Multiscale Data Analysis Using Binning, Tensor Decompositions, and Backtracking

Dimitri Leggas, Thomas S. Henretty, James Ezick, Muthu Baskaran, Brendan von Hofe,
Grace Cimaszewski, M. Harper Langston, Richard Lethin
Reservoir Labs, Inc.
New York, NY
Email: {leggas, henretty, ezick, baskaran, vonhofe, grace, langston, lethin}@reservoir.com

***Abstract*—Large data sets can contain patterns at multiple scales (spatial, temporal, etc.). In practice, it is useful for data exploration techniques to detect patterns at each relevant scale. In this paper, we develop an approach to detect activities at multiple scales using tensor decomposition, an unsupervised high-dimensional data analysis technique that finds correlations between different features in the data. This method typically requires that the feature values are discretized during the construction of the tensor in a process called "binning." We develop a method of constructing and decomposing tensors with different binning schemes of various features in order to uncover patterns across a set of user-defined scales. While binning is necessary to obtain interpretable results from tensor decompositions, it also decreases the specificity of the data. Thus, we develop backtracking methods that enable the recovery of original source data corresponding to patterns found in the decomposition. These technique are discussed in the context of spatiotemporal and network traffic data, and in particular on Automatic Identification System (AIS) data.**

***Index Terms*—Tensor Decomposition, Geospatial, Spatiotemporal, Multiscale**

I. Introduction

Tensor decompositions are an unsupervised, high-dimensional data analysis technique used for pattern discovery [1]. Decomposing tensors gives a low-rank approximation of the original data in terms of components that reveal correlations between points along different dimensions of the tensor. Each dimension (i.e., mode) of the tensor corresponds to some feature of the data, so the multilinear relationships described by the tensor decomposition components indicate latent structure found in the data. Tensor decompositions have successfully extracted patterns from a multitude of data sources including spatiotemporal [2], network traffic [3], genomics [4], and consumer data [5].

While tensor decompositions are a powerful technique for exploratory data analysis, some pre-processing of raw data is typically needed to build a suitable tensor for analysis. Raw data in many domains is often tabular, meaning it has a row-column form in which rows are data entries and columns are features. Before performing a decomposition, the high-dimensional structure must be explicitly represented as a tensor by choosing a subset of the columns to correspond to the modes of the tensor. Each index along each mode represents a unique value of that feature observed in the data. Then as each index in the tensor encodes a distinct combination of feature values, it is usually necessary to discretize the data in each mode before constructing the tensor. This process can be understood as placing similar data into a bin, so we use the term "binning" to refer to the discretization. Each bin is named by an appropriate label and corresponds to an index along one mode of the tensor.

Not performing binning can result in constructing a superdiagonal tensor. This can lead to a degenerate case and make it impossible to find relevant structure in the data. Additionally, the binning technique chosen for each mode of the tensor has a direct effect on the results of the tensor decomposition. In particular, the binning technique controls the scale of the discovered patterns. For example, for a tensor in which one mode corresponds to time, binning timestamps by the minute or hour is useful for extracting patterns that happen within a day, whereas binning timestamps by the day is useful for extracting patterns that manifest over the course of a week. A binning decision must be made for each mode in the tensor, and each choice will emphasize certain relationships between features in the data. While binning is necessary to ensure interpretable results and can be useful for determining the scale of patterns extracted, it also coarsens the data clustered by the decomposition components resulting in a loss of information. For example, if minute-granularity time is binned to the day in order to extract patterns happening at week scale, decomposition components cannot indicate exactly when discovered patterns occur.

In this paper, we develop techniques for extracting patterns from data at multiple scales and for recovering the full information associated with data points in decomposition components. Our contributions are:

- We describe a systematic method of binning and decomposing tensors at multiple scales in which different relevant binning schemes are chosen for each mode resulting in a grid of tensors.
- We develop general methods referred to as "backtracking" that map decomposition components to data entries. They recover the full specificity of the data associated with a pattern and the other features (columns) associated with the data but not used in the tensor. We describe two forms, one suitable for data files and another for databases.

We show the application of these techniques to spatiotemporal data with activities occurring on multiple scales and discuss its use on network traffic data. Our experiments use Automatic Identification System (AIS) data, which describe the position and movement of vessels in North American waters. The resulting workflow extracts more patterns than would be found by a single decomposition and provides additional information to the user about patterns that is not found in the decomposition components themselves.

II. Related Work

Tensor decompositions have been used for data exploration in a variety of domains. Henretty et al. [2] show that tensor decompositions provide insight into spatiotemporal data, cluster the data into coherent patterns, and are a viable approach for performing anomaly detection. Likewise in network analysis, Baskaran et al. [3] show that tensor decompositions not only isolate malicious network traffic and possible intrusions, but also provide actionable insights for eliminating threats. While these works demonstrate that tensor decompositions are useful for gaining insight into the types of data discussed in this paper, they do not explicitly address the issue of scale of the patterns occurring. In this work, we manipulate the tensor construction via binning scheme selection in order to extract patterns at relevant scales.

Matsui et al. [5] use tensor decompositions to study consumer expenditure at multiple timescales. Their approach encodes the different timescales as separate modes in the tensor (Entry (i, j, k) is the number of items bought by consumer i on day j of week k). This tensor construction approach differentiates the expenditure patterns of different socioeconomic groups at both intra- and inter-week timescales. In this paper, we extend the multiscale analysis to arbitrary features through a binning-based approach.

While tensor decompositions themselves reduce the cognitive load placed on the data analyst by isolating coherent patterns in the data, further processing and analysis of components is a possibly tedious task that can be partially automated. Multiple machine learning approaches are available to help post-process tensor decompositions. Henretty et al. [6] propose using topic models to find a baseline of expected components in spatiotemporal data. This approach also enables the user to find components that appear in multiple data sets. Ezick et al. [7] combine tensor decompositions and graph analytics by using decomposition components to construct graph queries that associate identified patterns with other available information such as additional log fields. In this paper, we propose a general approach to gathering additional available information called "backtracking." Backtracking can take the form of enumerating data lines associated with components or constructing queries to retrieve associated data from any database. In addition to providing more contextual information, backtracking restores any resolution lost in the data due to binning.

III. Description of Approach

A. Tensor Decompositions

Tensor decompositions are higher-dimensional analogs to the singular value decomposition that explain the variance in arrays with order greater than two. Tensors are a suitable format for representing high-dimensional data because each dimension, also called a mode, can be aligned with a different feature in the data. For example, in a three-dimensional tensor encoding spatiotemporal data of ships, the modes could encode observed times, the vessel identifier known as Maritime Mobile Service Identity (MMSI), and locations. In network traffic analysis, a four-dimensional tensor might encode timestamps, the source and destination IP addresses, and the destination port of network connections. These two examples are used throughout the rest of the paper.

In this paper, the tensor decomposition algorithm used is the CANDECOMP/PARAFAC (CP) decomposition illustrated in Fig. 1. The CP decomposition represents the original tensor as a sum of rank-1 components, that is tensors that can be written as an outer product of as many vectors as there are modes. The decomposition is performed by choosing the desired number of components, known as the rank of the decomposition, and minimizing the distance between the original tensor and a tensor reconstructed by the components per some distance metric that varies depending on the exact CP decomposition used. For the decompositions in this paper, we use a variant known as CP Alternating Poisson Regression (CP-APR) [8], which assumes the tensor entries are drawn from Poisson distributions and enforces non-negativity constraints on the components. The appropriateness of this choice will be made evident in the following discussion of binning.

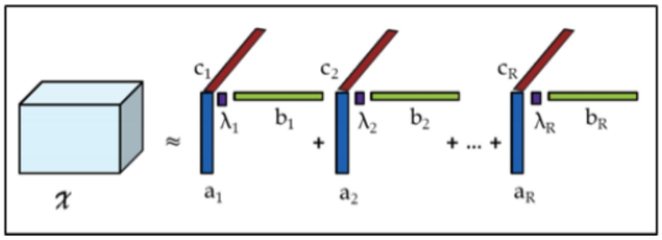


Fig. 1. CP tensor decomposition. The tensor is decomposed as a sum of rank-1 components.

As a list of vectors, each component in a decomposition can be represented as a collection of histograms, one for every mode. Each graph enumerates the labels, i.e. the observed bin values, of the mode along the x-axis and a score on the y-axis (see Fig. 2). Scores are normalized so that they sum to one and are indicators of the strength of the representation of the indices in that component. Indices with a large score in each mode indicate one way that the features are correlated in the data, i.e. a pattern in the data. Expressing these multidimensional relationships, the components found by CP decompositions reveal latent structure in the tensor. The scale of these patterns depend on the semantics of the indices of the tensor. Therefore, the construction of the tensor with appropriate data bins is a crucial step toward extracting desirable patterns from the data.

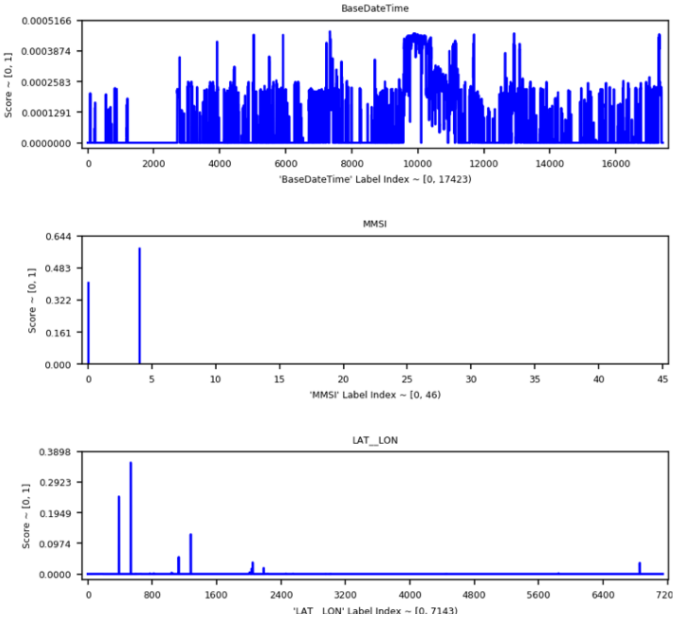


Fig. 2. Representation of a rank-1 tensor decomposition component by a collection of histograms. This component has three modes, BaseDateTime, MMSI, and LAT_Lon, describing the date and time, the identifier, and position of vessels, respectively.

B. Binning

For a tensor, each index along each mode corresponds to a value of the feature that the mode represents. Because the CP decomposition treats each index as semantically distinct, it is important to discretize certain features so that similar though non-equal values are treated as the same. For example, depending on the application, an event that takes place at 10:23:04 may be considered to happen at essentially the same time as one at 10:23:15. If such is the case, then it may make sense to bin the time feature to the minute, so that those two events occur at an equivalent time.

The process of constructing tensors to encode tabular data consists of selecting which features to use as modes and a binning scheme for each mode. It is also possible not to perform any binning along a mode if it is not necessary. Some common binning schemes in spatiotemporal and network traffic data are: time binned by day, hour, minute, second, etc.; geographic coordinates binned by rounding the latitude and longitude values; and IP addresses binned by subnet masks.

A sample network log with the features of Time, Source IP, Destination IP, and Destination Port is found in Table I. Table II shows the same data where Time has been binned by the minute and the other features have not been binned. These binning choices cause the second and third entry in Table I to have the same values for every feature, so the corresponding entry in the tensor has a value of two. Note that two log lines must have the same binned value in all features to correspond with the same tensor entry. For example, lines one and four in Table I have different binned times, so they contribute to distinct entries. The tensor represented in Table II has size $3 \times 2 \times 2 \times 2$ and three entries. The sizes of each mode reflect the number of unique values in each feature, and the entries correspond to the unique rows after binning.

TABLE I
RAW NETWORK DATA

Network Log (2020-05-27)			
Time	**Source IP**	**Dest. IP**	**Dest. Port**
14:28:23	10.0.0.1	10.0.0.3	53
15:33:57	10.0.0.2	10.0.0.1	80
15:33:12	10.0.0.2	10.0.0.1	80
18:55:39	10.0.0.1	10.0.0.3	53

TABLE II
NETWORK ACTIVITY TENSOR

Binned Tensor Entries				
Time	**Source IP**	**Dest. IP**	**Dest. Port**	**Value**
14:28:00	10.0.0.1	10.0.0.3	53	1
15:33:00	10.0.0.2	10.0.0.1	80	2
18:55:00	10.0.0.1	10.0.0.3	53	1

Constructing tensors from data tables by binning and counting the number of rows in the same bin makes CP-APR a natural choice for the decomposition because each entry is a positive integer value. The non-negativity constraints improve the quality and interpretability of the components as the scores in the modes represent the degree of presence of the corresponding labels in the pattern.

C. Multiscale Tensor Analysis

The mode binning schemes chosen at the time of tensor construction emphasize patterns at certain scales. For example, by binning location modes at different spatial resolutions, tensor decompositions can discern patterns at the various scales. Thus, coarse spatial binning enables the detection of trajectories over large distances in the original data set. Similarly, coarse temporal binning can detect patterns in trajectories that occur relatively infrequently over a long period of time.

Large datasets can contain patterns that occur at multiple scales. For example, some patterns may occur within the context of a single day of activity, while others may occur at longer timescales, such as over the course of a week, month, or year. In order to capture patterns of activity at multiple scales, we propose the following approach. Select multiple binning schemes for each mode. Then for each combination of the binning schemes, construct the corresponding tensor and decompose it in order to extract patterns at that scale.

This procedure is akin to constructing and decomposing a grid of tensors, each of which will reveal patterns at different relevant scales. Fig. 3 gives an example of such a grid that uses different binning schemes for spatial and temporal modes. The location mode is binned by country, region and city, and the time mode is binned by week, month, and year. Suppose the data used to construct the tensors tracked individuals' locations. The tensor with city and week binning schemes, when decomposed, is more likely to highlight weekly patterns like commuting to work. On the other hand, the tensor with country and year binning, when decomposed, is more likely to highlight international travel patterns.

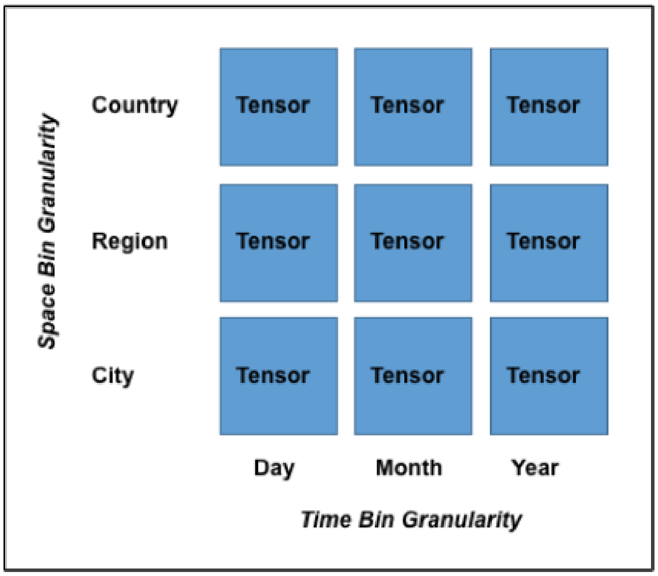


Fig. 3. Considering activities at multiple scales through various binning schemes.

A similar approach can also be applied to network traffic data. Patterns in network logs may also occur at various timescales. As location modes may be binned according to various spatial binning schemes, IP addresses may be binned by applying various subnet masks. This binning scheme allows for blocks of IP addresses to be treated as equivalent. For example, an organization may employ a network architecture that separates network services onto a different subnet than personal computers. At a different scale, subnets may differentiate different offices of the company. The patterns found in the tensor decomposition will represent activities at the "scale" of the subnet. For example, if IP addresses are binned according to a finer grain subnet, a pattern might reveal that personal computers communicate with DNS servers in a specific manner, whereas binning at a coarser grain may reveal patterns of traffic between multiple offices.

D. Backtracking

Backtracking provides a means to recover the data elements that contribute to patterns discovered by the decomposition process. This is useful to recover any loss of precision due to binning or the process of decomposing the tensor and to gain access to any additional features associated with those data elements. We provide two approaches to backtracking. The first approach for constructing mappings from tensor decomposition components to data entries occurs in two stages. *(i)* While grouping entries together by their binning, in addition to tracking the count for use as the index's value in the tensor, also record the row numbers of the associated data lines. *(ii)* Compute the Cartesian product of the indices of non-zeros in each mode in order to determine which tensor entries the component contributes to. In practice, if the the original tensor is sparse, components may produce entries at indices not appearing in it, and a single component may reproduce more entries than appear in the tensor. In this case, it is faster to iterate over the tensor entries and check whether or not each one appears in the component.

Consider the data in Table III recording the position of ships at various times. The features in the data are Time, MMSI, fused latitude and longitude, and vessel type. A tensor, which appears in Table IV is constructed using the first three features with time binned to the hour, and latitude and longitude binned to single decimal places. Note that lines one and two are the same after binning, so the corresponding tensor entry has value two and the row numbers are recorded. Now, after decomposing the tensor, if a component contains tensor entry one, it implies data lines one and two contribute to that pattern. The additional features in the data can now be used to analyze the component. For example, the vessel type associated with both of those rows is "Ferry," so the data analyst can conclude the activities involve ferries. Additionally, the actual positions and times of the vessels are recovered from the data lines.

TABLE III
Raw AIS Data

AIS records (2020-05-27)				
Row	**Time**	**MMSI**	**Lat_Lon**	**Type**
0	12:54:57	368657000	43.000241_-82.542321	Fishing
1	13:33:12	366990580	45.349522_-73.919873	Ferry
2	13:55:39	366990580	45.319297_-73.947413	Ferry
3	14:12:23	368657000	43.041229_-82.415294	Fishing

TABLE IV
AIS Tensor

Binned Tensor Entries					
Entry	**Time**	**MMSI**	**Lat_Lon**	**Value**	**Rows**
0	12:00:00	368657000	43.0_-82.5	1	{0}
1	13:00:00	366990580	45.3_-73.9	2	{1, 2}
2	14:00:00	368657000	43.0_-82.4	1	{3}

The second approach to backtracking consists of generating selection criteria that can be used to construct queries for pulling additional information from databases. This method involves converting each bin into a criterion that describes data in the bin. For example, the time label of 13:00:00 in Table IV that is binned to the hour is represented by the criterion "Time >= 13:00:00 AND Time < 14:00:00." Likewise, the Lat_Lon label 43.0_-82.4 is equivalent to the criterion "Lat >= 43.0 AND Lat < 43.1 AND Lon >= -82.4 and Lon < 82.3." In order to search a database for data associated with tensor entry one in Table IV, one could construct the following WHERE clause as part of a query: WHERE Time

>= 13:00:00 AND Time < 14:00:00 AND Lat >= 45.3 AND Lat < 45.4 AND Lon >= 73.9 and Lon < 73.8 AND MMSI == 366990580. Once tensor entries that contribute to a component are enumerated, the associated data can be pulled by constructing these queries for each entry. These criteria are platform-agnostic pseudo-queries that can easily be translated into different languages for any query system, such as Splunk [9] and Neo4j [10].

These two approaches to backtracking introduce a small amount of extra bookkeeping in exchange for a system to pull additional information associated with components. The techniques are useful for different methods of storing data. The first method, which produces a list of line numbers associated with a component, is useful when the data exists in a file in which each line is an entry. On the other hand, the second technique, which constructs general queries, is useful when data is stored in a graph or relational database and can be coupled with any existing infrastructure.

IV. Experimental Evaluation

We apply the multiscale tensor analysis to Automatic Identification System (AIS) data, which are collected by onboard navigation safety devices. These vessel traffic data record the location and characteristics of ships in US and international waters. In particular, we analyze a subset of records from the Connecticut, New York, and New Jersey area between the years 2015 and 2017 [11].

For each tensor constructed, we use three modes: timestamp, latitude / Longitude, and MMSI. The grid used in this experiment have three spatial and three temporal binning schemes. The timestamps are binned by day, month, and year, and latitude / longitude are binned to one, two, and three decimal places. For each decomposition, the rank is 100.

Different patterns are more clearly isolated with specific combinations of spatial and temporal binning schemes. Typically, patterns are extracted in a subset of the tensors in the grid, usually in a row or column of the grid. That is, for a fixed spatial or temporal binning scheme (and allowing the other to vary), the pattern can be extracted. Taken together, all of the decompositions provide a more complete description of the activities in the data set than any individual decomposition. By using the backtracking information, we gain additional explanatory insight into the components. Figure 7 shows the components discussed in this section according to which decompositions included them.

One pattern found when month binning along with any spatial binning are used is shown in Fig. 4. The latitude and longitude are binned to three and one decimal places in the two images, which are generated by finding the latitude and longitude associated with the scores in the mode encoding location. The size of the marker corresponds to the score of the associated index. The components from the different decompositions have the same corresponding time mode, which suggests that the timescale of months is appropriate for isolating this pattern. While the same pattern appear for multiple spatial binning choices, it is more interpretable at one spatial scale. The coarser spatial binning only suggests the area of the activity, but the fine spatial binning allows for the tracks of the vessel to be visible. Inspecting the backtracking data for this component reveals that the activity in this component corresponds to a fishing vessel, the behavior of which explains the multiple parallel tracks. The backtracking data also provides a ship heading for each point recorded by the component allowing for the path to be reconstructed entirely.

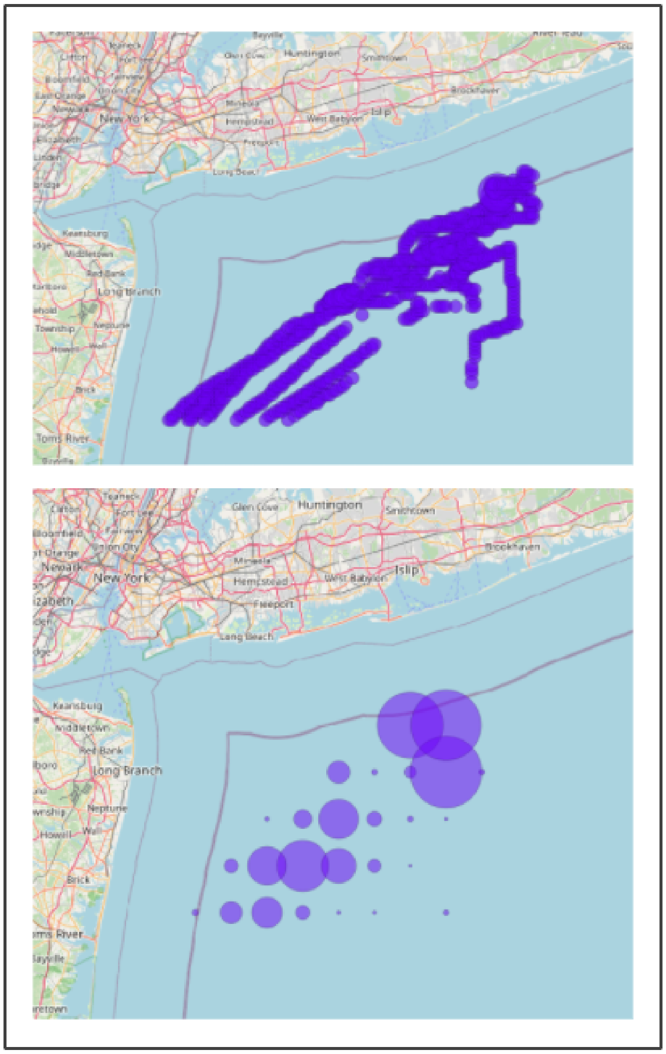


Fig. 4. A fishing vessel in the Atlantic. The pattern is more cleanly isolated with finer spatial binning.

On the other hand, a pattern that appears at a fixed spatial resolution but at multiple temporal scales is vehicle traffic in the Long Island Sound during the summer months. Every possible time binning scheme is suitable for isolating this pattern, while fine spatial binning (rounding to three decimal places) is necessary to isolate the activity in the narrow sound. Again, the latitude and longitude are extracted from the component and plotted, and these are show in Fig. 5. The different timescales lead to different levels of interpretability. The time mode with binning by day shows varying activity over the course of the summer and gives specific information about peak days of activity. Mapping the time mode indices to their labels reveals that the busiest days (based on the scores in the time mode) are Fourth of July and Memorial Day holidays. At the other extreme, when the time mode is binned to the year, each label is represented as every year recreational vehicles traverse the sound, but all specificity of when the activity occurred is lost and it is impossible to interpret the pattern as taking place during summer months without backtracking. This loss of resolution is made clear in Fig. 6. The backtracking information indicates that these components consist of recreational vehicle traffic, which explains the fact that activity in these components peaked on public holidays.

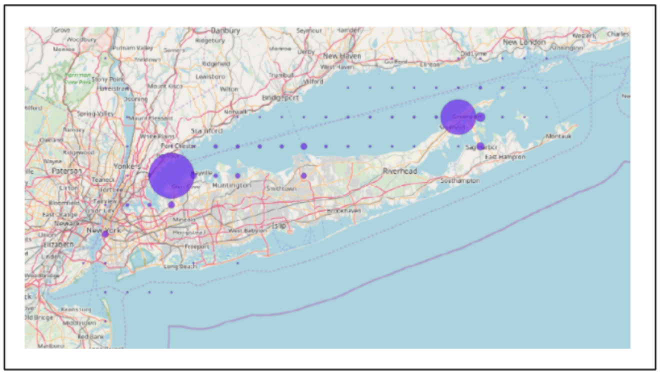


Fig. 5. Recreational vehicles during the summer in the Long Island Sound.

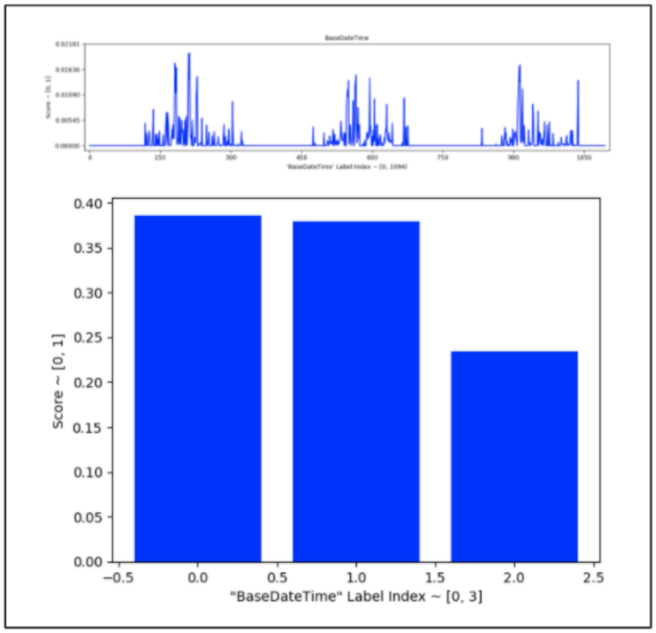
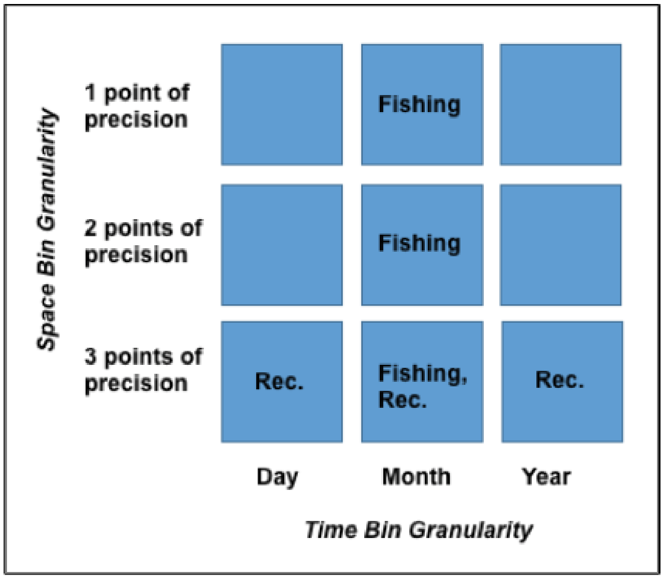


Fig. 6. Alternative time modes for the pattern in Fig. 5 under different binning schemes. The x-axis of the time mode plots are indices for the labels, which are days and years, respectively.



	Day	Month	Year
1 point of precision		Fishing	
2 points of precision		Fishing	
3 points of precision	Rec.	Fishing, Rec.	Rec.

Fig. 7. Activities found in decompositions of tensors with different binning in the time and space mode. Backtracking information indicates that the two patterns discussed involve fishing and recreational vehicles.

Binning at multiple scales in multiple modes allows for the extraction of a greater number of distinct patterns. In general, a given pattern in the AIS data appears in decompositions for fixed time mode binning or a fixed space mode binning. While the binning scheme of the other mode can vary and still reveal the pattern, the pattern has greater interpretability for one of the binning schemes. Further explanation of components is gained by using backtracking data to recover other features in the data.

V. Conclusion

Binning is a crucial transformation performed during tensor construction. The choices made at this stage determine the quality and resolution of the components found during the tensor decomposition. When using tensor decompositions for data exploration, it is necessary to choose binning schemes in all modes that help to reveal patterns at an appropriate scale. Applying multiple binning schemes is a viable strategy for detecting patterns at multiple spatial and temporal scales. Indeed, in our experiments, different binning choices reveal different patterns. Varying the binning schemes along the modes in the tensor not only extracts more patterns but also allows the patterns in the data to be understood at multiple scales and help the user find a suitable interpretation of the components.

Backtracking is a straightforward methodology that supplies the user with additional information by performing extra bookkeeping during the construction and decomposition of the tensor. This technique increases the explanatory power of decomposition components by providing additional features in the data and previously binned data at full resolution.

Acknowledgment

This material is based upon work supported by the NGA under Contract No. HM0476-18-C-0038.

References

[1] T. G. Kolda and B. W. Bader, “Tensor Decompositions and Applications,” *SIAM Review*, vol. 51, no. 3, pp. 455–500, September 2009.

[2] T. Henretty, M. Baskaran, J. Ezick, D. Bruns-Smith, and T. A. Simon, “A quantitative and qualitative analysis of tensor decompositions on spatiotemporal data,” in *High Performance Extreme Computing Conference (HPEC), 2017 IEEE*. IEEE, 2017, pp. 1–7.

[3] M. M. Baskaran, T. Henretty, J. Ezick, R. Lethin, and D. Bruns-Smith, “Enhancing network visibility and security through tensor analysis,” *Future Generations Computer Systems*, vol. 96, 2 2019.

[4] V. Hore, A. Viñuela, A. Buil, J. Knight, M. I. McCarthy, K. Small, and J. Marchini, “Tensor decomposition for multiple-tissue gene expression experiments,” *Nature genetics*, vol. 48, no. 9, p. 1094, 2016.

[5] A. Matsui, T. Kobayashi, D. Moriwaki, and E. Ferrara, “Detecting multitimescale consumption patterns from receipt data: A non-negative tensor factorization approach,” *arXiv preprint arXiv:2004.13277v1*, 2020.

[6] T. Henretty, M. H. Langston, M. Baskaran, J. Ezick, and R. Lethin, “Topic Modeling for Analysis of Big Data Tensor Decompositions,” in *SPIE Disruptive Technologies in Information Science*, Orlando, FL, Apr. 2018.

[7] J. Ezick, T. Henretty, M. Baskaran, R. Lethin, J. Feo, T.-C. Tuan, C. Coley, L. Leonard, R. Agrawal, B. Parsons, and W. Glodek, “Combining Tensor Decompositions and Graph Analytics to Provide Cyber Situational Awareness at HPC Scale,” in *IEEE High Performance Extreme Computing Conference*, Waltham, MA, Sep. 2019.

[8] E. C. Chi and T. G. Kolda, “On Tensors, Sparsity, and Nonnegative Factorizations,” arXiv:1304.4964 [math.NA], December 2011. [Online]. Available: http://arxiv.org/abs/1112.2414

[9] “Siem, aiops, application management, log management, machine learning, and compliance — splunk,” https://www.splunk.com/, accessed: 2020-06-02.

[10] “The scalable, enterprise graph data platform — neo4j,” https://neo4j.com/product/, accessed: 2020-06-02.

[11] “Vessel traffic data,” https://marinecadastre.gov/ais/, accessed: 2020-06-02.