Application of Data Fusion via Canonical Polyadic Decomposition in Risk Assessment of Musculoskeletal Disorders in Construction: Procedure and Stability Evaluation

Amrita Dutta[1]; Scott P. Breloff[2]; Fei Dai, M.ASCE[3]; Erik W. Sinsel[4]; Christopher M. Warren[5]; Robert E. Carey[6]; and John Z. Wu[7]

Abstract: Missing data is a common problem in data collection for work-related musculoskeletal disorder (WMSD) risk-assessment studies. It can cause incompleteness of risk indicators, leading to erroneous conclusion on potential risk factors. Previous studies suggested that data fusion is a potential way to solve this issue. This research evaluated the numerical stability of a data fusion technique that applies canonical polyadic decomposition (CPD) for WMSD risk assessment in construction. Two knee WMSD risk-related data sets—three-dimensional (3D) knee rotation (kinematics) and electromyography (EMG) of five knee postural muscles—collected from previous studies were fused for the evaluation. By comparing the consistency performance with and without data fusion, it revealed that for all low to high proportion of missing data (10%–70%) from both kinematics and EMG data sets, the WMSD risk assessment using fused data sets outperformed using unfused kinematics data sets. For large proportions of missing data (>50%) from both kinematics and EMG data sets, better performance was observed by using fused data sets in comparison with unfused EMG data sets. These findings suggest that data fusion using CPD generates a more reliable risk assessment compared with data sets with missing values and therefore is an effective approach for remedying missing data in WMSD risk evaluation. **DOI: 10.1061/(ASCE)CO.1943-7862.0002106.** *© 2021 American Society of Civil Engineers.*

Author keywords: Tensor decomposition; Risk assessment; Data fusion; Construction safety.

Introduction

Missing data is a common problem associated with data collection in work-related musculoskeletal disorder (WMSD) risk assessment in the field of construction ergonomics (Young et al. 2011). Data collected for assessing risk, either from the laboratory or construction site, often suffer from irregular spatial and temporal resolution. Among multiple trials of experiments, data of a specific trial can be missing due to an instrument failure, human-induced error, or at random. Therefore, missing data can occur in almost all research that involves data collection even when the study is well-designed and conducted in a controlled environment (Brown and Kros 2003). Missing data can cause incompleteness and misrepresentation of the risk indicators, leading to less statistical power and biased and invalid conclusions on the effects of the potential risk factors on the assessed WMSD risk.

Data fusion has been found to be a potential solution in dealing with incomplete data sets used in WMSD risk assessment by replacing the missing data reserving the interrelation among different factors and the risk indicators across multiple data sets. In a previous study, a data fusion method was proposed to treat imperfect WMSD risk-assessment data using a canonical polyadic decomposition (CPD) tensor decomposition (Dutta et al. 2020a). However, the numerical stability of the CPD fusion has not been evaluated comprehensively. When WMSD risks need to be characterized and quantified by multiple risk indicators, how these risk indicators will interact with each other for risk assessment if some of their proportions become missing is still unknown. Also, whether the CPD fusion is numerically stable to reconstruct a new data set by fusing those risk indicators and imputing the missing values to provide more accurate and reliable risk-assessment results compared with the unfused incomplete data sets has not been studied.

This research evaluates the numerical stability of the data fusion via CPD for WMSD risk assessment in construction when the quality of risk-related data sets becomes degraded due to the missing data in multiple inputs. To this end, this research compared the risk-assessment results resulting from the fused data sets with those resulting from multiple unfused incomplete data sets. CPD was

[1]Graduate Research Assistant, Dept. of Civil and Environmental Engineering, West Virginia Univ., P.O. Box 6103, Morgantown, WV 26506. Email: amdutta@mix.wvu.edu

[2]Biomedical Research Engineer, National Institute for Occupational Safety and Health, 1095 Willowdale Rd., Morgantown, WV 26505. Email: sbreloff@cdc.gov

[3]Associate Professor, Dept. of Civil and Environmental Engineering, West Virginia Univ., P.O. Box 6103, Morgantown, WV 26506 (corresponding author). ORCID: https://orcid.org/0000-0002-8868-2821. Email: fei.dai@mail.wvu.edu

[4]Computer Scientist, National Institute for Occupational Safety and Health, 1095 Willowdale Rd., Morgantown, WV 26505. Email: ESinsel@cdc.gov

[5]Mechanical Engineer, National Institute for Occupational Safety and Health, 1095 Willowdale Rd., Morgantown, WV 26505. Email: cpw4@cdc.gov

[6]Mechanical Engineer, National Institute for Occupational Safety and Health, 1095 Willowdale Rd., Morgantown, WV 26505. ORCID: https://orcid.org/0000-0002-4648-4464. Email: ohn7@cdc.gov

[7]Senior Research Biomechanical Engineer, National Institute for Occupational Safety and Health, 1095 Willowdale Rd., Morgantown, WV 26505. Email: ozw8@cdc.gov

Note. This manuscript was submitted on September 16, 2020; approved on February 26, 2021; published online on May 26, 2021. Discussion period open until October 26, 2021; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Construction Engineering and Management***, © ASCE, ISSN 0733-9364.

applied to fuse the incomplete data sets with different portions of subsets intentionally removed from two real WMSD risk-related data sets, referred to as original complete data sets, collected from previous experimental studies. The risk-assessment results obtained from the fused complete and the unfused incomplete data sets were then compared with those obtained from the original complete data sets to elucidate the performance of the present technique.

Background

Problems with Missing Data

The problem associated with missing data is relatively common in almost all research. Studies that involve human subjects and technologies for data collection can have missing data for various reasons including technology-induced errors (e.g., disconnection of sensors, errors in communicating with the database server, or instrument failures), human-induced errors (e.g., accidental human omission), or other unknown reasons (Data 2016). In the field of construction ergonomics, optical motion-capture technology is a prevalent technique for tracking and recording human motion data digitally that is used for ergonomic risk assessment and analysis. In an optical motion-capture system, multiple video cameras are utilized to track the human movements by estimating the three-dimensional (3D) positions of a set of reflective markers that are strategically attached to human body. A common caveat with this technology is the occasional occurrence of missing data due to occlusions caused by props, limbs, bodies, or other markers. Some other contributing factors are type of postural movements and recording conditions such as distance from the object, environmental artifacts, line-of-site issues, and extra reflection (Liu and McMillan 2006).

Surface electromyography (EMG) system is another technique that records the electrical activity produced by muscle contraction and is used for muscle activation measurement (Brandt et al. 2017; Jebelli and Lee 2019; Jia et al. 2011). EMG signals become missing due to either the disconnection of electrodes, artifacts, or during the collection of very-low-amplitude signals (Akmal et al. 2019). Sometimes, positions of some markers or electrodes can be missing for a long period of time during data collection due to the dynamic nature of the motion and extended occlusions (Liu and McMillan 2006), which can have significant effects on the conclusions drawn from the data.

Missing data associated with data collection process of these technologies is a common phenomenon that has made knowledge discovery in ergonomics a complicated process because most methodologies of knowledge extraction such as data mining, machine learning, pattern recognition, and statistical analysis, are sensitive to incomplete observations or sample instances (Young et al. 2011). The absence of data reduces statistical power, which refers to the possibility that the analysis results will reject the null hypothesis when it is false (Kang 2013). In WMSD studies, missing data can cause incompleteness and misrepresentation of the risk indicators, leading to inconsistent, biased estimation of the effects of the potential risk factors on the risk indicators, which can lead to erroneous conclusions. For this reason, missing data should be dealt with carefully to ensure reliable risk assessment.

Research Works Addressing Missing Data Problems

There is a diversity of methods for handling missing data in the existing literature. Some common methods include simply omitting observations with the missing data (Kang 2013), regression imputation (Button et al. 2013), mean substitution (Malhotra 1987), replacing the missing data with the last observed values (Hamer and Simpson 2009), and estimating the missing data using conditional distribution of the other variables (Gelman and Raghunathan 2001). These methods have been proven useful in handling missing data in many research specialties. However, they have some limitations for application to impute missing data in WMSD risk-assessment studies. In methods of omitting observations, the observations with missing data are simply removed, which decreases the sample size and adversely affects the statistical power. The regression imputation method does not add any new information except increasing the sample size and compromising the standard error. Mean substitution methods may cause inconsistent biases if there is a great inequality in the proportion of missing values of different variables. Methods replacing missing data with the last observed values assume that there will be no changes in the outcome. For a large proportion of missing data, methods using conditional distribution of the other variables may provide inconsistent bias because the relationship among the variables needs to be properly computed.

Moreover, the high-dimensionality or multimodality nature of WMSD risk-related data, where risk exposures can be obtained in form of different risk indicators collected at different working settings (often referred to as risk factors) and from different subjects, limit the applicability of the aforementioned methods in missing values imputation in that the latent relationship among the risk indicators and the factors cannot be well-captured (Gebregziabher and DeSantis 2010). Quite often, multiple risk-indicator data are collected for in-depth understanding of risks and workers' risk-taking behaviors. Data fusion can be an efficient method in this regard because by this method, risk-related information is integrated from multiple data sets based on the latent relationship among the risk factors and risk indicators such that a more complete representation of the measurements of the risk indicators can be produced, compared with what can be represented in incomplete and unfused data sets (Lahat et al. 2015).

Tensor Decomposition–Based Data Fusion

Tensor decomposition can accurately extract the interrelationships among various features from multiple data sets and learn the latent structures and collaborative relationships among those features to approximate the pattern of the data during fusion. Therefore, tensor decomposition can be a useful tool for fusion of high-dimensional data sets (Kolda and Bader 2009). Tensors are generalizations of matrices to higher dimensions and are a very powerful tool to model multidimensional data (De Lathauwer et al. 2000). Tensor decomposition was previously used for anomaly detection (Xie et al. 2017), object profiling (Charlier et al. 2018), pattern recognition (Xiong et al. 2010), filling in missing values (Dauwels et al. 2012), evolution prediction (Dunlavy et al. 2011), and posture recognition (Chen et al. 2017).

Canonical Polyadic Decomposition

CPD is one of the most widely used tensor decomposition techniques, which is generally suggested to estimate latent relationships among different features within a data set (Rabanser et al. 2017). CPD can effectively be used for imputing missing data because it captures the interactions among various features of a high-dimensional data sets during imputation (Dauwels et al. 2012). CPD has also been found to be effective in fusing multiple incomplete data sets and filling out missing data in signal processing and machine learning (Sidiropoulos et al. 2017).

CPD factorizes a tensor into a sum of component rank-one tensors (Carroll and Chang 1970; Hitchcock 1927). For example, in CPD, a given three-dimensional tensor $\mathbf{X} \in \mathbb{R}^{I \times J \times K}$ can be written

$$\mathbf{X} \approx \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \equiv [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] \tag{1}$$

where R = positive integer, which is referred to as the rank of a tensor. The rank of tensor $\mathbf{X} = R$ can be defined as the smallest number of a rank-one tensor that is required to represent the tensor $\mathbf{X}$ as their sum (Rabanser et al. 2017) (Fig. 1); a rank-one tensor is defined as a decomposition of an M-dimensional tensor into one outer product of M vectors. R is also called latent factor. I, J, and K are the sizes of the dimensions of the tensor $\mathbf{X}$ where, $\mathbf{a}_r \in \mathbb{R}^I$, $\mathbf{b}_r \in \mathbb{R}^J$, and $\mathbf{c}_r \in \mathbb{R}^K$ for $r = 1, \ldots, R$; r represents the indices of rank-one tensors; $\circ$ indicates the vector outer product; $\mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \in \mathbb{R}^{I \times J \times K}$ is an outer product of three vectors $\mathbf{a}_r$, $\mathbf{b}_r$, and $\mathbf{c}_r$ and is referred to as a rank-one tensor. $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are the factor matrices obtained for each dimension after decomposition of the tensor $\mathbf{X}$ and are referred to as the combination of the vectors from the rank-one components (Kolda and Bader 2009). Here, the factors matrices are $\mathbf{A} \in \mathbb{R}^{I \times R}$, $\mathbf{B} \in \mathbb{R}^{J \times R}$, and $\mathbf{C} \in \mathbb{R}^{K \times R}$.

The CPD of a three-dimensional tensor is illustrated in Fig. 1.

Applications of Data Fusion in Construction

In construction, data fusion has been used to improve the quality of load measurement during building cooling of automated systems and for efficient and effective management knowledge discovery (Huang et al. 2009; Soibelman et al. 2004). Other applications of data fusion in construction include automated identification, location estimation, dislocation detection of construction materials in jobsites (Razavi and Haas 2010, 2011), automated progress tracking of construction projects (Shahi et al. 2012), structural health monitoring (Soman et al. 2018), damage identification of civil structures (Anaissi et al. 2018), and construction productivity monitoring (Pradhan et al. 2011).

In construction ergonomics, a computationally efficient approach using tensor decomposition was developed to recognize construction workers' awkward postures (Chen et al. 2017). Fusion of data from continuous remote monitoring of construction workers' location and physiological status was used to identify safe and unsafe behaviors of construction workers (Cheng et al. 2012). Fusion of spatiotemporal and workers' thoracic posture data has been done for understanding the worker's activity type for productivity assessment (Cheng et al. 2013). A position and posture data fusion method was proposed for evaluation of construction workers' behavioral risks (Chen et al. 2019). A framework for location data fusion and pose estimation of excavators have been developed by Soltani et al. (2018).

Application of CPD in Treating Missing Data in WMSD Research

Recently, Dutta et al. (2020a) proposed a method of fusing multiple incomplete risk-related data sets that applied CPD to fuse and fill in missing data by leveraging the correlation among multiple risk indicators within those data sets (Dutta et al. 2020a). Multiple incomplete risk-related data sets were structured together to form a multidimensional tensor, which was then decomposed by CPD into factor matrices that extrapolated the latent structures and collaborative relationships between the risk factors and risk indicators. The factor matrices were subsequently used to reconstruct the tensor that had all the missing risk indicator values imputed and existing risk indicator values fused, readily available for risk assessment.

Problem Statement and Research Objective

Missing data can lead to misrepresentation of the risk indicators and underpowered analyses, which can provide erroneous conclusions on the contribution of potential risk factors on the imposed WMSD risk. In prior work, CPD has been shown to be promising in fusing multiple incomplete high-dimensional risk-related data sets and filling out missing data within those data sets based on the latent correlations among the risk factors and risk indicators for WMSD risk assessment (Dutta et al. 2020a). For validation, the fusion performance of the proposed method was analyzed by comparing the reconstructed complete data set by fusion with the original complete data set collected from the controlled lab experiment to demonstrate the reconstruction performance of the fusion process. For different proportions of missing values, the consistency of risk-assessment results obtained from the reconstructed complete data set by fusion to those obtained from the original complete data set was also analyzed to see if the fusion performance was consistent.

Although the previous work validated the proposed method in terms of consistency and reliability and showed the significance of the method in risk assessment, there remained a gap in the validation process. That is, regarding the numerical stability of CPD in fusion and given a different portion of missing data in multiple incomplete risk-related data sets, it is still unknown whether, at what portion of missing data, and to what extent, CPD is numerically stable in generating a set of fused risk indicators that outperform the ones with no fusion in the incomplete data sets in risk assessment. To fill in this gap, the objective of this research was to evaluate the numerical stability of the data fusion via CPD in improving and ensuring reliable WMSD risk assessment in construction when the quality of risk-related data sets becomes degraded due to the missing data in multiple inputs.

In other words, this work aims to investigate if, and to what extent, data fusion via CPD could reduce the inconsistencies in WMSD risk assessment made by unfused incomplete data sets and thus could enhance the risk-assessment accuracy. To do so, this research compared the consistencies of the risk-assessment results obtained from the fused complete and unfused incomplete data sets with those obtained from the original complete data sets collected from the experiment. The resulting knowledge would further validate the numerical stability of tensor decomposition–based data fusion and thereby revitalize the significance of using the method

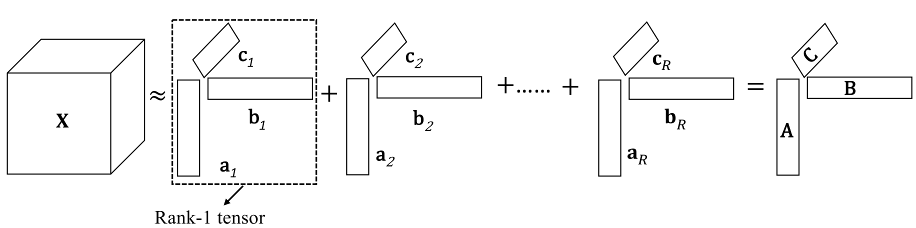

Fig. 1. CPD tensor decomposition.

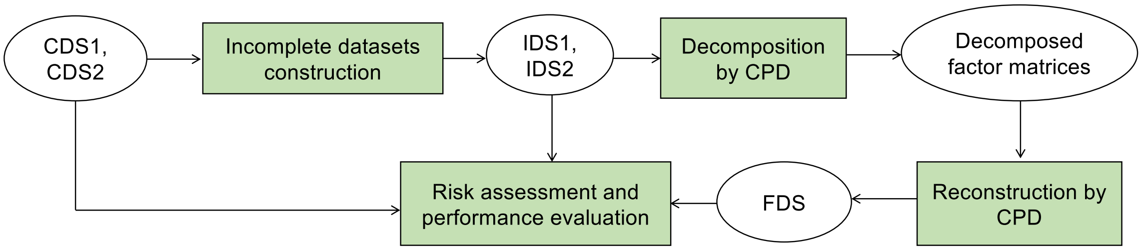

Fig. 2. Procedures of the experimental design.

as an alternative way of risk assessment when data collection process becomes affected by missing data.

Experimental Design and Implementation

The experimental design included the procedures depicted in Fig. 2. First, two risk-related data sets were collected from previous human subject laboratory experimental studies (Breloff et al. 2019; Dutta et al. 2020b; Kingston et al. 2016). The first one [Complete data set 1 (CDS1)] contained calculated knee rotation (kinematics) data, representing five knee rotational angles: flexion, abduction, adduction, and internal and external rotation collected at different roof slopes and working postures. The second one [Complete data set 2 (CDS2)] contained calculated EMG data describing muscle activation of five knee postural muscles: biceps femoris, rectus femoris, semitendinosus, vastus lateralis, and vastus medialis collected at the same roof slopes and working postures. Because awkward knee rotations and heightened muscle activation have been identified as two quantitative risk indicators of WMSDs in lower extremities and the roof slope and working posture are two WMSD risk factors, these data sets were representatives of WMSD risk-related data sets and therefore used in this study (Breloff et al. 2019; Dutta et al. 2020b; Kingston et al. 2016).

Then, a certain proportion of data were intentionally removed from both CDS1 and CDS2 to construct two incomplete data sets: Incomplete data set 1 (IDS1) and Incomplete data set 2 (IDS2). Next, CPD was applied to IDS1 and IDS2 to fuse the knee kinematics and EMG data and filling out the missing data within those data sets. Before the application of CPD, those data sets were structured as a multidimensional tensor. CPD first decomposed those data sets into factor matrices based on the latent relationship among different risk factors and the risk indicators. The factor matrices were then used to reconstruct the rotation (kinematics) and EMG data through fusion, leading to a new data set, referred to as the fused data set (FDS). In FDS, all the risk indicator (rotation and EMG) values were fused. Next, data sets IDS1, IDS2, and FDS were analyzed to produce risk-assessment results. The risk-assessment results obtained from FDS, IDS1, and IDS2 were then compared with those obtained from CDS1 and CDS2. This process was repeated with each time that IDS1 and IDS2 had a different portion of missing values removed from CDS1 and CDS2 to collectively evaluate whether the fused data sets outperformed the incomplete data sets in risk assessment.

Risk-Related Experimental Data Collection

Risk-related data were collected from prior human subject laboratory experimental studies. Nine healthy male volunteers [26.1 years (5.6 years), 180.2 cm (6.1 cm), and 99.7 kg (27.6 kg)] participated in this experiment. They simulated the shingle installation roofing task at three roof slopes (0°, 15°, and 30°) with two kneeling postures (static and dynamic). Each participant performed the task for five trials.

The knee kinematic data (segment endpoint data from motion capture system) were collected using a VICON optical motion analysis system with 14 MX Vicon cameras (Oxford, United Kingdom); 42 retroreflective motion capture markers were placed bilaterally on the participants' hip joints, thighs, knee joints, shanks, ankles, toes, heels, and feet, following the approach used in a similar study done by Pollard et al. (2011). The collected 3D coordinates of these markers were used to calculate the five knee rotation angles that represent the WMSD risk associated with awkward and extreme kneeling postures. In the knee kinematics data set (CDS1), each data point represents the maximum knee rotation angle value of a knee rotation for a certain subject at a specific slope, posture, and trial. Filters were used to smoothen these trajectory data for noise attenuation. More details on the knee kinematics data collection and data processing are available from Breloff et al. (2019).

Muscular activities of the 10 bilateral thigh muscles were recorded using a surface EMG system (Noraxon Desktop Direct Transmission System with myoMUSCLE Master software version 3.10, Scottsdale, Arizona). Surface EMG Ag/AgCl electrodes were placed on the palpated muscle bellies of each of the specified muscles according to the instructions outlined by Reichert and Stelzenmueller (2011). The EMG data represent the WMSD risk associated with heightened activation of knee postural muscles that might cause knee joint overloading. In the EMG data set (CDS2), each data point represents the maximum normalized EMG value of a knee postural muscle of a certain subject at a specific slope, posture, and trial. Raw EMG signals were also filtered and rectified for removing short-term fluctuation and noise attenuation. More details on the knee muscle EMG data collection and data processing have been given by Dutta et al. (2020b).

Incomplete Data Set Construction (Data Removal)

A proportion of data was intentionally removed from both CDS1 and CDS2 to make them incomplete. The removed data were randomly selected from the data sets. The proportions of the missing data from both data sets are provided in Table 1. This way, 14 incomplete data sets, seven for knee kinematics and EMG data each, were constructed with different portions of missing values.

Fusion of the Data Sets Using CPD Tensor Decomposition

The resulting incomplete kinematics and EMG data sets were then used for fusion. Because the knee rotation (kinematics) and EMG data were collected from the same subjects at the same work settings (i.e., roof slope, working posture, and trial), they were first combined by the early integration data fusion method (Žitnik and Zupan 2014). In this method, the column vectors of these two types of risk indicators were concatenated to form a single set of

Table 1. Proportion of missing data

Type	Data set	
	Knee kinematics (IDS1)	Knee EMG (IDS2)
Missing proportion (%)	10	10
	20	20
	30	30
	40	40
	50	50
	60	60
	70	70

length 10 vectors, which was referred as risk indicators in this study. The following 49 incomplete data sets were considered for tensor construction and fusion (Table 2).

Because the risk indicators were measured from different subjects at different working postures, roof slopes, and multiple trials, the data were then represented as a five-dimensional incomplete tensor $\mathbf{T} \in \mathbb{R}^{I_1 \times I_2 \times I_3 \times I_4 \times I_5}$, where, I_1 = number of subjects (total = 9), I_2 = number of risk indicators [kinematics and EMG: total = 10 (five knee rotation angles and five knee muscle EMG)], I_3 = number of trials (total = 5), I_4 = number of postures [total = 2 (static and dynamic)], and I_5 = number of slopes [total = 3 (0°, 15°, and 30°)]. In short, the tensor $\mathbf{T}$ maps the knee rotation and EMG measurements of nine subjects obtained from five trials performed on three roof slopes using two postures. Tensor $\mathbf{T}$ is a multidimensional representation of certain IDS1 and IDS2.

In the similar way, the complete data sets collected from the experiment were also constructed to a five-dimensional tensor $\mathbf{Y} \in \mathbb{R}^{I_1 \times I_2 \times I_3 \times I_4 \times I_5}$ as ground truth, which is a multidimensional representation of CDS2 and CDS2 and was used later for validation of the fusion method.

Once a tensor $\mathbf{T}$ was constructed, CPD tensor decomposition was applied to decompose the tensor $\mathbf{T}$ into five factor metrics, representing information in each mode or dimension. The factor matrices were represented as matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, and $\mathbf{E}$, providing latent information in modes of subjects, risk indicators, trials, postures, and slopes, respectively, where factor metrics are $\mathbf{A} \in \mathbb{R}^{I_1 \times R}$, $\mathbf{B} \in \mathbb{R}^{I_2 \times R}$, $\mathbf{C} \in \mathbb{R}^{I_3 \times R}$, $\mathbf{D} \in \mathbb{R}^{I_4 \times R}$, and $\mathbf{E} \in \mathbb{R}^{I_5 \times R}$. Each factor matrix represented the latent relationship between the corresponding dimension and all the other dimensions captured by latent factor R.

Proper R values ensures a unique mapping of the original tensor to the decomposed tensor. In this study, the R value was computed by applying the core consistency diagnostic method (CORCONDIA) method described by Bro and Kiers (2003), where a parameter named core consistency was measured to ensure an optimal representation of the original tensor as a sum of rank-one tensors. To compute the smallest number of the rank-one tensor, a one-component model ($R = 1$) was attempted first. Then, the values of R were gradually increased until the core consistency value reached close to 100%, which indicates an appropriate CPD model that properly explains the data variability. R values with the highest core consistencies were considered as the smallest number of rank-one tensors.

Table 2. Datasets with missing values for tensor construction

% of missing data from kinematics dataset	% of missing data from EMG dataset						
10	10	10	10	10	10	10	10
20	20	20	20	20	20	20	20
30	30	30	30	30	30	30	30
40	40	40	40	40	40	40	40
50	50	50	50	50	50	50	50
60	60	60	60	60	60	60	60
70	70	70	70	70	70	70	70

Factor matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, and $\mathbf{E}$ were then computed by solving the following optimization equation:

$$\arg\min \left\| \mathbf{T} - \sum_{r=1}^{R} \mathbf{A}_r \circ \mathbf{B}_r \circ \mathbf{C}_r \circ \mathbf{D}_r \circ \mathbf{E}_r \right\| \tag{2}$$

To solve Eq. (2), the alternating least-squares algorithm was applied using the Tensor Toolbox from MATLAB version 9.3 (Bader and Kolda 2019).

The reconstructed tensor $\mathbf{T}' \in \mathbb{R}^{I_1 \times I_2 \times I_3 \times I_4 \times I_5}$ was then computed using the following equation:

$$\mathbf{T}' = \sum_{r=1}^{R} \mathbf{A}_r \circ \mathbf{B}_r \circ \mathbf{C}_r \circ \mathbf{D}_r \circ \mathbf{E}_r \tag{3}$$

This reconstructed tensor was the FDS that approximately mapped the tensor $\mathbf{T}$ and is represented as $\mathbf{T}' \in \mathbb{R}^{I_1 \times I_2 \times I_3 \times I_4 \times I_5}$. In this tensor, CPD imputed all the missing values and reconstructed the existing risk indicators minimizing the reconstruction error. For this reconstruction, Tensorlab version 3.0 (a MATLAB package for tensor computation) was used (Debals et al. 2019). The reconstructed fused rotation and EMG data were then extracted from FDS to a spreadsheet to generate the risk-assessment results.

Performance Evaluation of Data Fusion

To examine the performance of fused complete and unfused incomplete data sets in risk assessment, the following four consistency measures were computed as performance metrics: (1) consistency of the effects of roof slope and working posture on five knee rotation angles between CDS1 and FDS (Con_{Rot_f}), (2) consistency of the effects of roof slope and working posture on five knee postural muscles' EMG between the CDS2 and FDS (Con_{EMG_f}), (3) consistency of the effects of roof slope and working posture on five knee rotation angles between the CDS1 and IDS1 (Con_{Rot_i}), and (4) consistency of the effects of roof slope and working posture on five knee postural muscles' EMG between CDS2 and IDS2 (Con_{EMG_i}).

Con_{Rot_f} was computed using Eq. (4)

$$Con_{Rot\_f} = \frac{\sum_{i=1}^{P} f(m_i, m_i')}{P} \times 100 \tag{4}$$

where P = total number of possible effects of roof slope, working posture, and their interactions on five knee rotation angles [total 15 = 3 factors (slope, posture, and slope–posture interaction) ×5 knee rotation angles]; m_i = effects of the factors (slope and posture) on the response variables (five knee rotation angles) obtained from CDS1; m_i' = effects of the factors (slope and posture) on the response variables (five knee rotation angles) obtained from FDS; and

$$f(m_i, m_i') = \begin{pmatrix} 1, & \text{if } m_i = m_i' \\ 0, & \text{if } m_i \neq m_i' \end{pmatrix}$$

Con_{EMG_f} was computed using Eq. (5)

$$Con_{EMG\_f} = \frac{\sum_{i=1}^{Q} f(n_i, n_i')}{Q} \times 100 \tag{5}$$

where Q = total number of possible effects of roof slope, working posture, and their interactions on EMG values of five knee postural muscles [total $15 = 3$ factors (slope, posture, and slope–posture interaction $\times 5$ knee postural muscles]; n_i = effects of the factors (slope and posture) on the response variables (EMG values of five knee postural muscles) obtained from the CDS2; n_i' = effects of the factors (slope and posture) on the response variables (EMG values of five knee postural muscles) obtained from FDS; and

$$f(n_i, n_i') = \begin{pmatrix} 1, & \text{if } n_i = n_i' \\ 0, & \text{if } n_i \neq n_i' \end{pmatrix}$$

Con_{Rot_i} was computed using Eq. (6)

$$Con_{Rot\_i} = \frac{\sum_{i=1}^{P} f(m_i, k_i)}{P} \times 100 \tag{6}$$

where P = total number of possible effects of roof slope, working posture, and their interactions on five knee rotation angles [total $15 = 3$ factors (slope, posture, and slope–posture interaction) $\times 5$ knee rotation angles]; m_i = effects of the factors (slope and posture) on the response variables (five knee rotation angles) obtained from CDS1; k_i = effects of the factors (slope and posture) on the response variables (five knee rotation angles) obtained from the IDS1; and

$$f(m_i, k_i) = \begin{pmatrix} 1, & \text{if } m_i = k_i \\ 0, & \text{if } m_i \neq k_i \end{pmatrix}$$

Con_{EMG_i} was computed using Eq. (7)

$$Con_{EMG\_i} = \frac{\sum_{i=1}^{Q} f(n_i, l_i)}{Q} \times 100 \tag{7}$$

where Q = total number of possible effects of roof slope, working posture, and their interactions on EMG values of five knee postural muscles [total $15 = 3$ factors (slope, posture, and slope–posture interaction $\times 5$ knee postural muscles]; n_i = effects of the factors (slope and posture) on the response variables (EMG values of five knee postural muscles) obtained from CDS2; l_i = effects of the factors (slope and posture) on the response variables (EMG values of five knee postural muscles) obtained from IDS2; and

$$f(n_i, l_i) = \begin{pmatrix} 1, & \text{if } n_i = l_i \\ 0, & \text{if } n_i \neq l_i \end{pmatrix}$$

Results

Consistency Curves of Rotation for Different Proportions of Missing Values

Fig. 3 shows the consistency curves of rotation for different proportions of missing values in IDS1 and IDS2. Each plot shows two consistency curves. The continuous line is for the fused rotation values in FDS. The dashed line is for the unfused rotation values in IDS1. In general, for different proportions of missing data in IDS1 and IDS2, the fused rotation values in FDS provided more consistent risk-assessment results than the unfused ones in IDS1, although for some proportions of missing values (e.g., 10% and 20% missing data from IDS1 with 60% and 70% missing data from IDS2 and 30% missing data from IDS1 with 30%–50% missing data from IDS2), both unfused and fused-based assessment results demonstrated the similar consistencies. These findings suggested better performance in risk assessment by the risk indicators of the fused data set that might provide more consistent risk-assessment results than the unfused ones in the incomplete data set.

Consistency Curves of EMG for Different Proportions of Missing Values

Fig. 4 shows the consistency curves of EMG for different proportions of missing values in IDS1 and IDS2. Similarly, each plot shows two consistency curves. The continuous line is for the fused EMG values in FDS. The dashed line is for the unfused EMG values in IDS2. For 10%–40% missing data from IDS1 and 10%–60% missing data from IDS2, the performance of the fusion method was poor where the unfused EMG values of IDS2 provided better consistency results than the fused EMG values in FDS. However, with the increase of the missing data in IDS1 (from 50% to 70%) and for a large percentage of missing data in IDS2 (50%–70%), the fused EMG data of FDS demonstrated better performance. In general, for EMG data, the fused EMG values demonstrated higher consistencies when large proportions of data (>50%) were missing in IDS1 and IDS2. These findings indicated that the fused risk indicators in the fused data set performed better in risk assessment than the unfused ones in the incomplete data set even when a large proportion of data were missing.

In Fig. 5, a summary matrix of the risk-assessment performance by the fused versus unfused rotation and fused versus unfused EMG data have been provided. Here, the equals sign indicates that the fused risk indicators provide similar results to the unfused ones, a plus sign indicates that fused risk indicators outperform the unfused ones; and finally, a minus sign indicates that fused risk indicators underperform the unfused ones.

Discussion

Data collection for WMSD risk assessment either from laboratory experiments or from construction sites can potentially have missing data due to any technology or human-induced error. Quite often, risks cannot be fully quantified with a single risk indicator, and thus multiple heterogeneous risk indicators data are collected for risk assessment. As mentioned, a previous study indicated that data fusion by applying CPD can be efficient to fill in missing data in such data sets because the latent relationship among different risk factors and risk indicators available in those data sets are reserved.

This research evaluated the numerical stability of data fusion in WMSD risk assessment that applies CPD to fuse multiple risk-related data sets and fills in missing data. More specifically, despite using malformed experimental data with missing values as input, if and to what extent fusion through CPD was numerically stable to provide more reliable risk-assessment result compared with incomplete data sets was analyzed in this study. For that purpose, two risk-related data sets containing roofers' knee kinematics and EMG data collected from previous experimental studies were used for fusion. The effects of roof slope and working posture on five knee rotation angles and five knee postural muscles' EMG were measured as risk-assessment results. Different proportions of data were intentionally removed from those data sets to make them incomplete. CPD was thus applied to fuse those incomplete data sets.

Risk-assessment results were computed from the original complete, unfused incomplete, and fused complete data sets. The performance of the fusion was evaluated based on how consistent the risk-assessment results obtained from the fused complete and the unfused incomplete data sets were with those obtained from the original complete data sets that were collected from experimental studies. For this purpose, four consistency measures, namely consistency of the effects of roof slope and working posture on five knee rotation angles between the original complete and fused data sets, consistency of the effects of roof slope and working posture on five knee postural muscles' EMG between the original complete and fused data sets,

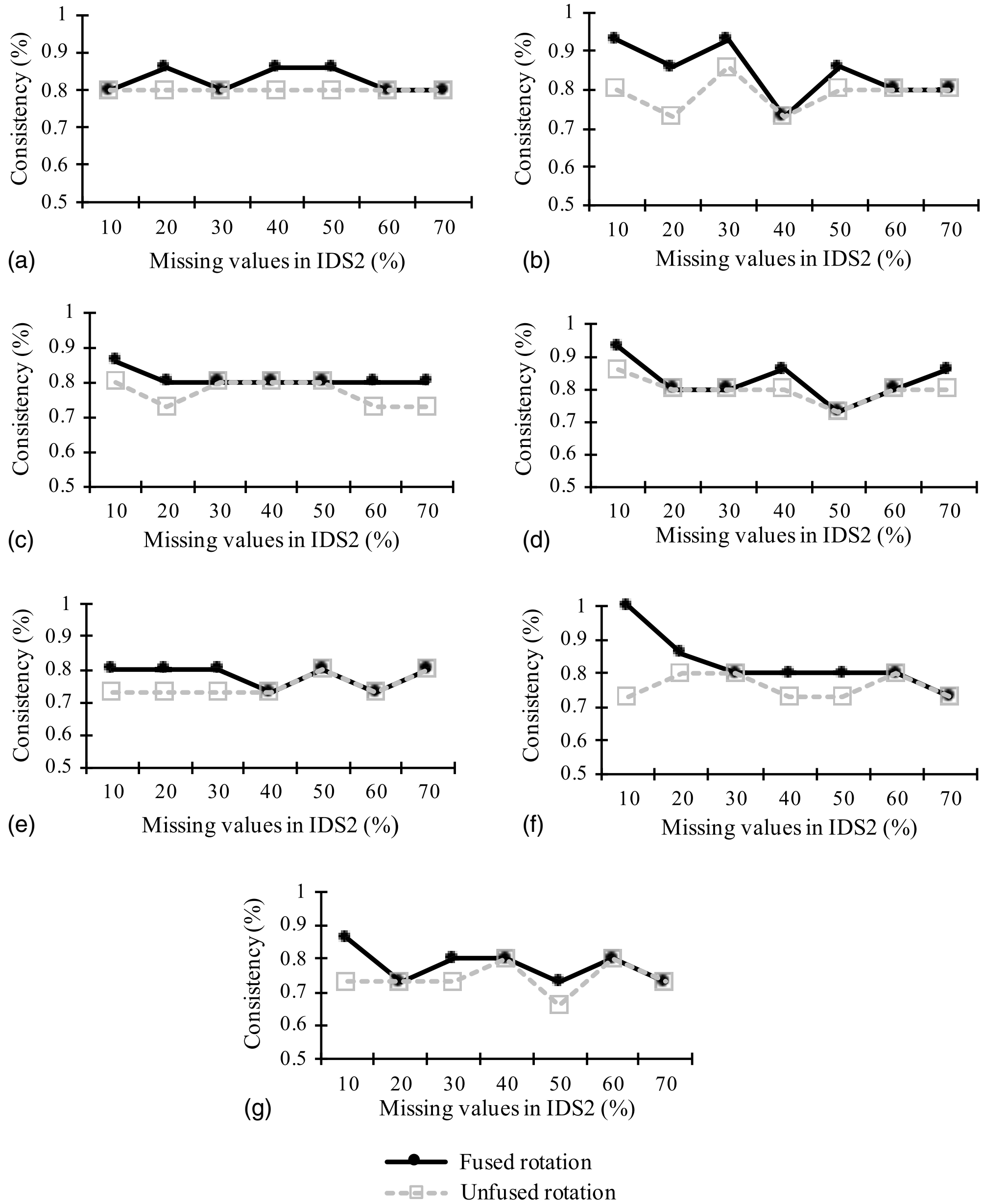

Fig. 3. Consistency curves of rotation for different proportions of missing values in IDS1: (a) 10%; (b) 20%; (c) 30%; (d) 40%; (e) 50%; (f) 60%; and (g) 70%.

consistency of the effects of roof slope and working posture on five knee rotation angles between the original complete and unfused incomplete data sets, and consistency of the effects of roof slope and working posture on five knee postural muscles' EMG between the original complete and unfused incomplete data sets, were computed as performance metrics.

For simulating incomplete data sets, this study adopted a random data removal process. In ergonomic risk-assessment studies that involve human subjects and technologies, the common reasons for data dropout are technology-induced errors (e.g., disconnection of sensors or instrument failures) and experimental conditions (e.g., nature of posture, poor line of sight, extra reflections, or occlusions). They generally represent a random process (Dziura et al. 2013; Kang 2013; Kaushal 2014), and there is no specific pattern of data dropout that leads to missing data. Therefore, a random data removal process was used to simulate incomplete data set and to keep the methodology aligned with the scope of this study. However, further studies are required to analyze the performance of data fusion if missing data do not follow a random pattern.

The consistency results demonstrated that for significant portions of missing values from both data sets, the fused risk indicators of the fused data sets could provide more reliable risk-assessment results than the unfused risk indicators in the incomplete data sets. The risk-assessment results obtained from the fused data sets were more consistent with those obtained from the original complete data sets compared with the unfused incomplete data sets. For example, for 10%–70% missing data in both kinematics and EMG data sets, more consistent risk effects on the knee rotations were obtained when rotation and EMG data were fused rather than from the unfused rotation data that were in the incomplete rotation data set. The probable reason for better performance by fusion was the

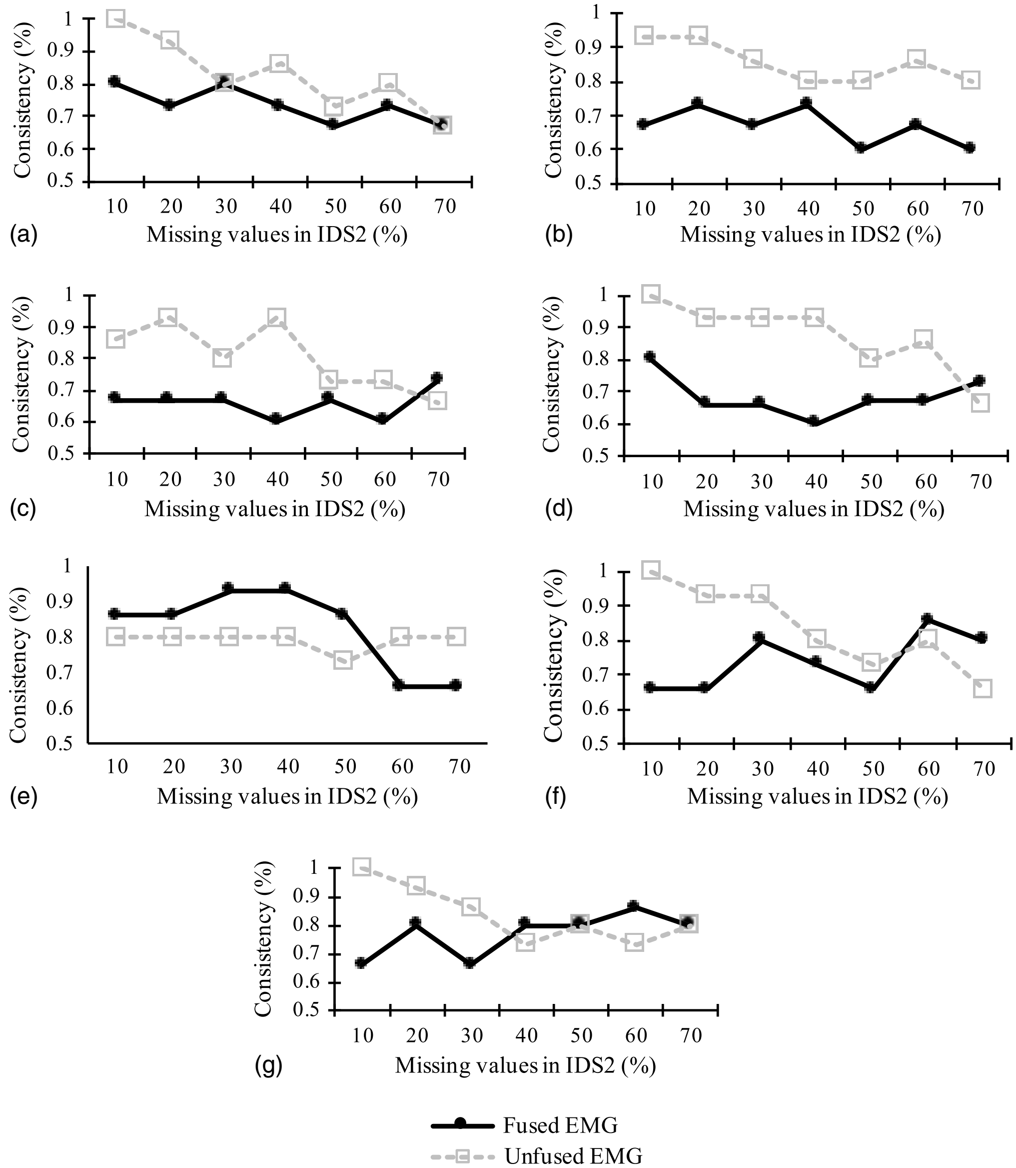

Fig. 4. Consistency curves of EMG for different proportions of missing values from IDS1: (a) 10%; (b) 20%; (c) 30%; (d) 40%; (e) 50%; (f) 60%; and (g) 70%.

Missing data from EMG dataset (%)	Missing data from rotation dataset (%): 10	20	30	40	50	60	70
10	=	+	+	+	+	+	+
20	+	+	+	=	+	+	=
30	=	+	=	=	+	=	+
40	+	=	=	+	=	+	=
50	+	+	=	=	=	+	+
60	=	=	+	=	=	=	=
70	=	=	+	+	=	=	=

(a)

Missing data from EMG dataset (%)	Missing data from rotation dataset (%): 10	20	30	40	50	60	70
10	−	−	−	−	−	−	−
20	−	−	−	−	−	−	−
30	=	−	−	−	−	−	−
40	−	−	−	−	−	−	+
50	−	−	−	−	−	−	=
60	−	−	−	−	+	+	+
70	=	−	+	+	+	+	=

(b)

Fig. 5. Summary matrix of the risk-assessment performance of (a) fused versus unfused rotation; and (b) fused versus unfused EMG.

small reconstruction error ensured by the factor matrices generated in CPD process. CPD could minimize the reconstruction error during fusion, which improved the performance of the fused data in risk assessment despite a significant proportion of data being missing. The better consistency by fusion is an indicator of good fusion performance of the CPD tensor decomposition–based data fusion. A previous study also found that fusion through CPD could provide reliable risk-assessment results 87%–70% consistent to those provided by the original experimental data sets for up to 40% missing data (Dutta et al. 2020a).

For 10%–40% missing data from rotation data set and 10%–60% missing data from EMG data set, the fused data did not outperform the unfused data in the incomplete data sets when the risk effects were analyzed on muscles' EMG. However, with further increase of the missing data from 50% to 70% in both data sets, the fused EMG data started showing better consistency than the unfused ones in the incomplete EMG data set, although the previous study showed that for more than 50% missing data, the performance of the fusion started to decrease (Dutta et al. 2020a).

Despite the degradation of the fusion performance, in this study, the fused EMG performed better compared with unfused ones in the incomplete EMG data sets with 50%–70% of missing data. The possible reason can be attributed to the low variance of the EMG data in the incomplete data sets, which helped capture the effects of the risk factors on the imposed risk using the incomplete data set itself unless a significant proportion of data were missing. But due to the comparatively small reconstruction errors in the fusion, the fused data performed better in risk assessment despite the significant proportion of missing data.

Overall, for the specific data sets used in this study, the fusion-based risk assessment outperformed the one with no fusion, even when a large proportion of data (>50%) were missing. The findings of this study showed the numerical stability of fusion via CPD, which can minimize the inconsistencies in the risks assessment results obtained from the incomplete data sets. The findings also suggested that applying data fusion through CPD can be a better alternative and effective approach for WMSD risk assessment in construction rather than using data sets with missing values when the data collection is badly affected due to the presence of missing data.

Nevertheless, there are some limitations in this study. First, ground-truth data were collected from an experimental study performed in a controlled laboratory setting, not from a real construction site. Second, only aggregated metrics, maximum knee rotation, and normalized muscle activity measurement, were considered as summary data for fusion; detailed time-series features were not analyzed. Missing data were simulated by removing different proportions of data from the aggregated metric–based summary data, not from the time-series data initially collected in the experiment. Even after careful experimental design and implementation, about 2% and 20% data were missing in the rotation and EMG summary data, respectively, because some of the trial data were entirely missing. But for the sake of obtaining complete data sets for validation and comparison, observations with missing data were intentionally removed from the summary data in this study. The synchronization of multiple time-series data sets with different time stamps and sampling rates and the corresponding data processing during fusion are worthwhile for further investigation.

Third, the raw EMG data were filtered using traditional well-established approaches where high-pass and low-pass filters were used to remove the low-frequency motion artifacts and to smooth the signal, respectively. Also, a moving-average filter was used to remove short-term fluctuations that may generate from intrinsic noise sources such as inherent noises from the electronic equipment or noises generated in the skin–electrode interface. However, some other extrinsic sources of noises, such as power line noises, surrounding uncontrolled electromagnetic environments, and their effects on the EMG, were not studied because they were out of scope of this study. Several new algorithms recently proposed for intrinsic and extrinsic noise removal, including but are not limited to independent components analysis (Jiang et al. 2019), recurrent neural network (Machado et al. 2020), independent vector analysis (Wang et al. 2020), and noise reference signal–based denoising method (Lee et al. 2020), may be further investigated for processing and quality improvements of EMG data collected from construction sites.

Conclusion and Future Extension

This research evaluated the numerical stability of fusion through CPD for WMSD risk assessment in construction when the data collected suffer from missing data. The findings of this study suggest that CPD is numerically stable to generate a set of fused risk indicators based on the interrelations among the risk factors, and the risk indicators that outperformed the unfused risk indicators of the incomplete data sets in WMSD risk assessment. This signifies the effectiveness of using data fusion through CPD as an alternative approach of filling out missing data in WMSD risk-assessment studies.

The fused data set was found to be more consistent to the complete experimental data set than the incomplete data set in terms of generating risk-assessment results. Even for a large proportion (>50%) of missing data, the fused data provided more reliable risk-assessment results than the unfused ones in the incomplete data sets. These findings ensure that data fusion through CPD can help enable accurate and reliable assessment of work-related risk factors for WMSDs among construction workers when the data collection is affected by missing data.

Despite the fact that fusing incomplete ergonomic data may yield few practical values for daily work on a construction site, it is indispensable for many research studies to make most use of construction ergonomic data collected from lab experiments or field operations for reliable risk assessment of different constriction tasks by facilitating knowledge discovery from a complete and more accurate data set without losing any latent pattern or valuable information of the data. Adequately addressing missing values will ensure reasonable support of research hypotheses, which is very important for proper risk identification and intervention development for alleviating the risk. By proper imputation of missing data preserving the latent relationship among different features of the data set, the assessed method may also help in more accurate risk prediction, awkward posture recognition, worker's productivity assessment, automated ergonomic risk monitoring, and construction activity recognition using machine learning models in the field of occupational safety and health.

In the future, the method of fusion of multiple WMSD risk indicators by representing them as a high-dimensional tensor and then decomposing them for fusion can be expanded to explore similar studies related to other body parts, such as low back, feet, and shoulder injuries of construction workers, including roofers. Moreover, the experiment will be conducted in a real construction site involving the participation of professional roofers, and data will be collected onsite using an inertial measurement units (IMU)–based motion capture system. This will enable to explore the performance of data fusion to handle possible challenges involved in data collection in a real construction site including missing data, noise, occlusion, and sunlight issues for accurate and reliable risk assessment. The framework of applying data fusion in handling missing data may also be extended to other similar use cases in the fields of medicine, clinical research, sports science, human factors research, and human–computer interaction where motion capture technology is used to track and record human motion.

Data Availability Statement

Some or all data, models, or code generated or used during the study are available from the corresponding author by request (experimental data).

Acknowledgments

The authors acknowledge the support of the National Institute for Occupational Safety and Health (NIOSH), who funded this research. The findings and conclusions in this research are those of the authors and do not necessarily represent the official position of the National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention.

References

Akmal, M., S. Zubair, M. Jochumsen, E. N. Kamavuako, and I. K. Niazi. 2019. "A tensor-based method for completion of missing electromyography data." *IEEE Access* 7: 104710–104720. https://doi.org/10.1109/ACCESS.2019.2931371.

Anaissi, A., M. Makki Alamdari, T. Rakotoarivelo, and N. Khoa. 2018. "A tensor-based structural damage identification and severity assessment." *Sensors* 18 (2): 111. https://doi.org/10.3390/s18010111.

Bader, B., and T. Kolda. 2019. "MATLAB tensor toolbox version 2.4." Accessed May 14, 2019. https://www.sandia.gov/~tgkolda/TensorToolbox/index-2.6.html.

Brandt, M., L. L. Andersen, A. Samani, M. D. Jakobsen, and P. Madeleine. 2017. "Inter-day reliability of surface electromyography recordings of the lumbar part of erector spinae longissimus and trapezius descendens during box lifting." *BMC Musculoskeletal Disord.* 18 (1): 1–16. https://doi.org/10.1186/s12891-017-1872-y.

Breloff, S. P., A. Dutta, F. Dai, E. W. Sinsel, C. M. Warren, X. Ning, and J. Z. Wu. 2019. "Assessing work-related risk factors for musculoskeletal knee disorders in construction roofing tasks." *Appl. Ergon.* 81 (Nov): 102901. https://doi.org/10.1016/j.apergo.2019.102901.

Bro, R., and H. A. Kiers. 2003. "A new efficient method for determining the number of components in PARAFAC models." *J. Chemom. J. Chemom. Soc.* 17 (5): 274–286. https://doi.org/10.1002/cem.801.

Brown, M. L., and J. F. Kros. 2003. "Data mining and the impact of missing data." In *Industrial management & data systems*. Bingley, UK: Emerald Publishing.

Button, K. S., J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò. 2013. "Power failure: Why small sample size undermines the reliability of neuroscience." *Nat. Rev. Neurosci.* 14 (5): 365–376. https://doi.org/10.1038/nrn3475.

Carroll, J. D., and J.-J. Chang. 1970. "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition." *Psychometrika* 35 (3): 283–319. https://doi.org/10.1007/BF02310791.

Charlier, J., R. State, and J. Hilger. 2018. "Non-negative paratuck2 tensor decomposition combined to LSTM network for smart contracts profiling." In *Proc., 2018 IEEE Int. Conf. on Big Data and Smart Computing (BigComp)*, 74–81. New York: IEEE.

Chen, H., X. Luo, Z. Zheng, and J. Ke. 2019. "A proactive workers' safety risk evaluation framework based on position and posture data fusion." *Autom. Constr.* 98 (Feb): 275–288. https://doi.org/10.1016/j.autcon.2018.11.026.

Chen, J., J. Qiu, and C. Ahn. 2017. "Construction worker's awkward posture recognition through supervised motion tensor decomposition." *Autom. Constr.* 77 (May): 67–81. https://doi.org/10.1016/j.autcon.2017.01.020.

Cheng, T., G. C. Migliaccio, J. Teizer, and U. C. Gatti. 2012. "Data fusion of real-time location sensing and physiological status monitoring for ergonomics analysis of construction workers." *J. Comput. Civ. Eng.* 27 (3): 320–335. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000222.

Cheng, T., J. Teizer, G. C. Migliaccio, and U. C. Gatti. 2013. "Automated task-level activity analysis through fusion of real time location sensors and worker's thoracic posture data." *Autom. Constr.* 29 (Jan): 24–39. https://doi.org/10.1016/j.autcon.2012.08.003.

Data, M. C. 2016. *Secondary analysis of electronic health records*. New York: Springer.

Dauwels, J., L. Garg, A. Earnest, and L. K. Pang. 2012. "Tensor factorization for missing data imputation in medical questionnaires." In *Proc., 2012 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2109–2112. New York: IEEE.

Debals, O., F. Van Eeghem, N. Vervliet, and L. De Lathauwer. 2019. "Tensor computations using Tensorlab." Accessed May 19, 2019. https://www.tensorlab.net/demos/tutorial.pdf.

De Lathauwer, L., B. De Moor, and J. Vandewalle. 2000. "On the best rank-1 and rank-(r 1, r 2, . . . , rn) approximation of higher-order tensors." *SIAM J. Matrix Anal. Appl.* 21 (4): 1324–1342. https://doi.org/10.1137/S0895479898346995.

Dunlavy, D. M., T. G. Kolda, and E. Acar. 2011. "Temporal link prediction using matrix and tensor factorizations." *ACM Trans. Knowl. Discovery Data (TKDD)* 5 (2): 1. https://doi.org/10.1145/1921632.1921636.

Dutta, A., S. P. Breloff, F. Dai, E. W. Sinsel, R. E. Carey, C. M. Warren, and J. Z. Wu. 2020a. "Fusing imperfect experimental data for risk assessment of musculoskeletal disorders in construction using canonical polyadic decomposition." *Autom. Constr.* 119 (Nov): 103322. https://doi.org/10.1016/j.autcon.2020.103322.

Dutta, A., S. P. Breloff, F. Dai, E. W. Sinsel, C. M. Warren, R. E. Carey, and J. Z. Wu. 2020b. "Effects of working posture and roof slope on activation of lower limb muscles during shingle installation." *Ergonomics* 63 (9): 1182–1193. https://doi.org/10.1080/00140139.2020.1772378.

Dziura, J. D., L. A. Post, Q. Zhao, Z. Fu, and P. Peduzzi. 2013. "Strategies for dealing with missing data in clinical trials: From design to analysis." *Yale J. Biol. Med.* 86 (3): 343.

Gebregziabher, M., and S. M. DeSantis. 2010. "Latent class based multiple imputation approach for missing categorical data." *J. Stat. Plann. Inference* 140 (11): 3252–3262. https://doi.org/10.1016/j.jspi.2010.04.020.

Gelman, A., and T. E. Raghunathan. 2001. "Using conditional distributions for missing-data imputation." *Stat. Sci.* 15 (Jun): 268–269.

Hamer, R. M., and P. M. Simpson. 2009. "Last observation carried forward versus mixed models in the analysis of psychiatric clinical trials." *Am. J. Psychiatry* 166 (6): 639–641. https://doi.org/10.1176/appi.ajp.2009.09040458.

Hitchcock, F. L. 1927. "The expression of a tensor or a polyadic as a sum of products." *J. Math. Phys.* 6 (1–4): 164–189. https://doi.org/10.1002/sapm192761164.

Huang, G., S. Wang, F. Xiao, and Y. Sun. 2009. "A data fusion scheme for building automation systems of building central chilling plants." *Autom. Constr.* 18 (3): 302–309. https://doi.org/10.1016/j.autcon.2008.09.003.

Jebelli, H., and S. Lee. 2019. "Feasibility of wearable electromyography (EMG) to assess construction workers' muscle fatigue." In *Advances in informatics and computing in civil and construction engineering*, 181–187. New York: Springer.

Jia, B., S. Kim, and M. A. Nussbaum. 2011. "An EMG-based model to estimate lumbar muscle forces and spinal loads during complex, high-effort tasks: Development and application to residential construction using prefabricated walls." *Int. J. Ind. Ergon.* 41 (5): 437–446. https://doi.org/10.1016/j.ergon.2011.03.004.

Jiang, X., G.-B. Bian, and Z. Tian. 2019. "Removal of artifacts from EEG signals: A review." *Sens. Multidiscip. Digital Publ. Inst.* 19 (5): 987. https://doi.org/10.3390/s19050987.

Kang, H. 2013. "The prevention and handling of the missing data." *Korean J. Anesthesiol* 64 (5): 402–406. https://doi.org/10.4097/kjae.2013.64.5.402.

Kaushal, S. 2014. "Missing data in clinical trials: Pitfalls and remedies." *Int. J. Appl. Basic Med. Res.* 4 (3): 6.

Kingston, D. C., L. M. Tennant, H. C. Chong, and S. M. Acker. 2016. "Peak activation of lower limb musculature during high flexion kneeling and transitional movements." *Ergonomics* 59 (9): 1215–1223. https://doi.org/10.1080/00140139.2015.1130861.

Kolda, T. G., and B. W. Bader. 2009. "Tensor decompositions and applications." *SIAM Rev.* 51 (3): 455–500. https://doi.org/10.1137/07070111X.

Lahat, D., T. Adali, and C. Jutten. 2015. "Multimodal data fusion: An overview of methods, challenges, and prospects." *Proc. IEEE* 103 (9): 1449–1477. https://doi.org/10.1109/JPROC.2015.2460697.

Lee, G., B. Choi, H. Jebelli, C. Ryan Ahn, and S. Lee. 2020. "Noise reference signal–based denoising method for EDA collected by multimodal biosensor wearable in the field." *J. Comput. Civ. Eng.* 34 (6): 4020044. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000927.

Liu, G., and L. McMillan. 2006. "Estimation of missing markers in human motion capture." *Vis. Comput.* 22 (9–11): 721–728. https://doi.org/10.1007/s00371-006-0080-9.

Machado, J., M. C. Tosin, L. B. Bagesteiro, and A. Balbinot. 2020. "Recurrent neural network for contaminant type detector in surface electromyography signals." In *Proc., 42nd Annual Int. Conf. of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 3759–3762. New York: IEEE.

Malhotra, N. K. 1987. "Analyzing marketing research data with incomplete information on the dependent variable." *J. Market. Res.* 24 (1): 74–84. https://doi.org/10.1177/002224378702400107.

Pollard, J. P., W. L. Porter, and M. S. Redfern. 2011. "Forces and moments on the knee during kneeling and squatting." *J. Appl. Biomech.* 27 (3): 233–241. https://doi.org/10.1123/jab.27.3.233.

Pradhan, A., B. Akinci, and C. T. Haas. 2011. "Formalisms for query capture and data source identification to support data fusion for construction productivity monitoring." *Autom. Constr.* 20 (4): 389–398. https://doi.org/10.1016/j.autcon.2010.11.009.

Rabanser, S., O. Shchur, and S. Günnemann. 2017. "Introduction to tensor decompositions and their applications in machine learning." Preprint, submitted November 29, 2017. https://arxiv.org/abs/arXivpreprint.

Razavi, S. N., and C. T. Haas. 2010. "Multisensor data fusion for on-site materials tracking in construction." *Autom. Constr.* 19 (8): 1037–1046. https://doi.org/10.1016/j.autcon.2010.07.017.

Razavi, S. N., and C. T. Haas. 2011. "Reliability-based hybrid data fusion method for adaptive location estimation in construction" *J. Comput. Civ. Eng.* 26 (1): 1–10. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000101.

Reichert, B., and W. Stelzenmueller. 2011. *Palpation techniques: Surface anatomy for physical therapists*. New York: Thieme.

Shahi, A., J. M. Cardona, C. T. Haas, J. S. West, and G. L. Caldwell. 2012. "Activity-based data fusion for automated progress tracking of construction projects." In *Proc., Construction Research Congress 2012: Construction Challenges in a Flat World*, 838–847. Reston, VA: ASCE.

Sidiropoulos, N. D., L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos. 2017. "Tensor decomposition for signal processing and machine learning." *IEEE Trans. Signal Process.* 65 (13): 3551–3582. https://doi.org/10.1109/TSP.2017.2690524.

Soibelman, L., L. Y. Liu, and J. Wu. 2004. "Data fusion and modeling for construction management knowledge discovery." In *Proc., Int. Conf. on Computing in Civil and Building Engineering*. Weimar, Germany. Weimar, Germany: Bauhaus-Universität Weimar.

Soltani, M. M., Z. Zhu, and A. Hammad. 2018. "Framework for location data fusion and pose estimation of excavators using stereo vision." *J. Comput. Civ. Eng.* 32 (6): 04018045. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000783.

Soman, R., M. Kyriakides, T. Onoufriou, and W. Ostachowicz. 2018. "Numerical evaluation of multi-metric data fusion based structural health monitoring of long span bridge structures." *Struct. Infrastruct. Eng.* 14 (6): 673–684. https://doi.org/10.1080/15732479.2017.1350984.

Wang, K., X. Chen, L. Wu, X. Zhang, X. Chen, and Z. J. Wang. 2020. "High-density surface EMG denoising using independent vector analysis." *IEEE Trans. Neural Syst. Rehabil. Eng.* 28 (6): 1271–1281. https://doi.org/10.1109/TNSRE.2020.2987709.

Xie, K., X. Li, X. Wang, G. Xie, J. Wen, J. Cao, and D. Zhang. 2017. "Fast tensor factorization for accurate internet anomaly detection." *IEEE/ACM Trans. Networking (TON)* 25 (6): 3794–3807. https://doi.org/10.1109/TNET.2017.2761704.

Xiong, L., X. Chen, T.-K. Huang, J. Schneider, and J. G. Carbonell. 2010. "Temporal collaborative filtering with bayesian probabilistic tensor factorization." In *Proc., 2010 SIAM Int. Conf. on Data Mining*, 211–222. University City, PA: Society for Industrial and Applied Mathematics.

Young, W., G. Weckman, and W. Holland. 2011. "A survey of methodologies for the treatment of missing values within datasets: Limitations and benefits." *Theor. Issues Ergon. Sci.* 12 (1): 15–43. https://doi.org/10.1080/14639220903470205.

Žitnik, M., and B. Zupan. 2014. "Data fusion by matrix factorization." *IEEE Trans. Pattern Anal. Mach. Intell.* 37 (1): 41–53. https://doi.org/10.1109/TPAMI.2014.2343973.