231654

OFFICE OF THE PRESIDENT
RECEIVED
Date: JUL 18 2023
Time: 6:20 pm
By:

SLSU OFFICE OF RESEARCH SERVICES

Office of
Research Services

18 July 2023

Dr. DORACIE B. ZOLETA-NANTES
University President

Thru: **Dr. MARISSA C. ESPERAL**
Vice President for Research, Extension, Production,
Development and Innovation

Dear Mesdames,

The CHED-DBM Joint Circular 3, s. 2023 (commonly known as the new instrument for Faculty Re-Classification) requires that Peer Reviewer engagement of faculty members in academic journals receive proper authorization from the President or the concerned Vice President. However, these guidelines were issued towards the end of the coverage period of the 1st Cycle of the Joint Circular (July 1, 2019–July 31, 2023).

As you may be aware, peer review requests from academic journals are normally directly communicated by editors to the peer reviewer and not through the institution where s/he may be affiliated. In consultation with the Institutional Evaluation Committee, I was informed that the CHED provides leeway for additional evidence for Peer Reviewer engagement – that a list of institutionally-recognized peer reviewer engagement would be enough as additional evidence for this cycle.

In this regard, I wish to respectfully seek your **approval in principle** of the participation of faculty members listed in the attached file. Rest assured that the ORS thoroughly screened these reported Peer Reviewer engagement of our faculty members to include only those done with reputable journal publications and book publishers.

We look forward to your usual support on this matter as this will contribute greatly to the career development of our dedicated faculty researchers.

Thank you very much!

Very truly yours,

NICANOR L. GUINTO, PhD
Director, Office of Research Services

Recommending Approval:

MARISSA C. ESPERAL, PhD
Vice President for Research, Extension,
Production, Development and Innovation

APPROVED / DISAPPROVED

Doracie B. Zoleta-Nantes, PhD
University President

JUL 19 2023

SOUTHERN LUZON STATE UNIVERSITY
Office of Research Services

C E R T I F I C A T I O N

This is to certify that the **peer reviewer engagement** of the personnel named below are approved in principle as they have been invited to review journal articles and/or book proposals while being affiliated with the University. For having been directly contacted by Editors of reputable journals and book publishers, their recognized expertise and leadership in their respective areas of research specialization contributed significantly to building the good name of Southern Luzon State University in local and international academic circles.

Name	Academic Rank	College/ Campus	Area of Research Specialization	Journal Name/ Book Publisher that made the request	Coverage/ Readership	Indexed/ Published by:	Tentative Title of the Article/ Book Proposal reviewed	Date when the invitation is received:	Date when the review was sent back to the editor:
AGUDILLA, MARY ANN R.	ASSOCIATE PROFESSOR 4	CAG	BIODIVERSITY, INSECTS, ECOSYSTEM VALUATION	PHILIPPINE JOURNAL OF SCIENCE	International	Scopus	SETTING THE INITIAL CARBON TAX RATE FOR THE CARBON TAX POLICY IN THE PHILIPPINES THROUGH THE SOCIAL COSTS OF CARBON AND WILLINGNESS TO PAY METHODS, AND THE CORRESPONDING BENEFIT-COST ANALYSIS	12/11/2022	1/2/2023
AGUDILLA, MARY ANN R.	ASSOCIATE PROFESSOR 4	CAG	BIODIVERSITY, INSECTS, ECOSYSTEM VALUATION	ACADEMIA-BIOLOGY	International	Academia Publishing	TREE HEIGHT, CANOPY COVER AND LEAF LITTER PRODUCTION OF RHIZOPHORA APICULATA IN BAGANGA, DAVAO, ORIENTAL, PHILIPPINES	1/11/2023	1/27/2023
Alinea, Jess Mark L.	Assistant Professor I	Lucena Campus	TVET, Technical Teacher Education, Curriculum and Instruction	Journal of Technical Education and Training	International	Scopus	The Role of Al-Balqa Applied University in Developing Vocational Education in Jordan	10/26/2021	11/2/2021
Alinea, Jess Mark	Assistant Professor I	Lucena Campus	TVET, Technical Teacher Education,	Journal of Technical	International	Scopus	Training-based Assessment of Employees Performance: A Case Study of Bahir Dar	12/27/2021	1/5/2022

			Communication	Applied Linguistics					
Guinto, Nicanor L.	Associate Professor III	College of Arts and Sciences	Sociolinguistics, Discourse Analysis, Communication	rEFLections	International	Scopus/ King Mongkut's University, Thailand	Filipino Non-Native English-Speaking Teachers and the Bias in Their Own Backyard	07/10/2023	07/19/2023
Maaliw, Renato III R.	Associate Professor II	CEN	Computer Vision, Machine Learning, Data Analytics	Cogent Engineering	International	Scopus, Web of Science, ASEAN Citation Index	Integrating Video Feedback Into Architectural Design Education to Engage Diverse Learning Styles	3/27/2023	4/20/2023
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Machine Learning, Computer Vision, Data Analytics	Healthcare Analytics (Elsevier)	International	Scopus, Web of Science, ASEAN Citation Index	Prediction of Systolic and Diastolic Blood Pressures Using Machine Learning	5/4/2023	5/16/2023
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Computer Vision, Machine Learning, Data Analytics	Engineering (MDPI)	International	Scopus, Web of Science, ASEAN Citation Index	Using ARIMA to Predict the Growth in the Subscriber Data Usage	11/4/2022	11/14/2022
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Computer Vision, Machine Learning, Analytics	Sensors (MDPI)	International	Scopus, Web of Science, ASEAN Citation Index	Missing Traffic Data Imputation with a Linear Model Based on Probabilistic Principal Component Analysis	12/2/2022	12/10/2022
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Computer Vision, Machine Learning, Data Analytics, Computer Engineering	Sensors (MDPI)	International	Scopus, Web of Science, ASEAN Citation Index	Using Machine Learning on V2X Communications Data for VRU's Collisions Predictions	12/23/2022	12/26/2022
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Computer Vision, Machine Learning, Data Analytics	Applied Science (MDPI)	International	Scopus, Web of Science, ASEAN Citation Index	Performance Predictions of Sci-Fi Films via Machine Learning	1/31/2023	2/5/2023
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Computer Vision, Machine Learning, Data Analytics, Computer Engineering	Sustainability (MDPI)	International	Scopus, Web of Science, ASEAN Citation Index	Thermal Images Classifications of Solid Wastes with Deep Convolutional Neural Networks	2/15/2023	2/25/2023
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Computer Vision, Machine Learning, Data Analytics, Computer Engineering	Sustainability (MDPI)	International	Scopus, Web of Science, ASEAN Citation Index	Static Evaluation of a Midimew Connected Torus Network for Next Generation Supercomputers	3/2/2023	3/13/2023
Maaliw,	Associate	College of	Computer Vision,	Journal of	International	Scopus, Web of	Machine-Learning-Based Composition	3/23/2023	4/1/2023

Renato III R.	Professor II	Engineering	Machine Learning, Data Analytics, Computer Engineering	Nuclear Engineering (MDPI)		Science, ASEAN Citation Index	Analysis of the Stability of V–Cr–Ti Alloys		
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Computer Vision, Machine Learning, Data Analytics, Computer Engineering	Mathematics (MDPI)	International	Scopus, Web of Science, ASEAN Citation Index	A Federated Personal Mobility Service in Autonomous Transportation	5/19/2023	5/29/2023
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Computer Vision, Machine Learning, Data Analytics, Computer Engineering	IJERPH (MDPI)	International	Scopus, Web of Science,	Machine Learning in Predicting Severe Acute Respiratory Infection	6/6/2023	6/11/2023
Maaliw, Renato III R.	Associate Professor II	College of Engineering	Computer Vision, Machine Learning, Data Analytics, Computer Engineering	Journal of Theoretical and Applied Electronic Commerce Research	International	Scopus, Web of Science, ASEAN Citation Index	Unveiling the Power of ARIMA, Support Vector Machine and Random Forest Regressors for the Future of Dutch Employment Market	6/14/2023	6/23/2023
Mabunga, Zoren P.	Instructor 1	College of Engineering	Artificial Intelligence, Electronics and Communication Engineering, Internet of Things	2022 IEEE 18th International Colloquium on Signal Processing & Applications (CSPA 2022)	International	Scopus	Semi Autonomous Detection of Bite Points for a Surgical Needle	2/24/2022	3/7/2022
Mabunga, Zoren P.	Instructor 1	Engineering	Artificial Intelligence, Electronics and Communication Engineering, Internet of Things	IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC-2021)	International	Scopus	1. A Survey of Vulnerability Management Using Machine Learning Techniques, 2. An Adaptive Algorithm based on Interference Aware Cooperative Energy Efficiency Maximization for 5G UltraDense Networks, 3. GRAMIN GENIE-A SMART KIOSK, 4. An Automated Deep Learning Model for Detecting Sarcastic Comments,	7/2/2021	8/12/2021

YAO, CLAIRE ANN M.	ASSISTANT PROFESSOR IV	CABHA MAIN	BUSINESS ENTREPRENEURSHIP, PRODUCT DEVELOPMENT, TOURISM, LEISURE, AND HOSPITALITY	PATHWAY TO REFEREED JOURNAL PUBLICATION IN THE FIELD OF BUSINESS	Local	INSTITUTIONAL	PROBLEMS ENCOUNTERED BY MSME'S IN TAGUIG CITY AND THE ACTION TO COUNTER THE POSSIBLE EFFECTS OF ASEAN INTEGRATION: A SITUATION ANALYSIS	3/24/2020	4/4/2020
YAO, CLAIRE ANN M.	ASSISTANT PROFESSOR IV	CABHA MAIN	BUSINESS ENTREPRENEURSHIP, PRODUCT DEVELOPMENT, TOURISM, LEISURE, AND HOSPITALITY	PATHWAYS TO REFEREED JOURNAL IN THE FIELD OF BUSINESS	Local	INSTITUTIONAL	MANYAMAN MANGAN QUENI (DELICIOUS TO EAT HERE):SUCCESS FACTORS OF SELECTED RESTAURANT ENTREPRENEURS IN PAMPANGA	4/16/2020	4/21/2020

Issued this 19th day of July 2023 at Southern Luzon State University, Lucban, Quezon.

NICANOR L. GUINTO, Ph.D.
Director, Office of Research Services

MARISSA C. ESPERAL, Ph.D.
Vice President, REPDI

DORACIE B. ZOLETA-NANTES, Ph.D.
University President

RM

Renato Racelis Maaliw

(Maaliw III, Renato R. R.)

Southern Luzon State University

Web of Science ResearcherID: EXW-3524-2022

Published names	Maaliw, Renato R., III Maaliw, Renato R. Maaliw, Renato Racelis, III Maaliw Iii, R. R. Maaliw Iii, Renato R. R.
Published Organizations	Southern Luzon State University, Southern Luzon Univ
Subject Categories	Computer Science; Telecommunications; Engineering; Environmental Sciences & Ecology; Materials Science

Documents | Peer Review

Verified peer reviews

4	Sensors
3	Sustainability
2	Applied Sciences
2	Eng
2	Journal of Theoretical and Applied Electronic Commerce Research
1	Cogent Engineering
1	International Journal of Environmental Research and Public Health
1	International Journal of Retail and Distribution Management
1	Journal of Nuclear Engineering
1	Mathematics

Metrics

Open dashboard

Profile summary

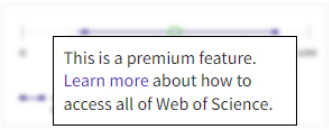
20	Total documents
15	Web of Science Core Collection publications
0	Preprints
18	Verified peer reviews
0	Verified editor records

Web of Science Core Collection metrics

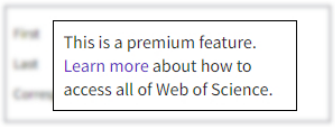
3 H-Index	**15** Publications in Web of Science
17 Sum of Times Cited	**10** Citing Articles
0 Sum of Times Cited by Patents	**0** Citing Patents

View citation report

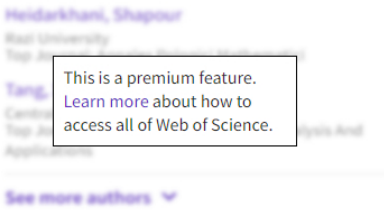
Author Impact Beamplot Summary



Author Position



You might be interested in...



See more authors

Co-authors

6 of many

[Sensors] Manuscript ID: sensors-2093775 - Review Request

External | Research Reviews (Journals) ×

Sensors Editorial Office <sensors@mdpi.com> Fri, Dec 2, 11:08 AM (7 days ago)
to me, Sensors, Marcus

Dear Dr. Maaliw,

We have received the following paper, submitted to Sensors (https://www.mdpi.com/journal/sensors/).

Type of manuscript: Article
Title: Missing Traffic Data Imputation with A Linear Model Based on Probabilistic Principal Component Analysis

We kindly invite you to review this paper and evaluate its suitability for publication in Sensors. The article abstract is available at the end of this message.

If you choose to accept this invitation, we would appreciate receiving your comments within 1 week. Please let us know if you are likely to need more time to complete your review.

Please click on the link below to let us know if you will be able to provide a review and access the full manuscript and review report form.

https://susy.mdpi.com/user/review/review/33661884/AaKnFw6e

In recognition of the contribution of reviewers, for thorough and timely review reports we provide discount vouchers for Article Processing Charges (APCs) applicable for manuscripts accepted for publication after peer review in any MDPI journal. Advice for completing your review can be found at: https://www.mdpi.com/reviewers.

Please disclose any potential conflicts of interest you might have concerning the manuscript's contents or the authors.

If you are not able to review this manuscript, we kindly ask you to decline by clicking on the above link such that we can continue processing this submission. We would also appreciate any feedback you can provide at this time (i.e., your general impression regarding the quality of this manuscript) and any suggestions for alternative expert reviewers.

Sensors is an open access journal of MDPI. Thank you very much for your consideration and we look forward to hearing from you.

Kind regards,
Mr. Marcus Yang
Assistant Editor
E-Mail: marcus.yang@mdpi.com

Sensors
(http://www.mdpi.com/journal/sensors/)

Sensors Impact Factor 2021: 3.847
Sensors CiteScore 2021: 6.4

New section | Sensors Development - Open for Submission
https://www.mdpi.com/journal/sensors/sections/sensors_development

Special Issue | "Women's Special Issue Series: Sensors" -Open for Submission https://www.mdpi.com/journal/sensors/special_issues/WoS

Special Issue | "Sensors Young Investigators' Contributions Collection" -Open for Submission
https://www.mdpi.com/journal/sensors/special_issues/young_investigators_contributions_collection

Editor's Choice Articles - Call for Reading
https://www.mdpi.com/journal/sensors/editors_choice

Sensors News
https://www.mdpi.com/journal/sensors/announcements

Guest Editor Recruitment
https://www.mdpi.com/journalproposal/sendproposalspecialissue/sensors
Topical Advisory Panel - Open for Application
https://www.mdpi.com/journal/sensors/topical_advisory_panel_application

Twitter: https://twitter.com/sensors_mdpi
LinkedIn: www.linkedin.com/company/sensors-mdpi/
Facebook: https://www.facebook.com/SensorsMDPI/

MDPI Branch Office, Beijing
Sensors Editorial Office
E-mail: sensors@mdpi.com

E-mail: sensors@mdpi.com
http://www.mdpi.com/journal/sensors

MDPI
St. Alban-Anlage 66, 4052 Basel
Switzerland
http://www.mdpi.com/

Manuscript details:
Journal: Sensors
Manuscript ID: sensors-2093775
Type of manuscript: Article
Title: Missing Traffic Data Imputation with A Linear Model Based on Probabilistic Principal Component Analysis
Authors: Liping Huang *, Zhenghuan Li, Ruikang Luo, Rong Su *
Submitted to section: Vehicular Sensing,
https://www.mdpi.com/journal/sensors/sections/vehicular

Abstract: Even with the ubiquitous sensing data in intelligent transportation systems, such as the mobile sensing of vehicle trajectories, traffic estimation is still faced with the data missing problem due to the detector faults or limited number of probe vehicles as mobile sensors. Such data missing issue poses an obstacle for many further explorations, e.g., the link-based traffic status modeling. Although many studies have focused on tackling this kind of problem, existing studies mainly focus on the situation that data is missing at random and ignore the distinction between links of missing data. In the practical scenario, traffic speed data is always missing not at random (MNAR). The distinction for recovering missing data on different links has not been studied yet. In this paper, we propose a general linear model based on probabilistic principal component analysis (PPCA) for solving MNAR traffic speed data imputation. Further we propose a metric, i.e., Pearson score (p-score), for distinguishing links and investigate how the model performs on links with different p-score values. Experimental results show that the new model outperforms the typically used PPCA model, and missing data on links with higher p-score values can be better recovered.

Keywords: Missing Data; Urban Traffic Sensing; Probabilistic; Principal Component Analysis

[Sensors] Manuscript ID: sensors-2093775 - Review Request Accepted

External | Research Reviews (Journals) ×

sensors@mdpi.com
to me, Marcus

Sat, Dec 3, 3:49 PM (6 days ago)

Dear Dr. Maaliw,

Thank you very much for agreeing to review this manuscript:

Manuscript ID: sensors-2093775
Type of manuscript: Article
Title: Missing Traffic Data Imputation with A Linear Model Based on Probabilistic Principal Component Analysis
Authors: Liping Huang *, Zhenghuan Li, Ruikang Luo, Rong Su *
Submitted to section: Vehicular Sensing,
https://www.mdpi.com/journal/sensors/sections/vehicular

The review report form can be found here:
https://susy.mdpi.com/user/review/review/33661884/AaKnFw6e

The review report due date is: 10 December 2022

To ensure your anonymity throughout the peer review process, please do not include any identifying information in your review report either in the comments or in the metadata of any files that you upload. Please check the Guidelines for Reviewers: https://www.mdpi.com/reviewers

We look forward to receiving your valuable comments.

Kind regards,
Mr. Marcus Yang
Assistant Editor
E-Mail: marcus.yang@mdpi.com

Sensors
(http://www.mdpi.com/journal/sensors/)

Sensors Impact Factor 2021: 3.847
Sensors CiteScore 2021: 6.4

Special Issue "Sensors in 2023" - New Year Special Issue Series
A collection of high-quality reviews; Submissions Deadline: 31 March 2023
https://www.mdpi.com/journal/sensors/special_issues/552DDDN525

New section | Sensors Development - Open for Submission
https://www.mdpi.com/journal/sensors/sections/sensors_development
---
Editor's Choice Articles - Call for Reading
https://www.mdpi.com/journal/sensors/editors_choice

Sensors News
https://www.mdpi.com/journal/sensors/announcements

Guest Editor Recruitment
https://www.mdpi.com/journalproposal/sendproposalspecialissue/sensors
Topical Advisory Panel - Open for Application
https://www.mdpi.com/journal/sensors/topical_advisory_panel_application
---
Twitter: https://twitter.com/sensors_mdpi
LinkedIn: www.linkedin.com/company/sensors-mdpi/
Facebook: https://www.facebook.com/SensorsMDPI/

MDPI Branch Office, Beijing
Sensors Editorial Office
E-mail: sensors@mdpi.com
http://www.mdpi.com/journal/sensors

MDPI
St. Alban-Anlage 66, 4052 Basel
Switzerland
http://www.mdpi.com/

*** This is an automatically generated email ***

[Sensors] Manuscript ID: sensors-2093775 - Acknowledgement - Review Received

External Inbox ×

sensors@mdpi.com 10:16 AM (2 minutes ago)
to me, Sensors, Marcus

Dear Dr. Maaliw,

Thank you for submitting your review of the following manuscript:

Manuscript ID: sensors-2093775
Title: Missing Traffic Data Imputation with A Linear Model Based on Probabilistic Principal Component Analysis
Authors: Liping Huang *, Zhenghuan Li, Ruikang Luo, Rong Su *

Our Editorial Office and Academic Editors will contact you if they have any questions about your review report. We ask that you remain available, as far as possible, during the peer-review process in case of follow-up questions. To help us improve our services, we kindly ask you to fill in our online survey on the peer-review process at
https://www.surveymonkey.com/r/reviewerfeedbackmdpi

We encourage you to register an account on our submission system and bind your ORCID account (https://susy.mdpi.com/user/edit). You are able to deposit the review activity to your ORCID account manually via the below link:
https://susy.mdpi.com/user/reviewer/status/finished

We also invite you to contribute to Encyclopedia (https://encyclopedia.pub), a scholarly platform providing accurate information about the latest research results. You can adapt parts of your paper to provide valuable reference information for others in the field.

Kind regards,
Sensors Editorial Office
Postfach, CH-4020 Basel, Switzerland
Office: St. Alban-Anlage 66, CH-4052 Basel
Tel. +41 61 683 77 34 (office)
E-mail: sensors@mdpi.com
https://www.mdpi.com/journal/sensors/

*** This is an automatically generated email ***

Review Report Form

Journal	Sensors (ISSN 1424-8220)
Manuscript ID	sensors-2093775
Type	Article
Title	Missing Traffic Data Imputation with A Linear Model Based on Probabilistic Principal Component Analysis
Authors	Liping Huang * , Zhenghuan Li , Ruikang Luo , Rong Su *
Section	Vehicular Sensing
Abstract	Even with the ubiquitous sensing data in intelligent transportation systems, such as the mobile sensing of vehicle trajectories, traffic estimation is still faced with the data missing problem due to the detector faults or limited number of probe vehicles as mobile sensors. Such data missing issue poses an obstacle for many further explorations, e.g., the link-based traffic status modeling. Although many studies have focused on tackling this kind of problem, existing studies mainly focus on the situation that data is missing at random and ignore the distinction between links of missing data. In the practical scenario, traffic speed data is always missing not at random (MNAR). The distinction for recovering missing data on different links has not been studied yet. In this paper, we propose a general linear model based on probabilistic principal component analysis (PPCA) for solving MNAR traffic speed data imputation. Further we propose a metric, i.e., Pearson score (p-score), for distinguishing links and investigate how the model performs on links with different p-score values. Experimental results show that the new model outperforms the typically used PPCA model, and missing data on links with higher p-score values can be better recovered.

Thank you for contributing to the review process, your comments have been successfully submitted.

- see your review history
- download a letter confirming your review activity

Review Report Form

Reviewer's Information (will not be revealed to authors)

Name	Dr. Renato Racelis Maaliw
Email	rmaaliw@slsu.edu.ph
Website	https://www.researchgate.net/profile/Renato-Maaliw-Iii
Affiliation	Southern Luzon State University
Research Keywords	big data; data mining; machine learning; Analytics; Computer Vision, traffic data

Report 1 Hide Report and Author Response [-]

	High	Average	Low	No Answer
Originality / Novelty	( )	(x)	( )	( )
Significance of Content	( )	(x)	( )	( )
Quality of Presentation	( )	(x)	( )	( )
Scientific Soundness	( )	(x)	( )	( )
Interest to the readers	( )	(x)	( )	( )
Overall Merit	( )	(x)	( )	( )

Overall Recommendation

( ) Accept in present form
( ) Accept after minor revision (corrections to minor methodological errors and text editing)
(x) Reconsider after major revision (control missing in some experiments)
( ) Reject (article has serious flaws, additional experiments needed, research not conducted correctly)

English language and style

( ) English very difficult to understand/incomprehensible
( ) Extensive editing of English language and style required
( ) Moderate English changes required
(x) English language and style are fine/minor spell check required
( ) I don't feel qualified to judge about the English language and style

	Yes	Can be improved	Must be improved	Not applicable
Does the introduction provide sufficient background and include all relevant references?	( )	(x)	( )	( )
Are all the cited references relevant to the research?	(x)	( )	( )	( )
Is the research design appropriate?	( )	(x)	( )	( )
Are the methods adequately described?	( )	( )	(x)	( )
Are the results clearly presented?	( )	(x)	( )	( )
Are the conclusions supported by the results?	( )	( )	(x)	( )

Comments and Suggestions for Authors

This is a good work, however, I will raise the following issues:

1. The model is too simple (Linear), wherein realistically most data are not in the real world. Consider using non-linear models for data imputation techniques.

2. Consider other evaluation metrics other than RMSE, MAE, SMAPE or R-squared.

3. How do you account for different road links? Does this study can generalized for all of the scenarios?

4. I have to recommend for the revision of the methodology to not used a general linear method.

Less...

Yes No

Do you have any potential conflict of interest with regards to this paper?	( )	(x)
Did you detect plagiarism?	( )	(x)
Did you detect inappropriate self-citations by authors?	( )	(x)
Do you have any other ethical concerns about this study?	( )	(x)

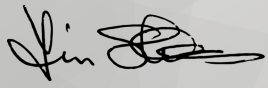
REVIEW CONFIRMATION CERTIFICATE

We are pleased to confirm that

Renato Racelis Maaliw

has reviewed 2 papers for the following MDPI journals in 2022:

Sensors, Eng



Dr. Shu-Kun Lin, Publisher and President
Basel, 9 December 2022

MDPI is a publisher of open access, international, academic journals. We rely on active researchers, highly qualified in their field to provide review reports and support the editorial process. The criteria for selection of reviewers include: holding a doctoral degree or having an equivalent amount of research experience; a national or international reputation in the relevant field; and having made a significant contribution to the field, evidenced by peer-reviewed publications.

sensors

MDPI

Type of the Paper (Article, Review, Communication, etc.)

Missing Traffic Data Imputation with A Linear Model Based on Probabilistic Principal Component Analysis

Liping Huang [1], Zhenghuan Li [1] , Ruikang Luo [1] and Rong Su [1,*]

[1] School of Electrical and Electronic Engineering, Nanyang Technological University
* Correspondence: rsu@ntu.edu.sg;

Abstract: Even with the ubiquitous sensing data in intelligent transportation systems, such as the mobile sensing of vehicle trajectories, traffic estimation is still faced with the data missing problem due to the detector faults or limited number of probe vehicles as mobile sensors. Such data missing issue poses an obstacle for many further explorations, e.g., the link-based traffic status modeling. Although many studies have focused on tackling this kind of problem, existing studies mainly focus on the situation that data is missing at random and ignore the distinction between links of missing data. In the practical scenario, traffic speed data is always missing not at random (MNAR). The distinction for recovering missing data on different links has not been studied yet. In this paper, we propose a general linear model based on probabilistic principal component analysis (PPCA) for solving MNAR traffic speed data imputation. Further we propose a metric, i.e., Pearson score (p-score), for distinguishing links and investigate how the model performs on links with different p-score values. Experimental results show that the new model outperforms the typically used PPCA model, and missing data on links with higher p-score values can be better recovered.

Keywords: Missing Data; Urban Traffic Sensing; Probabilistic; Principal Component Analysis

1. Introduction

Traffic data generated by loop detectors or floating cars in urban road networks serve as the foundation for various data-driven applications in intelligent transportation systems, including traffic forecasting and traffic control [1-3]. However, even with ubiquitous sensing data, missing data problem is almost inevitable due to either detector faults or a limited number of probe vehicles operating as mobile sensors in road networks, which means not each road in the network is covered by a detector or traveled by a probe vehicle in each minute [4-6]. Such issue of missing traffic data poses obstacles for many further data-driven explorations in both academic and industrial fields, e.g., the link-based traffic status modeling, and network-wise traffic dynamics capturing [7-8]. Hence accurate and reliable imputation is a basic need for such kind of incomplete data for the downstream explorations.

Many efforts have been done for estimating the missing traffic data on multiple traffic datasets, resulting the generative probabilistic model [9], the matrix decomposition and tensor factorization models [10-12], the autoencoder model [13], the fusion models [14]. Some basic mathematical models are also adopted, including the autoregressive integrated moving average (ARIMA) model, the Bayesian networks (BNs) method, the Markov chain Monte Carlo (MCMC) method, and the K-nearest neighbors (KNN) model, which are all tested in [15] for traffic missing data imputation.

The studies in [16] have validated that the matrix decomposition-based method is not capable for recovering missing data when the missing percentage large. The tensor models are based on the global structure capturing, and it is faced with challenging to permutation in the spatial and temporal dimension [17]. The probabilistic principal

Citation: To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Received: date
Accepted: date
Published: date

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

component analysis (PPCA) model [18], also plays a major role in missing data completion, due to its generative feature [19]. Observations are assumed to be generated from a low dimensional space, with which the missing data can be recovered by optimizing the generative parameters using the observations [20].

Although many studies have focused on tackling this kind of problem, existing studies focus on the situation that the data is missing at random. Specifically, missing data can be classified into missing at random, missing completely at random, and missing not at random (MNAR) [16]. MNAR always exist in the practical scenarios, and it is more challenging to estimate the missing values, which is the target of this paper.

The studies in [15] demonstrate that the PPCA model yields best performance among ARIMA, BNs, MCMC, and the KNN models, and in the research in [21], it has been certified that the PPCA model outperforms the basic tensor decomposition method. Hence in this study, we set the PPCA model as a basis and further improve the PPCA model for tackling the MNAR traffic data. Additionally, the missing data on different links or sensors may be of different levels of challenges for data completion. Hence, there is also a need to distinguish different scenarios that missing data is on different links or sensors. Instead of the centrality of a sensor in the network, we utilize the time series correlations to define a metric for distinguishing the role of a link in the traffic network. Such a metric is adaptive to the scenario that sensors or links are anonymous. Contributions of this work are summarized as below:

- We design a metric, p-score, to denote the relative importance of links in terms of time series observations, which is used to distinguish the links with missing values.
- We propose a linear model for the MNAR traffic data imputation, which is based on the probabilistic principal component analysis.
- We conduct experiments on a real-world traffic dataset using the model and the proposed metric. Experimental results show missing data on links with higher p-score values can be better recovered. Moreover, testing on the real-world dataset, the results of the proposed model on links with the lowest p-score value also outperforms the typically used PPCA model.

The remainder of the paper is structured as follows. Section 2 presents the problem statement of the missing traffic data imputation. Section 3 the details of the proposed model. Section 4 shows the outcome of the experimental evaluation results, Section 5 presents a short discussion of the potential application scenarios of the proposed method, and finally, Section 6 gives the conclusions of this paper and the directions for future studies.

2. Problem Statement

Let $\boldsymbol{Y} \in \mathbb{R}^{n \times p}$ be a traffic data organization matrix with each element $\boldsymbol{Y}_{ij}$ denoting the i^{th} observation of a link j.

We assume that the traffic data is missing and links with missing values are organized as

$$\boldsymbol{Y}_{\cdot m_1}, \boldsymbol{Y}_{\cdot m_2}, \dots, \boldsymbol{Y}_{\cdot m_d}$$

which is indexed by $\mathcal{M} := \{m_1, m_2, \dots, m_d\} \subset \{1, \dots, p\}$ with $d < p$ are supposed to have missing values. Here $\mathcal{M}$ is the link set that has missing values. Other values in $\boldsymbol{Y}$ are observed.

We label the missing status of $\mathbf{Y}_{ij}$ with another variable, written as

$$\boldsymbol{\Omega}_{ij} = \begin{cases} 0, & \boldsymbol{Y}_{ij} \; is \; missing \\ 1, & otherwise \end{cases}$$

Traffic sensing in urban road networks is faced with the missing data, or data sparsity problem. We construct the traffic matrix $\boldsymbol{Y}$ with all missing values in columns $\mathcal{M}$. The

missing data imputation problem is to estimate these missing values, i.e., $\widehat{\boldsymbol{Y}}_{im}, m \in \mathcal{M}$, where $\boldsymbol{\Omega}_{im} = 0$.

3. Methodology

3.1. PPCA

Assuming that the target variable is organized as a matrix $\boldsymbol{Y}$, and it can be drawn from $\boldsymbol{X}$ of a low rank by linear combination, written as

$$\boldsymbol{Y} = \mathbf{1}\boldsymbol{\alpha} + \boldsymbol{X}\boldsymbol{A} + \boldsymbol{\epsilon}$$

Here, $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$, where n is the sample number and p is the number of variables in the determination system. Specifically, in our link-based missing data imputation problem, p is the total number of links.

$\boldsymbol{X} = (\boldsymbol{X}_{1\cdot}|\cdots|\boldsymbol{X}_{n\cdot})^T$ is the latent variable. $\boldsymbol{X} \in \mathbb{R}^{n\times r}$, and the row is drawn from a Gaussian distribution with zero mean, i.e., $\boldsymbol{X}_{i\cdot} \sim \mathcal{N}(\mathbf{0}_r, \mathbf{I}d_{r\times r})$. Here $r < \min\{n, p\}$, indicating a lower dimension. $\boldsymbol{A}$ is the loading matrix of rank $r, \boldsymbol{A} \in \mathbb{R}^{r\times p}$. $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_{1\cdot}|\cdots|\boldsymbol{\epsilon}_{n\cdot})^T$ is a model error, and each row $\boldsymbol{\epsilon}_{i\cdot} \sim \mathcal{N}(\mathbf{0}_p, \sigma^2\mathbf{I}d_{p\times p}) \in \mathbb{R}^p$, which also has a zero mean. $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_p) \in \mathbb{R}^p$. Given the linear expression above, the mean value of each sample of $\boldsymbol{Y}$ is $\boldsymbol{\alpha}$.

3.2. Missing Variables Differentiation based on Time Series

Assume that we have missing data on two different variables $\boldsymbol{Y}_{\cdot j}, \boldsymbol{Y}_{\cdot k}$ with the same percent, the imputation accuracy can be different due to the variable's role in the whole variable set. In the traffic missing data imputation problem, two links $\boldsymbol{Y}_{\cdot j}, \boldsymbol{Y}_{\cdot k}$ may have different correlations to other links. In this section, we prose a metric to differentiate the role of each link.

The observation of each link is also a time series. We first adopt the Pearson correlation coefficient to estimate the correlation between each pair of time series, which is calculated as

$$\rho\left(\boldsymbol{Y}_{\cdot j}, \boldsymbol{Y}_{\cdot k}\right) = \frac{Cov(\boldsymbol{Y}_{\cdot j}, \boldsymbol{Y}_{\cdot k})}{\sigma_{\boldsymbol{Y}_{\cdot j}}\sigma_{\boldsymbol{Y}_{\cdot k}}}$$

By calculate the Pearson correlation among each pair of variables, we can obtain a correlation matrix, written as

$$P = \begin{bmatrix} 1 & \cdots & \rho(\boldsymbol{Y}_{\cdot 1}, \boldsymbol{Y}_{\cdot n}) \\ \vdots & \ddots & \vdots \\ \rho(\boldsymbol{Y}_{\cdot n}, \boldsymbol{Y}_{\cdot 1}) & \cdots & 1 \end{bmatrix}$$

We define a Pearson score (p-score) for each variable to differentiate the variables in $\mathcal{M}$, which is calculated as

$$P_{score}(\boldsymbol{Y}_{\cdot j}) = \sum_{k\in\{1,\ldots,n\}} P_{jk}$$

The variable $\boldsymbol{Y}_{\cdot j}$ that obtains a higher p-score value than $\boldsymbol{Y}_{\cdot k}$ denotes it has higher correlation to other links. Such a metric can differentiate the variables in terms of the time series observations. When $P_{score}\left(\boldsymbol{Y}_{\cdot j}\right) > P_{score}(\boldsymbol{Y}_{\cdot k})$, and the two links have the same missing data percentage, the imputation accuracy for $\boldsymbol{Y}_{\cdot j}$ should be higher than that of $\boldsymbol{Y}_{\cdot k}$.

3.3. Preliminaries and Assumptions

Assume that we have missing data on two different variables $\boldsymbol{Y}_{\cdot j}, \boldsymbol{Y}_{\cdot k}$ with the same percent, the imputation accuracy can be different due to the variable's role in the whole variable set. In the traffic missing data imputation problem, two links $\boldsymbol{Y}_{\cdot j}, \boldsymbol{Y}_{\cdot k}$ may have different correlations to other links. In this section, we prose a metric to differentiate the role of each link.

Assume that we have a D dimensional Gaussian distribution, written as

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{u},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{u})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{u})\right)$$

where $\boldsymbol{u}$ is a D-dimensional mean vector, $\boldsymbol{\Sigma}$ is a $D \times D$ covariance matrix, $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$. Then we partition the D-dimensional vector into two parts, written as

$$\boldsymbol{x} = \begin{pmatrix}\boldsymbol{x}_a \\ \boldsymbol{x}_b\end{pmatrix}$$

Correspondingly, the mean vector and the covariance matrix are respectively partitioned into two parts and four parts, written as below.

$$\boldsymbol{u} = \begin{pmatrix}u_a \\ u_b\end{pmatrix},\ \boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{pmatrix}$$

We further utilize another variable $\boldsymbol{\Lambda}$ to denote the inverse matrix of the covariance matrix, defined as

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1} = \begin{pmatrix}\boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb}\end{pmatrix}$$

Note that, we have the theory of matrix inverse as

$$\begin{pmatrix}\boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D}\end{pmatrix}^{-1} = \begin{pmatrix}\boldsymbol{M} & -\boldsymbol{MBD}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{CM} & \boldsymbol{D}^{-1}\boldsymbol{CMBD}^{-1}\end{pmatrix}$$

$$\boldsymbol{M} = (\boldsymbol{A}-\boldsymbol{BD}^{-1}\boldsymbol{C})^{-1}$$

Hence, for the inverse of the covariance matrix, we have

$$\begin{pmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{pmatrix}^{-1} = \begin{pmatrix}\boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb}\end{pmatrix}$$

where we care about the expression of $\boldsymbol{\Lambda}_{aa}$ and $\boldsymbol{\Lambda}_{ab}$, written as

$$\boldsymbol{\Lambda}_{aa} = (\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}{\boldsymbol{\Sigma}_{bb}}^{-1}\boldsymbol{\Sigma}_{ba})^{-1}$$

$$\boldsymbol{\Lambda}_{ab} = -(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}{\boldsymbol{\Sigma}_{bb}}^{-1}\boldsymbol{\Sigma}_{ba})^{-1}\ \boldsymbol{\Sigma}_{ab}\ {\boldsymbol{\Sigma}_{bb}}^{-1}$$

For the Gaussian distribution, the exponent parts can be expanded as

$$-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{u})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{u}) = -\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{u} + const$$

When we partition the D-dimensional vector into two parts $\boldsymbol{x} = (\boldsymbol{x}_a, \boldsymbol{x}_b)^T$, then the exponent part of the Gaussian distribution can be expanded into

$$-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{u})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{u})$$

$$= -\frac{1}{2}(\boldsymbol{x}_a-\boldsymbol{u}_a)^T\boldsymbol{\Lambda}_{aa}(\boldsymbol{x}_a-\boldsymbol{u}_a) - \frac{1}{2}(\boldsymbol{x}_a-\boldsymbol{u}_a)^T\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b-\boldsymbol{u}_b)$$

$$-\frac{1}{2}(\boldsymbol{x}_b-\boldsymbol{u}_b)^T\boldsymbol{\Lambda}_{ba}(\boldsymbol{x}_a-\boldsymbol{u}_a) - \frac{1}{2}(\boldsymbol{x}_b-\boldsymbol{u}_b)^T\boldsymbol{\Lambda}_{bb}(\boldsymbol{x}_b-\boldsymbol{u}_b)$$

We assume that $\boldsymbol{x}_b$ is known in advance, then it can be regarded as a constant. Hence, the first order of $\boldsymbol{x}_a$ is written as

$$\boldsymbol{x}_a{}^{\boldsymbol{T}}\{\boldsymbol{\Lambda}_{aa}\boldsymbol{u}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{u}_b)\} \quad (165)$$

which should have the same expression of the original expression for the first order part written as $\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{u}$. For $\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{u}$, when we consider the $\boldsymbol{x}_b$ is known, then $\boldsymbol{\Sigma}^{-1}\boldsymbol{u}$ can be written as $\boldsymbol{\Sigma}_{a|b}{}^{-1}\boldsymbol{u}_{a|b}$, which should be equal to $\boldsymbol{\Lambda}_{aa}\boldsymbol{u}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{u}_b)$, written as (166–168)

$$\boldsymbol{\Lambda}_{aa}\boldsymbol{u}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{u}_b) = \boldsymbol{\Sigma}_{a|b}{}^{-1}\boldsymbol{u}_{a|b} \quad (169)$$

Hence, we have the expression the estimated value of $\boldsymbol{u}_{a|b}$ written as conditional Gaussian distribution (170–171)

$$\boldsymbol{u}_{a|b} = \boldsymbol{\Sigma}_{a|b}\{\boldsymbol{\Lambda}_{aa}\boldsymbol{u}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{u}_b)\} \quad (172)$$

where $\boldsymbol{\Lambda}_{aa}$ and $\boldsymbol{\Lambda}_{ab}$ are already known as above. (173)

Based on the conditional Gaussian distribution, we replace the $\boldsymbol{x}_b$ part with $((\boldsymbol{Y}_{\cdot k})_{k\in\overline{\mathcal{M}}})$, which is assumed to known observations, and replace the $\boldsymbol{x}_a$ part as the unknown part $\boldsymbol{Y}_{\cdot m}$, which is to be estimated because that the data is missing. Then we have the expectation of the estimation as (174–177)

$$\mathbb{E}[\boldsymbol{Y}_{\cdot m}|((\boldsymbol{Y}_{\cdot k})_{k\in\overline{\mathcal{M}}})] = \boldsymbol{\alpha}_m + \boldsymbol{\Sigma}_{m,\overline{\mathcal{M}}}\boldsymbol{\Sigma}_{\overline{\mathcal{M}},\overline{\mathcal{M}}}^{-1}\left(Y_{\cdot\overline{\mathcal{M}}}^{T} - \boldsymbol{\alpha}_{\overline{\mathcal{M}}}\right) \quad (178)$$

Then the estimation of the missing data is calculated as (179)

$$\widehat{\boldsymbol{Y}}_{im} = \widehat{\boldsymbol{\alpha}}_m + \widehat{\boldsymbol{\Sigma}}_{m,\overline{\mathcal{M}}}\widehat{\boldsymbol{\Sigma}}_{\overline{\mathcal{M}},\overline{\mathcal{M}}}^{-1}\left(Y_{\cdot\overline{\mathcal{M}}}^{T} - \boldsymbol{\alpha}_{\overline{\mathcal{M}}}\right) \quad (180)$$

Hence, the missing data estimations depend on the estimations of $\widehat{\boldsymbol{\alpha}}_m$, $\widehat{\boldsymbol{\Sigma}}_{m,\overline{\mathcal{M}}}$ and $\widehat{\boldsymbol{\Sigma}}_{\overline{\mathcal{M}},\overline{\mathcal{M}}}^{-1}$. Below are assumptions for estimating the model parameters. (181–182)

Assumption A1: $\forall m \in \mathcal{M}, \forall j \in \mathcal{I}, \left(\boldsymbol{A}_{\cdot m}\left(\boldsymbol{A}_{\cdot j'}\right)_{j'\in\mathcal{I}_{-j}}\right)$ is invertible. $\mathcal{I}_{-j} = \mathcal{I}\backslash\{j\}$. (183)

Assumption A2: $\forall m \in \mathcal{M}, \forall j \in \mathcal{I}, \boldsymbol{Y}_{\cdot j} \perp \boldsymbol{\Omega}_{\cdot m}|\left((\boldsymbol{Y}_{\cdot k})_{k\in\overline{\{j\}}}\right)$. (184)

A1 denotes that the matrix $\left(\boldsymbol{A}_{\cdot m}\left(\boldsymbol{A}_{\cdot j'}\right)_{j'\in\mathcal{I}_{-j}}\right)$ is of full rank. A**2** denotes that, given the values in $(\boldsymbol{Y}_{\cdot k})_{k\in\overline{\{j\}}}$, the column $\boldsymbol{Y}_{\cdot j}$ is independent with the column $\boldsymbol{\Omega}_{\cdot m}$. (185–186)

The missing data imputation for MNAR is to estimate the value of $\boldsymbol{Y}_{im}$ with $m \in \mathcal{M}$ for i such that $\boldsymbol{\Omega}_{im} = 0$. Such an assumption **A2** leads to (187–188)

$$\mathbb{E}\left[\boldsymbol{Y}_{\cdot j}\middle|\left((\boldsymbol{Y}_{\cdot k})_{k\in\overline{\{j\}}}\right)\right] = \mathbb{E}\left[\boldsymbol{Y}_{\cdot j}\middle|\left((\boldsymbol{Y}_{\cdot k})_{k\in\overline{\{j\}}}\right), \boldsymbol{\Omega}_{im} = 1\right] \quad (189)$$

3.4 Estimation of $\boldsymbol{\alpha}$ (190)

We first define the regression coefficients of $\boldsymbol{Y}_{\cdot j}$ on $\boldsymbol{Y}_{\cdot m}$ and $\boldsymbol{Y}_{\cdot k}$, for $k \in \mathcal{I}_{-j}$ in the complete case, that will be used to express the mean of a variable with MNAR values. (191–192)

Considering the model $\boldsymbol{Y} = \boldsymbol{1\alpha} + \boldsymbol{XA} + \boldsymbol{\epsilon}$, with an assumption that matrix $\boldsymbol{A} \in \mathbb{R}^{r\times p}$ is of full rank r. Therefore, the expression of $\boldsymbol{Y}_{\cdot j}$ can be reduced to the following linear system. (193–195)

$$\left(\boldsymbol{Y}_{\cdot m}\ \left(\boldsymbol{Y}_{\cdot j'}\right)_{j'\in\mathcal{I}_{-j}}\right) = \boldsymbol{1\alpha}_{|r} + (\boldsymbol{X}_{\cdot 1}, \cdots, \boldsymbol{X}_{\cdot r})\boldsymbol{A}_{|r} + \boldsymbol{\epsilon}_{|r} \quad (196)$$

Here, $\boldsymbol{\alpha}_{|r}$ and $\boldsymbol{\epsilon}_{|r}$ are the reduced matrix of $\boldsymbol{\alpha}$ and $\boldsymbol{\epsilon}$. $\boldsymbol{A}_{|r} \in \mathbb{R}^{r\times r}$ denotes the reduced matrix of $\left(\boldsymbol{A}_{\cdot m}\ \left(\boldsymbol{A}_{\cdot j'}\right)_{j'\in\mathcal{I}_{-j}}\right)$. (197–198)

Given the assumption **A1**, the $\boldsymbol{A}_{|r}$ is invertible, and the inverted matrix is denoted as $\acute{\boldsymbol{A}}^{-1}$. The latent matrix of full rank r can be written as (199–200)

$$(\boldsymbol{X}_{\cdot 1}, \cdots, \boldsymbol{X}_{\cdot r}) = \left(\left(\boldsymbol{Y}_{\cdot m}\ \left(\boldsymbol{Y}_{\cdot j'}\right)_{j'\in\mathcal{I}_{-j}}\right) - \boldsymbol{1\alpha}_{|r} - \boldsymbol{\epsilon}_{|r}\right)\acute{\boldsymbol{A}}^{-1}\ . \quad (201)$$

Using the original model $\boldsymbol{Y} = \mathbf{1}\boldsymbol{\alpha} + \boldsymbol{XA} + \boldsymbol{\epsilon}$, the expression of $\boldsymbol{Y}_{\cdot j}$ is then can be written as

$$\boldsymbol{Y}_{\cdot j} = \mathbf{1}\boldsymbol{\alpha}_j + \left(\left(\boldsymbol{Y}_{\cdot m}\ \left(\boldsymbol{Y}_{\cdot j'}\right)_{j' \in \mathcal{I}_{-j}}\right) - \mathbf{1}\boldsymbol{\alpha}_{|r} - \boldsymbol{\epsilon}_{|r}\right)\hat{\boldsymbol{A}}^{-1}\boldsymbol{A}_{j\cdot} + \boldsymbol{\epsilon}_{\cdot j} =$$

$$\sum_{k \in \{m\} \cup \mathcal{I}_{-j}} \left(\sum_{l \in m \cup \mathcal{I}_{-j}} \hat{\boldsymbol{A}}^{-1}{}_{lk}\boldsymbol{A}_{jl}\right)\boldsymbol{Y}_{\cdot k} - \sum_{k \in \{m\} \cup \mathcal{I}_{-j}} \left(\sum_{l \in m \cup \mathcal{I}_{-j}} \hat{\boldsymbol{A}}^{-1}{}_{lk}\boldsymbol{A}_{jl}\right)(\mathbf{1}\boldsymbol{\alpha}_k + \boldsymbol{\epsilon}_{\cdot k}) + \mathbf{1}\boldsymbol{\alpha}_j + \boldsymbol{\epsilon}_{\cdot j}$$

where we can get the intercept and the coefficients of $\boldsymbol{Y}_{\cdot j}$ on $\left(\boldsymbol{Y}_{\cdot m}, (\boldsymbol{Y}_{\cdot k})_{k \in \mathcal{I}_{-j}}\right)$.

For $j \in \mathcal{I}$, $k \in \mathcal{I}_{-j}$, let $A^c_{j \to m, \mathcal{I}_{-j}[0]}$ be the intercept, and $\boldsymbol{A}^c_{j \to m, \mathcal{I}_{-j}[m]}$ $\boldsymbol{A}^c_{j \to m, \mathcal{I}_{-j}[k]}$be the coefficients standing for the effects of $\boldsymbol{Y}_{\cdot j}$ on $\left(\boldsymbol{Y}_{\cdot m}, (\boldsymbol{Y}_{\cdot k})_{k \in \mathcal{I}_{-j}}\right)$ in the complete case, i.e., when $\boldsymbol{\Omega}_{\cdot m} = 1$. Then we have

$$\left(\boldsymbol{Y}_{\cdot j}\right)_{|\boldsymbol{\Omega}_{\cdot m}=1} = \boldsymbol{A}^c_{j \to m, \mathcal{I}_{-j}[0]} + \sum_{j' \in \mathcal{I}_{-j}} A^c_{j \to m, \mathcal{I}_{-j}[j']}\boldsymbol{Y}_{\cdot j'} + A^c_{j \to m, \mathcal{I}_{-j}[m]}\boldsymbol{Y}_{\cdot k} + \boldsymbol{\zeta}^c$$

where the coefficients are calculated as below equations.

$$A^c_{j \to m, \mathcal{I}_{-j}[j']} = \sum_{l \in \{m\} \cup \mathcal{I}_{-j}} \hat{\boldsymbol{A}}^{-1}{}_{lj'}\boldsymbol{A}_{jl}\ , j' \in \mathcal{I}_{-j}$$

$$A^c_{j \to m, \mathcal{I}_{-j}[m]} = \sum_{l \in \{m\} \cup \mathcal{I}_{-j}} \hat{\boldsymbol{A}}^{-1}{}_{lm}\boldsymbol{A}_{jl}$$

$$A^c_{j \to m, \mathcal{I}_{-j}[0]} = \mathbf{1}\boldsymbol{\alpha}_j - \sum_{k \in \{m\} \cup \mathcal{I}_{-j}} \left(\sum_{l \in \{m\} \cup \mathcal{I}_{-j}} \hat{\boldsymbol{A}}^{-1}{}_{lk}\boldsymbol{A}_{jl}\right)\mathbf{1}\boldsymbol{\alpha}_k$$

$$\boldsymbol{\zeta}^c = -\sum_{k \in \{m\} \cup \mathcal{I}_{-j}} A^c_{j \to m, \mathcal{I}_{-j}[k]}\boldsymbol{\epsilon}_{\cdot k} + \boldsymbol{\epsilon}_{\cdot j}$$

Here the arrow $j \to m, \mathcal{I}_{-j}$ indicates the regression model of $\boldsymbol{Y}_{\cdot j}$ on $\boldsymbol{Y}_{\cdot (m, \mathcal{I}_{-j})}$, and the squared bracket $[k]$ indicates the coefficient. Based on the model setting, we have $\mathbb{E}[\boldsymbol{\epsilon}_{\cdot k}] = 0$, hence $\mathbb{E}[\boldsymbol{\zeta}^c] = 0$.

The assumption **A2** leads to

$$\mathbb{E}\left[\boldsymbol{Y}_{\cdot j}\middle|\left((\boldsymbol{Y}_{\cdot k})_{k \in \overline{\{j\}}}\right)\right] = \mathbb{E}\left[\boldsymbol{Y}_{\cdot j}\middle|\left((\boldsymbol{Y}_{\cdot k})_{k \in \overline{\{j\}}}\right), \boldsymbol{\Omega}_{im} = 1\right]$$

$$= \mathbb{E}\left[A^c_{j \to m, \mathcal{I}_{-j}[0]} + \sum_{k \in \{m\} \cup \mathcal{I}_{-j}} \boldsymbol{A}^c_{j \to m, \mathcal{I}_{-j}[k]}\boldsymbol{Y}_{\cdot k} + \zeta^c\middle|\left((\boldsymbol{Y}_{\cdot k})_{k \in \overline{\{j\}}}\right)\right]$$

$$= \boldsymbol{A}^c_{j \to m, \mathcal{I}_{-j}[0]} + \sum_{k \in \{m\} \cup \mathcal{I}_{-j}} A^c_{j \to m, \mathcal{I}_{-j}[k]}\boldsymbol{Y}_{\cdot k} + \quad \mathbb{E}\left[\boldsymbol{\zeta}^c\middle|\left((\boldsymbol{Y}_{\cdot k})_{k \in \overline{\{j\}}}\right)\right]$$

By taking the expectation of the left and right parts of the equality above given $\mathbb{E}[\boldsymbol{\epsilon}_{\cdot k}] = 0$ for $\forall k \in \{m\} \cup \mathcal{I}_{-j}$, we have

$$Left = \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{Y}_{\cdot j}\middle|\left((\boldsymbol{Y}_{\cdot k})_{k \in \overline{\{j\}}}\right)\right]\right] = \mathbb{E}\left[\boldsymbol{Y}_{\cdot j}\right] = \boldsymbol{\alpha}_j$$

$$Right = \mathbb{E}\left[\boldsymbol{A}^c_{j \to m, \mathcal{I}_{-j}[0]} + \sum_{k \in \{m\} \cup \mathcal{I}_{-j}} A^c_{j \to m, \mathcal{I}_{-j}[k]}\boldsymbol{Y}_{\cdot k} + \ \mathbb{E}\left[\boldsymbol{\zeta}^c\middle|\left((\boldsymbol{Y}_{\cdot k})_{k \in \overline{\{j\}}}\right)\right]\right]$$

$$= \boldsymbol{A}^c_{j \to m, \mathcal{I}_{-j}[0]} + \sum_{j' \in \mathcal{I}_{-j}} A^c_{j \to m, \mathcal{I}_{-j}[j']}\boldsymbol{\alpha}_k + \ \boldsymbol{A}^c_{j \to m, \mathcal{I}_{-j}[m]}\boldsymbol{\alpha}_m + \mathbb{E}[\boldsymbol{\zeta}^c]$$

Above two equalities are identical. So, we have

$$\widehat{\boldsymbol{\alpha}}_m = \frac{\boldsymbol{\alpha}_j - \boldsymbol{A}^c_{j\to m,\mathcal{J}_{-j}[0]} - \sum_{j'\in\mathcal{J}_{-j}} \boldsymbol{A}^c_{j\to m,\mathcal{J}_{-j}[j']}\boldsymbol{\alpha}_k}{\boldsymbol{A}^c_{j\to m,\mathcal{J}_{-j}[m]}}$$

3.5 Estimation of Variance and Covariance

Let $\boldsymbol{Z} = (\boldsymbol{Y}_{\cdot k})_{k\in\overline{\{j\}}}$, for the variance $\boldsymbol{\Sigma}_{\overline{\mathcal{M}},\overline{\mathcal{M}}}$, we have

$$\boldsymbol{\Sigma}_{\overline{\mathcal{M}},\overline{\mathcal{M}}} = Var\left(\boldsymbol{Y}_{\cdot j}\right) = \mathbb{E}\left[Var\left(\boldsymbol{Y}_{\cdot j}\middle|\boldsymbol{Z}\right)\right] + Var(\mathbb{E}[\boldsymbol{Y}_{\cdot j}|\boldsymbol{Z}]).$$

The assumption **A2** leads to $Var\left(\boldsymbol{Y}_{\cdot j}\middle|\boldsymbol{Z}\right) = Var\left(\boldsymbol{Y}_{\cdot j}\middle|\boldsymbol{Z}, \boldsymbol{\Omega}_{\cdot m} = 1\right)$. According to the conditional variance for a Gaussian vector, we have

$$Var\left(\boldsymbol{Y}_{\cdot j}\middle|\boldsymbol{Z}\right) = Var\left(\boldsymbol{Y}_{\cdot j}\right) - Cov\left(\boldsymbol{Z},\boldsymbol{Y}_{\cdot j}\right)Var(\boldsymbol{Z})^{-1}Cov\left(\boldsymbol{Z},\boldsymbol{Y}_{\cdot j}\right)^T.$$

Then we have the first term of $Var\left(\boldsymbol{Y}_{\cdot j}\right)$ as

$$\mathbb{E}\left[Var\left(\boldsymbol{Y}_{\cdot j}\middle|\boldsymbol{Z}\right)\right] = Var\left(\boldsymbol{Y}_{\cdot j}\right) - Cov\left(\boldsymbol{Z},\boldsymbol{Y}_{\cdot j}\right)Var(\boldsymbol{Z})^{-1}Cov\left(\boldsymbol{Z},\boldsymbol{Y}_{\cdot j}\right)^T|\boldsymbol{\Omega}_{\cdot m} = 1$$

For the second term of $Var\left(\boldsymbol{Y}_{\cdot j}\right)$, we have

$$Var\left(\mathbb{E}\left[\boldsymbol{Y}_{\cdot j}\middle|\boldsymbol{Z}\right]\right) = Var\left(\mathbb{E}\left[\boldsymbol{Y}_{\cdot j}\middle|\boldsymbol{Z}, \boldsymbol{\Omega}_{\cdot m} = 1\right]\right)$$

$$= Var\left(\sum_{k\in\{m\}\cup\mathcal{J}_{-j}} A^c_{j\to m,\mathcal{J}_{-j}[k]}\boldsymbol{Y}_{\cdot k} - \sum_{k\in\{m\}\cup\mathcal{J}_{-j}} A^c_{j\to m,\mathcal{J}_{-j}[k]}\mathbb{E}[\boldsymbol{\epsilon}_{\cdot k}|\boldsymbol{Z}] + +\boldsymbol{A}^c_{j\to m,\mathcal{J}_{-j}[0]} + \mathbb{E}[\boldsymbol{\epsilon}_{\cdot j}]\right)$$

where $\mathbb{E}[\boldsymbol{\epsilon}_{\cdot k}|\boldsymbol{Z}] = \sigma^2(Var(Z)^{-1})_{k\cdot}(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}])$.

For the covariances $\widehat{\boldsymbol{\Sigma}}_{m,\overline{\mathcal{M}}}$ between $\boldsymbol{Y}_{\cdot m}$, $\boldsymbol{Y}_{\cdot k}$, $k \in \mathcal{J}$, let $\boldsymbol{Z} = (\boldsymbol{Y}_{\cdot l})_{l\in\overline{\{k\}}}$, we have

$$\widehat{\boldsymbol{\Sigma}}_{m,\overline{\mathcal{M}}} = Cov\left(\boldsymbol{Y}_{\cdot m}\,\boldsymbol{Y}_{\cdot k}\right) = \mathbb{E}\left[\boldsymbol{Y}_{\cdot m}\,\boldsymbol{Y}_{\cdot k}\right] - \mathbb{E}\left[\boldsymbol{Y}_{\cdot m}\right]\mathbb{E}\left[\,\boldsymbol{Y}_{\cdot k}\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{Y}_{\cdot m}\,\boldsymbol{Y}_{\cdot k}|\boldsymbol{Z}\right]\right] - \mathbb{E}[\boldsymbol{Y}_{\cdot m}]\mathbb{E}[\,\boldsymbol{Y}_{\cdot k}] = \mathbb{E}\left[\boldsymbol{Y}_{\cdot m}\mathbb{E}[\boldsymbol{Y}_{\cdot k}|\boldsymbol{Z}]\right] - \mathbb{E}[\boldsymbol{Y}_{\cdot m}]\mathbb{E}[\,\boldsymbol{Y}_{\cdot k}].$$

For the first term, we have

$$\mathbb{E}\left[\boldsymbol{Y}_{\cdot m}\mathbb{E}[\boldsymbol{Y}_{\cdot k}|\boldsymbol{Z}]\right] = \mathbb{E}\left[\boldsymbol{Y}_{\cdot m}\mathbb{E}[\boldsymbol{Y}_{\cdot k}|\boldsymbol{Z}, \boldsymbol{\Omega}_{\cdot m} = 1]\right]$$

$$= \mathbb{E}\left[\boldsymbol{Y}_{\cdot m}\left(\boldsymbol{A}^c_{j\to m,\mathcal{J}_{-j}[0]} + \sum_{j'\in\{m\}\mathcal{J}_{-j}} A^c_{j\to m,\mathcal{J}_{-j}[j']}\boldsymbol{Y}_{\cdot j'} + \mathbb{E}(\boldsymbol{\zeta}^c|Z)\right)\right]$$

$$= \boldsymbol{A}^c_{j\to m,\mathcal{J}_{-j}[0]}\mathbb{E}[\boldsymbol{Y}_{\cdot m}] + A^c_{j\to m,\mathcal{J}_{-j}[m]}\mathbb{E}[\boldsymbol{Y}^2_{\cdot m}] + \sum_{j'\in\mathcal{J}_{-j}} A^c_{j\to m,\mathcal{J}_{-j}[j']}\mathbb{E}[\boldsymbol{Y}_{\cdot m}\boldsymbol{Y}_{\cdot j'}] + \mathbb{E}[\boldsymbol{Y}_{\cdot m}\mathbb{E}(\boldsymbol{\zeta}^c|Z)]$$

According to the derivation in [22], $\mathbb{E}[\boldsymbol{Y}_{\cdot m}\mathbb{E}(\boldsymbol{\zeta}^c|Z)]$ is calculated as

$$-\sigma^2\left(\sum_{l\in\mathcal{J}_{-k}}\sum_{s\in\mathcal{J}_{-k}} Var(\boldsymbol{Z})^{-1}\,A^c_{j\to m,\mathcal{J}_{-j}[l]}Cov(\boldsymbol{Y}_{\cdot m}\boldsymbol{Y}_{\cdot l}) + A^c_{j\to m,\mathcal{J}_{-j}[m]}\right).$$

Note that for the second term, $\mathbb{E}[\boldsymbol{Y}_{\cdot m}]\mathbb{E}[\,\boldsymbol{Y}_{\cdot k}]$, it can be directly calculated.

4. Experiment

4.1. Dataset and Preprocessing

We utilize a road traffic speed dataset published by [23]. road segments are anonymous, covering the main urban expressways within two months from August 1 to September 30, 2016, (a total of 61 days).

The time interval is 10-minute. From the original dataset, we select twenty links whose speed are generated in the morning rush hours (i.e., 7:00 am to 9:00 am) for evaluating the proposed method. The speed of each link is transformed to the congestion index, which is calculated as $\boldsymbol{v}_{ij} / \max(\boldsymbol{v}_{.j})$, $\boldsymbol{v}_{ij}$ denotes the i^{th} speed value of link j and $\boldsymbol{v}_{.j}$ denotes all observations of link j. For each link j, the time series length of speed observations is 732 (12 observations in two hours 61 days). Hence, the dimension of $\boldsymbol{Y}$ is $n = 732, p$is the number of links.

The basic assumption of the proposed model is that the observations of each link are drawn from a Gaussian distribution, Hence, we adopt the quantile-quantile plot (QQ Plot) to display the quantiles of the data (after normalizing) versus the theoretical quantile values from a normal distribution. If the distribution of the data is normal, then the data plot is linear. As shown in the below figure, the plot closely follows the straight lines, suggesting that the data after normalizing the congestion data has an approximately normal distribution.

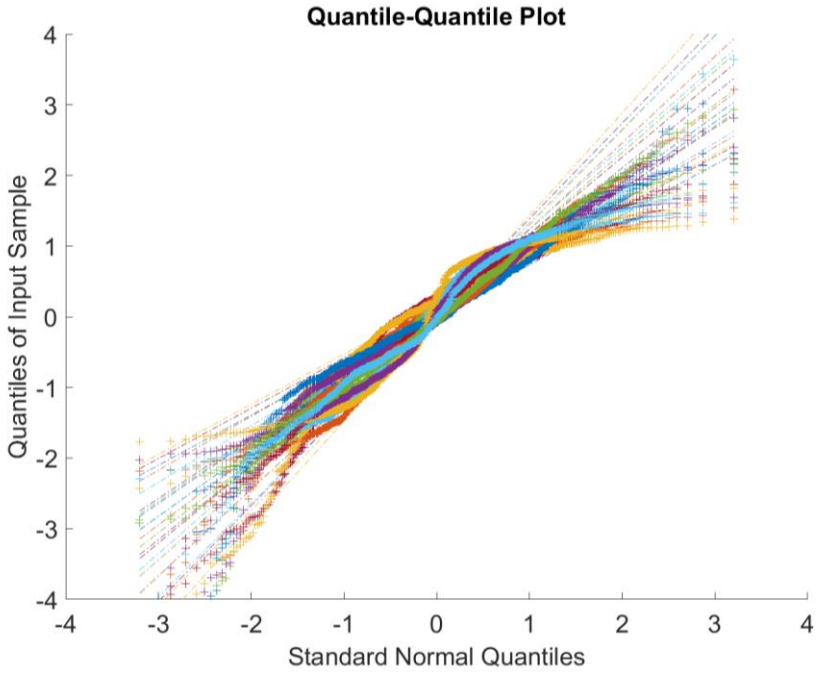


Figure. 1. This is a figure. Schemes follow the same formatting.

4.2. Metrics for Missing Data Imputation Accuracy

For evaluating the performance of missing data imputation, we adopt the below four metrics, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Symmetric Mean Absolute Percentage Error (SMAPE), and R^2. Note that a higher R^2 value denotes better accuracy.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$\text{SMAPE} = 100\% * \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2} \quad (282)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (283)$$

4.3. Benchmark and Experiment Settings (284)

We compare the new model with the typically used PPCA model, where σ^2 and $\boldsymbol{A}$ (285) is estimated by an expectation-maximum (EM) algorithm. We name it ppca-em in this (286) section. We further use the estimated σ^2 by EM as the known inputs of the new model in (287) this study. As to the rank r in the model, the best value is determined by the cross-validation (288) on the dataset. In this part, we further detail the experiment settings in terms of the (289) MNAR data generating and the settings of link set $\mathcal{M}$. (290)

4.3.1 Generating MNAR (291)

Note that the model targets at solving the imputation for MNAR data. We utilize the (292) mechanism of generating MNAR in [22]. Specifically, a logistic regression function is (293) adopted as $f(x) = 1/(-a(x-b))$, where x is an observation, and (a, b) is set for selecting (294) different missing percentage. The function transforms the observation x to a value in (295) $(0,1)$. The observation x with $f(x) > \mu$, is set to be the MNAR data, where μ is a random (296) threshold. We set the parameters (a, b) as below Table 1, which is corresponding to a (297) specific missing percentage. (298)

(299)

Table 1. Settings for Generating MNAR Data in the Experiments (300)

(301)

$\boldsymbol{a}$	$\boldsymbol{b}$	Missing Percentage
-1	-1.3	25%
3	0	50%
1	-1.3	75%

(302)

4.3.2 Settings of Link Set $\mathcal{M}$ (303)

Missing data on different links may obtain different recover accuracy, even with the (304) same missing percentage. For evaluating this proposition, we first test the missing data (305) imputation accuracy with different p-score values of the links. When a link observation (306) $\boldsymbol{Y}_{\cdot j}$ is set to be $\mathcal{M} = \{j\}$, all other links are set to be $\mathcal{I}_{-j} = \mathcal{I}\backslash\{j\}$, where $\mathcal{I} = \{1,2,\dots,20\}$. (307) Further, we test the missing data imputation accuracy of several select links (or link com- (308) bination) compared with the ppca-em model, to evaluate the advantage of the new model. (309)

4.4 Results and Analysis (310)

We first examine the relationship between the missing data imputation accuracy and (311) the proposed metric, i.e., p-score value. We select two links with the highest p-score value (312) and the lowest p-score value in the dataset. The p-score values of two selected links, i.e., (313) link 6 and link 18, are shown in Fig.2, where the color map denotes the ρ values between (314) the selected link and all links in the link set $\mathcal{I}$. (315)

(316)

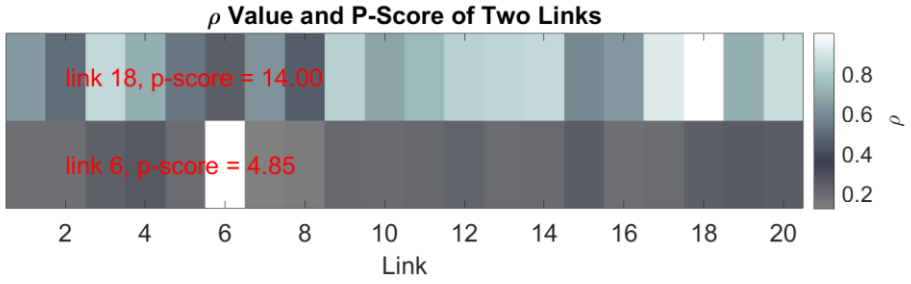


Figure. 2. ρ Value of Two Links with the Highest and Lowest P-Score

Accordingly, we calculate the absolute errors of the model on these two selected links. Fig. 3 shows the results missing data imputation results on these links in terms of different missing data percentages. We can see that missing data on the link 18, which is with a higher p-score value than that of link 6, is better recovered regarding all scenarios of missing data percentages (25%, 50%, 75%).

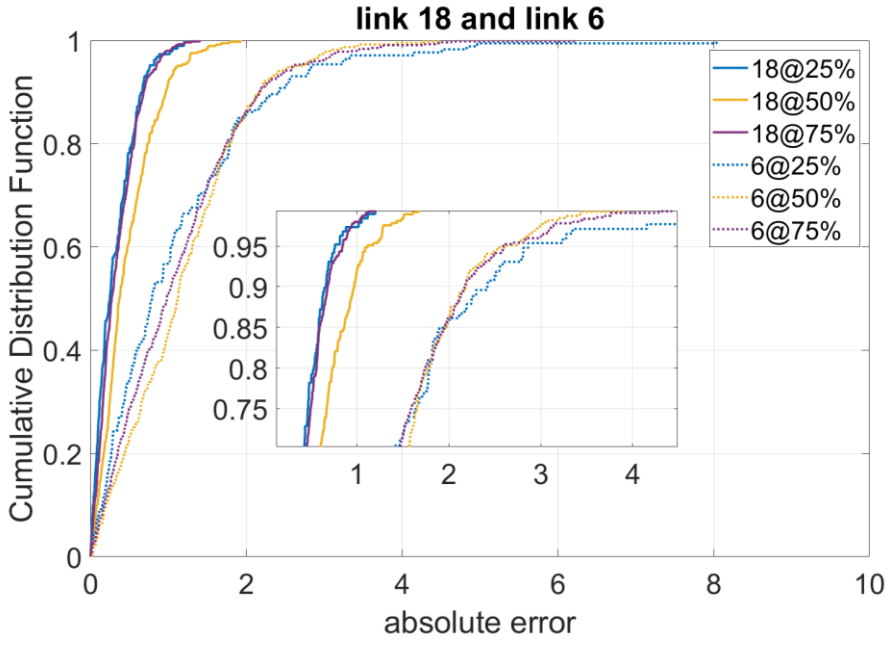


Figure. 3. Performance of the Algorithm for Links with Highest P-Score and Lowest P-Score

We further examine the relationship between the p-score value and the accuracy metrics on the traffic dataset, which is shown in below Fig.4. The accuracy measured by four metrics presents a positive correlation with the p-score value on different links, meaning that missing data on the links with higher p-score values can be better recovered.

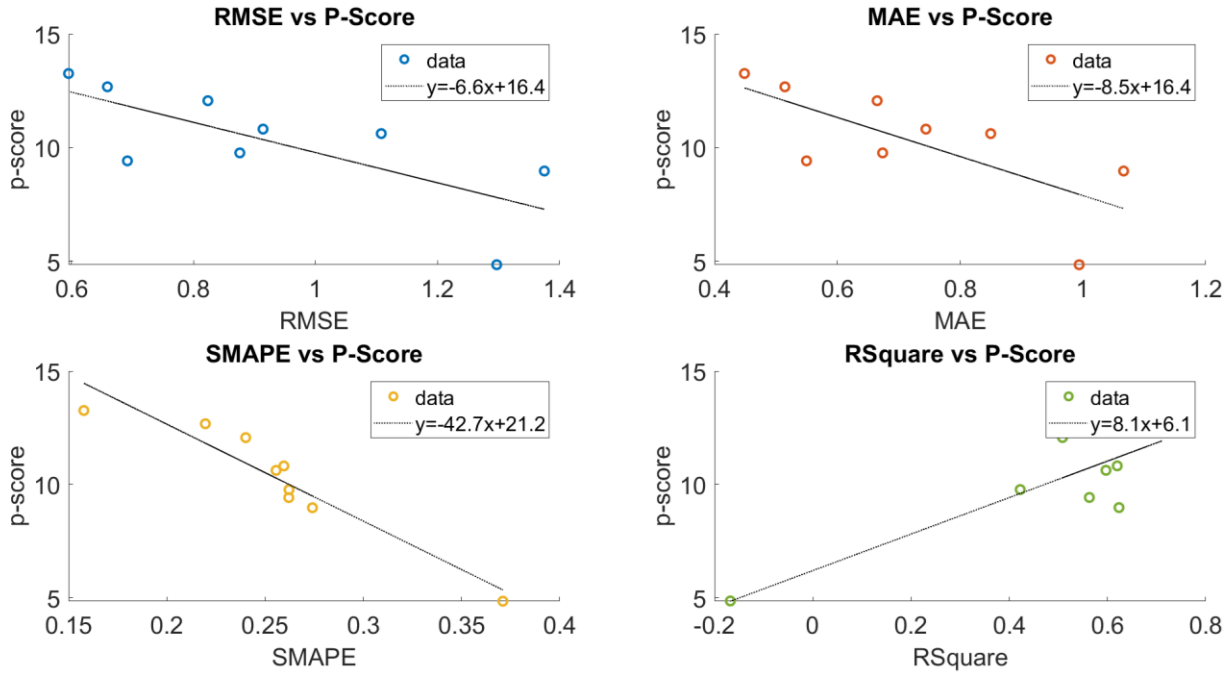


Figure. 4. Scatter Plot between the Accuracy Metrics and the p-score Values

We also compare the new model with the ppca-em model on other links or link combinations in the dataset. The settings and corresponding missing data imputation results measured by the four metrics are shown in Table 2.

Table 2. Experiment Setting and performance of the algorithms with different Percent of MNAR Data on Links

Experiment Setting										
Missing Rate (%) @ $\mathcal{M}$	50 @{1}		50 @{3}		75 @{1}		75 @{1,3}		75 @{3,5}	
p-score	10.62@{1}		13.26@{3}		10.62@{1}		–		9.42@{5}	
Performance Comparison										
Metrics	PPCA-EM	New	PPCA-EM	New	PPCA-EM	New	PPCA-EM	New	PPCA-EM	New
RMSE	0.992	0.746	0.559	0.595	1.069	0.746	0.835	0.871	0.942	0.627
MAE	0.81	0.564	0.458	0.448	0.789	0.564	0.598	0.625	0.665	0.468
SMAPE	0.34	0.223	0.216	0.157	0.289	0.223	0.231	0.228	0.253	0.201
R^2	0.15	0.688	0.595	0.681	0.545	0.688	0.208	0.677	0.115	0.740
Computing Time (Second)	6.54	2.03	6.29	2.03	6.73	2.64	6.06	4.06	11.32	4.11

Note that, Fig. 3 already shows that the new model obtains the worst accuracy on link 6. Hence, we further compare two models on this link to compare the new model with

the ppca-em model. The results are shown in Fig. 5. It shows that even on link 6, the absolute errors of the new model are still lower than the ppca-em model for three missing ratios.

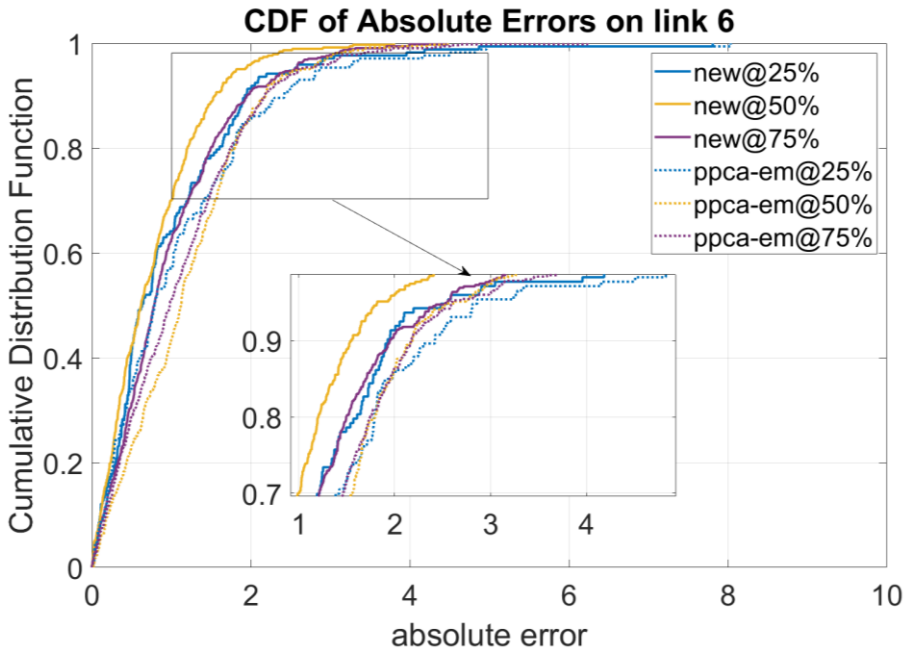


Figure. 5. Performance of Models on the Link with Lowest P-Score Values

Experiment results in Table 2 and Fig. 5 demonstrate that the new model performs better than the typcally used ppca-em model in terms of four accuracy metrics and computing time. It indicates that the new model is more effective and efficient for the MNAR traffic data imputation problem, which is the target of this study. The typical ppca-em method is usually used for imputation of data missing at random, whereas the new model is more general and is capable of MNAR data imputation.

5. Discussion

Our improved linear probabilistic principal component analysis method can be applied to a variety of missing traffic data imputation applications such as missing traffic speed estimation, or other traffic indicators. Especially, because the proposed missing data imputation method is a linear and interpretable model, which is naturally of high computing efficiency, thus, it can be utilized in the systems where real time missing data estimation is required. Additionally, the time-series based metric, P-Score value, is proposed to distinguish variables, e.g., links with missing traffic speed data, for estimating the missing values. Such a method can be applied to the applications of traffic surveillance systems, to identify which sensors should be of high priority to maintained in the systems to ensure the full surveillance, or which links should be equipped with sensor for traffic surveillance.

6. Conclusion

In this study, we propose a general linear model based on the PPCA to tackle the MNAR traffic data imputation problem. We also propose a time series-based metric, i.e., the p-score, to distinguish links that are of missing data. Experimental results on a real-world traffic dataset show that the proposed model performs better than the typically

used PPCA-EM model in terms of missing data imputation accuracy and computing time. Furthermore, we test the model on links with different p-score values. Experiment results show that the missing data on links with higher p-score values is better recovered. Such an observation helps us understand the data recovering distinction for different links in the road network, which has not been studied in any research to our best knowledge. In future work, we will further compare the model with other methods on more traffic datasets.

Author Contributions: Conceptualization, Liping Huang and Rong Su; methodology, Liping Huang; software, Zhenghuan Li.; validation, Ruikang Luo; formal analysis, Rong Su.; investigation, Liping Huang; resources, Rong Su.; data curation, Zhenghuan Li.; writing—original draft preparation, Lipng Huang; writing—review and editing, Liping Huang and Rong Su.; visualization, Liping Huang.; supervision, Rong Su; project administration, Rong Su; funding acquisition, Rong Su. All authors have read and agreed to the published version of the manuscript.

Funding: This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s), and A*STAR under its Industry Alignment Fund (LOA Award I1901E0046).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.
Data Availability Statement: the dataset used in this paper is published by [23], which can be found at https://doi.org/10.5281/zenodo.1205229.
Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yuan H.; Li G. A survey of traffic prediction: from spatio-temporal data to intelligent transportation. Data Science and Engineering, **2021**, 6(1), 63-85.
2. Neelakandan S.; Berlin M. A.; Tripathi S., et al. IoT-based traffic prediction and traffic signal control system for smart city. Soft Computing, **2021**, 25(18), 12241-12248.
3. Tan, H. C.; Wu Y. K.; Feng J. S. ; Wang W. H. and Ran B. Traffic missing data completion with spatial-temporal correlations, In Proceedings of 93rd Annual Meeting of the Transportation Research Board, Washington, DC. 12-16 Jan 2014.
4. Li H. P.; Wang Y. H. and Li M. Modified GAN Model for Traffic Missing Data Imputation. *CICTP* 2020. **2020**, 3013-3023.
5. Yang F, Liu G, Huang L, et al. Tensor Decomposition for Spatial—Temporal Traffic Flow Prediction with Sparse Data. Sensors, **2020**, 20(21), 6046.
6. Huang L. P.; Zhao S. D. and Luo R. K., et al. An incremental map matching approach with speed estimation constraints for high sampling rate vehicle trajectories, In proceedings of IEEE 17th International Conference on Control & Automation (ICCA). IEEE, 2022, 758-765, Italy, Naples, 27030 June 2022.
7. Huang L. P.; Yang Y. J.; Chen H. C. et al. Context aware road travel time estimation by coupled tensor decomposition based on trajectory data. *KBS*, **2022**, 245, 108596.
8. Huang L, Li Z, Zhao S, et al. Coupling Urban Road Travel Time and Traffic Status from Vehicle Trajectories by Gaussian Distribution. In Proceedings of IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2022, 4056-4061, Macau, China, 8-12 Oct 2022.
9. Huang L. P.; Yang Y. J. and Zhao X. H., et al. Sparse data-based urban road travel speed prediction using probabilistic principal component analysis. IEEE Access, **2018**, 6, 44022-44035.
10. Asif, M. T.; Mitrovic N.; Garg L., et al. Low-dimensional models for missing data imputation in road networks. In Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada, 26-31 May 2013.
11. Jia X.; Dong X.; Chen M., et al. Missing data imputation for traffic congestion data based on joint matrix factorization. Knowledge-Based Systems, **2021**, 225, 107114.
12. Asif M. T.; Mitrovic N., Dauwels J., et al. Matrix and tensor-based methods for missing data estimation in large traffic networks. IEEE Transactions on intelligent transportation systems, **2016**, 17, 1816-1825.
13. Jiang B., Siddiqi M D, Asadi R, et al. Imputation of missing traffic flow data using denoising autoencoders. Procedia Computer Science, **2021**, 184, 84-91.
14. Shang Q.; Yang Z.; Gao S., et al., An imputation method for missing traffic data based on FCM optimized by PSO-SVR, Journal of Advanced Transportation, **2018**, Article ID 2935248. DOI: https://doi.org/10.1155/2018/2935248
15. Li Y. B.; Li Z. H., and Li L. Missing traffic data: comparison of imputation methods, IET Intelligent Transport Systems, **2018**, 8, 51-57.

16. Wu P.; Xu L.; Huang Z. Imputation methods used in missing traffic data: a literature review. In Proceedings of International Symposium on Intelligence Computation and Applications, Guangzhou, China, 20-21 Nov 2019.
17. Chen X.; Lei M.; Saunier N., et al. Low-rank autoregressive tensor completion for spatiotemporal traffic data imputation. IEEE Transactions on Intelligent Transportation Systems, **2022**, 23(8), 12301-12310.
18. Tipping M. E.; , Bishop C. M. Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology), **1999**, 61(3), 611-622.
19. Ilin A., and Raiko T. Practical approaches to principal component analysis in the presence of missing values. The Journal of Machine Learning Research, **2010**, 11, 1957-2000.
20. Audigier B.; Husson F. and Josse J. Multiple imputation for continuous variables using a Bayesian principal component analysis. Journal of statistical computation and simulation, **2016**, 86, 2140-2156.
21. Qu L., Li L. Zhang Y., et al. PPCA-based missing data imputation for traffic flow volume: A systematical approach. IEEE Transactions on intelligent transportation systems, **2009**, 10, 512-522.
22. Sportisse, A.; Boyer C., and Josse J. Estimation and imputation in probabilistic principal component analysis with missing not at random data. Advances in Neural Information Processing Systems, **2020**, 33 ,7067-7077.
23. Chen X.; Yang J; Sun L. A nonconvex low-rank tensor completion model for spatiotemporal traffic data imputation. Transportation Research Part C: Emerging Technologies, **2020**, 117, 102673.

IMPACT FACTOR 3.847

Indexed in: PubMed

sensors

Submit to *Sensors*

Review for *Sensors*

Journal Menu

- *Sensors* Home
- Aims & Scope
- Editorial Board
- Reviewer Board
- Topical Advisory Panel
- Instructions for Authors
- Special Issues
- Topics
- Sections & Collections
- Article Processing Charge
- Indexing & Archiving
- Editor's Choice Articles
- Most Cited & Viewed
- Journal Statistics
- Journal History
- Journal Awards
- Society Collaborations
- Conferences
- Editorial Office

Journal Browser

volume

issue

Go

Forthcoming issue

Current issue

Vol. 22 (2022) Vol. 11 (2011)
Vol. 21 (2021) Vol. 10 (2010)
Vol. 20 (2020) Vol. 9 (2009)
Vol. 19 (2019) Vol. 8 (2008)
Vol. 18 (2018) Vol. 7 (2007)
Vol. 17 (2017) Vol. 6 (2006)
Vol. 16 (2016) Vol. 5 (2005)
Vol. 15 (2015) Vol. 4 (2004)
Vol. 14 (2014) Vol. 3 (2003)
Vol. 13 (2013) Vol. 2 (2002)
Vol. 12 (2012) Vol. 1 (2001)

Editorial Board

- Biosensors Section
- Chemical Sensors Section
- Physical Sensors Section
- Intelligent Sensors Section
- Sensor Networks Section
- Remote Sensors Section
- Optical Sensors Section
- Electronic Sensors Section
- Sensor Materials Section
- Internet of Things Section
- Biomedical Sensors Section
- Communications Section
- Fault Diagnosis & Sensors Section
- Wearables Section
- Nanosensors Section
- Sensing and Imaging Section
- Sensors and Robotics Section
- Vehicular Sensing Section
- Radar Sensors Section
- Navigation and Positioning Section
- Smart Agriculture Section
- Environmental Sensing Section
- Industrial Sensors Section
- Sensors Development Section

Members (1760)

Search by first name, last name, affiliation, interest...