing rates. For 80% missing rate, our method exhibited an improvement of 35% compared to the closest performance for Porto traffic data.

E. Discussion

Experiments showed that our urban and time aware CP tensor completion approach is efficient in recovering missing traffic information with both random and structured missing values. The other algorithms, although achieved competitive performance with low missing values rate, they failed to recover the traffic tensor with very high imputation. On overall, the proposed technique performed better on T-Drive data compared to Porto Taxi. This has been also the case for all techniques used for comparison.

It is worth noting that performance of our approach depends on the regularization parameter λ and β . We analyze the variation of the Relative Error with respect to different λ and β to recover Porto data under structured missing values. The results are depicted in Fig. 11. We notice that high regularization of the matrix norm, i.e. high λ , value leads to higher error. In addition, we notice that the optimal choice for parameter β is in the range of 0.01 to 0.02. Higher or lower values result in higher RE. Automatic tuning of these parameters is an open research question. For this work, λ and β parameters are empirically chosen.

We further compare the performance on tensors constructed using the source and destination sub-regions only and using all locations visited in the journey from source to destination. Without loss of generality, we conduct this experiment on T-Drive data with $R = 6$ where we attempt to recover the tensors under different corruption levels with random and structure missing values and evaluate the RE. We refer to the first tensor as Beijing-S2D and the second one as Beijing-All. To quantify the sparsity of each tensor, we use the $S_{l_2}^{l_1}$ [48]:

$$S_{log}(x) = - \sum_i \log(1 + x_i^2) \quad (43)$$

Computation of the sparsity measure shows that Beijing-S2D is more sparse than Beijing-All with 5 order of magnitude where $S_{log} = -1.210^6$ and $S_{log} = -6.510^6$ for Beijing-S2D and Beijing-All respectively. Results, depicted in Figures 13 and 14 show that the proposed approach achieved similar performances on both tensors and under both missing values scenarios. Hence, we can conclude that it is not affected by the sparsity of the input tensor. Our design choice for the traffic tensor is motivated by the fact that using all trajectories' locations would result in constructing a traffic tensor that better reflects the traffic information in the area of study.

Finally, we analyze the time complexity of each completion approach. We run each algorithm until the minimization of its objective function is less than 10^{-6} . We illustrate in Fig. 12 the results of the experiment with time in log scale. The proposed approach achieved relatively high time complexity as the algorithm requires computing multiple operations and applying Moore-Penrose inverse. Therefore it is important to

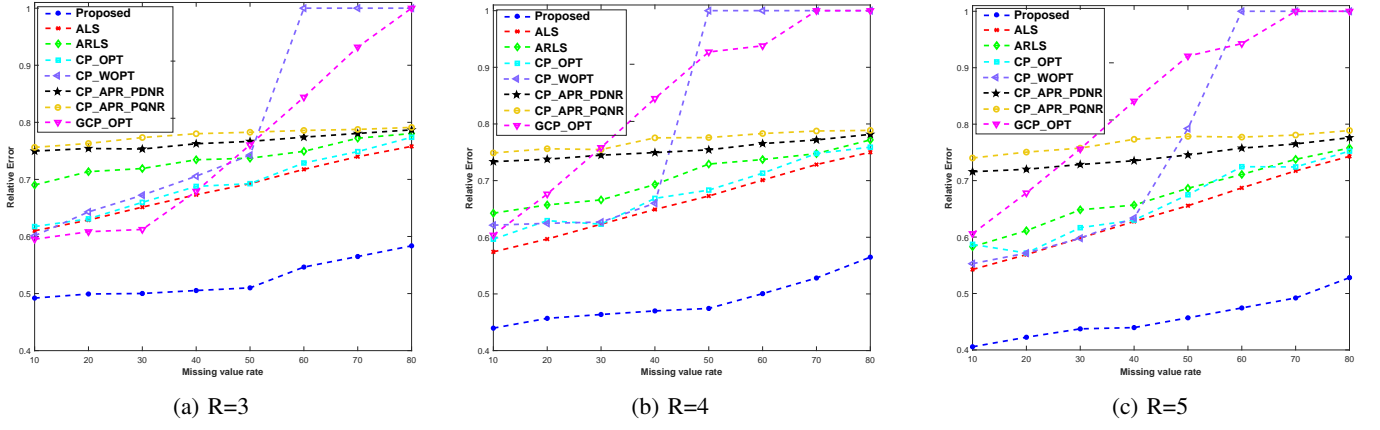


Fig. 5: Recovering Porto traffic tensor with random missing values: Relative Error results

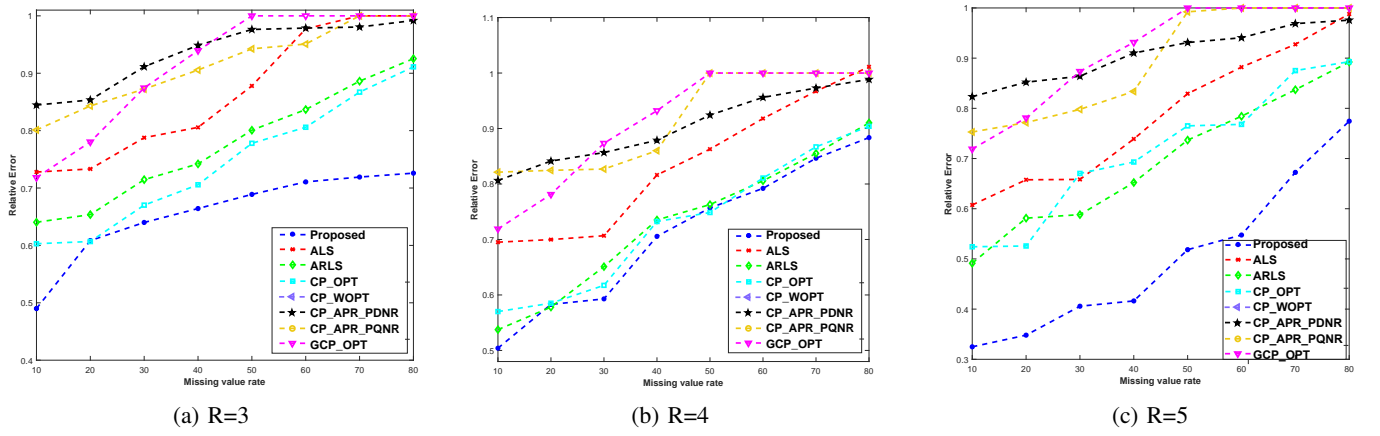


Fig. 6: Recovering Beijing traffic tensor with random missing values: Relative Error results

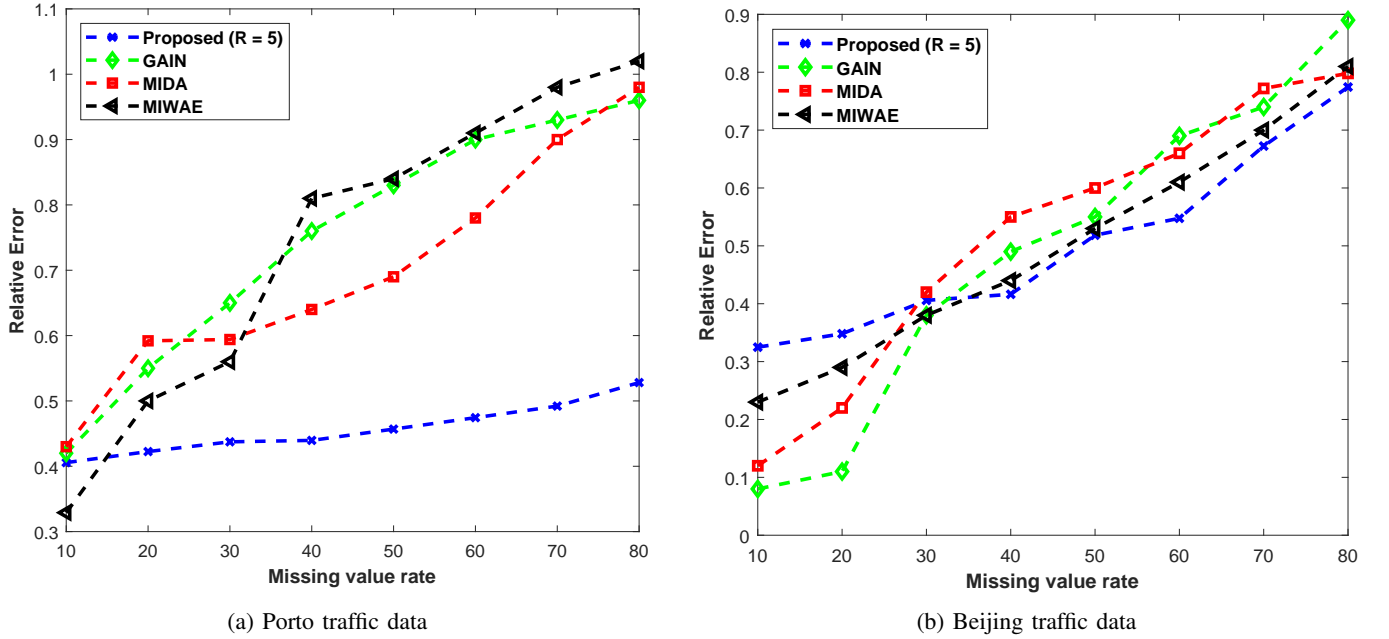


Fig. 7: Comparison with generative models

achieve a tradeoff between performance and execution time.

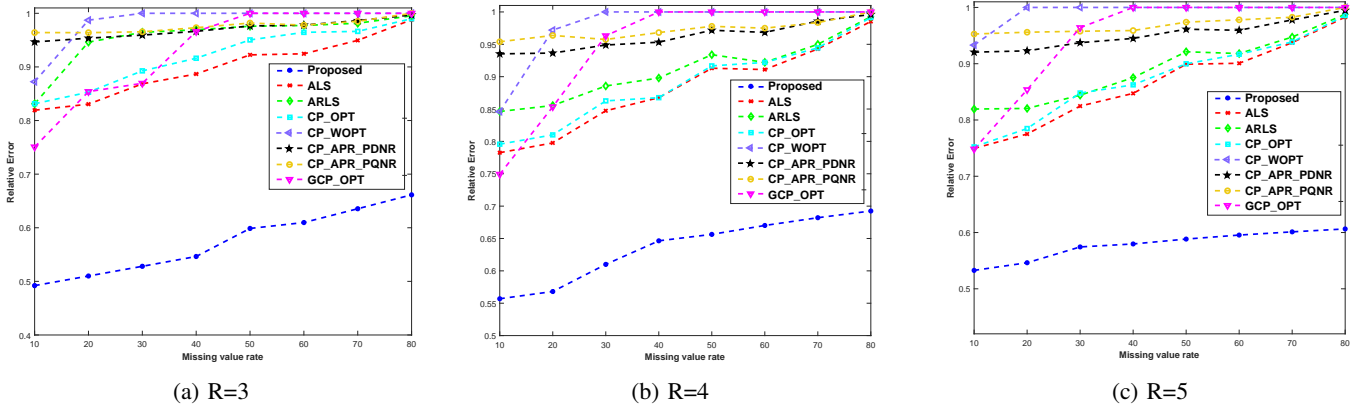


Fig. 8: Recovering Porto traffic tensor with structured loss of values: Relative Error results

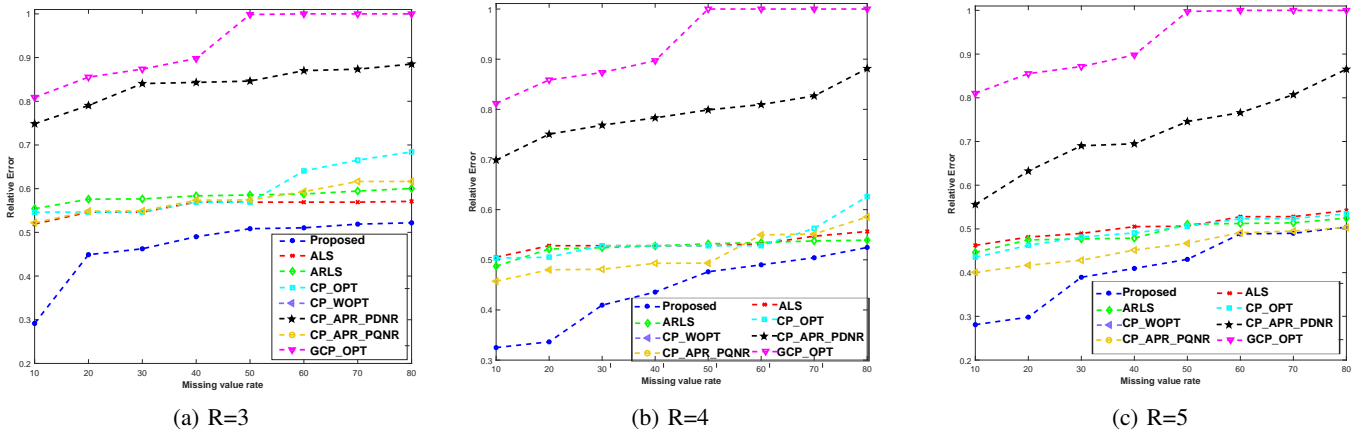


Fig. 9: Relative Error results for recovering Beijing traffic tensor with structured missing values

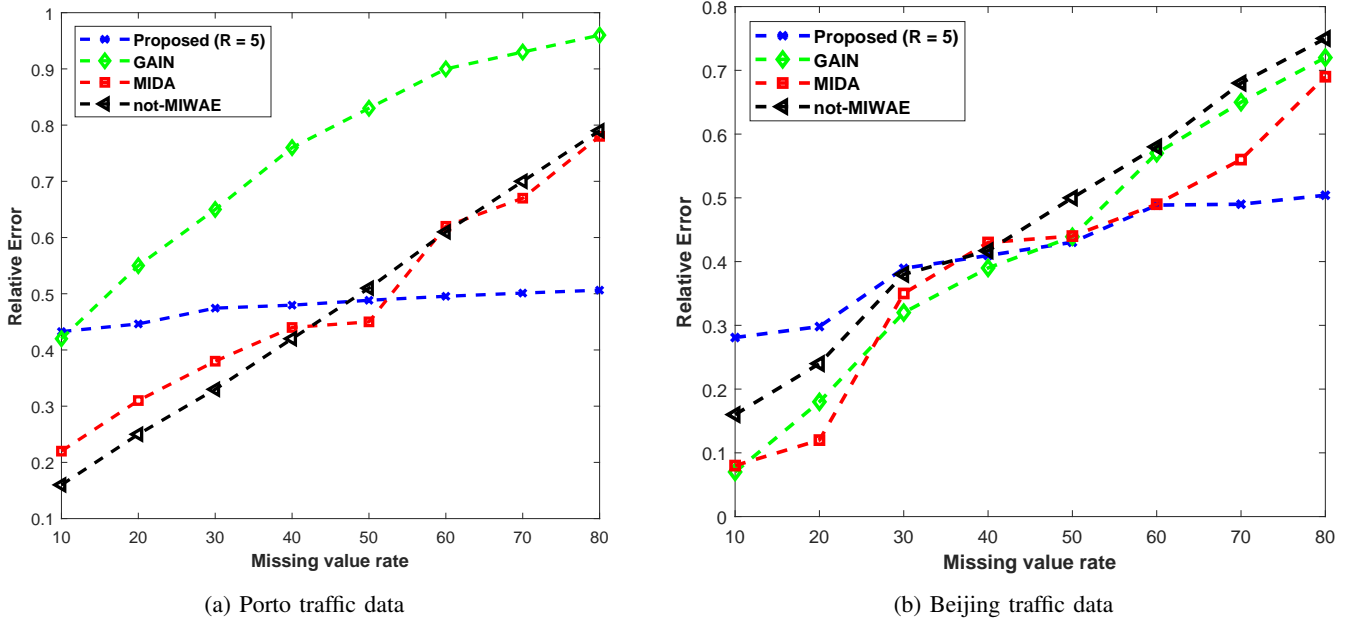


Fig. 10: Comparison with generative models and structured missing values

VI. CONCLUSION

We proposed a CP based completion approach to recover the missing values from traffic tensor. We augmented the CP algo-

rithm with additional information related to the urban context of the area of study. This includes several biodiversity-inspired characteristics related to the richness, diversity, concentration

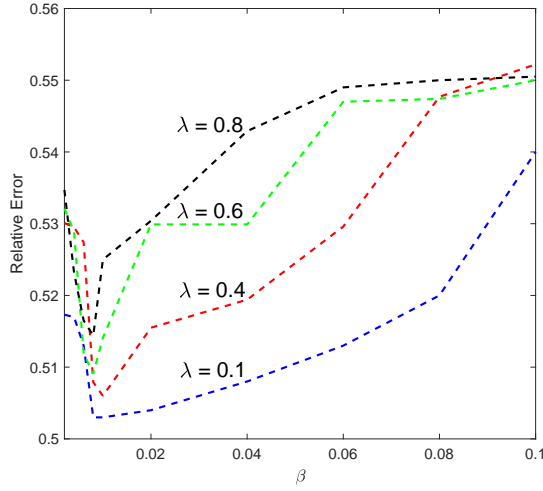


Fig. 11: Recovering Porto traffic data under structured missing values: Variation of the Relative Error with respect to λ and β

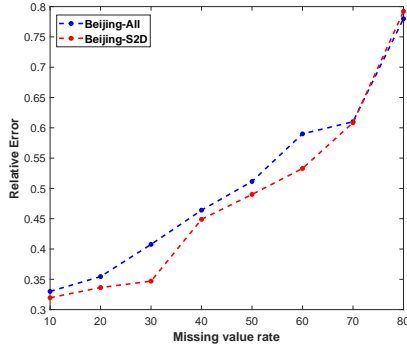


Fig. 12: Recovering T-Drive data constructed with source and destination only and all trajectories' locations: Random missing values

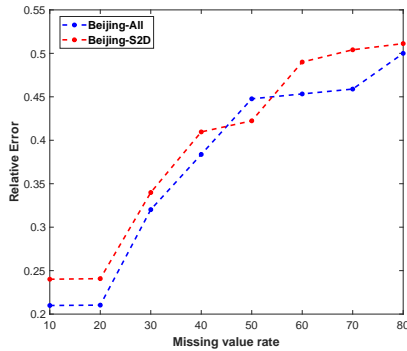


Fig. 13: Recovering T-Drive data constructed with source and destination only and all trajectories' locations: Structured missing values

and traffic convenience. In addition, we take into account the temporal information by considering the periodicity of the traffic data. We established two comparison scenarios and analyzed the proposed approach from time complexity perspective. Our findings showed that the CP completion

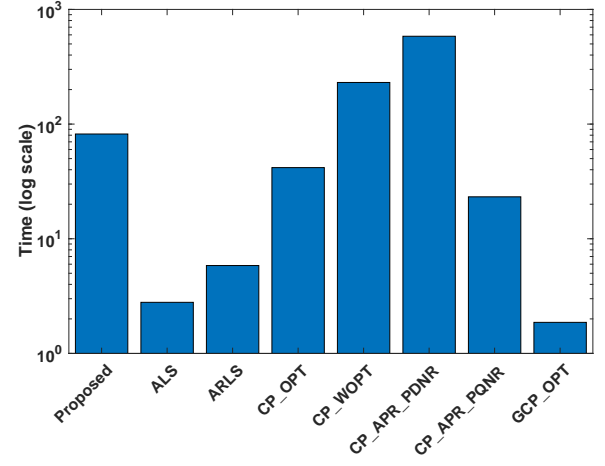


Fig. 14: Time complexity

approach augmented with the proposed urban and time information achieved competitive recovery performance. In future work, we will focus on alleviating the time complexity. We will also address the choice of the regularization parameters and propose a solution for automatic tuning

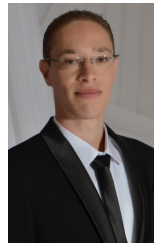
ACKNOWLEDGEMENT

This research was made possible by NPRP 9-224-1-049 grant from the Qatar National Research Fund (a member of The Qatar Foundation). The statements made herein are solely the responsibility of the authors.

REFERENCES

- [1] Junbo Zhang, Yu Zheng and Dekang Qi. 2017. Deep spatio-temporal residual networks for citywide crowd flows Prediction, Thirty-First AAAI Conference on Artificial Intelligence, 1655-1661
- [2] Minh X. Hoang, Yu Zheng and Ambuj K. Singh. 2016. FCCF: Forecasting citywide crowd flows, ACM SIGSPATIAL, 6:1-6:10 Based on Big Data
- [3] Wenhao Huang, Guojie Song, Haikun Hong and Kunqing Xie. 2014. Deep architecture for traffic flow prediction: deep belief networks with multitask learning, IEEE Transactions on Intelligent Transportation Systems, 2191-2201
- [4] Y. Zhang, M. Roughan, W. Willinger and L. Qiu, "Spatio-temporal compressive sensing and internet traffic matrices", ACM SIGCOMM, pp. 267-278, 2009
- [5] D. Donoho, "Compressed sensing, IEEE Transactions on Information Theory", vol. 52, no. 4, pp. 1289-1306, 2006
- [6] M. Roughan, Y. Zhang, W. Willinger and L. Qiu, "Spatio-temporal compressive sensing and internet traffic matrices (Extended Version)", IEEE/ACM Transactions on Networking, vol. 20, no. 3, pp. 662-676, 2012
- [7] Z. Wen, W. Yin and Y. Zhang, "Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm", Mathematical Programming Computation, vol. 4, no. 4, pp. 333-361, 2012
- [8] Z. Long, Y. Liu, L. Chen and Ce Zhu, "Low rank tensor completion for multiway visual data", Signal Processing, vol. 155, pp. 301-316, 2019
- [9] J. Liu, P. Musialski, P. Wonka and J. Ye, "Tensor completion for estimating missing values in visual data", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 208-220, 2012
- [10] E. Acar, D. Dunlavy, T. Kolda and M. Mørup, "Scalable tensor factorizations for incomplete data", Chemometrics and Intelligent Laboratory Systems, vol. 106, no. 1, pp. 41-56, 2011
- [11] N. Li, S. Kindermann and C. Navasca, "Some convergence results on the Regularized Alternating Least-Squares method for tensor decomposition", Linear Algebra and its Applications, vol. 438, no. 2, pp. 796-812, 2013

- [12] Y. Li, Z. Li and L. Li, "Missing traffic data: comparison of imputation methods" IET Intelligent Transport Systems, vol. 8, no. 1, pp. 51-57, 2014
- [13] D. Ni, J. D. Leonard, A. Guin and C. Feng, "Multiple imputation scheme for overcoming the missing values and variability issues in ITS data", Journal of Transportation Engineering, vol. 131, no. 12, pp. 931-938, 2005
- [14] X. Chena, Z. Hea and L. Sun, "A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation", Transportation Research Part C, vol. 98, pp. 73-84, 2019
- [15] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using Markov chain Monte Carlo", Proceedings of the 25th international conference on Machine learning, pp. 880-887, 2008
- [16] J. Wang, J. Wu, Z. Wang, F. Gao and Z. Xiong, "Understanding urban dynamics via context-aware tensor factorization with neighboring regularization", arXiv:1905.00702
- [17] H. Tan, Y. Wu, B. Shen, P. J. Jin, B. Ran, "Short-Term Traffic Prediction Based on Dynamic Tensor Completion", IEEE Transactions on Intelligent Transportation Systems, vol. 17, pp. 2123-2133, 2016
- [18] Z. Li, N. D. Sergin, H. Yan, C. Zhang and F. Tsung, "Tensor completion for weakly-dependent data on Graph for metro passenger flow prediction", arXiv:1912.05693
- [19] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world", In The 17th ACM SIGKDD international conference on Knowledge Discovery and Data mining, pp. 316-324, 2011
- [20] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang "T-drive: driving directions based on taxi trajectories", In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 99-108, 2010
- [21] J. Yoon, J. Jordon and M. van der Schaar, "GAIN: Missing data imputation using Generative Adversarial Nets", arXiv:1806.02920
- [22] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Networks", arXiv:1406.2661
- [23] S. Cheng-Xian Li, B. Jiang and B. Marlin, MisGAN: Learning from Incomplete Data with Generative Adversarial Networks, 7th International Conference on Learning Representations, 2019
- [24] G. Boquet, A. Morell, J. Serrano and J. Lopez Vicario, "A variational autoencoder solution for road traffic forecasting systems: Missing data imputation, dimension reduction, model selection and anomaly detection", Transportation Research Part C: Emerging Technologies, vol. 115, pp. 102622, 2020
- [25] D. P. Kingma and M. Welling, "Auto-Encoding variational Bayes", arXiv:1312.6114
- [26] V. Fortuin, D. Baranchuk, G. Rätsch and S. Mandt, "GP-VAE: Deep Probabilistic Time Series Imputation", arXiv:1907.04155
- [27] P.-A. Mattei and J. Frellsen, "MIWAE: Deep generative modelling and imputation of incomplete Data", arXiv:1812.02633
- [28] Y. Burda, R. Grosse and R. Salakhutdinov, "Importance Weighted Autoencoders", arXiv:1509.00519
- [29] N. Bruun Ipsen, P.-A. Mattei and J. Frellsen, "not-MIWAE: deep generative modelling with missing not at random data", arXiv:2006.12871
- [30] L. Gondara and K. Wang, "MIDA: multiple imputation using denoising autoencoders", arXiv:1705.02737
- [31] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford and X. Chen, "Improved Techniques for Training GANs", arXiv:1606.03498.
- [32] N. J. Yuan, Y. Zheng and X. Xie "Segmentation of urban areas using road networks", Microsoft, Albuquerque, NM, USA, Tech. Rep. MSR-TR-2012-65, 2012
- [33] Y. Zheng, Y. Liu, J. Yuan and X. Xie, "Urban computing with Taxicabs", Proceedings of the 13th international conference on Ubiquitous computing, pp.89-98, 2011
- [34] L.-K. Lim and P. Comon, "Nonnegative approximations of nonnegative tensors", Journal of Chemometrics, vol. 23, pp. 432-441, 2009
- [35] J. Yuan, Y. Zheng and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs", international conference on Knowledge discovery and data mining, pp. 186-194, 2012
- [36] N. J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng and H. Xiong, "Discovering urban functional zones using latent activity trajectories", IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 3, pp. 712-725, 2015
- [37] H. Zhou, D. Zhang, K. Xie and Y. Chen, "Spatio-temporal tensor completion for imputing missing internet traffic data", IEEE International Conference on Performance, Computing and Communications (IPCCC), pp. 1-7, 2015
- [38] Z. Li, J. Wang and J. Han, "Mining event periodicity from incomplete observations", Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 444-452, 2012
- [39] ZheZ.nhui Li, J. Wang and J. Han, "ePeriodicity: mining event periodicity from incomplete observations", IEEE Transactions on Knowledge and Data Engineering, Vol. 27, No. 5, pp. 1219-1232, 2015
- [40] J. S. Richman, and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy". American Journal of Physiology-Heart and Circulatory Physiology, vol. 278, no. 6, pp. H2039-H2049, 2000
- [41] X. Dong, C. Chen, Q. Geng, Z. Cao, X. Chen, J. Lin, Y. Jin, Z. Zhang, Y. Shi and X. Douglas Zhang, "An Improved method of handling missing values in the analysis of sample entropy for continuous monitoring of physiological signals", Entropy, vol. 21, no. 274., 2019
- [42] L. Jost. "Entropy and diversity", Oikos, vol. 113, no 2, pp. 363-375, 2006
- [43] J. D. Scargle, "Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data", Astrophysical Journal, vol. 263, pp. 835-853, 1982
- [44] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications", SIAM Review, vol. 51, pp. 455-500, 2009
- [45] E. Acar D. M. Dunlavy and T. G. Kolda, "A scalable optimization approach for fitting canonical tensor decompositions", Journal of Chemometrics, vol. 25, pp. 67-86, 2011
- [46] E. C. Chi, T. G. Kolda, "On Tensors, Sparsity, and Nonnegative Factorizations", arXiv:1112.2414
- [47] D. Hong, T. G. Kolda and J. A. Dueresch, "Generalized Canonical Polyadic Tensor Decomposition", arXiv:1808.07452
- [48] S. Rickard, M. Fallon, "The Gini index of speech", in: Proceedings of the 38th Conference on Information Science and Systems (CISS'04), 2004



in urban computing and mobile health systems.

Ahmed Ben Said received the Ph.D. degree in computer Science from the University of Burgundy, France, in 2015. He was a Research Assistant with Qatar University on several projects, including the simulation of a surgical cutting operation using 3-D modeling, the usage of multispectral image for face recognition, and the development of reliable mHealth system for remote patient diagnosis. He currently holds a postdoctoral position at Qatar University. His research interests include machine learning and computer vision. He is also interested



Abdelkarim Erradi is an Associate Professor in the Computer Science and Engineering Department at Qatar University. His research and development activities and interests focus on service-oriented computing, cloud Services composition and mobile crowdsensing. He leads several funded research projects in these areas. He has authored several scientific papers in international conferences and journals. He received his Ph.D. in computer science from the University of New South Wales, Sydney, Australia. Besides his academic experience, he possesses 12 years professional experience as a Designer and a Developer of large scale enterprise applications.

Abdelkarim Erradi is an Associate Professor in the Computer Science and Engineering Department at Qatar University. His research and development activities and interests focus on service-oriented computing, cloud Services composition and mobile crowdsensing. He leads several funded research projects in these areas. He has authored several scientific papers in international conferences and journals. He received his Ph.D. in computer science from the University of New South Wales, Sydney, Australia. Besides his academic experience, he possesses 12 years professional experience as a Designer and a Developer of large scale enterprise applications.