GSTSM Package: Finding Frequent Sequences in Constrained Space and Time

Antonio Castro, Heraldo Borges, Cassio Souza, Jorge Rodrigues, Fábio André Machado Porto, Esther Pacitti, Rafaelli Coutinho, Eduardo Ogasawara

GSTSM Package: Finding Frequent Sequences in Constrained Space and Time

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

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

Cassio Souza
CEFET/RJ
cassio.souza@eic.cefet-rj.br

Jorge Rodrigues
CEFET/RJ
jorge.rodrigues@eic.cefet-rj.br

Fabio Porto
LNCC
fporto@lncc.br

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

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

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

Data mining tools have been used to find interesting patterns in different areas of knowledge in various problems [1]. The sequence mining knowledge area is a specialization of data mining, focused on finding sequences or series of events in datasets [12, 15]. Such sequences may form patterns, a set of frequent attributes that appear persistently among the dataset. It means that its frequency exceeds a user-defined minimum threshold [3].

Several types of events involve both temporal and spatial data, such as financial to understand sales patterns over time and space [1], and hydrological data for river water quality monitoring in different points over time [1, 2]. They correspond to Time-stamped Sequence (TS) events distributed in space [11]. Mining sequences related to space and time enables to find knowledge related to phenomena that involve both spatial and temporal components, trying to find all sequences of significant, useful, interesting, and non-trivial events [3, 13, 1].

However, spatio-temporal sequential patterns may have low support if considered the entire dataset, but they can be frequent if considered only a period and region [8]. The Generalized Spatial-Time Sequence Miner (*GSTSM*) package can find these patterns, being able to efficiently discover the region and period where they occur. This way, *GSTSM* would be the right tool to find time-localized patterns.

This work describes the process, structure, and usage of the *GSTSM* package using a synthetic (but still complete) example. However, we also provide a glimpse of applicability in a real-world dataset. *GSTSM* was able to found sequential patterns in seismic data. They correspond to seismic horizons, which are important elements in the application domain.

2 Related Work

There are different methods for spatio-temporal data mining. Some use only data mining, searching for frequent patterns, considering only time [10]. Others combine techniques by seeking in time and then grouping in space [7]. Furthermore, there is a diversity in how constraints are handled. Some use global support, a value that is valid for the entire dataset [2]. Others consider local support, using predefined windows of time and space [9].

This work differs by seeking frequent sequences in time that occur in spatial groups. Instead of using predefined constraints for time and space, three density parameters are established: a minimum frequency to be achieved within the period, a maximum distance that a position can be from any other in the group, and a lower limit of distinct positions in the group. Thus, the formalization presented in this work can find different sizes of sequences, time intervals, and spatial regions where a sequence is frequent, based on the concepts of RG, KRG, and SRG introduced in [5].

As far as the conducted research has reached, the only work with a similar approach found in the literature is proposed by [4], which considers one-dimensional space. The present work is a generalization that presents a formalization considering space in its three-dimensional form.

3 Demonstration overview

GSTSM is a package which provides polymorphic functions that let the user extends its functionalities, as it is based on *R* [14] language *S3* classes. The source code can be found at GitHub [6]. The package has a main class named *GSTSM*, that needs the parameters D, P, γ, β, and σ to instantiate an object, explained as follows:

- `D` and `P` represents the TS dataset with their respective positions. Each TS must be associated with one position. It means that the number of timestamped sequences (columns) in `D` must be equal to the number of positions (rows) in `P`.
- for the user defined thresholds values in the range $]0, 1]$ for γ, values starting from 2 for β, and integer values starting from 1 for σ.

A *GSTSM* object is an instance of S3 Class built as a list with all this information. Furthermore, it generates an adjacency matrix that informs each position which other positions are at a maximum distance of σ.

The *GSTSM* package has the *mine()* method that implements the entire process of finding frequent sequences. It receives as input a *GSTSM* object and provides as output a list of the SRGs of all sizes found. The user does not need to call any other method to get the results. This method calls and passes all the necessary parameters to make the entire process transparent to the user.

The other methods used in each process step are polymorphic and can be extended by the user. It gives the user the ability to try its implementation. These are described as follows:

- *find()* has two input parameters: a *GSTSM* object and a set of candidate sequences of size k. It provides as output the KRGs for each candidate.
- *merge()* has also two input parameters: a *GSTSM* object and a set of candidate sequences of size k containing information about the KRG of each one. The method returns the SRGs with the candidate sequences of size k.
- *generate_candidates()* has two input parameters: a *GSTSM* object and a set of SRGs of size k. There are no SRGs to pass to generate candidates of size one, a `NULL` value can be used. The method provides the candidate sequences of size $k + 1$.

An illustrative example shows the use of the *GSTSM* package functions. To start, the first action is installing and loading *GSTSM* package and then setting all the inputs for the package: D, P, γ, β, and σ. For D, we use a simple dataset. For P, positions in a row are used, with one unit distance. Each position is associated with a time series, such as p_1 to t_1 and p_2 to t_2. The values for the user-defined thresholds are: $\gamma = 0.8$, $\beta = 2$, and $\sigma = 1$. After setting the input parameters, we can instantiate the *GSTSM* object and execute the *mine()* method. Listing 1 shows the code using the R command line.

Listing 1: R example of the use of *GSTSM*

```
# loading the GSTSM package
> library("gstsm")
```

```
# loading Spatial Timestamped Sequence
> path <-
"https://eic.cefet-rj.br/~dal/wp-content/uploads/2023/05/"
> load(url(paste(path, "dataset.rdata"))) # dataset D
> load(url(paste(path, "positions.rdata"))) # positions P
# mining dataset
> gstsm_object <- gstsm(D, P, gamma=0.8, beta=2, sigma=1)
> result <- mine(gstsm_object)
```

4 Conclusion

GSTSM is the first tool for mining sequences in spatial time-stamped sequences datasets able to discover constrained patterns in time and space with all three dimensions. The package discovers patterns that may not be frequent over an entire dataset but are grouped in space and frequent in a time interval. It would not be easy to find these patterns without this tool. The results can differ from conventional data mining tools and give different insights about data behavior. The patterns are groups of positions and periods where the sequences are frequent according to the input parameters. The package is also extensible, enabling users to incorporate heuristics and optimizations to drive the discovery of patterns.

Acknowledgements

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

References

[1] H. Alatrista-Salas, J. Azé, S. Bringay, F. Cernesson, N. Selmaoui-Folcher, and M. Teisseire. A knowledge discovery process for spatiotemporal data: Application to river water quality monitoring. *Ecological Informatics*, 26(P2):127–139, 2015.

[2] H. Alatrista-Salas, S. Bringay, F. Flouvat, N. Selmaoui-Folcher, and M. Teisseire. Spatio-sequential patterns mining: Beyond the boundaries. *Intelligent Data Analysis*, 20(2):293–316, 2016.

[3] B. Aydin and R. Angryk. Spatiotemporal event sequence mining from evolving regions. In *Proceedings - International Conference on Pattern Recognition*, volume 0, pages 4172–4177, 2016.

[4] R. Campisano, H. Borges, F. Porto, F. Perosi, E. Pacitti, F. Masseglia, and E. Ogasawara. Discovering tight space-time sequences. In *Lecture Notes in Computer Science*, volume 11031, pages 247–257, 2018.

[5] A. Castro, H. Borges, R. Campisano, E. Pacitti, F.Porto, R. Coutinho, and E. Ogasawara. Generalização de Mineração de Sequências Restritas no Espaço e no Tempo. In *Anais do Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 313–318. SBC, oct 2021.

[6] A. Castro, C. Souza, J. Rodrigues, E. Pacitti, F. Masseglia, R. Coutinho, and E. Ogasawara. GSTSM Package. Technical report, https://cran.rstudio.com/web/packages/gstsm/index.html, CRAN, 2022.

[7] C. Flamand, M. Fabregue, S. Bringay, V. Ardillon, P. Quénel, J. Desenclos, and M. Teisseire. Mining local climate data to assess spatiotemporal dengue fever epidemic patterns in French Guiana. *Journal of the American Medical Informatics Association : JAMIA*, 21(e2):e232–240, 2014.

[8] Y. Huang, L. Zhang, and P. Zhang. A framework for mining sequential patterns from spatio-temporal event data sets. *IEEE Transactions on Knowledge and Data Engineering*, 20(4):433–448, 2008.

[9] B. Koseoglu, E. Kaya, S. Balcisoy, and B. Bozkaya. ST Sequence Miner: visualization and mining of spatio-temporal event sequences. *Visual Computer*, 36(10-12):2369–2381, 2020.

[10] K. Li and Y. Fu. Prediction of human activity by discovering temporal sequence patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1644–1657, 2014.

[11] C. Mooney and J. Roddick. Sequential pattern mining - Approaches and algorithms. *ACM Computing Surveys*, 45(2), 2013.

[12] S. Parthasarathy, M. Zaki, M. Ogihara, and S. Dwarkadas. Incremental and interactive sequence mining. In *International Conference on Information and Knowledge Management, Proceedings*, pages 251–258, 1999.

[13] G. Sunitha and A. Rama Mohan Reddy. Mining frequent patterns from spatiotemporal data sets: A survey. *Journal of Theoretical and Applied Information Technology*, 68(2):265–274, 2014.

[14] R. C. Team. R: A Language and Environment for Statistical Computing. Technical report, https://www.R-project.org/, Vienna, Austria, 2020.

[15] M. J. Zaki. Sequence mining in categorical domains: incorporating constraints. In *Proceedings of the ninth international conference on Information and knowledge management*, CIKM '00, pages 422–429, New York, NY, USA, nov 2000. Association for Computing Machinery.