GSTSM Package: Finding Frequent Sequences in Constrained Space and Time

Heraldo Borges, Antonio Castro, Fábio Porto, Rafaelli Coutinho, Esther Pacitti, Eduardo Ogasawara

GSTSM Package: Finding Frequent Sequences in Constrained Space and Time

Heraldo Borges
CEFET/RJ
heraldo.borges@cefet-rj.br

Antonio Castro
CEFET/RJ
antonio.castro@eic.cefet-rj.br

Fabio Porto
LNCC
fporto@lncc.br

Rafaelli Coutinho
CEFET/RJ
rafaelli.coutinho@cefet-rj.br

Esther Pacitti
University of Montpellier & INRIA
Esther.Pacitti@lirmm.fr

Eduardo Ogasawara
CEFET/RJ
eogasawara@ieee.org

Abstract

Spatial time-stamped sequences have information about time and space where events occur. Mining such sequences can bring important insights. However, not all sequences are frequent over an entire dataset. Some are only common in subsets of time and space. This article explains the first tool for mining these sequences in constrained space and time: the *GSTSM* R package. It allows users to search for spatio-temporal patterns that are not frequent in the entire database, but are dense in restricted time-space intervals. Thus, making it possible to find non-trivial patterns that would not be found using common data mining tools.

***K*eywords** Data Mining · Spatial-Temporal · Time Series · Sequential Mining

1 Introduction

Pattern discovery has become increasingly challenging in sensor and IoT data [12, 8]. In this context, a motif is a particular pattern that we can understand as a subsequence that occurs a significant number of times in time series [11, 6]. Relevant time series phenomena present different behaviors when analyzed in space and time. These phenomena are best modeled as space-time series, where each time series is associated with a position in space [10]. In this context, the motifs are specified in space and time and might not be discovered when we only analyze the temporal dimension. Discovering motifs becomes challenging when we look at the spatiotemporal series [2]. This problem is challenging for many reasons:

Traditional approaches are not effective for spatiotemporal data. To find out spatiotemporal motifs, it is necessary to identify regions of space and time in which they are frequent. Traditional approaches are inefficient as they identify motifs only in the temporal dimension, excluding those more spatially distributed. Finding and analyzing spatiotemporal motifs may enable us to understand how the phenomenon behaves [2].

Lack of spatiotemporal motif discovery tools. Some tools, such as VizTree [5] and GrammarViz [9], were developed for the discovery and visualization of motifs in time series. These tools are not suitable for working with spatiotemporal data. Besides, they are not modular, *i.e.*, they do not enable the development, execution, and comparison of other approaches in the same environment.

Spatiotemporal motif visualization tools. Discovered spatial-time motifs visualization and their distributions over space and time can bring new insights. Some initiatives work with the spatiotemporal approach to identify motifs in the trajectory data [7] but focus on analyzing moving objects. It excludes works investigating phenomena that occur at each position throughout time, where the sensors are fixed.

It is essential to create tools enabling data scientists to interact with spatial-temporal motif discovery systems to address the abovementioned challenges. In this demo, we present *STMotif Explorer*, a visual tool we developed. In addition to

enabling interaction with discovered motifs, the tool provides the user with an interactive process where it is possible to register discovery algorithms following a canonical data structure and execute these with registered data. These features open up several opportunities, enabling us to compare and view motif discovery functions and see the results in the same environment.

Besides this introduction, the paper is organized into four more sections. Section 2 provides the background. Section 3 describes the *STMotif Explorer*. Section 4 presents the demonstration scenario. Finally, Section 5 provides the conclusion.

2 Spatiotemporal Motif Discovery and Visualization

Borges et al. [2] proposed the Combined Series Approach (*CSA), i.e.*, an approach to discover and rank motifs in spatial-time series. CSA is organized in three main steps: (i) normalization & SAX indexing; (ii) discovery of spatial-time motifs; (iii) ranking of spatial-time motifs. CSA is summarized in Algorithm 1. It takes as input a spatial-time series dataset S, a word size w, an alphabet size a, sb and tb corresponding to spatial and temporal block sizes, and spatiotemporal constraints σ and κ.

Algorithm 1 Combined Series Approach

1: **function** $CSA(S, w, a, sb, tb, \sigma, \kappa)$
2: $S \leftarrow normSAX(S, a)$
3: $STMotifs \leftarrow discoverSTMotifs(S, w, sb, tb, \sigma, \kappa)$
4: $rSTMotifs \leftarrow rankSTMotifs(STMotifs)$
5: return $rSTMotifs$

Visual time series exploration has been extensively studied [4]. However, we find a very restricted bibliography for approaches to motif visualization in space-time series. From them, some missing features are: (i) rank and view space-time motifs to shed light on the most important ones; (ii) compare different approaches over different datasets; (iii) evaluate the motifs found compared to ground truths.

3 STMotif Explorer

We developed *STMotif Explorer* based on the following objectives: (i) spatial-time motifs visualization; (ii) modularity; and (iii) comparison of results. Regarding visualization, the goal is to provide an interactive environment where users can view and explore the motifs discovered in a spatiotemporal dataset. Such interactivity enables a deeper investigation of discovered patterns.

The modularity enables users to register and execute their spatial-time motif detection and ranking algorithms using a canonical data structure. The tool can view and save the results obtained from this processing. Besides the algorithms, it will also be possible to register new spatiotemporal databases, even those with ground-truth results, which the motif discovery algorithms can use. Finally, the comparison feature lets us view Spatial-Temporal Motif Discovery (STMD) algorithms relative to ground truth data. It can be used for performance comparison in a single environment. The tool provides data visualization and statistics of the results.

To provide all these features, we designed *STMotif Explorer* into two main parts depicted in Figure 1.a. The API provides the means to register the algorithms and data to the system. Besides enabling the execution of the registered algorithms, the interface provides functionalities for visual interaction with the data and the results. The interface also provides ways to compare the obtained results. The core algorithms behind the tool are implemented using the R language.

3.1 *STMotif Explorer*: API

The API enables users to interact with the infrastructure through three components.

Register Algorithm and Rank Functions. The system architecture is modular, enabling the addition of new algorithms for execution through the interface. The user can register new functions for motif discovery and ranking. The canonical data structure *STMotifDS* was defined to ease the parameterization of the functions and subsequent comparison of the obtained results. This structure is made up of *Motifs* information, with the positions of its occurrences *MotifPositions*, and other information regarding each motif *MotifInformation*. This information is related to the result of the processing by the algorithm (such as the distance among the occurrences and entropy). Besides these, it is

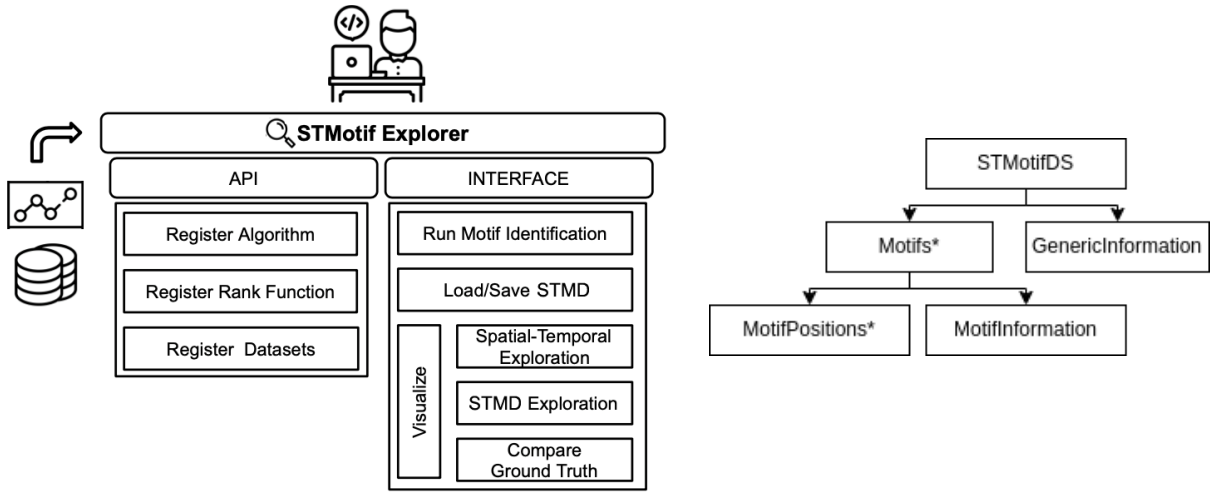

Figure 1: STMotif Explorer Architecture (a); STMotifDS Data Structure (b)

composed of the *GenericInformation* field, where information used by the algorithm is stored (such as the parameters and dataset reference). The structure is summarized in Figure 1.b.

Register Datasets. An important task for verifying and validating an approach is execution across multiple domains. Users can also register new spatiotemporal datasets that can be used in the execution of registered algorithms through the interface. Besides, it is possible to register files containing the results of other algorithms and with the ground truth results, which can be used through the interface. These files are in `RData` format and follow the canonical data structure *STMotifDS*. All files need to be registered, with their signature of the functions, in the file `CONFIG.xml` to become available by the tool.

3.2 *STMotif Explorer*: Interface

The graphical web-based interface (depicted in Figure 2), also implemented in R and JavaScript, provides how users can run the register algorithms and interact with the data. It has options to run motif discovery algorithms (*Run Motif Identification* component). After execution, the *Load/Save STMD* component is invoked to store the results within the system architecture.

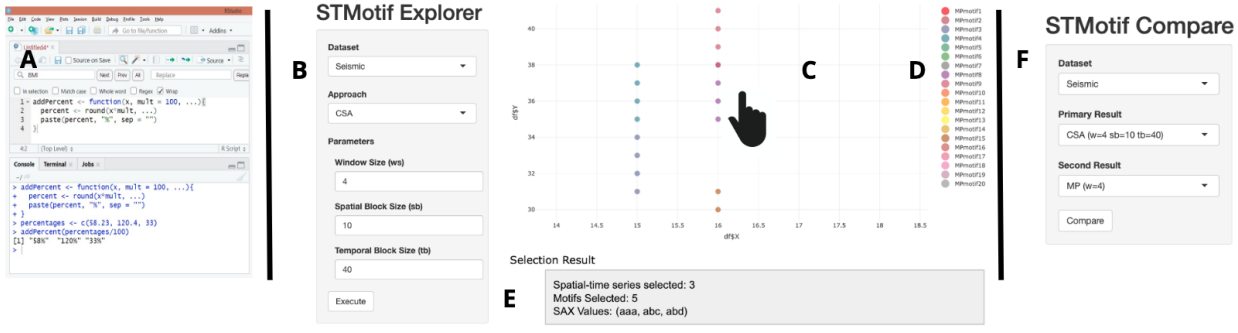

Figure 2: Screenshots of *STMotif Explorer* API and user interface

Through *Exploration* features, the user can interact with the processing results. These features provide data visualization and a panel with updated dynamic information as iterations with the data. The user can view (i) the spatiotemporal dataset (*Spatial-Temporal Exploration*); (ii) the complete set of discovered motifs, with all their occurrences (*STMD Exploration*); (iii) only one motif and its occurrences distributed by the data; (iv) a region of space and time, with the occurrences contained therein. Finally, the tool enables the comparison of results from different approaches in a single space providing visual and statistical information about the data (*Compare Ground Truth*).

4 Demonstration Scenario

This section presents the entire process of applying *STMotif Explorer* in identifying, ranking, and visualizing spatiotemporal motifs in scenarios with real data. Our cases also present the flexibility of the tool, which enables the inclusion of new motif discovery and ranking functions. Also, we present visualization and analysis scenarios of the discovered motifs, comparing different approaches applied in the same domain. This demo uses the seismic spatial-time dataset [3]. *STMotif Explorer* has open source code and a demonstration video that can be obtained through the GitHub repository[1].

Scenario 1: API. In this scenario, we present the modularity of the tool by receiving new functions and data. Figure 2A presents an example of the $MPMotifDiscovery()$ function written in the R language for motif discovery. This function uses the *TSMP* R package [1], which implements the Matrix Profile [12]. The Matrix Profile (MP) approach is based on computing the distance of a sequence with the most similar subsequence present in the time series. Two other functions are already provided by the architecture, one for STMD, $SearchSTMotif()$ and the $RankSTMotif()$ ranking function, both using the STMotif R Package, which reflects the CSA approach [2]. Then the ground truth results of the seismic dataset motifs are recorded to compare the results.

Scenario 2: Run Motif discovery. Once the functions and datasets are registered, the user can use the functionalities available in the interface to execute the algorithms. Figure 2B shows the menu to select the dataset, the STMD, and ranking functions. When the user selects the approach to be executed, fields for the parameter values are presented. In this scenario, we execute the CSA approach using the default parameter values defined in the API. It uses the $RankSTMotif()$ ranking function with the seismic dataset. At the end of execution, the tool informs the user about the end of the process. It saves the results in the architecture in an `RData` format file, following the canonical data structure, and is now available for use in the other features.

Scenario 3: Data Visualization. Figure 2C shows the arrangement of occurrences of the discovered motifs. The user also can iterate with the data at different levels of detail by magnifying the image. It is possible to select a region of space and view only the motifs found in this area. In the menu shown in Figure 2D, the discovered motifs are listed, and sorted according to the previous ranking, starting with the best ranking. Each motif represents a set of the same occurrences and is presented in a different color. The user can select a subset of motifs to view through this menu. We choose the information related to the best-ranked motif, its occurrences, and its signal in the space-time series (Figure 2E).

Scenario 4: Comparison. It offers users a tool to compare the results of different approaches and the result of one approach with its corresponding ground truth (Figure 2F). To obtain the result of the MP approach, we performed the motif discovery process of the *MPMotifDiscovery* function. Then, we compare it with the ground truth. Users can check that the approach is inaccurate, given the statistical results presented in the information table, even with many occurrences discovered. The visualization analysis confirms this argument since it even returns multiple occurrences. They are visibly distant from the correct occurrences.

5 Conclusion

This paper introduces a Spatial-Temporal Motif Visualization tool, the *STMotif Explorer*, that effectively addresses the gap of visualized constrained space-time motifs. It provides a comprehensive system for interactive discovery, visualization, and comparative analysis of motifs. The tool enables filtering and ranking of motifs during the visualization and provides ways to explore motifs with ground truth data. This feature helps researchers and practitioners evaluate different algorithms and the quality of discovered motifs in various domains.

Acknowledgements

The authors thank CNPq, CAPES, and FAPERJ for partially sponsoring this research.

References

[1] F. Bischoff and P. Rodrigues. tsmp: An R Package for Time Series with Matrix Profile. *R Journal*, 12(1):76–86, 2020.

[1] Available at `https://github.com/cefet-rj-dal/STMotifexplorer`.

[2] H. Borges, M. Dutra, A. Bazaz, R. Coutinho, F. Perosi, F. Porto, F. Masseglia, E. Pacitti, and E. Ogasawara. Spatial-time motifs discovery. *Intelligent Data Analysis*, 24(5):1121–1140, 2020.

[3] dgbes. Netherlands Offshore F3 Block - Complete. Technical report, https://opendtect.org/osr/Main/NetherlandsOffshoreF3BlockComplete4GB, 2018.

[4] P. Eichmann, N. Tatbul, F. Solleza, and S. Zdonik. Visual exploration of time series anomalies with metro-viz. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 1901–1904, 2019.

[5] J. Lin, E. Keogh, S. Lonardi, J. P. Lankford, and D. M. Nystrom. VizTree: a tool for visually mining and monitoring massive time series databases. In *Proceedings of the Thirtieth international conference on Very large data bases - Volume 30*, VLDB '04, pages 1269–1272, Toronto, Canada, aug 2004. VLDB Endowment.

[6] M. Linardi, Y. Zhu, T. Palpanas, and E. Keogh. Matrix profile goes MAD: variable-length motif and discord discovery in data series. *Data Mining and Knowledge Discovery*, 34(4):1022–1071, 2020.

[7] T. Oates, A. Boedihardjo, J. Lin, C. Chen, S. Frankenstein, and S. Gandhi. Motif discovery in spatial trajectories using grammar inference. In *International Conference on Information and Knowledge Management, Proceedings*, pages 1465–1468, 2013.

[8] E. Ramanujam and S. Padmavathi. Comprehensive review on time series motif discovery using evolutionary techniques. *International Journal of Advanced Intelligence Paradigms*, 23(1-2):155–170, 2022.

[9] P. Senin, J. Lin, X. Wang, T. Oates, S. Gandhi, A. Boedihardjo, C. Chen, S. Frankenstein, and M. Lerner. GrammarViz 2.0: A tool for grammar-based pattern discovery in time series. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8726 LNAI(PART 3):468–472, 2014.

[10] S. Shekhar, S. Feiner, and W. Aref. Spatial computing. *Communications of the ACM*, 59(1):72–81, 2016.

[11] S. Torkamani and V. Lohweg. Survey on time series motif discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(2), 2017.

[12] C.-C. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. Dau, Z. Zimmerman, D. Silva, A. Mueen, and E. Keogh. Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile. *Data Mining and Knowledge Discovery*, 32(1):83–123, 2018.