A Spatial Missing Value Imputation Method for Multi-view Urban Statistical Data

Yongshun Gong\(^1,\ast\), Zhibin Li\(^1\), Jian Zhang\(^1,\ast\), Wei Liu\(^1\), Bei Chen\(^2\) and Xiangjun Dong\(^3,\ast\)

\(^1\)Faculty of Engineering and IT, University of Technology Sydney, Sydney, Australia
\(^2\)Microsoft Research, Beijing, China
\(^3\)School of Computer, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

\{yongshun.gong, zhibin.li\}@student.uts.edu.au, \{jian.zhang, wei.liu\}@uts.edu.au,
beichen@microsoft.com, d-xj@163.com

Abstract

Large volumes of urban statistical data with multiple views imply rich knowledge about the development degree of cities. These data present crucial statistics which play an irreplaceable role in the regional analysis and urban computing. In reality, however, the statistical data divided into fine-grained regions usually suffer from missing data problems. Those missing values hide the useful information that may result in a distorted data analysis. Thus, in this paper, we propose a spatial missing data imputation method for multi-view urban statistical data. To address this problem, we exploit an improved spatial multi-kernel clustering method to guide the imputation process cooperating with an adaptive-weight non-negative matrix factorization strategy. Intensive experiments are conducted with other state-of-the-art approaches on six real-world urban statistical datasets. The results not only show the superiority of our method against other comparative methods on different datasets, but also represent a strong generalizability of our model.

1 Introduction

Urban statistic data connect social sciences, urban computing, administrative management, transportation, and regional planning that are significant for city development [Murgante and Danese, 2011; Zheng \textit{et al.}, 2014; Gong \textit{et al.}, 2020]. These statistical data usually include multi-fold views (e.g., views of Population and Economy) to reveal the growth gaps among different administrative regions from various perspectives. For example, the economy view records the key economic indicators for fine-grained regions, such as the number of industries and employee statistics; and the population view consists of detailed population information of all age groups in each region.

The statistic data provide key statistics to governments, business and the community on social science, for the benefit of all aspects of human life.

However, in some places, statistical data are hard to be entirely acquired due to document defacement, error recordings, and statistician misplay. Such missing data hide useful information which may cause distorted results for further analysis. To the best of our knowledge, it is still a blank field concerning this specific problem, but the real demand appears. Hence, the missing value imputation for urban statistical data is a vital task for reliable urban computing and government services.

In this paper, we study the problem of missing-data imputation for the Australian Bureau of Statistics (ABS), which has some unique challenges:

- **Missing temporal information.** In the real-world data from ABS, almost all the missing values in the current year were also missing in the past years, which may be caused by the region restriction and complicated human-made errors. This violates the basic assumption of matrix completion [Candès and Recht, 2009] that the unobserved entries are sampled uniformly at random. Thus matrix completion-based approaches may not work in this case.

- **Multi-view problem.** The complicated underlying interactions suggest that simply recovering the missing information without considering the correlations among attributes and multi-modes will end up with a poor performance. For example, the economy view has strong correlations with the income and population views, so that a high-quality economy in a region usually goes along with a better income and a larger population; and a low-level economy in a region has a high probability of being connected with a lower income and a smaller population.

- **Spatial correlation mining problem.** As illustrated in Fig-
imputation, they have disadvantages in nature, i.e., they are
not designed for this spatial problem. [Chen and Liu, 2012]
used the inverse distance weighting (IDW) method to inter-
polate the spatial rainfall distribution. [Wu and Li, 2013] uti-
lized the spatial information as inputs in a residual kriging
method to estimate the average monthly temperature. Unlike
the spatial model, some successful spatio-temporal models
were proposed for use with time stream data [Yi et al., 2016;
Cheng and Lu, 2017; Zhou and Huang, 2018; Atluri et al.,
2018]. However, they focused on filling missing entries by
considering both spatial and temporal properties, and would
not perform well on the static spatial data without the tempo-
ral information. Furthermore, these discussed methods lever-
aged the spatial guidance but did not consider the problem on
multi-view datasets.

2.2 Multi-view Learning
Multi-view learning methods involved the diversity of differ-
ent views that can jointly optimize functions based on various
feature subsets [Singh and Gordon, 2008; Li et al., 2018].
[Xu et al., 2015] proposed a matrix co-factorization based
method (MVL-IV) to embed different views into a shared
domain that the incomplete views can be estimated
by the information on observed views. To connect multi-
ple views, MVL-IV assumes that different views have dis-
crete 'feature' matrices (i.e., \{\mathbf{H}_i\}_{i=1}^d), but correspond to the
same coefficient matrix (i.e., \mathbf{W}). However, it does not ex-
plain the spatial correlations and may suffer from the imbal-
ance problem, i.e., if there is a substantial missing ratio gap
between views, the coefficient matrix \mathbf{W} is mostly learned from
the dense view. In our method, we have addressed
this weakness by introducing guidance matrices. Another
widely used strategy for solving the multi-view problem is
tensor factorization [Rendle et al., 2009; Xiong et al., 2010],
but this restricts a regular tensor that requires the number of
dimensions per view to be the same. Moreover, multiple
kernel learning with incomplete views [Trivedi et al., 2010;
Liu et al., 2017] only focuses on completing missing kernels
instead of filling missing values. To the best of our knowl-
edge, none of the above studies considered both spatial and
multi-view problems. Hence, in this paper, we proposed an
effective missing value imputation model for multi-view ur-
ban statistical data.

3 The Proposed Method
3.1 Problem Description and Preliminary
As illustrate in Figure 2, this research focuses on completing
the missing values in the urban statistical data, where one ur-
ban dataset contains multiple views, e.g., Income, Population,
Economy views, etc. For a dataset with n regions (r_1,...,r_n)
and d views, the dimension of attributes in the p-th view is
m_p (1 \leq p \leq d). Our method aims to impute the missing
values with a high accuracy.

Multi-view NMF
The multi-view NMF aims to learn a latent subspace \mathbf{W} \in \mathbb{R}_+^{n \times k}
by multiple views \{\mathbf{X}_1,...,\mathbf{X}_d\} through the multi-view
generation matrices \mathbf{H}_p \in \mathbb{R}_+^{k \times m_p}. The basic missing data

3.2 The Proposed Method
To handle the multi-view problem with spatial charac-
teristic, we propose a Spatially related Multi-Kernel K-
Means (S-MKKM) method to identify the underlying
relationships among multiple views and capture the re-
gional similarities.

We propose an adaptive-weight non-negative matrix fac-
torization approach to leverage the information learned
above to tackle the multi-view missing data imputation
problem. Besides, the proposed method also takes the
guidance from the single-view and the real geographic
information with KNN strategy into consideration.

A spatial multi-view missing data imputation method
for urban statistical data based on non-negative matrix
factorization is proposed, called SMV-NMF. SMV-NMF
does not rely on the temporal information but achieves a
great performance only using spatial information.

Our experiments on six real-world datasets verify the
effectiveness of our method. All the empirical results
show that the proposed method SMV-NMF outperforms
all the other state-of-the-art approaches. Furthermore,
SMV-NMF shows strong generalizability and can trans-
fer the constructed model from one urban dataset to an-
other well.

2 Related Work
2.1 Spatial Missing Data Imputation
Missing data imputation is a significant task for data anal-
ysis [Van Buuren, 2018]. In the spatially related problem,
neighborhood and collaborative filtering [Su and Khoshgof-
taar, 2009; Yi et al., 2016] based methods are two kinds of
dominant approaches in missing data filling. Although some
classical methods (e.g., zero-filling, mean value filling, re-
gression models) can be applied to the spatial missing data
imputation, they have disadvantages in nature, i.e., they are
not designed for this spatial problem. [Chen and Liu, 2012]
The imputation model can be described as the following optimization objective:

\[
\arg \min_{W \geq 0, H_p \geq 0} J_0 = \sum_{p=1}^{d} \|Y_p \odot (X_p - WH_p)\|_F^2, \tag{1}
\]

where \(Y_p\) are indicator matrices whose entry \(Y_p(i, j)\) is one if \(X_p(i, j)\) has been recorded (for observed values) and zero otherwise (for missing values); and \(\odot\) is the Hadamard product operator.

**Multiple Kernel K-means (MKKM)**

Let \(\{x_i\}_{i=1}^n\) be a collection of \(n\) samples (region), \(x_i\) represents the statistical features of the \(i\)-th region, and \(\phi_p(\cdot)\) be the \(p\)-th view mapping that maps \(x\) onto the \(p\)-th reproducing kernel Hilbert space. In this case, each sample has multiple feature representations defined by a group of feature mappings \(\phi_\beta(x_i) = [\beta_1 \phi_1(x_i)^T, \ldots, \beta_d \phi_d(x_i)^T]^T\), where \(\beta\) consists of the coefficients of the \(d\) base kernels. A kernel function can be expressed as \(\kappa_\beta(x_i, x_j) = \phi_\beta(x_i)^T \phi_\beta(x_j) = \sum_{p=1}^{d} \beta_p^2 \kappa_p(x_i, x_j)\). And a kernel matrix \(K_\beta\) is then calculated by applying the kernel function \(\kappa_\beta(\cdot, \cdot)\) to \(\{x_i\}_{i=1}^n\). Based on the kernel matrix \(K_\beta\), the objective of MKKM can be written as:

\[
\min_{V, \beta} \text{Tr}(K_\beta(I_n - V V^T)) \tag{2}
\]

\[
\text{s.t. } V \in \mathbb{R}^{n \times l}, V^T V = I_1, \beta^T 1_d = 1, \beta_p \geq 0, \forall p,
\]

where \(V\) is the clustering matrix; \(1_d \in \mathbb{R}^d\) is a column vector with all 1 elements; \(I_n\) and \(I_l\) are identity matrices with size \(n\) and \(l\); \(l\) is the number of clusters.

### 3.2 Multi-view Spatial Similarity Guidance

As discussed in Section 2.2, multi-view matrix factorization based methods suffer from the imbalance problem. In this paper, we build the similarity guidance \(X_p^{mv}\) for the \(p\)-th view \(X_p\) to address this problem. Accordingly, we propose an approach to obtain regional similarities via the spatially related MKKM model, called S-MKKM. The basic idea is that the development of a city gradually fosters different functional groups, such as educational and business districts, where the regions belonging to the same group would have strong connections with each other [Zheng et al., 2014]. S-MKKM utilizes the MKKM clustering algorithm combined with a graph Laplacian dynamics strategy (an effective smoothing approach for finding spatial structure similarity [Deng et al., 2016; Gong et al., 2018]) to cluster regions into the functional groups. Specifically, we construct a graph Laplacian matrix \(L\), defined as \(L = D - M\), where \(M\) is a graph proximity matrix that is constructed from the regional physical topology (i.e., \(M_{i,j} = 1\) if and only if the region \(x_i\) is contiguous to \(x_j\)), and \(D\) is a diagonal matrix \(D_{i,i} = \sum_j (M_{i,j})\). With this constraint, the S-MKKM model is expressed as follows:

\[
\min_{V, \beta} \text{Tr}(K_\beta(I_n - V V^T)) + \alpha \text{Tr}(V^T LV) \tag{3}
\]

\[
\text{s.t. } V \in \mathbb{R}^{n \times l}, V^T V = I_1, \beta^T 1_d = 1, \beta_p \geq 0, \forall p,
\]

where \(\alpha\) is the regularization parameter; \(V\) is the consensus clustering matrix.

To get the complete kernels, we initially impute the missing data for each view by a simple method, such as KNN or MF. After that, Eq. (3) can be solved by alternately updating \(V\) and \(\beta\): i) With the kernel coefficients \(\beta\) fixed, \(V\) can be obtained by choosing the \(l\) smallest eigenvectors of \((-K_\beta + \alpha L)\). ii) With \(V\) fixed, \(\beta\) can be optimized via solving the quadratic programming with linear constraints [Liu et al., 2017].

The objective of the S-MKKM is to discover the regions with similar properties and build the guidance matrices \(X_p^{mv}\). After having gotten \(V\), \(X_p^{mv}\) can be built. Figure 3 shows an example of this process. The construction process of \(X_p^{mv}\) is that i) for the unknown entry \(x_{ij}\), and the region \(x_i\) in \(c\)-th cluster, we use its corresponding value \(x_{ij}\) from the centroid region to impute \(x_{ij}\); ii) if the corresponding value of centroid region is also missed, a greedy strategy will be used to find the nearest observed value for imputation.

### 3.3 Adaptive-Weight NMF

To learn the knowledge from \(X_p^{mv}\) more reliably, we propose an adaptive weighting strategy in the NMF imputation process. The adaptive-weight matrix of the \(p\)-th view is denoted as \(Z_p \in \mathbb{R}_+^{n \times n}\), which is built by an exponential function as shown in Eq. (4) and (5).

\[
z_{p(i)} = e^{-\text{Dist}(v_i, v_{c(i)})}, \tag{4}
\]
where $D_{ist}(\cdot, \cdot)$ is the Euclidean distance calculating from the geo-location ($v_i$) and its corresponding centroid region ($v_c(i)$), here we use the latent embedding $v_i$ to represent the geo-location of region $i$, and $v_c(i)$ represents the centroid of the $c$-th cluster which contains region $v_i$; $z_p \in \mathbb{R}^n$ is a column vector and $1_m$ is all-ones vector with size $m$. It is not a straight way for imputation, but the adaptive-weight matrix $Z_p$ controls how much information can be extracted. $Z_p$ adjusts the penalty of each estimated entry. As emphasised in the First Law of Geography [Tobler, 1970], the near things have more spatial correlations than distant things. If the distance between $x_c$ and $x_v(i)$ is small, we want a high penalty to guide the imputation process.

Combining the above strategy, our model can be described as the following optimization function:

$$
\arg\min_{W \geq 0, H_p \geq 0} \mathcal{J}_1 = \mathcal{J}_0 + \lambda_1 \sum_{p=1}^{d} \left\| \tilde{Y}_p \otimes Z_p \odot (X_p^{mv} - W H_p) \right\|_F^2,
$$

where $\tilde{Y}_p = 1 - Y_p$, 1 is an all one matrix that has the same size as $Y_p$; $X_p^{mv}$ is a homomorphic matrix of $X_p$; and $\lambda_1$ is the regularization parameter to control the learning rate of $X_p^{mv}$.

### 3.4 Improved by Single-view and KNN Guidances

S-MKKM aims to find the regional groups by considering multiple views simultaneously. However, it is obvious that each view has its characteristics, and the relationships between regions in one specific view are also critical for imputing missing entries. To consider the above knowledge, we assume the spatially related kernel k-means (S-KKM) to capture the similarities among regions of each view. It is essentially analogous to the learning process of S-MKKM as discussed in section 3.2, but considering each view, respectively. For one view $X_p$, the S-KKM model is expressed as follows:

$$
\begin{align*}
\min_{V_p} \text{Tr}(K_p(I_n - V_p V_p^T)) + \alpha \text{Tr}(V_p^T L V_p) \\
\text{s.t. } V_p \in \mathbb{R}^{n \times l}, V_p^T V_p = I_l,
\end{align*}
$$

where $K_p$ is one separate kernel and $V_p$ represents the $p$-th clustering matrix based on $X_p$.

In fact, to reduce the complexity of our model, we assume that the physical location affects the clustering performance with the same degree and the number of clusters is the same as that in S-MKKM, i.e., $l$ and $\alpha$ are the same as used in Eq. (3). The reason behind this assumption is that most cities have the same functional regions, such as the residential region and business region. Thus, it is reasonable that we choose the same $\alpha$ and $l$ in this practical task. Besides, $\alpha$ and $l$ are very stable due to the intrinsic property of the urban statistical data, and we fixed them in the experiments. The single view guidance matrix $X_p^{mv}$ and adaptive-weight matrix $Z_p$ can be constructed by the same strategy of building $X_p^{mv}$ and $Z_p$.

Furthermore, for each region, its $k$-nearest spatial neighbors imply rich information that should be considered in our model. Even though the regional physical topology is already involved in multi-view and single-view learning processes, the KNN is a more flexible method. After structuring $X_p^{knn}$ which is an imputed matrix with the average value of $k$-nearest neighbors, our final optimization function is shown as follows:

$$
\begin{align*}
\arg\min_{W \geq 0, H_p \geq 0} \mathcal{J}_1 = \mathcal{J}_0 + \lambda_2 \sum_{p=1}^{d} \left\| \tilde{Y}_p \otimes Z_p \odot (X_p^{mv} - W H_p) \right\|_F^2 + \lambda_3 \sum_{p=1}^{d} \left\| \tilde{Y}_p \odot (X_p^{knn} - W H_p) \right\|_F^2,
\end{align*}
$$

where $\lambda_2$ and $\lambda_3$ are the regularization parameters to control the learning rate of $X_p^{mv}$ and $X_p^{knn}$, respectively.

Given the estimated factor matrices $W$ and $H_p$ based on the above update equations, the filled data are given by:

$$
\tilde{X}_p = Y_p \odot X_p + \tilde{Y}_p \odot (W H_p).
$$

### 3.5 Learning Algorithm

As Eq. (8) is a non-convex problem, we use the multiplicative update strategy [Lee and Seung, 2001] to ensure the convergence under the following update rules. We first initialize latent space matrices ($W$ and $H_p$) by decomposing data matrices $\{X_1...X_d\}$. The update rules for $W$ and $H_p$ are presented in Eq. (10) - (11).

$$
\begin{align*}
W &= W \odot \\
&\sum_{p=1}^{d} (Y_p \odot X_p + \tilde{Y}_p \odot (\lambda_1 Z_p \odot X_p^{mv} + \lambda_2 Z_p \odot X_p^{mv} + \lambda_3 X_p^{knn})) H_p^T \\
&\sum_{p=1}^{d} ((Y_p + \tilde{Y}_p \odot (\lambda_1 Z_p + \lambda_2 Z_p + \lambda_3 1)) \odot (W^T H_p) H_p^T)
\end{align*}
$$

$$
\begin{align*}
H_p &= H_p \odot \\
W(Y_p \odot X_p + \tilde{Y}_p \odot (\lambda_1 Z_p \odot X_p^{mv} + \lambda_2 Z_p \odot X_p^{mv} + \lambda_3 X_p^{knn})) \\
&/ W(Y_p + \tilde{Y}_p \odot (\lambda_1 Z_p + \lambda_2 Z_p + \lambda_3 1)) \odot (W^T H_p)
\end{align*}
$$

The above two multiplicative update rules guarantee to be non-negative if the initialization is positive. Without this constraint, the matrices $W$ and $H_p$ could be negative, thus the imputation results could be negative too, which is a contradiction to the facts. The process of SMV-NMF is summarized in Algorithm 1.

**Time complexity and convergence.** We discuss the time complexity and convergence of SMV-NMF here. The time complexity of guidance matrices $X_p^{mv}$ and $X_p^{knn}$ is mainly affected by MKKM. Even though MKKM has a high computational complexity (O(n^3)), it is not involved in update loop of variables ($W$ and $H_p$). Eq. (10) and Eq. (11) present that the time complexity of our final function is governed by matrix multiplication operations in each iteration. Therefore, the
time complexity per iteration is dominated by \( O(nk^2) \). Due to the pursing of pinpoint accuracy, we sacrifice efficiency to some degree in this real-world problem. In terms of convergence, Algorithm 1 is guaranteed to converge when \( W \) or \( H_p \) is fixed, because the second-order derivatives regarding \( W \) or \( H_p \) are positive semi-definite. Thus, the objective function can achieve its optimal value by optimizing \( W \) and \( H_p \) alternately.

4 Experiments

In this paper, we have conducted complete experiments to demonstrate the effectiveness of our method\(^1\).

4.1 Datasets

There are six real-world urban statistical datasets (Sydney, Melbourne, Brisbane, Perth, SYD-large, and MEL-large), where -large datasets contain much more fine-grained regions from Australian Bureau of Statistics (2017). Each dataset contains four views, i.e., Economy, Family, Income, and Population. The size of the six datasets are 174, 284, 220, 130, 2230, 1985 respectively. The designation of regions is based on the Australian Statistical Geography Standard for the best practical value. The scales of different views are normalized into the same range \([0, 10]\) so that we can evaluate the results together. The numbers of the dimension of the four views are 43, 44, 50, 97, respectively. We choose Australian cities mostly because the Australian Bureau of Statistics provides enough data for our study, while such data from other countries is inaccessible to us. However, our method is general enough and can be applied to other cities with administrative areas and statistical census data. To guarantee the diversity of testing, for each missing ratio, we randomly select the test columns and repeat the experiment 20 times and report average results.

\(^1\)The strict proof, resource code, and parameters used to achieve the best performance on different datasets are shown in the https://github.com/SMV-NMF/SMV-NMF.

4.2 Baselines & Measures

Baselines

We compare the proposed method SMV-NMF with the following 12 baselines. All parameters of the proposed method and baselines are optimized by the grid search method.

- **sKNN**: A classical method that uses the average values of its \( k \) nearest spatial neighbors as an estimate \((k=6)\).
- **MKKM\(^a\)**, **MKKM\(^b\)**: A MKKM based method to handle the incomplete views [Liu et al., 2017]. We modified it to adapt to the spatially related data, then interpolated a missing value by its \( k \) nearest spatial neighbors \((k=6)\); Utilize the mean value of each cluster to fill the missing data.
- **NMF**: Fill the missing data by NMF.
- **IDW**: A global spatial learning method compared in many works [Chen and Liu, 2012; Cheng and Lu, 2017].
- **UCF**: The Local spatial learning method based on collaborative filtering [Su and Khoshgoftaar, 2009; Yi et al., 2016].
- **IDW+UCF**: The average result of IDW and UCF.
- **MVL-IV**: A state-of-the-art multi-view learning method based on matrix co-factorization, which learns a same coefficient matrix to connect multiple views [Xu et al., 2015].
- **ST-MVL**: A state-of-the-art method to impute spatio-temporal missing data [Yi et al., 2016]. We only use its spatial part due to the problem of missing temporal information.
- **SMV-MF**, **MV-NMF\(^a\)**, **MV-NMF\(^b\)**: Remove the non-negativity constraint in SMV-NMF; Remove the graph Laplacian dynamics strategy in SMV-NMF when building the \( X^{mv} \) and \( X^{sv} \); Remove the KNN guidance in SMV-NMF.

**Measures.** We utilized the most widely used evaluation metrics in our paper, namely Mean Relative Error (MRE) and Root Mean Square Error (RMSE).

\[
MRE = \frac{\sum_{i=1}^{Q} |u_i - \hat{u}_i|}{\sum_{i=1}^{Q} u_i}, \quad RMSE = \sqrt{\frac{\sum_{i=1}^{Q} (u_i - \hat{u}_i)^2}{Q}},
\]

where \( \hat{u}_i \) is a prediction for missing value and \( u_i \) is the ground truth; \( Q \) is the number of prediction values.

4.3 Results on Urban Statistical Datasets

The first set of experiments is designed to assess performance on each dataset. We pick up half of statistical fields (properties) in each urban dataset randomly as the validation set, and the other half as the test set. In the test set, we randomly select missing ratios from 10% to 70% to evaluate the imputation accuracy.

Table 1 presents the average errors of all missing ratios across different test methods. It is clear show that our approaches (SMV-MF, MV-NMF\(^a\), MV-NMF\(^b\), SMV-NMF) perform much better than other baselines across different missing ratios on six real-world datasets, where SMV-NMF achieves the best results. Without the non-negativity constraint, SMV-MF performs worse than SMV-NMF, which demonstrates the effectiveness of this constraint. MVL-IV yields better results than ST-MVL, MKKM\(^a\), IDW+UCF, and NMF because it considers the multi-view problem.

To represent our results more clearly, we pick the top eight methods varying different missing ratios on the Sydney dataset, which is shown in Figure 4. It is apparent that NMF
Table 2: Generalizability test. We report the average MRE and RMSE of all missing ratios and best results are bold.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Sydney MRE</th>
<th>Sydney RMSE</th>
<th>Melbourne MRE</th>
<th>Melbourne RMSE</th>
<th>Brisbane MRE</th>
<th>Brisbane RMSE</th>
<th>Perth MRE</th>
<th>Perth RMSE</th>
<th>SYD-large MRE</th>
<th>SYD-large RMSE</th>
<th>MEL-large MRE</th>
<th>MEL-large RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>sKNN</td>
<td>0.3302</td>
<td>1.5319</td>
<td>0.3108</td>
<td>1.3181</td>
<td>0.3534</td>
<td>1.4787</td>
<td>0.3701</td>
<td>1.5754</td>
<td>0.2998</td>
<td>1.2543</td>
<td>0.2635</td>
<td>1.155</td>
</tr>
<tr>
<td>MKKM-MF</td>
<td>0.3281</td>
<td>1.5507</td>
<td>0.3462</td>
<td>1.4635</td>
<td>0.3773</td>
<td>1.5934</td>
<td>0.3986</td>
<td>1.6992</td>
<td>0.3413</td>
<td>1.6112</td>
<td>0.3067</td>
<td>1.4552</td>
</tr>
<tr>
<td>IDW</td>
<td>0.3321</td>
<td>1.5183</td>
<td>0.3187</td>
<td>1.3188</td>
<td>0.3517</td>
<td>1.4663</td>
<td>0.3724</td>
<td>1.5574</td>
<td>0.3273</td>
<td>1.4992</td>
<td>0.3081</td>
<td>1.2587</td>
</tr>
<tr>
<td>UCF</td>
<td>0.3566</td>
<td>1.6631</td>
<td>0.3380</td>
<td>1.4635</td>
<td>0.3626</td>
<td>1.5928</td>
<td>0.3757</td>
<td>1.6554</td>
<td>0.3327</td>
<td>1.4230</td>
<td>0.3321</td>
<td>1.5093</td>
</tr>
<tr>
<td>IDW+UCF</td>
<td>0.3300</td>
<td>1.4604</td>
<td>0.3141</td>
<td>1.3048</td>
<td>0.3408</td>
<td>1.3967</td>
<td>0.3591</td>
<td>1.4924</td>
<td>0.3045</td>
<td>1.2236</td>
<td>0.2970</td>
<td>1.2111</td>
</tr>
<tr>
<td>MKKM-MF</td>
<td>0.3073</td>
<td>1.4393</td>
<td>0.2833</td>
<td>1.2264</td>
<td>0.3167</td>
<td>1.3479</td>
<td>0.3546</td>
<td>1.5066</td>
<td>0.2915</td>
<td>1.2808</td>
<td>0.3019</td>
<td>1.2300</td>
</tr>
<tr>
<td>NMF</td>
<td>0.2189</td>
<td>1.3841</td>
<td>0.1990</td>
<td>1.1557</td>
<td>0.2225</td>
<td>1.3048</td>
<td>0.2469</td>
<td>1.2866</td>
<td>0.2385</td>
<td>1.1996</td>
<td>0.2032</td>
<td>1.0660</td>
</tr>
<tr>
<td>ST-MVL</td>
<td>0.2948</td>
<td>1.3137</td>
<td>0.2833</td>
<td>1.1796</td>
<td>0.3117</td>
<td>1.2932</td>
<td>0.3325</td>
<td>1.3949</td>
<td>0.2948</td>
<td>1.0772</td>
<td>0.2829</td>
<td>1.1453</td>
</tr>
<tr>
<td>MV-L-IV</td>
<td>0.1948</td>
<td>1.0603</td>
<td>0.1744</td>
<td>0.8185</td>
<td>0.1970</td>
<td>0.9698</td>
<td>0.2252</td>
<td>1.0676</td>
<td>0.1792</td>
<td>0.8959</td>
<td>0.1824</td>
<td>0.9223</td>
</tr>
<tr>
<td>SMV-MF</td>
<td>0.2032</td>
<td>0.7051</td>
<td>0.1574</td>
<td>0.6770</td>
<td>0.1847</td>
<td>0.7703</td>
<td>0.2239</td>
<td>1.0095</td>
<td>0.1681</td>
<td>0.8046</td>
<td>0.1763</td>
<td>0.8124</td>
</tr>
<tr>
<td>MV-NMF</td>
<td>0.1806</td>
<td>0.9257</td>
<td>0.1640</td>
<td>0.7296</td>
<td>0.2170</td>
<td>0.9721</td>
<td>0.2170</td>
<td>0.9721</td>
<td>0.1714</td>
<td>0.8246</td>
<td>0.1858</td>
<td>0.8613</td>
</tr>
<tr>
<td>MV-NMF</td>
<td>0.1829</td>
<td>0.9609</td>
<td>0.1738</td>
<td>0.8048</td>
<td>0.1647</td>
<td>0.7703</td>
<td>0.2239</td>
<td>1.0095</td>
<td>0.1681</td>
<td>0.8046</td>
<td>0.1763</td>
<td>0.8124</td>
</tr>
<tr>
<td>SMV-NMF</td>
<td>0.1773</td>
<td>0.9084</td>
<td>0.1667</td>
<td>0.7471</td>
<td>0.1574</td>
<td>0.7051</td>
<td>0.2097</td>
<td>0.9347</td>
<td>0.1620</td>
<td>0.7753</td>
<td>0.1692</td>
<td>0.7911</td>
</tr>
</tbody>
</table>

Figure 4: Average RMSE with the variation of missing ratios.

Figure 5: The average RMSE in generalizability tests.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Dataset Melbourne MRE</th>
<th>Dataset Melbourne RMSE</th>
<th>Dataset Brisbane MRE</th>
<th>Dataset Brisbane RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDW+UCF</td>
<td>0.3311</td>
<td>1.4026</td>
<td>0.3659</td>
<td>1.5624</td>
</tr>
<tr>
<td>IDW</td>
<td>0.3324</td>
<td>1.3374</td>
<td>0.3697</td>
<td>1.4934</td>
</tr>
<tr>
<td>IDW+UCF</td>
<td>0.3182</td>
<td>1.3061</td>
<td>0.3518</td>
<td>1.4552</td>
</tr>
<tr>
<td>MKKM-MF</td>
<td>0.2827</td>
<td>1.2018</td>
<td>0.3137</td>
<td>1.3020</td>
</tr>
<tr>
<td>ST-MVL</td>
<td>0.2794</td>
<td>1.1391</td>
<td>0.3123</td>
<td>1.2698</td>
</tr>
<tr>
<td>NMF</td>
<td>0.1538</td>
<td>0.9067</td>
<td>0.1781</td>
<td>0.9196</td>
</tr>
<tr>
<td>MV-L-IV</td>
<td>0.1510</td>
<td>0.7879</td>
<td>0.1636</td>
<td>0.8089</td>
</tr>
<tr>
<td>SMV-NMF</td>
<td>0.1506</td>
<td>0.7202</td>
<td>0.1493</td>
<td>0.6718</td>
</tr>
</tbody>
</table>

4.4 Generalizability Test

We conduct experiments on testing the generalizability in this section. In detail, we choose the dataset Sydney as the validation set and two urban datasets (Melbourne and Brisbane) as the test sets. We report the experimental results on eight available algorithms. SMV-NMF is the most outstanding approach, as shown in Figure 5.

Our method represents strong generalizability which can transfer the constructed model from one urban dataset to another. This is because there are high correlations among cities. For example, the number of functional regions of each city is mostly the same, resulting in the same amount of clusters. The gap between SMV-NMF and MVL-IV narrows as the missing ratio increases, but the former is more robust than the latter because SMV-NMF achieves the best results across all missing ratios. Table 2 reveals the average errors using two evaluation metrics. The generality test demonstrates that our model SMV-NMF is a universal model that performs well crossing different urban statistical datasets.

5 Conclusion

In this paper, we propose a spatial missing data imputation method for multi-view urban statistical data, called SMV-NMF. To address the multi-view problem, an improved spatial multi-kernel method is designed to guide the imputation process based on the NMF strategy. Moreover, the spatial correlations among different regions are involved in our method from two perspectives. Firstly, the latent similarities are discovered by S-MKKN and S-KKM based on the idea of finding functional regions, and secondly, KNN is used for capturing the information of real geographical positions. We conduct intensive experiments on six real-world datasets to compare the performance of our model and other state-of-the-art approaches. The results not only show that our approach outperforms all other methods, but also represent strong generalizabilities crossing different urban datasets.

Table 1: The average MRE and RMSE of all missing ratios on four urban statistical datasets. Best results are bold.
References


