HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 4 (2023)

Retrieval of Missing Remotely Sensed Tropospheric NO_2 Data Using Tensor Completion

Rohan Shankar[1] and Ryan Solgi[#]

[1]Mountain View High School, USA
[#]Advisor

ABSTRACT

Missing values in remotely sensed satellite data present a significant challenge for accurate analysis and interpretation of environmental data. Factors such as dead pixel values or cloud coverage can lead to significant gaps in datasets, degrading its overall quality and value. Tensor completion utilizes high-dimensional arrays known as tensors and their low-rank decompositions to recover missing values. This paper demonstrates the application and value of applying Tensor Completion to enhance remotely sensed Nitrogen-Dioxide satellite data and proposes an algorithm for its use. An efficient and accurate recovery method is proposed by leveraging relationships within the tensor and by employing suitable tensor decomposition methods. The proposed algorithm will enhance the analysis, interpretation, and utilization of satellite data. The algorithm is validated on real-world satellite datasets and its superiority demonstrated against existing alternate data recovery methods such as IDW and Kriging.

Introduction

Air pollution is one of the most significant health and environmental issues of our time. Extensive studies have demonstrated the adverse effects of air pollution on respiratory and cardiovascular health, leading to increased mortality rates. Moreover, the emission of greenhouse gasses, pollutants, and other particulate matter exacerbate climate change, emphasizing the urgency to address air pollution as a pressing global concern [1, 2].

Among the many pollutants contributing to this issue, Nitrogen Dioxide (NO_2) holds significance due to its severe impact on respiratory health, with studies linking it to degenerative conditions such as lung cancer [3]. While not itself a greenhouse gas, it is considered a precursor to formation of other greenhouse gasses such as tropospheric ozone. For these reasons, NO_2 data is highly valued for environmental research.

While land-based instruments exist to measure NO_2, these are often concentrated in urban areas (and resultantly sparse in rural areas). Since satellites cover much larger swathes of land and ocean, remote sensing data sets are preferred for use in analysis and environmental studies [4].

Measured satellite data can be visualized to be stored in high-dimensional arrays, also known as tensors. However, satellite data is often non-uniform and incomplete due to missing values and can therefore be difficult to analyze. This can be due to a variety of reasons, such as the satellite being offline or cloud coverage interfering with its sensors [5]. This non-uniformity complicates data analysis and necessitates the use of sophisticated imputation techniques, increasing complexity and potentially introducing further uncertainty. For these reasons, methods to recover missing values are preferred. This research focuses on recovery of NO_2 data from the Sentinel-5P satellite.

Tensor completion is one of many methods to recover missing values. At its core, tensor completion exploits underlying relationships in the data to infer missing data point values. The more relationships in the data (i.e., the lower rank a tensor is), the better tensor completion works. Tensor Completion has previously been explored in areas such as computer vision and random graphs [6, 7]. It was proven that exact reconstruction

of a tensor is possible with only minimal data from the original tensor. Additionally, Tensor Completion has been successfully used in remote sensing to reconstruct Landsat images that were previously obscured by clouds, Tensor Completion showed better results than several state-of-the-art algorithms [5].

Despite being a powerful tool, tensor completion's application in the context of NO_2 data has been unexplored. This paper aims to address this research gap by introducing a tensor completion algorithm using CP (CANDECOMP/PARAFAC) Decomposition specifically designed for NO_2 data from the Sentinel-5P satellite. This algorithm seeks to improve the quality of the recovered data, decrease the complexity of data processing, and reduce uncertainty in the analysis, thereby advancing several environmental research methods. By evaluating the algorithm using several evaluation metrics, one can demonstrate its effectiveness and potential as a robust tool for environmental researchers and policymakers alike. This work attempts to contribute to ongoing efforts to mitigate air pollution and its effects on health and the climate.

Data and Methodology

Data Overview

The Sentinel-5P NO_2 data [8, 9] in this study is stored in a high-dimensional array, commonly known as a tensor. At a high level, tensors are a mathematical representation of data with multiple dimensions. Tensors can have an arbitrary number of dimensions, making them versatile. An important characteristic of all tensors is their order, which represents the number of dimensions a tensor has. This work uses a 3rd order tensor: two spatial dimensions (longitude and latitude), and one temporal dimension.

Since the complete dataset covers the entire world, and processing it would be extremely time intensive, the dataset was restricted to United States only, as shown below in Figure 1, and to the time period between Jan 1, 2019 and Jan 5, 2019.

Figure 1. Measured NO_2 data across the United States. Retained values are displayed in varying shades of blue. Missing values are shown in black.

CP-Decomposition

Tensor Decomposition is a key part of Tensor Completion and breaks up a high-order tensor into several low-rank factors. The CP (CANDECOMP/PARAFAC) Decomposition technique decomposes a higher-order tensor into several rank-1 tensors. Each rank-1 tensor is significantly smaller than the original tensor, reducing its memory requirement and processing complexity, while still retaining most of the patterns from the original

higher-rank tensor [10, 11, 12]. Figure 2 illustrates the process of decomposing the original tensor into several rank-1 tensors/factors.

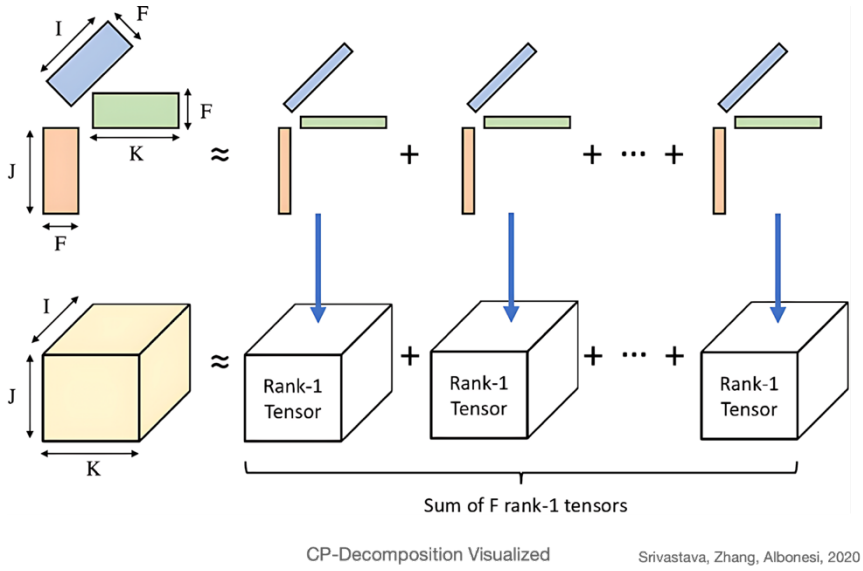

Figure 2. CP-Decomposition: A high rank tensor is decomposed into several rank-1 tensors, or factors. The more factors that original data is decomposed into, the better the representation [10].

Tensor Completion Explanation

Tensor Completion leverages inherent relationships in the data to fill in missing entries in a tensor and involves several key steps. An Alternating Least Squares (ALS) approach for Tensor Completion has been shown to be efficient, as detailed in papers such as [13].

1. The algorithm begins with an input tensor with missing values ($P_{\Omega}(T)$). The indices of missing values (Ω) rank (r), and iterations (τ) are provided.
2. Several vectors (u_0^l, σ_l, etc.) are initialized to starting estimates of the tensor's components. The vectors are generated through the Robust Tensor Power Method (RTPM) and provide an efficient starting point for the algorithm.
3. A nested for loop is then implemented, where the inner loop cycles through each factor (the number of factors represented by (represented by r), and the outer loop represents multiple complete passes through every factor (represented by τ).
4. Within the inner loop, the squared difference (arg min) between the observed and estimated tensor (excluding the current component q) is minimized, as shown in Equation 1.
5. After the estimate is normalized, every updated component is collected into a new tensor. Finally, the outer product of the final estimates form a completed tensor.

Step 4 is especially important. By iterating through one factor at a time, and not updating any other components, the step can be reduced to a simple least squares minimization problem. Additionally, since the original tensor, missing values, and number of factors remain constant, this can be further reduced to a convex optimization problem, greatly easing the complexity.

Equation 1: The minimization algorithm used in Step 4 is shown below. Tensor $P_{\Omega}(T)$ subtracts constructed tensor $P_{\Omega}(\hat{T})$ and the result is minimized to optimize the completion algorithm.

$$\underset{\hat{T},\, rank(\hat{T})=r}{\text{minimize}} \left\| P_\Omega(T) - P_\Omega(\hat{T}) \right\|_F^2$$

Experimental Setup

The Tensor Completion algorithm is applied to remotely sensed NO_2 data from the Sentinel-5P satellite.

The original dataset contained ~15% missing values. To ensure a source of ground truth while evaluating the performance of algorithm, an additional ~25% of values were intentionally removed from the dataset as shown in Figure 3. This is done for the purpose of evaluating the performance of the model against a known ground truth. Comparing the original values to the completed/predicted values can provide a better understanding of the model's performance. This performance comparison is shown in Table 1 for a subset of ranks.

The work was implemented using Python library *TensorLearn* [14] due to its ease of access and usability. Several CP-Decomposition and Tensor Completion methods were used to recover the missing values.

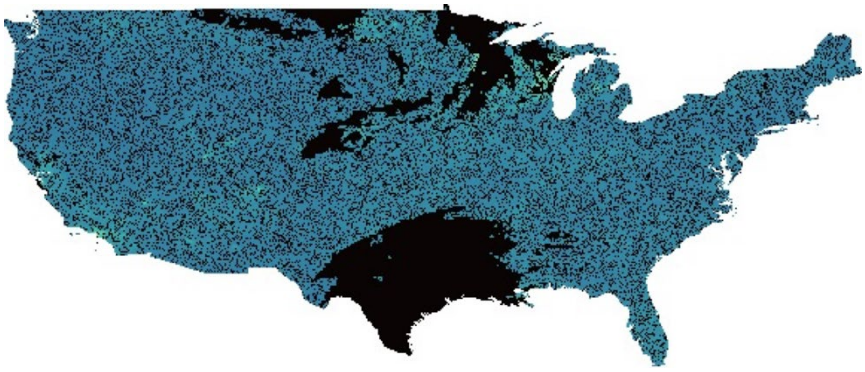

Figure 3. ~25% missing data was artificially introduced into the satellite data set.

Table 1. Tensor Completion Performance metrics versus rank. Intermediate ranks show the best correlation.

Rank	R^2	RMSE	R
10	0.671	0.035	0.852
20	0.753	0.031	0.899
30	0.785	0.029	0.917
40	0.798	0.028	0.923
50	0.804	0.027	0.927
60	0.807	0.027	0.930
70	0.806	0.027	0.930
80	0.803	0.027	0.929
90	0.793	0.028	0.924
100	0.777	0.029	0.916

Results and Discussion

Statistical Analysis

When running tensor completion on the Sentinel-5P dataset, various variables were varied - including rank, iterations, and the percentage of missing values added to the data. In this case, rank refers to the complexity of CP-Decomposition performed on the original tensor. A higher rank means that the algorithm better fit the factors to the original data. R^2, Root Mean Squared Error (RMSE), and Correlation Coefficient (R) were calculated using the ground truth. The results are displayed in Table 1.

The trends show a clear increase in R^2 and R with rank, while RMSE decreases with rank. However, at higher ranks (close to 90-100), the model begins to taper off, and worsen in performance. For example, the R^2 at a rank of 100 is 0.03 worse than a rank of 80. This trend is also seen in the other metrics. This is likely due to the model becoming overfit on the original data, and therefore not reconstructing the tensor correctly. These statistical trends are visualized in Figure 4, where it can be clearly seen that the performance starts to taper (worsens) at higher ranks.

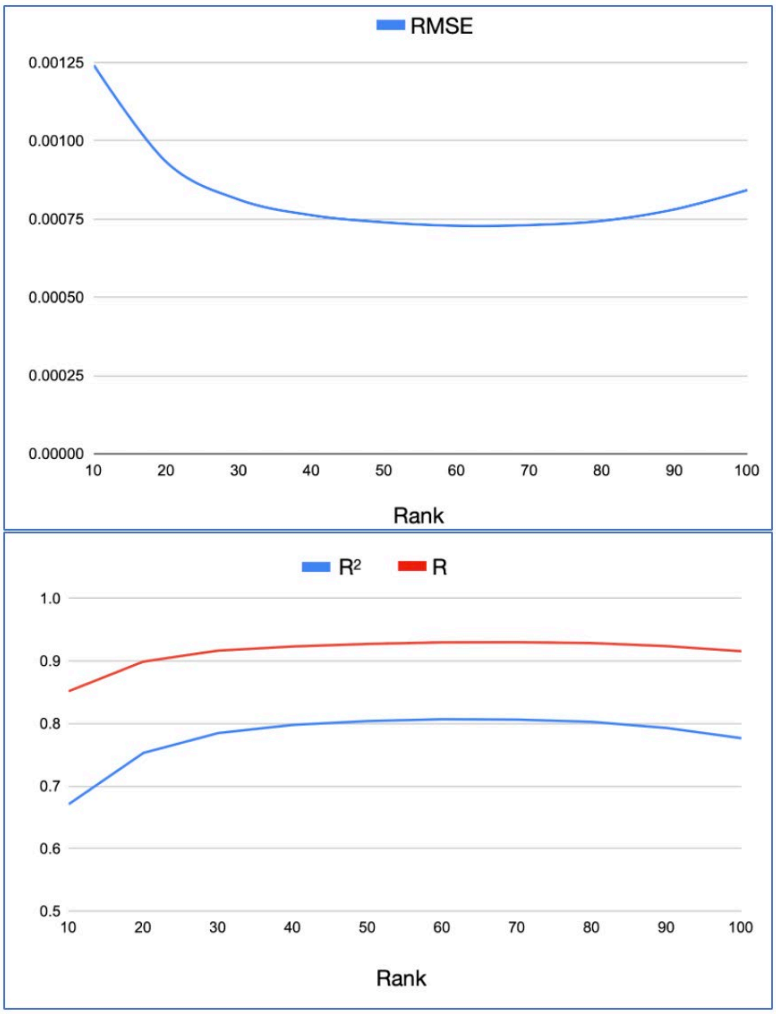

Figure 4. R^2, R, and RMSE performance vs Rank. Degradation of performance at higher ranks >80 is observed

Comparison to Existing Methods

Tensor completion using CP-Decomposition is compared to a variety of existing methods. The two most popular methods to retrieve missing values are Inverse Distance-Weighted Interpolation (IDW) and Kriging.

IDW and Kriging work on similar principles. IDW takes the n-nearest filled in data points and uses them to compute a missing value. The "Inverse" in IDW is due to the way IDW weights data points: closer data points are weighted higher. Kriging works on a similar principle, except it incorporates spatial geographical relationships to estimate the missing value (rather than simply choosing the n-nearest points). It then forms a variogram and models spatial correlation in the data. The improved accuracy is achieved, but it is accompanied with larger computational complexity. Both IDW and Kriging methods are widely used in the field of geographical data analysis.

Table 2 shows the performance of Tensor Completion at an optimal rank of 80 is compared benchmark techniques IDW and Kriging [15, 16, 17, 18]. In terms of performance metrics, Kriging improves R^2, R and Residual error over IDW at the cost of increased $O(n^3)$ complexity (i.e. increased processing time and resources). Tensor Completion maintains many of the advantages of performance delivered by Kriging, but maintains a O(n) complexity – i.e. it has a much faster processing time for very large datasets. Tensor Completion has a slightly degraded RMSE when compared to Kriging benchmark, but it is still significantly better than IDW benchmark, and still very low.

Table 2. Tensor Completion Performance metrics compared to IDW and Kriging at optimal rank 80.

	IDW [15, 16, 17] *(academic benchmark)*	**Kriging** [18] *(academic benchmark)*	**Tensor Completion** *(this paper, Sentinel-5P)*
Time Complexity n = # of samples	**O(n)**	$O(n^3)$	**O(n)**
Degree of Fit, R^2	0.30~0.40	0.61~0.7	**0.80**
Degree of Fit, R	~	**0.85~0.95**	**0.92**
Residual Error, RMSE	1.22~0.9	**0.05~0.01**	**0.025**

Conclusion

In this paper, Tensor Completion algorithm was successfully applied to remotely sensed Nitrogen-Dioxide data from the Sentinel-5P satellite. Missing values were introduced artificially, the data was then normalized, and existing patterns were assessed to fine tune the model's performance.

Tensor Completion was shown to be extremely promising, and its performance was shown to be equal or better than existing interpolation techniques. It can be a strong tool for researchers to retrieve missing values in many satellite datasets and has the potential to bring significant benefits to climate change research.

Limitations and Future Work

While this paper makes significant contribution to the research, there are some limitations and areas of further work and improvements are identified.

One limitation of the work is that the artificial missing data was randomly removed from the satellite dataset. A different method of removing data – such as to simulate cloud cover, can offer additional insight to the performance of the different data recovery methods.

Tensor Completion relies on "low rank" of a dataset to exploit patterns and dependencies in it. Its performance (vs IDW and Kriging) can therefore vary depending on the rank of the target data set.

One future work project can involve running IDW and Kriging on the Sentinel-5P dataset as well so that a more direct performance comparison to Tensor Completion can be made – instead of relying existing research benchmarks for IDW and Kriging. Due to their very high complexity of implementation $O(n^3)$ of these algorithms, this is unfortunately expected to be extremely time and processing power intensive.

An additional future work can implement other Tensor Decomposition methods for the Tensor Completion. CP-Decomposition was successfully implemented in this paper, but other methods such as Tucker and Tensor-Train Decomposition can be evaluated for potential additional improvements as well.

Acknowledgments

I would like to thank Dr. Lina Kim, Director of Academic Programs at University of California Santa Barbara for her untiring leadership at 1-1 Research Mentorship Program (RMP) which enabled this research work. The RMP program is a mentored research program performed for college course credit. I would also like to thank Pratyush Tripathy with the Department of Geography at University of California Santa Barbara for his guidance and mentorship.

References

[1] M. Kampa and E. Castanas, “Human health effects of Air Pollution,” Environmental Pollution, vol. 151, no. 2, pp. 362–367, 2008. doi:10.1016/j.envpol.2007.06.012

[2] B. Brunekreef and S. T. Holgate, “Air Pollution and health,” The Lancet, vol. 360, no. 9341, pp. 1233–1242, 2002. doi:10.1016/s0140-6736(02)11274-8

[3] L. Bai et al., “Exposure to ambient air pollution and the incidence of lung cancer and breast cancer in the Ontario Population Health and Environment Cohort,” International Journal of Cancer, vol. 146, no. 9, pp. 2450–2459, 2019. doi:10.1002/ijc.32575

[4] J. Fishman et al., “An investigation of widespread ozone damage to the soybean crop in the upper Midwest determined from ground-based and satellite measurements,” Atmospheric Environment, vol. 44, no. 18, pp. 2248–2256, 2010. doi:10.1016/j.atmosenv.2010.01.015

[5] T.-Y. Ji, N. Yokoya, X. X. Zhu, and T.-Z. Huang, “Nonlocal tensor completion for multitemporal remotely sensed images’ inpainting,” IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 6, pp. 3047–3061, 2018. doi:10.1109/tgrs.2018.2790262

[6] Y. Panagakis et al., "Tensor Methods in Computer Vision and Deep Learning," in Proceedings of the IEEE, vol. 109, no. 5, pp. 863-890, May 2021. doi: 10.1109/JPROC.2021.3074329

[7] Jain, P. and Oh, S., “Provable Tensor Factorization with Missing Data”, arXiv e-prints, 2014. doi:10.48550/arXiv.1406.2784

[8] Y. Lops, M. Ghahremanloo, A. Pouyaei, Y. Choi, J. Jung, S. Mousavinezhad, A. K. Salman, & D. Hammond, “Spatiotemporal estimation of TROPOMI NO_2 column with depthwise partial convolutional neural network,” *Neural Computing and Applications* 35, 15667–15678, 2023. doi: 10.1007/s00521-023-08558-1

[9] Y. Lops, M. Ghahremanloo, A. Pouyaei, Y. Choi, J. Jung, S. Mousavinezhad, A. K. Salman, & D. Hammond, “Spatiotemporal Estimation of TROPOMI NO2 Column with Depthwise Partial Convolutional Neural Network (1.0) [Data set]”, Zenodo, 2023. doi: 10.5281/zenodo.7770693

[10] N. Srivastava, “Design and Generation of Efficient Hardware Accelerators for Sparse and Dense Tensor Computations,” PhD Thesis, 2020. doi: 10.7298/5ksm-sm92

[11] J. H. de M. Goulart, M. Boizard, R. Boyer, G. Favier, and P. Comon, “Tensor CP decomposition with structured factor matrices: Algorithms and performance,” IEEE Journal of Selected Topics in Signal Processing, vol. 10, no. 4, pp. 757–769, 2016. doi:10.1109/jstsp.2015.2509907

[12] M. A. Veganzones, J. E. Cohen, R. Cabral Farias, J. Chanussot, and P. Comon, “Nonnegative tensor CP decomposition of hyperspectral data,” IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 5, pp. 2577–2588, 2016. doi:10.1109/tgrs.2015.2503737

[13] P. Comon, X. Luciani, and A. L. de Almeida, “Tensor decompositions, alternating least squares and other tales,” Journal of Chemometrics, vol. 23, no. 7–8, pp. 393–405, 2009. doi:10.1002/cem.1236

[14] R. Solgi, RMSOLGI/TensorLearn: A python package for advanced tensor learning. Retrieved August 31, 2023 from https://github.com/rmsolgi/TensorLearn

[15] P. A. Bostan, G. B. M. Heuvelink, and S. Z. Akyurek, “Comparison of regression and Kriging techniques for mapping the average annual precipitation of Turkey,” International Journal of Applied Earth Observation and Geoinformation, vol. 19, pp. 115–126, 2012. doi:10.1016/j.jag.2012.04.010

[16] K. Krivoruchko and A. Gribov, “Evaluation of empirical Bayesian kriging,” Spatial Statistics, vol. 32, p. 100368, 2019. doi:10.1016/j.spasta.2019.100368

[17] N. Theodossiou and P. Latinopoulos, “Evaluation and optimisation of groundwater observation networks using the Kriging methodology,” Environmental Modelling & Software, vol. 21, no. 7, pp. 991–1000, 2006. doi:10.1016/j.envsoft.2005.05.001

[18] B. I. Harman, H. Koseoglu, and C. O. Yigit, “Performance evaluation of IDW, Kriging and Multiquadric interpolation methods in producing noise mapping: A case study at the city of Isparta, Turkey,” Applied Acoustics, vol. 112, pp. 147–157, 2016. doi:10.1016/j.apacoust. 2016.05.024