Tensor Mining of Public Railroad Accident Reports

Evangelos E. Papalexakis, Jia Chen, Dawon Ahn, Miguel Gutierrez
University of California Riverside, USA

Raymond Jiang
Glen A. Wilson High School,
Hacienda Heights, CA

Ping Xu, Constantine Tarawneh
The University of Texas – Rio Grande Valley,
USA

IHHA 2025
13TH INTERNATIONAL HEAVY HAUL
ASSOCIATION CONFERENCE 2025

November 17-21, 2025 | The Broadmoor, Colorado Springs, CO, USA

Introduction

Given vast amounts of existing public accident reports, recording a wealth of accident report variables (spatial, temporal, weather conditions, etc), can we leverage tensor mining techniques in order to extract insightful patterns from public accident report data, and if so, what are the challenges and advantages of doing so compared to standard data mining techniques?

Proposed methodology

Our key methodological component revolves around the concept of a tensor, which is a multi-dimensional extension of a matrix or a data array and can naturally express relations across variables of interest in multiple dimensions. An excel sheet whose columns capture different accident report variables can be naturally represented as a tensor. A tensor decomposition is a data mining algorithm that allows us to extract meaningful hidden and emerging patterns from such data. In this work, we start from existing tensor decompositions and we adapt them for the purposes of this application. We face two major challenges: (1) the number of variables of interest is too large and not all of them may provide important and valuable insights. How can we identify "needles in the haystack" among those variables? and (2) Since we are conducting exploratory analysis, there is no ground truth or golden standard baseline that we can use to evaluate our findings.

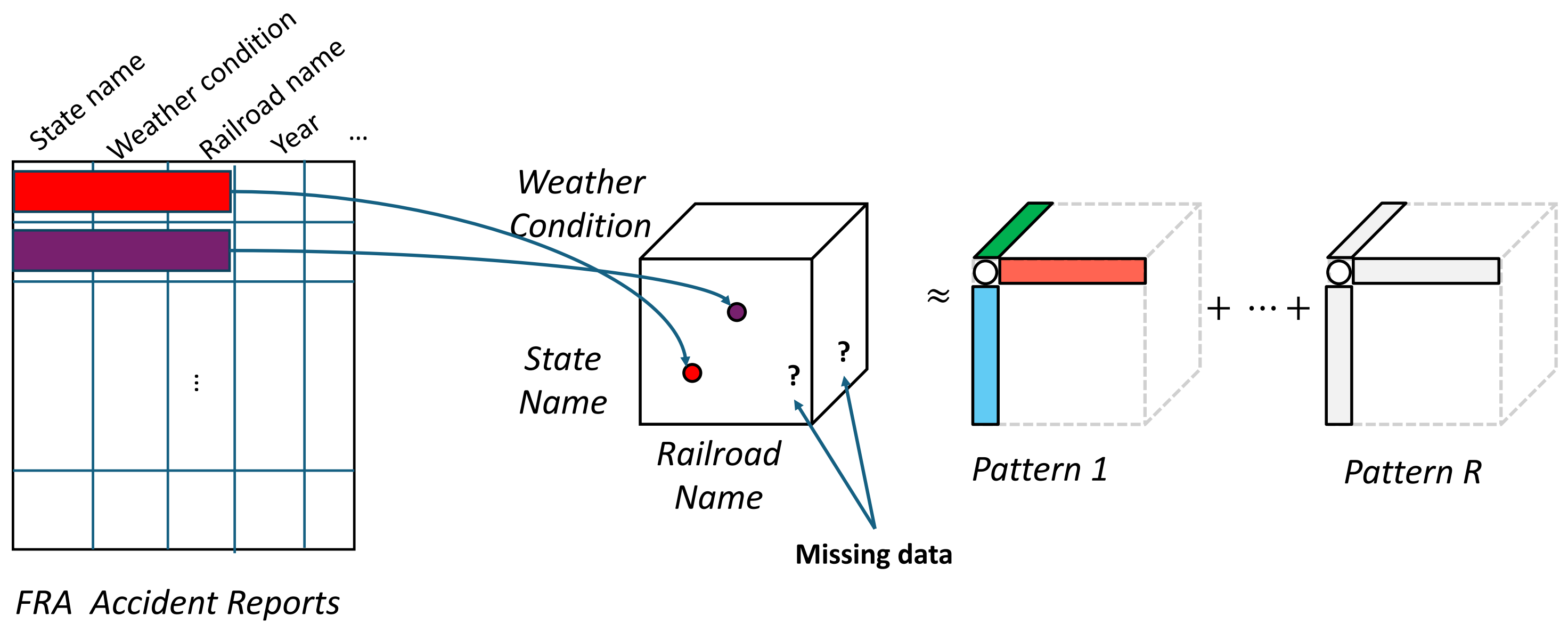

Figure 1: The Canonical Polyadic Decomposition (CPD) can analyze FRA Accident Reports in an interpretable latent space

Constructing Tensors with Meaningful Structure

In constructing tensor datasets from raw data, the challenge is that even though virtually any structured dataset (such as the FRA accident reports) can be readily and trivially seen as a tensor (i.e., a multi-dimensional matrix), we are particularly interested in tensors that can be mined for interpretable patterns, because tensor methods require data to have certain structure, typically so-called low-rank or low-dimensional structure. As a result, a first step here is to provide a method for identifying and forming as many well-structured tensors as possible from the raw data.

In our case, goodness of structure is measured by how well the data adhere to the CPD tensor model, as well as how "clusterable" are the data within the tensor. Thus, we are going to evaluate different automatically generated tensor datasets from the raw report data using a number of intrinsic measures of structure quality and then identify datasets which score high-enough on all measures, as shown in Fig 2.

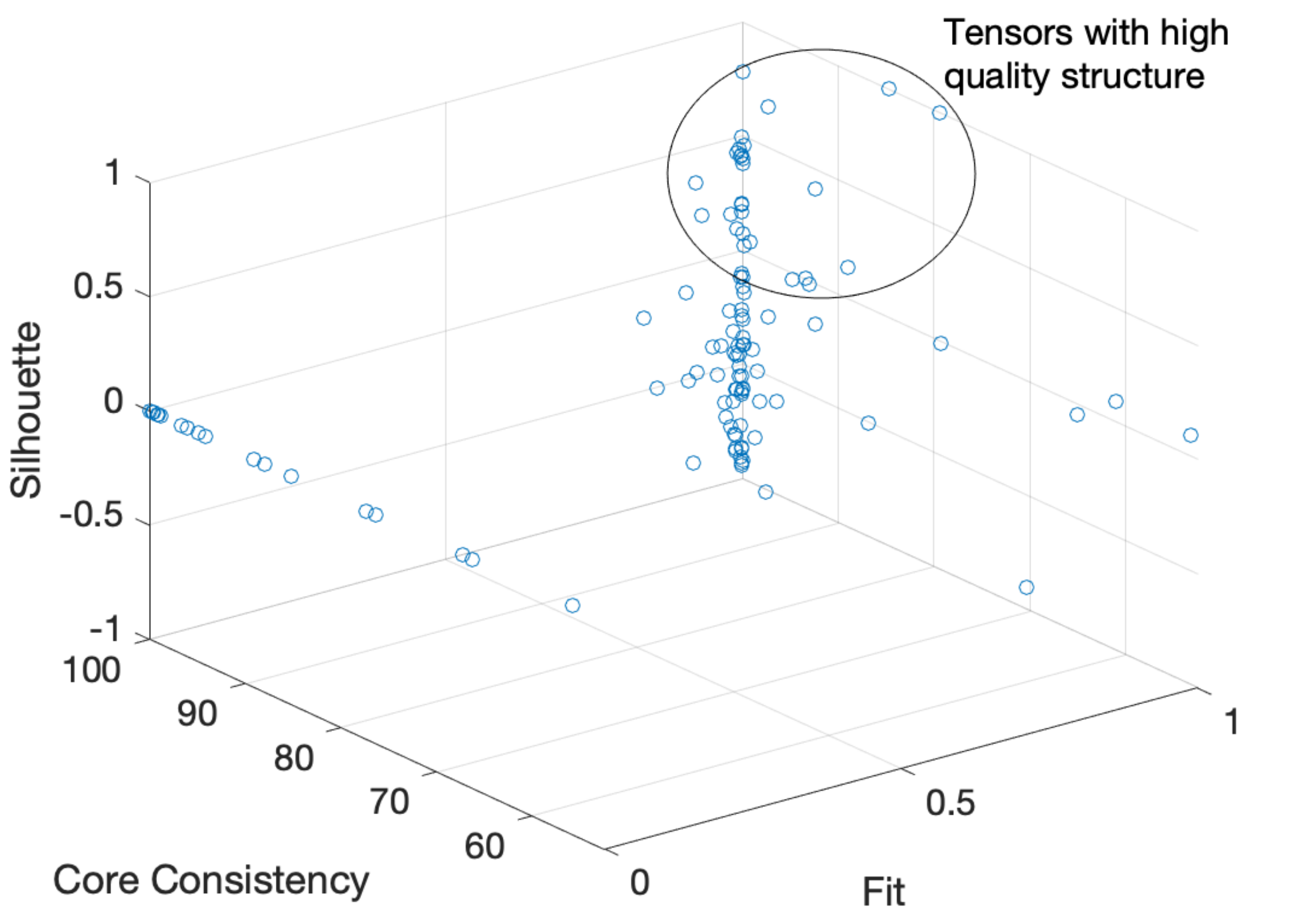

Figure 2: Scoring automatically generated tensors

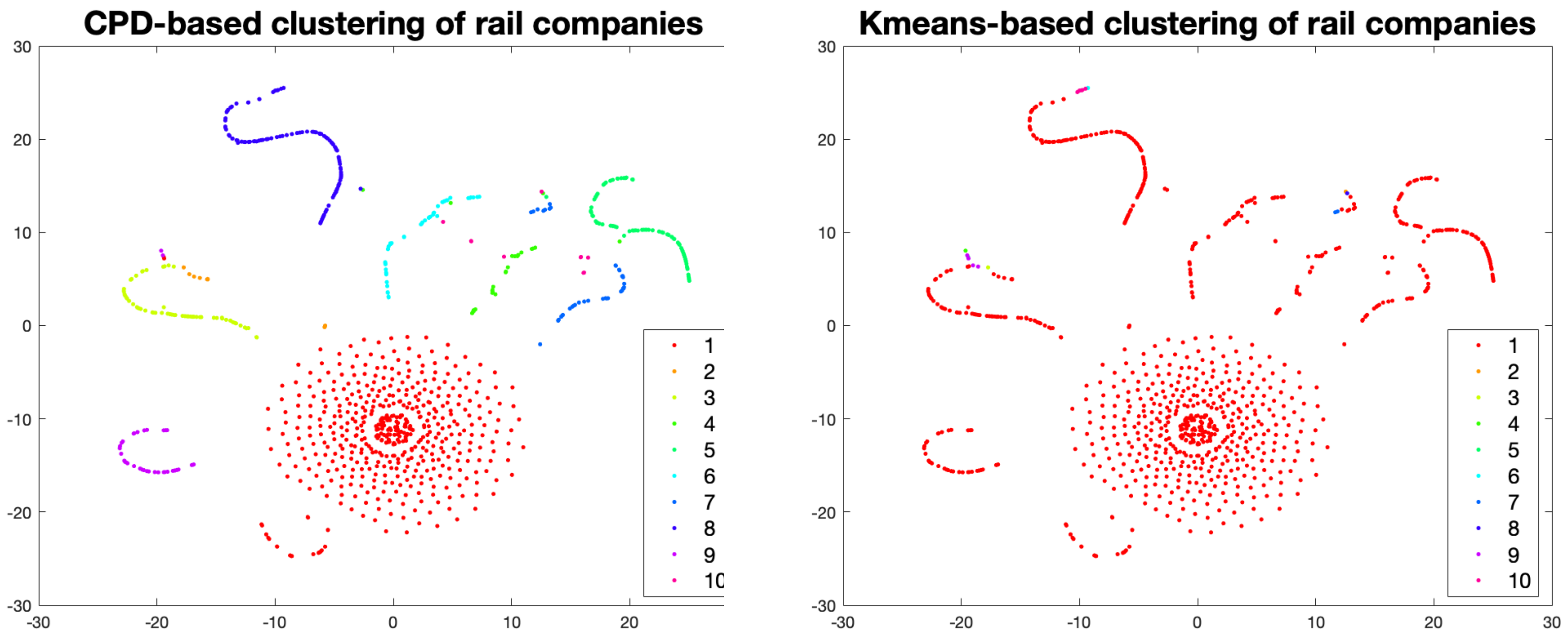

Figure 3. Left: Clustering rail companies using our proposed tensor method. Same color indicates same cluster membership. We observe that even though the clusters are not necessarily linear, our proposed method groups coherent regions of the space successfully. Right: Clustering of companies using traditional K-means clustering, which fails in capturing the non-linear patterns in the space.

Completing Missing Data

In case of missing entries from a given accident report, we can use tensor completion methods [3] in order to approximately fill-in the missing values. In Figure 4 we show indicative results where the CostCo neural tensor completion [3] can successfully recover missing report entries.

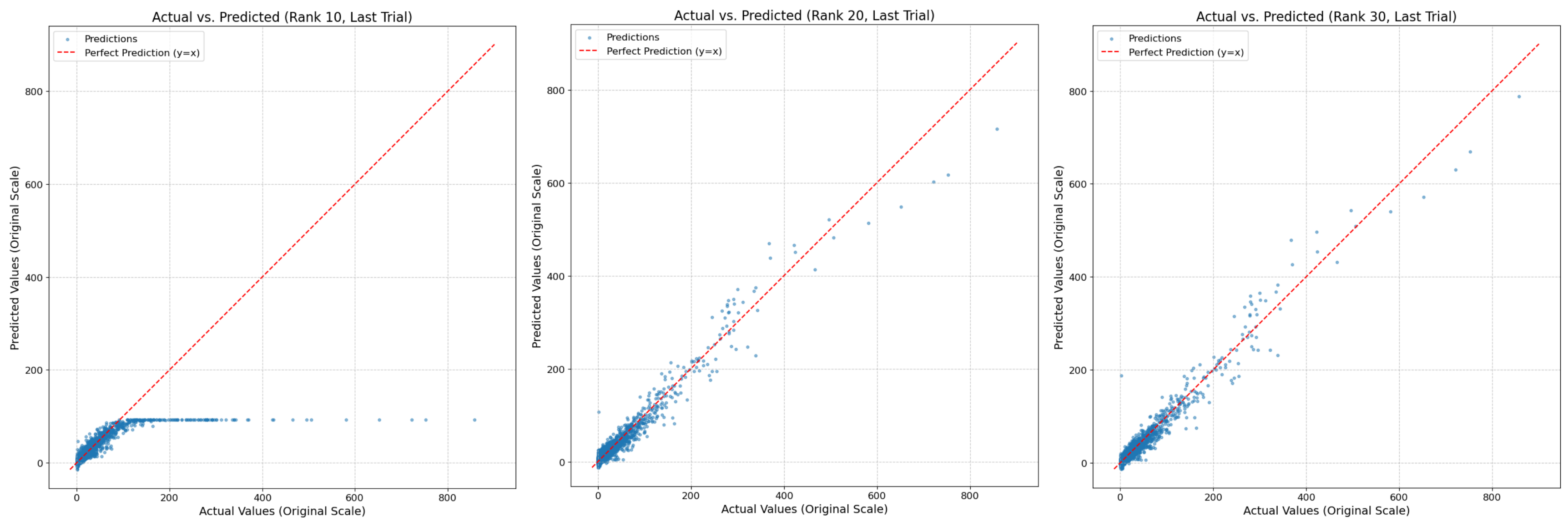

Figure 4: Predicted vs. Actual values for missing entry prediction for different low dimensionalities of tensor decomposition.

Conclusions

In this work, we are the first to demonstrate that tensor mining is a very viable technique for extracting actionable insights from public accident reports and can be used by practitioners and policy-makers in improving their understanding of potentially hidden risk factors and other interesting emerging patterns as well as handling incomplete data.

References

1. Sidiropoulos, Nicholas D., Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E. Papalexakis, and Christos Faloutsos. "Tensor decomposition for signal processing and machine learning." IEEE Transactions on signal processing 65, no. 13 (2017): 3551-3582.
2. US Department of Transportation. "US Highway Rail Grade Crossing Accident Dataset." https://www.kaggle.com/datasets/yogidsba/us-highway-railgrade-crossing-accident?resource= download. Data retrieved from US DOT
3. Liu, Hanpeng, Yaguang Li, Michael Tsang, and Yan Liu. "Costco: A neural tensor completion model for sparse tensors." In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 324-334. 2019.

Acknowledgments

This work was supported by the University Transportation Center for Railway Safety (UTCRS) at UTRGV through the USDOT UTC Program under Grant No. 69A3552348340 , NSF CREST Center for Multidisciplinary Research Excellence in Cyber-Physical Infrastructure Systems (MECIS) grant no. 2112650. and NSF CAREER grant no. IIS 2046086 Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding parties.

For further information please contact:
Evangelos Papalexakis , UC Riverside, epapalex@cs.ucr.edu
Jia Chen , UC Riverside, jiac@ucr.edu

INTERNATIONAL HEAVY HAUL ASSOCIATION