Science & Technology Development Journal – Engineering and Technology 2024, ():1-11

Open Access Full Text Article

Research Article

GPS trajectory imputation: A hybrid approach combined clustering and GAIN-based algorithm

Khang Nguyen Duy[1,2,*], Thanh Hoang Le Hai[1,2], Nguyen Tran Tho[2], Trung Dang Anh[1,2], Nam Thoai[1,2]

Use your smartphone to scan this QR code and download this article

ABSTRACT

The advancement of computing power and the proliferation of big data have opened unprecedented avenues for the Intelligent Transportation Systems (ITS) community to extract valuable insights from Global Positioning System (GPS) trajectory data. However, the reality of real-world GPS trajectory data often lacks complete information due to various factors (e.g. detector damage, transmission loss, ...), thus posing significant challenges for trajectory analysis and operational efficiencies within transportation systems. To address this issue, time series data imputation techniques have emerged as critical solutions to accurately fill in missing data points. Existing imputation approaches can be classified into statistical methods and deep generative models. Significantly, within the domain of deep generative models, Generative Adversarial Imputation Networks (GAIN) have exhibited promise in the realm of data imputation. Nonetheless, their limited capacity to effectively handle time series data represents a notable limitation. Additionally, GPS trajectories, particularly those of buses, exhibit a distinctive characteristic wherein each vehicle is assigned to one or more predetermined routes, adding complexity to the data imputation process.

In response to these challenges, this study proposes a novel hybrid imputation approach, Cluster-GRUI-GAIN, which integrates clustering techniques (e.g. KNN) with the enhanced generative adversarial imputation network, GRUI-GAIN. By combining the strengths of clustering and GAIN, our hybrid approach aims to enhance the accuracy of time series data imputation for GPS trajectories with diverse missing rates and significant gaps. Specifically, the GRUI-GAIN model within our proposed Cluster-GRUI-GAIN framework incorporates GRUI (GRU for Imputation) within the generator. This strategic integration enhances the model's ability to effectively handle missing data within time series, thereby bolstering the accuracy and reliability of imputations. Experimental evaluations on real-world dataset demonstrate that our proposed Cluster-GRUI-GAIN approach outperforms baseline methods in terms of time series imputation accuracy and offers robust and accurate imputations, making it well-suited for practical transportation applications.

Key words: GPS trajectory, data imputation, generative adver- sarial network, clustering, hybrid

[1]*High Performance Computing Laboratory, Faculty of Computer Science and Engineering (HPC Lab), Ho Chi Minh City University of Technology (HCMUT), Vietnam National University Ho Chi Minh City (VNU-HCM), Vietnam*

[2]*TIST Lab, Advanced Institute of Interdisciplinary Science and Technology, Ho Chi Minh City University of Technology (HCMUT), Vietnam National University Ho Chi Minh City (VNU-HCM), Vietnam*

Correspondence

Khang Nguyen Duy, High Performance Computing Laboratory, Faculty of Computer Science and Engineering (HPC Lab), Ho Chi Minh City University of Technology (HCMUT), Vietnam National University Ho Chi Minh City (VNU-HCM), Vietnam

TIST Lab, Advanced Institute of Interdisciplinary Science and Technology, Ho Chi Minh City University of Technology (HCMUT), Vietnam National University Ho Chi Minh City (VNU-HCM), Vietnam

Email: khang.nguyenndk3659@hcmut.edu.vn

INTRODUCTION

With the exponential growth of computing power and the abundance of big data, the Intelligent Transportation Systems (ITS) community now has an unprecedented opportunity to extract valuable insights from the vast amount of data available. GPS trajectory data which is time series data plays a crucial role in numerous applications and research endeavors within transportation systems. Whether it is facilitating route planning for individuals or aiding transportation management and control for researchers and governments, the availability of comprehensive GPS trajectory data is essential [1]. Unfortunately, actual GPS trajectory data obtained from sensors or other sources often suffer from incomplete information due to various factors. Numerous studies have highlighted the issue of missing data in various trajectory and transportation databases. For instance, Qu et al. [2] identified missing data ratios in Beijing typically around 10%, but occasionally reaching as high as 20% to 25% due to various factors. These data gaps pose significant challenges for trajectory analysis and other practical operations.

To address this issue, trajectory data imputation or more generally, time series data imputation emerges as a critical technique aimed at accurately filling in these missing data points. Given the ever-increasing richness of traffic data, trajectory data imputation remains a pressing and highly relevant area of investigation [3].

Existing techniques for handling missing data can be broadly classified into two main categories: statistical methods and deep generative models. Statistical approaches frequently rely on stringent assumptions concerning the nature of missing data patterns. For example, mean/median averaging [4], linear regression [5], MICE [6], and K-nearest neighbors [7] can only

Cite this article : Duy K N, Hai T H L, Tho N T, Anh T D, Thoai N. **GPS trajectory imputation: A hybrid approach combined clustering and GAIN-based algorithm**. *Sci. Tech. Dev. J. – Engineering and Technology* 2024; ():1-11.

History
- Received: 25-9-2023
- Accepted: 26-3-2024
- Published Online:

DOI :

handle data missing at random. Latent variables models with EM algorithm [8] can impute data missing not at random but are restricted to certain parametric models. The deep generative models offer a flexible framework for missing data imputation. For instance, several studies [9–11] develop variants of recurrent neural networks to impute time series. Luo et al. [12] leverage generative adversarial training (GANs) [13] to learn complex missing patterns.

Notably, Yoon et al. [14] introduced the Generative Adversarial Imputation Network (GAIN), a pioneering approach for addressing missing data imputation. This method has significantly propelled the field of data imputation by employing a generator that produces a completed vector based on the available observations, while a discriminator endeavors to discern between the entries in the completed dataset that originated from observations and those that were imputed. Nonetheless, a noteworthy limitation of GAIN lies in its relatively diminished capacity to effectively impute missing data within time series datasets.

Moreover, GPS trajectories, particularly GPS bus trajectories, possess the distinctive characteristic of each vehicle being assigned to one or more predetermined routes. Alabadla et al. [15] highlight the effectiveness of hybrid approaches that combine multiple machine learning methods, resulting in improved imputation performance. Building upon this insight, we propose a hybrid approach in this study that integrates clustering techniques with the enhanced generative adversarial imputation network (GRUI-GAIN). Our aim is to enhance the accuracy of GAIN when dealing with diverse missing rates and significant missing gaps in the GPS trajectory data. By leveraging the strengths of both clustering and GAIN-based, we anticipate achieving more accurate and robust imputations in scenarios where missing data is prevalent. In particular, we make the following technical contributions:

- We propose a hybrid approach called Cluster-GRUI-GAIN to improve the GPS trajectory imputation accuracy of clustering and GAIN under various missing values and large missing gaps by combining these two methods.
- We utilize the GAIN-based model, which incorporates GRUI (GRU for Imputation) within the generator, enhancing its ability to handle missing data in time series, thereby enhancing the imputation quality of GAIN [14] in time series. We refer to this improved model as GRUI-GAIN.
- We evaluate our model on real-world datasets. Experimental results show that our model outperforms the baselines in terms of the accuracy of time series imputation.

RELATED WORKS

A. Generative Adversarial Networks

Neural networks have significant advancements and have been widely employed across various practical applications. Numerous neural network models have been proposed to tackle different problem domains [16,17]. Notably, generative adversarial network (GANs) [13], a framework for constructing generative models approximating the target distribution, has emerged as a powerful approach and achieved state-of-the-art performance in diverse learning tasks [18–20]. GANs are characterized by their discriminator, which plays a pivotal role in discerning the discrepancy between the generated distribution and the target distribution. The GANs algorithm follows an iterative training process, where the discriminator progressively provides a more rigorous critique of the generator's outputs. This interplay between the generator and discriminator leads to the refinement and improvement of the overall model performance. GANs have proven to be highly effective in capturing complex data distributions, enabling the generation of realistic samples, and enhancing the quality of generated outputs in various domains.

B. Deep Generative Imputation Methods

Several imputation methods utilizing GAN frameworks have been introduced in the literature. Luo et al. [12] propose GRUI (GRU for Imputation), which effectively models the temporal information of incomplete time series data. In their GAN model, both the generator and discriminator are based on the GRUI architecture. Building upon this work, Luo et al. [21] present E2GAN, an end-to-end imputation method that offers improvements over the previous two-stage approach in [12]. E2GAN employs an auto-encoder based on GRUI as its generator, aiming to simplify model training difficulties and enhance imputation performance.

Moreover, in the realm of missing value imputation for multivariate time series data, Miao, Xiaoye, et al [22] introduce SSGAN, a novel semi-supervised generative adversarial network model, with a generator, discriminator, and classifier. By incorporating a temporal reminder matrix and a semi-supervised classifier, SSGAN achieves remarkable improvements in imputation and prediction performance when compared to existing methods, as demonstrated through extensive experiments on benchmark time series datasets.

In addition, Liu et al.[23] propose a non-autoregressive model named NAOMI for spatiotemporal sequence imputation. NAOMI comprises a bidirectional encoder and a multiresolution decoder, which work together to effectively handle missing data in spatiotemporal sequences. Adversarial training techniques are further incorporated to enhance the imputation performance of NAOMI.

These advancements in GAN-based imputation methods, such as GRUI, E2GAN, NAOMI, and SSGAN demonstrate the ongoing efforts to address the challenges of incomplete time series and spatiotemporal data imputation, leading to improved imputation performance in diverse domains.

C. Clustering-based Imputation

Clustering is a data partitioning technique that involves grouping a dataset into distinct classes or clusters based on specific criteria, such as a distance metric. The primary objective of clustering is to maximize the similarity among data objects within the same cluster while ensuring significant differences between objects belonging to different clusters.

Clustering finds applications in diverse fields, including data compression, information retrieval, pattern recognition, and bioinformatics. It also holds the potential for imputing missing data sets. In the context of imputation, clustering can be approached in two ways. One approach involves dividing the original dataset into complete and missing subsets. The complete dataset is then clustered to obtain distinct clusters. Subsequently, missing data objects are assigned to the most similar clusters based on a similarity measurement, and the information within the clusters is utilized to fill in the missing values. The other approach involves initializing the original dataset and directly clustering it, potentially redefining the similarity measure.

In recent developments, clustering – based approaches have begun incorporating temporal, spatial, global, and local perspectives. For example, Xiuwen et al.[24] employed a multi-view learning method based on temporal and spatial correlations to impute time series data. The primary objective of clustering techniques is to classify datasets into clusters by minimizing intra-cluster dissimilarity, thereby enabling effective data organization and analysis.

PROBLEM FORMULATION

In the context of GPS trajectory data, as depicted in Table 1, a fundamental format consists of timestamp, latitude, and longitude coordinates. Timestamps provide temporal context, indicating when the location was recorded, while latitude and longitude specify the vehicle's geographic position.

Let X represent the GPS trajectory data which can also be interpreted as time series data in a d-dimensional space and observed over n timestamps T = {t0, t_1, t_{n-1}}, is represented as: $X = \{x_0, x_1, \ldots, x_{n-1}\} \in R^{n\times d}$, where x_i is the i-th observation vector within X, and x_{ij} represents the j-th feature within the observation vector x_i.

In this study, the dimensionality, d, is set to 2, representing the two geographic coordinates (latitude and longitude). We refer to X as the data vector and also define the mask matrix, denoted as M, which serves the purpose of indicating which components of X are missing, and it is defined as follow:

$$m_{ij} = \begin{cases} 0, \; if\ x_{ij}\ is\ not\ observed \\ 1, \; otherwise \end{cases}$$

We define a matrix $\delta \in R^{n\times d}$ that records the time gap from the last observation to the current timestamp,

$$\delta_{ij} = \begin{cases} t_i - t_{i-1}, \; if\ m_{(i-1)j} = 1,\ i > 0 \\ \delta_{ij} + t_i - t_{i-1}, \; if\ m_{(i-1)j} = 0,\ i > 0 \\ 0, \; if\ i = 0 \end{cases}$$

METHOD: IMPUTATION BASED ON GAIN

A. Trajectory Part Clustering

The core idea of the hybrid imputation approach is to use the clustering technique to generate a small representative training dataset, which is applied to imputation in the GRUI-GAIN model. Figure 1 shows the whole framework of the proposed hybrid approach. Firstly, we divide the imputation into coarse and fine imputation. The original dataset is first imputed with the Last Observation Carried Forward (LOCF)[25] method. This step prevents the clustering algorithm from dealing with the missing dataset directly. Subsequently, the dataset X' is clustered using the K-Means clustering algorithm to generate different clustering results $\{X_1', X_2', \ldots, X_n'\}$. Finally, each cluster is finely imputed by using GRUI-GAIN. The structure of GRUI-GAIN model is shown at Figure 3. The new complete dataset Y is obtained by merging the clusters.

Table 1: Sample of GPS Trajectory Data.

Timestamp	Latitude	Longitude
2019-07-01 17:03:53	23.49468633	87.31687190
2019-07-01 17:03:54	23.49459298	87.31687570
2019-07-01 17:03:55	23.49455566	87.31686814
…	…	…

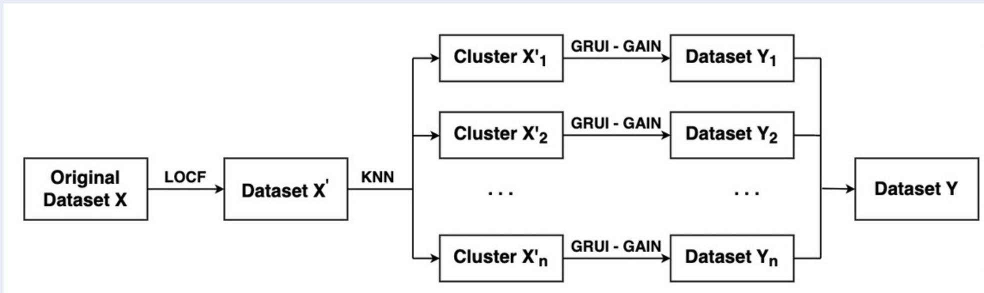

Figure 1: The hybrid approach that combines clustering techniques with GRUI- GAIN model.

B. The Review of GAIN

In the GAIN framework[14], the central components include the generator G and the discriminator D. An additional element, known as the hint H, plays a crucial role.

The generator, G, operates by observing a real data vector, which may contain missing values. It focuses on imputing these missing values while considering the information available in the observed data. Ultimately, it produces a completed vector as its output. The discriminator, D, takes this completed vector as input and is tasked with distinguishing between the components of the vector that were originally observed and those that have been imputed. The discriminator's role is to assess the authenticity of the imputed data. Importantly, the hint, H, plays a vital role in this process. It provides additional information to the discriminator regarding the missingness of the original sample. Essentially, the hint ensures that the generator, G, imputes the missing data in a manner consistent with the true underlying data distribution.

In particular, the output of the generator G and discriminator D in the GAIN framework can be represented as follows:

$$
\begin{aligned}
x_G &= G(X, M, (1-M)\odot Z)\\
m_D &= D(z_R, H)\\
x_R &= M\odot x + (1-M)\odot x_G
\end{aligned}
$$

where Z is a d-dimensional noise and x_R is the reconstructed sample.

The objectives of GAIN are structured as follows:

$$
\min_D \frac{1}{N}\sum_{k=1}^{N} L_D(M, m_D)
$$

$$
\min_G \frac{1}{N}\sum_{k=1}^{N} L_D(M, m_D) + \alpha L_R(X, x_R)
$$

where α is a weight parameter, L_D, L_G are a cross entropy loss and L_R is a reconstruction loss.

C. GRUI Cell for Generator

We have adopted the GRUI (GRU for Imputation), proposed in[12], to process the incomplete time series in the Generator G of GAIN. The GRUI is inspired by the GRUD[9]. Nevertheless, the GRUI is more simple than the GRUD. As Figure 2 illustrates, it follows the structure of GRUD with the removal of the input decay.

The key concept behind GRUI is the incorporation of a time decay vector β, which serves to reduce the memory retention of the GRU cell. The update functions of GRUI are outlined below.

$$
\begin{aligned}
\beta_{t_i} &= 1/e^{(0, W_\beta \delta_{t_i} + b_\beta)},\ h'_{t_{i-1}} = \beta_{t_i}\odot h_{t_{i-1}}\\
\mu_{t_i} &= \sigma\left(W_\mu\left[h'_{t_{i-1}}, x_{t_i}\right] + b_\mu\right)\\
r_{t_i} &= \sigma\left(W_r\left[h'_{t_{i-1}}, x_{t_i}\right] + b_r\right)\\
\widetilde{h}_{t_i} &= tanh\left(W_{\widetilde{h}}\left[r_{t_i}\odot h'_{t_{i-1}}, x_{t_i}\right] + b_{\widetilde{h}}\right)\\
h_{t_i} &= (1-\mu_{t_i})\odot h'_{t_{i-1}} + \mu_{t_i}\odot \widetilde{h}_{t_i}
\end{aligned}
$$

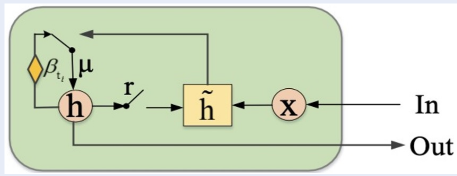

Figure 2: The GRUI cell [12].

where δ is the time lag matrix introduced in the "Problem Formulation" part, and W_β, W_r, W_μ, b_β, b_μ, b_r, $b_{\tilde{h}}$ are training parameters. The formulation of β guarantees that with the increase of time lags σ, the value of β decreases. The smaller the σ, the bigger the β. This formulation also makes sure that $\beta \in (0,1]$. While the primary focus of this paper does not revolve around the GRUI, it is worth mentioning that our research successfully leverages the GRUI within Generator G to effectively process incomplete time series.

The very first input of G is the random noise vector z (random values from a continuous uniform distribution, a common configuration is to use the interval ([-0.01,+0.01]) and every row of the σ of the fake sample is a constant value. For any incomplete time series x, we try to find the best vector z so that the generated sample x_G is most similar to z. Same as GRUI, we add a squared error loss to the loss function of the generator.

D. Discriminator Network Architecture

In contrast to the architecture of GAIN, in our method there is no Hint Generator and, consequently, no Hint Matrix is generated. So, the output of the Discriminator, D, is $m_D = D(x_D)$. Moreover, our Discriminator network adopts a slimmer architecture, consisting of only two layers, in contrast to GAIN's three-layered Discriminator.

Notably, the Discriminator D in the GRUI-GAIN model adopts the hyperbolic tangent activation function (tanh) in its output layers. This choice is motivated by two key reasons: firstly, the optimizer used in neural networks tends to converge faster when inputs are linearly transformed to have zero means, unit variances, and are decorrelated, as discussed in the study by LeCun et al. [26]; secondly, the tanh activation function's derivatives are larger than those of the sigmoid, leading to faster convergence for the optimizer when tanh is employed.

Furthermore, the GRUI-GAIN architecture involves dual Discriminators, one for real data and the other for fake data. This setup allows for a more comprehensive evaluation and comparison, ensuring the effectiveness of our imputation strategy.

EXPERIMENTAL RESULTS

A. Dataset

For our experimental dataset, we utilize the public bus GPS dataset in India [27]. As shown in Figure 4, this dataset was obtained from 6 volunteers who were instructed to travel within the sub-urban city of Durgapur, specifically along the route known as "54 Feet.". During their trips on intra-city buses, the volunteers recorded sensor logs using an Android application installed on commercially available smartphones. In this dataset, each round trip covered a total of 24km, and the total distance covered during this entire period is 720km. Following data processing, we selected 102 bus trajectories from the following date ranges: June 26 to July 06, 2019; September 03 to September 05, 2019; and September 12 to September 23, 2019. Table 2 presents a sample of GPS trajectory data from a bus journey on July 3, 2019, where GPS coordinates were recorded at 15-second intervals. Notably, the GPS coordinates were recorded at regular 15-second intervals.

Besides the spatial diversities like populous zones, and marketplaces, ... they also captured data across different timezones starting from 6 AM to 9 PM, each day. For this, they planned the data collection in different time intervals like – 6 AM to 9 AM – Early Morning, 9 AM to 1 PM – Morning, 1 PM to 5 PM – Afternoon, and 5 PM to 9 PM – Evening. Figure 5 illustrates the general data distribution across various time zones.

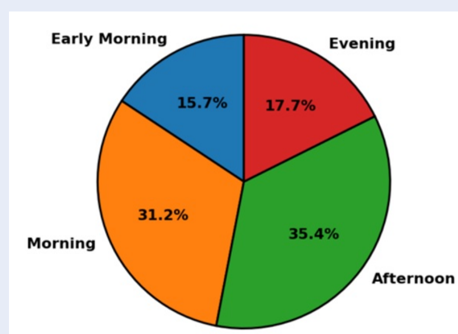

Figure 5: Temporal Distribution Analysis of the Dataset [27].

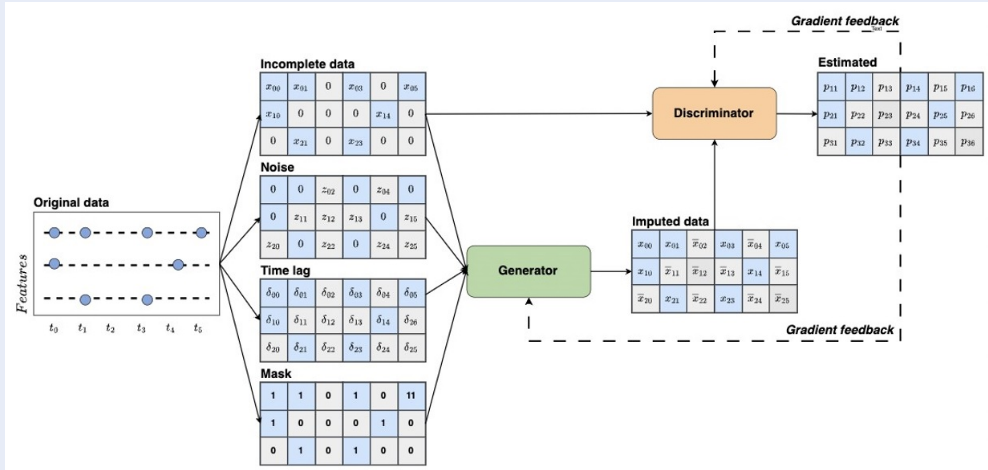

Figure 3: The structure of GRUI-GAIN.

Table 2: GPS Trajectory Data for a Bus Journey on July 3, 2019.

date	timestamp	latitude	longitude
2019-07-03	08:02:20	23.49456677	87.31685814
2019-07-03	08:02:35	23.49458654	87.31684882
2019-07-03	08:02:50	23.49445208	87.31695798
2019-07-03	08:03:05	23.49437571	87.31721719
…	…	…	…
2019-07-03	08:44:20	23.56413802	87.28326889

B. Compared Methods and Performance Indicator

In this study, we compare Cluster-GRUI-GAIN with a range of baseline methods, including:

- **Mean**: Missing values are replaced with the mean value of the available data [4].
- **Last observed value (LOCF)**: Missing values are replaced with the most recent observed value [25].
- **K – nearest neighbor (KNN)**: Missing values are imputed by using the values of the k nearest neighboring samples [7].
- **Multivariate Imputation by Chained Equations (MICE)**: Missing values are imputed using an iterative regression model that estimates the missing values based on the observed values of other variables [6].
- **GAIN**: GAN-based imputation method that utilizes a hint vector to impute missing values [14].
- **E2GAN**: Another GAN-based approach that employs an auto-encoder structure based on GRUI as the generator for imputation [21].

These baseline models serve as comparative approaches for evaluating the performance of the proposed hybrid imputation approach. By contrasting our hybrid approach with these established approaches, we can assess its effectiveness and advantages in handling missing values in the dataset.

Regardless of the specific imputation technique employed, the primary objective is to ensure that the imputed values closely approximate the true values. To evaluate the performance of our imputation approach in our experimental setup, we adopt the Root Mean Square Error (RMSE) as our metric. A smaller RMSE indicates superior results, highlighting the accuracy and effectiveness of the imputation process. By minimizing the RMSE, we aim to achieve the highest level of fidelity between the imputed values and the observed value.

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N} \left(X_{obs} - Y_{imp}\right)^2}{N}}$$

where x_{obs} is the observed value, Y_{imp} is the imputed value.

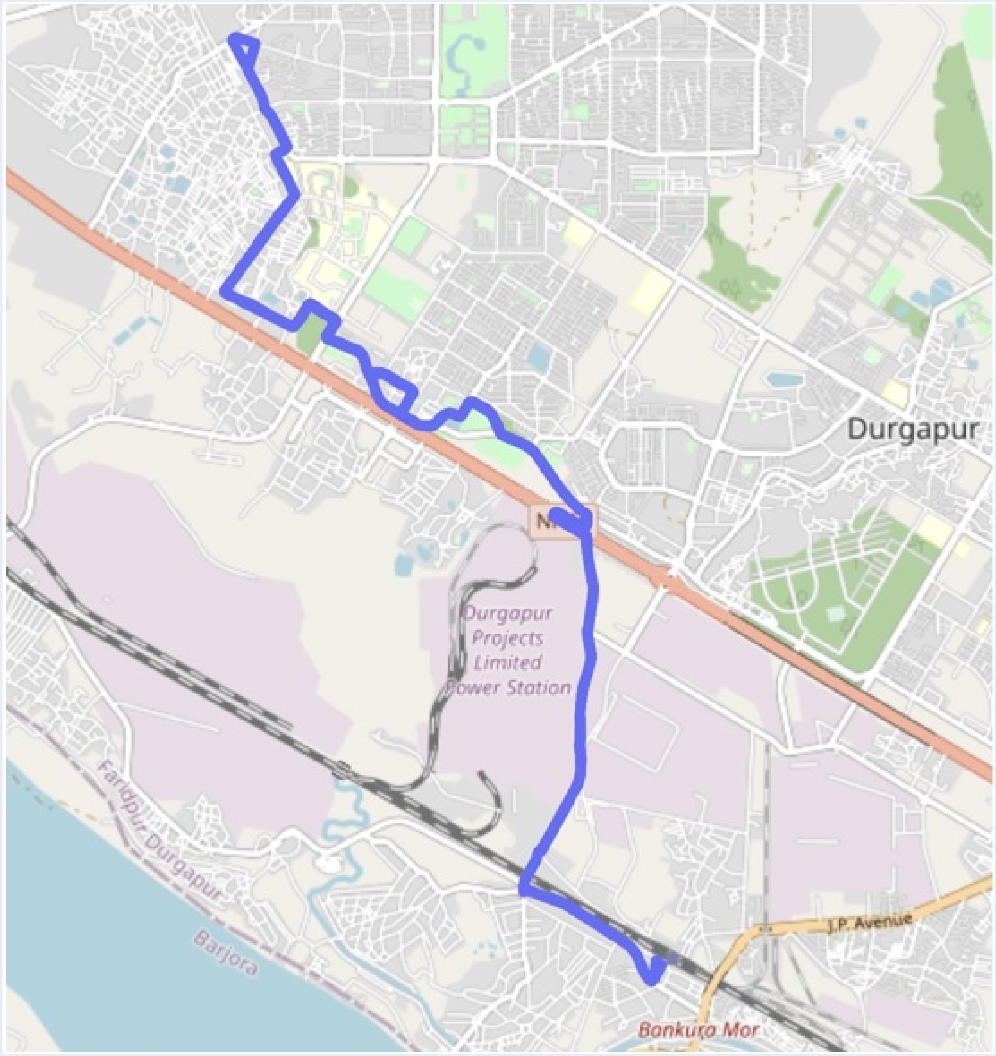

Figure 4: The GPS bus trajectory dataset utilized in this study was collected from the suburban city of Durgapur, located in India.

C. Details of Implementation

In our study using the dataset from [27], we systematically evaluate missing data imputation by randomly dropping between 10% and 80% of trajectory data. We then impute these missing values and assess accuracy using RMSE.

The GRUI cells employ 16 hidden units, a fixed 0.3 dropout rate, and incorporate batch normalization. We standardize input data to have zero mean and unit variance. We allocate 15% of the dataset each for validation and testing. Key parameters for this dataset include an epoch of 10, a batch size of 16, a learning rate of 0.002, and λ set to 2 for regularization.

D. Performance Comparison for GPS Trajectory Data

In Table 3, we present the RMSE results of the proposed hybrid approach and the baseline models. The missing rate, indicating the percentage of dropped values, is listed in the first column, while the subsequent columns display the corresponding RMSE values. Notably, the GAN-based methods consistently exhibit the highest imputation accuracies across all scenarios. The proposed hybrid approach, Cluster-GRUI-GAIN, emerges as one of the top-performing methods, outperforming other approaches in most cases. Additionally, the proposed hybrid approach demonstrates a significant advantage in handling large missing gaps, which will be further explored and discussed in the latter part of this paper.

E. Imputation Accuracy Under Clustering

Figure 6 illustrates the trends of RMSE for Cluster-GRUI-GAIN and the comparison algorithms (GRUI-GAIN without clustering, GAIN, and KNN) when imputing GPS trajectory data with varying missing rates. While both GAIN and GRUI-GAIN are affected by data sparsity, resulting in fluctuating imputation performance as the missing rate increases, Cluster-GRUI-GAIN effectively addresses the data sparsity challenge under high missing rates. This leads to improved robustness and enhanced accuracy for datasets with higher missing rates.

As a result, the hybrid approach is well-suited for handling datasets with higher missing rates or greater sparsity in practical applications.

Additionally, we investigate the impact of the number of clusters on the performance of our proposed approach. Figure 7 illustrates that, across various missing rates, Cluster-GRUI-GAIN consistently achieves better results when the number of clusters is set to 3. This finding holds true for most cases in the dataset, indicating the robustness and effectiveness of our approach in terms of imputation performance. Regardless of the missing rate, selecting K=3 yields favorable outcomes with our hybrid approach.

DISCUSSION

Different Gap Size Analysis

In order to assess the imputation accuracy of the hybrid imputation approach, we examine its performance under different gap sizes. Specifically, we randomly remove 15-minute, 30-minute, and 45-minute of data from random trajectories, creating gappy time series for analysis. As depicted in Figure 8, we observe a deterioration in imputation accuracy as the gap size increases. This decline can be attributed to the diminishing temporal correlation as the gap size expands.

However, the hybrid imputation approach, which leverages the clustering method to generate a representative training dataset, exhibits superior modeling capabilities compared to E2GAN. Consequently, the hybrid imputation approach is more suitable for handling datasets with a higher missing gap in practical applications.

CONCLUSIONS AND FUTURE WORK

In conclusion, this research introduces Cluster-GRUI-GAIN, a novel hybrid imputation approach designed to enhance the accuracy of imputing time series data, particularly GPS trajectory data. By combining clustering techniques with the improved generative adversarial imputation network, GRUI-GAIN, our approach addresses the challenge of missing data in transportation systems. Our extensive experiments on real-world datasets have demonstrated the superiority of Cluster-GRUI-GAIN over baseline methods. It consistently achieves higher imputation accuracy, making it especially well-suited for datasets with higher missing rates and significant gaps. Furthermore, our approach exhibits resilience when faced with data sparsity and outperforms other methods in handling large missing gaps. This research signifies a significant step forward in the field of data imputation for transportation systems, with the potential to impact various practical applications in the real world. Future work will explore broader applications and fine-tune clustering parameters to further optimize the approach.

This study highlights the potential of combining clustering and deep generative models to tackle complex data imputation tasks. In future research endeavors, we aspire to extend the utility of our hybrid imputation approach to diverse domains beyond GPS trajectories. Our objectives include exploring varied clustering methodologies and conducting performance evaluations on an even wider array of real-world datasets. Additionally, we plan to perform comprehensive experimental comparisons, including benchmarking against state-of-the-art methods such as SSGAN[22], in the context of multivariate time series data imputation within the GPS trajectory domain.

ACKNOWLEDGEMENT

This research was conducted within the ”Developing a data streaming platform used in urban and rural areas” project sponsored by TIST Lab.

We gratefully acknowledge the valuable support and resources provided by HPC Lab at HCMUT, VNU-HCM.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTION

Nam Thoai, Nguyen Tran Tho, Trung Dang Anh and Thanh Hoang Le Hai provided guidance and strategic direction and shaping research objectives.

Khang Nguyen Duy made sub- stantial contributions by actively engaging in data collection, model development, experimental work, and constructive discussions.

Figure 6: Imputation accuracy under clustering.

Table 3: The RMSE (the smaller, the better) results of Cluster-GRUI-GAIN and other baseline imputation methods on the GPS bus trajectory dataset.

Missing Rate (%)	Mean [4]	LOCF [25]	KNN [7]	MICE [6]	GAIN [14]	E2GAN [21]	Cluster-GRUI-GAIN
10	0.846	0.366	0.548	0.554	0.374	**0.286**	0.307
20	0.804	0.538	0.610	0.548	0.516	**0.448**	0.466
30	0.991	0.721	0.647	0.691	0.637	**0.572**	0.579
40	0.940	0.676	0.694	0.680	0.652	0.626	**0.624**
50	0.866	0.724	0.736	0.744	0.676	0.609	**0.604**
60	0.892	0.747	0.778	0.758	0.742	0.709	**0.696**
70	0.988	0.858	0.784	0.868	0.762	0.716	**0.712**
80	1.075	0.863	0.857	1.047	0.805	0.748	**0.735**

REFERENCES

1. Wang FY. Parallel control and management for intelligent transportation systems: Concepts, architectures, and applications. IEEE Trans Intell Transp Syst. 2010;11(3):630-638;Available from: https://doi.org/10.1109/TITS.2010.2060218.
2. Qu L, Li L, Zhang Y, Hu J. PPCA-based missing data imputation for traffic flow volume: A systematical approach. IEEE Trans Intell Transp Syst. 2009;10(3):512-522;Available from: https://doi.org/10.1109/TITS.2009.2026312.
3. Li Y, Li Z, Li L. Missing traffic data: comparison of imputation methods. IET Intell Transp Syst. 2014;8(1):51-57;Available from: https://doi.org/10.1049/iet-its.2013.0052.
4. Acuna E, Rodriguez C. The treatment of missing values and its effect on classifier accuracy. In: Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15-18 July 2004. Springer; 2004. pp. 639-647;Available from: https://doi.org/10.1007/978-3-642-17103-1_60.
5. Ansley CF, Kohn R. On the estimation of ARIMA models with missing values. In: Time Series Analysis of Irregularly Observed Data: Proceedings of a Symposium held at Texas A & M University, College Station, Texas February 10-13, 1983. Springer; 1984. pp. 9-37;Available from: https://doi.org/10.1007/978-1-4684-9403-7_2.
6. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. J Stat Softw. 2011;45:1-67;Available from: https://doi.org/10.18637/jss.v045.i03.
7. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. Springer; 2009. vol. 2;Available from: https://doi.org/10.1007/978-0-387-84858-7.
8. Nelwamondo FV, Mohamed S, Marwala T. Missing data: A comparison of neural network and expectation maximization techniques. Curr Sci. 2007;1514-1521;.

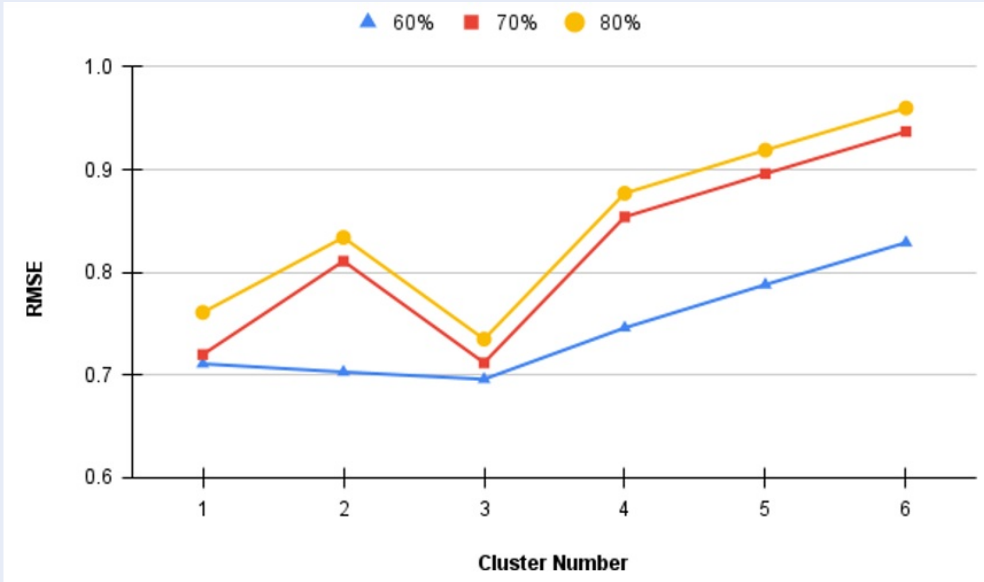

Figure 7: RMSE of Cluster-GRUI-GAIN with various number of clusters.

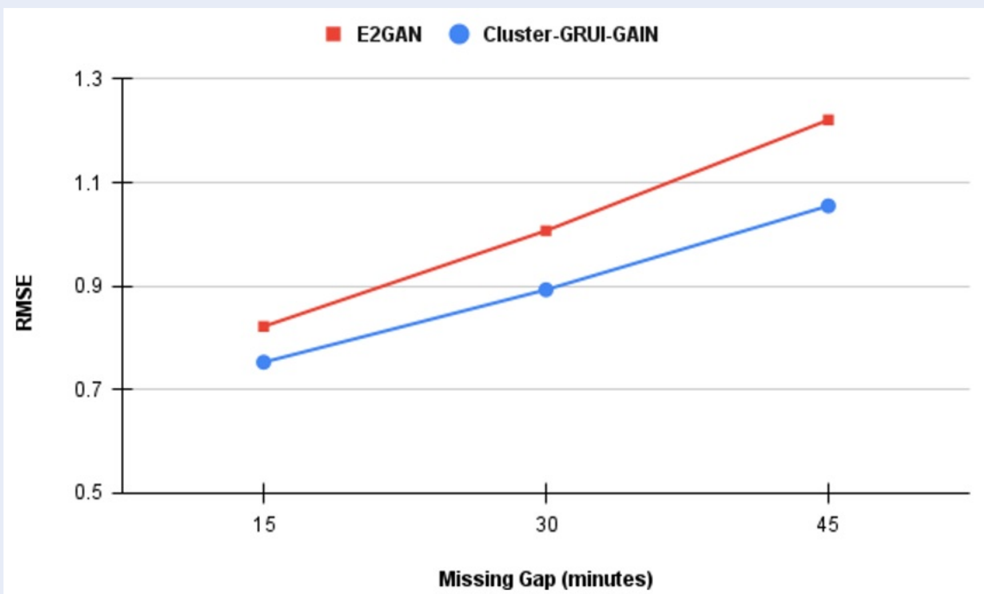

Figure 8: Comparisons of the imputed for different gaps.

9. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep. 2018;8(1):6085;PMID: 29666385. Available from: https://doi.org/10.1038/s41598-018-24271-9.
10. Yoon J, Zame WR, van der Schaar M. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. IEEE Trans Biomed Eng. 2018;66(5):1477-1490;PMID: 30296210. Available from: https://doi.org/10.1109/TBME.2018.2874712.
11. Cao W, Wang D, Li J, Zhou H, Li L, Li Y. BRITS: Bidirectional recurrent imputation for time series. In: Advances in neural information processing systems. 2018;.
12. Luo Y, Cai X, Zhang Y, Xu J, et al. Multivariate time series imputation with generative adversarial networks. In: Advances in neural information processing systems. 2018;.
13. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Advances in neural information processing systems. 2014;27;.
14. Yoon J, Jordon J, Schaar M. GAIN: Missing data imputation using generative adversarial nets. In: International conference on machine learning. PMLR. 2018;5689-5698;.
15. Alabadla M, Sidi F, Ishak I, Ibrahim H, Affendey LS, Ani ZC, Jabar MA, Bukar UA, Devaraj NK, Muda AS, et al. Systematic review of using machine learning in imputing missing values. IEEE Access. 2022;10:44 483-44 502;Available from: https://doi.org/10.1109/ACCESS.2022.3160841.
16. Gondara L, Wang K. MIDA: Multiple imputation using denoising autoencoders. In: Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III 22. Springer; 2018. pp. 260-272;Available from: https://doi.org/10.1007/978-3-319-93040-4_21.
17. Kiranyaz S, Ince T, Iosifidis A, Gabbouj M. Operational neural networks. Neural Comput Appl. 2020;32:6645-6668;Available from: https://doi.org/10.1007/s00521-020-04780-3.
18. Mukherjee S, Asnani H, Lin E, Kannan S. ClusterGAN: Latent space clustering in generative adversarial networks. In: Proceedings of the AAAI conference on artificial intelligence. 2019;33(01):4610-4617;Available from: https://doi.org/10.1609/aaai.v33i01.33014610.
19. Ortac¸ G, Dog˘an Z, Orman Z, S¸ AMLI R. Baby face generation with generative adversarial neural networks: a case study. Acta Infologica. 2020;4(1):1-9;.
20. Xu L, Veeramachaneni K. Synthesizing tabular data using generative adversarial networks. arXiv preprint arXiv:1811.11264. 2018;.
21. Luo Y, Zhang Y, Cai X, Yuan X. E2GAN: End-to-end generative adversarial network for multivariate time series imputation. In: Proceedings of the 28th international joint conference on artificial intelligence. 2019;3094-3100;Available from: https://doi.org/10.24963/ijcai.2019/429.
22. Miao X, Wu Y, Wang J, Gao Y, Mao X, Yin J. Generative semi-supervised learning for multivariate time series imputation. In: Proceedings of the AAAI conference on artificial intelligence. 2021;35(10):8983-8991;Available from: https://doi.org/10.1609/aaai.v35i10.17086.
23. Liu Y, Yu R, Zheng S, Zhan E, Yue Y. Naomi: Non-autoregressive multiresolution sequence imputation. In: Advances in neural information processing systems. 2019;32;.
24. Yi X, Zheng Y, Zhang J, Li T. ST-MVL: Filling missing values in geo-sensory time series data. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence. 2016;.
25. Woolley SB, Cardoni AA, Goethe JW. Last-observation-carried-forward imputation method in clinical efficacy trials: review of 352 antidepressant studies. Pharmacotherapy. 2009;29(12):1408-1416;PMID: 19947800. Available from: https://doi.org/10.1592/phco.29.12.1408.
26. LeCun Y, Bottou L, Orr G, Mu¨ller K. Efficient backprop in: Neural networks: Tricks of the trade, 9-48. Springer. 2012;10:3-540;Available from: https://doi.org/10.1007/978-3-642-35289-8_3.
27. Mandal R, Karmakar P, Chatterjee S, Das Spandan D, Pradhan S, Saha S, Chakraborty S, Nandi S. Exploiting multi-modal contextual sensing for city-bus's stay location characterization: Towards sub-60 seconds accurate arrival time prediction. ACM Trans Internet Things. 2023;4(1):1-24;Available from: https://doi.org/10.1145/3549548.

Tạp chí Phát triển Khoa học và Công nghệ – Engineering and Technology 2024, ():1-1

Open Access Full Text Article

Bổ khuyết lộ trình di chuyển GPS: Phương pháp tiếp cận kết hợp thuật toán phân cụm và giải thuật dựa trên GAIN

Nguyễn Duy Khang[1,2,*], Hoàng Lê Hải Thanh[1,2], Trần Thọ Nguyên[2], Đặng Anh Trung[2], Thoại Nam[1,2]

Use your smartphone to scan this QR code and download this article

[1]*Phòng Thí nghiệm Tính toán Hiệu năng cao, Khoa Khoa học và Kỹ thuật Máy tính (HPC Lab), Trường Đại học Bách Khoa (HCMUT), Đại học Quốc gia Thành phố Hồ Chí Minh (VNU-HCM), Việt Nam*

[2]*TIST Lab, Viện Khoa học và Công nghệ Tiên tiến Liên ngành, Trường Đại học Bách Khoa (HCMUT), Đại học Quốc gia Thành phố Hồ Chí Minh (VNU-HCM), Việt Nam*

Liên hệ

Nguyễn Duy Khang, Phòng Thí nghiệm Tính toán Hiệu năng cao, Khoa Khoa học và Kỹ thuật Máy tính (HPC Lab), Trường Đại học Bách Khoa (HCMUT), Đại học Quốc gia Thành phố Hồ Chí Minh (VNU-HCM), Việt Nam

TIST Lab, Viện Khoa học và Công nghệ Tiên tiến Liên ngành, Trường Đại học Bách Khoa (HCMUT), Đại học Quốc gia Thành phố Hồ Chí Minh (VNU-HCM), Việt Nam

Email: khang.nguyenndk3659@hcmut.edu.vn

Lịch sử

- Ngày nhận: 25-9-2023
- Ngày chấp nhận: 26-3-2024
- Ngày đăng:

DOI :

TÓM TẮT

Với Sự tiến bộ về sức mạnh tính toán và sự phát triển của dữ liệu lớn đã mở ra những cơ hội chưa từng có cho cộng đồng Hệ thống Giao thông Thông minh (ITS) để trích xuất những thông tin quý giá từ dữ liệu quỹ đạo dữ liệu quỹ đạo Hệ thống Định vị Toàn cầu (GPS). Tuy nhiên, thực tế của dữ liệu quỹ đạo GPS trong thế giới thực thường thiếu thông tin đầy đủ do nhiều yếu tố khác nhau (ví dụ: hỏng cảm biến, mất truyền, ...), từ đó đặt ra những thách thức đáng kể cho việc phân tích quỹ đạo và hiệu quả hoạt động trong các hệ thống giao thông. Để giải quyết vấn đề này, các kỹ thuật bổ khuyết dữ liệu chuỗi thời gian đã xuất hiện như những giải pháp quan trọng để điền vào các điểm dữ liệu bị thiếu một cách chính xác. Các phương pháp bổ khuyết hiện có có thể được phân loại thành các phương pháp thống kê và mô hình tạo sinh sâu. Đáng chú ý, trong lĩnh vực của các mô hình tạo sinh sâu, Mạng Bổ Khuyết Đối Nghịch Tạo Sinh Dữ Liệu (GAIN) đã thể hiện tiềm năng trong lĩnh vực bổ khuyết dữ liệu. Tuy nhiên, khả năng hạn chế của chúng trong việc xử lý hiệu quả dữ liệu chuỗi thời gian là một hạn chế đáng chú ý. Ngoài ra, các quỹ đạo GPS, đặc biệt là của các xe buýt, có đặc điểm độc đáo khi mỗi phương tiện được gán vào một hoặc nhiều tuyến đường đã được xác định trước, tạo ra sự phức tạp trong quá trình bổ khuyết dữ liệu.

Để đáp ứng những thách thức này, nghiên cứu này đề xuất một phương pháp bổ khuyết kết hợp mới, Cluster-GRUI-GAIN, kết hợp các kỹ thuật phân cụm (ví dụ: KNN) với mạng bổ khuyết đối nghịch tạo sinh được cải thiện, GRUI-GAIN. Bằng cách kết hợp những ưu điểm của phân cụm và GAIN, phương pháp kết hợp của chúng tôi nhằm mục tiêu nâng cao độ chính xác của việc bổ khuyết dữ liệu chuỗi thời gian cho các quỹ đạo GPS với các tỷ lệ thiếu khác nhau và các khoảng trống đáng kể. Cụ thể, mô hình GRUI-GAIN trong Cluster-GRUI-GAIN mà chúng tôi đề xuất tích hợp GRUI (GRU cho Imputation) vào bộ tạo sinh. Sự tích hợp chiến lược này cải thiện khả năng của mô hình trong việc xử lý dữ liệu thiếu trong chuỗi thời gian, từ đó tăng cường độ chính xác và đáng tin cậy của các giá trị bổ khuyết. Đánh giá thực nghiệm trên bộ dữ liệu thế giới thực cho thấy rằng phương pháp Cluster-GRUI-GAIN mà chúng tôi đề xuất vượt qua các phương pháp cơ sở về độ chính xác của việc bổ khuyết dữ liệu chuỗi thời gian và cung cấp các bổ khuyết mạnh mẽ và chính xác, làm cho nó phù hợp cho các ứng dụng.

Từ khoá: lộ trình GPS, bổ khuyết dữ liệu, mạng đối nghịch tạo sinh, gom cụm, kết hợp

Trích dẫn bài báo này: Khang N D, Thanh H L H, Nguyên T T, Trung D A, Nam T. **Bổ khuyết lộ trình di chuyển GPS: Phương pháp tiếp cận kết hợp thuật toán phân cụm và giải thuật dựa trên GAIN**. *Sci. Tech. Dev. J. - Eng. Tech.* 2024; ():1-1.