Comparing Motif Discovery Techniques with Sequence Mining in the Context of Space-Time Series

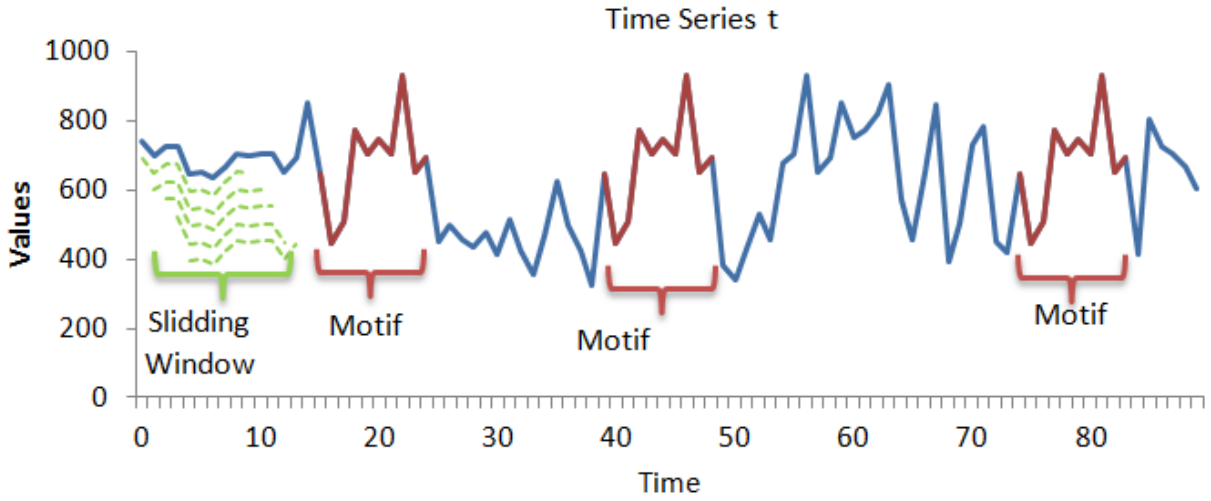
CEFET/RJ

Eduardo Ogasawara
eogasawara@ieee.org
http://eic.cefet-rj.br/~eogasawara

Discovering Space-Time Motifs

Discovering motifs in time series

- Data deluge scenario pushes us for new ways for collecting, storing, processing a large amount of data
- Many phenomena can be observed and organized as time series (sequences of observations)
- A relevant area that is being explored in time series analysis is finding patterns
- A particular pattern that occurs a significant number of times in time series is denominated motif [1]
- Enables the understanding of some specific behaviors observed in time series, in many areas of knowledge, such as weather prediction, wind generation, image recognition, seismic amplitude
- A vast number of motifs discovery techniques, methods, and algorithms have been developed [2,3]
  - They include discovering motifs of a variable length, without constraints (parameter-free) , multivariate time series



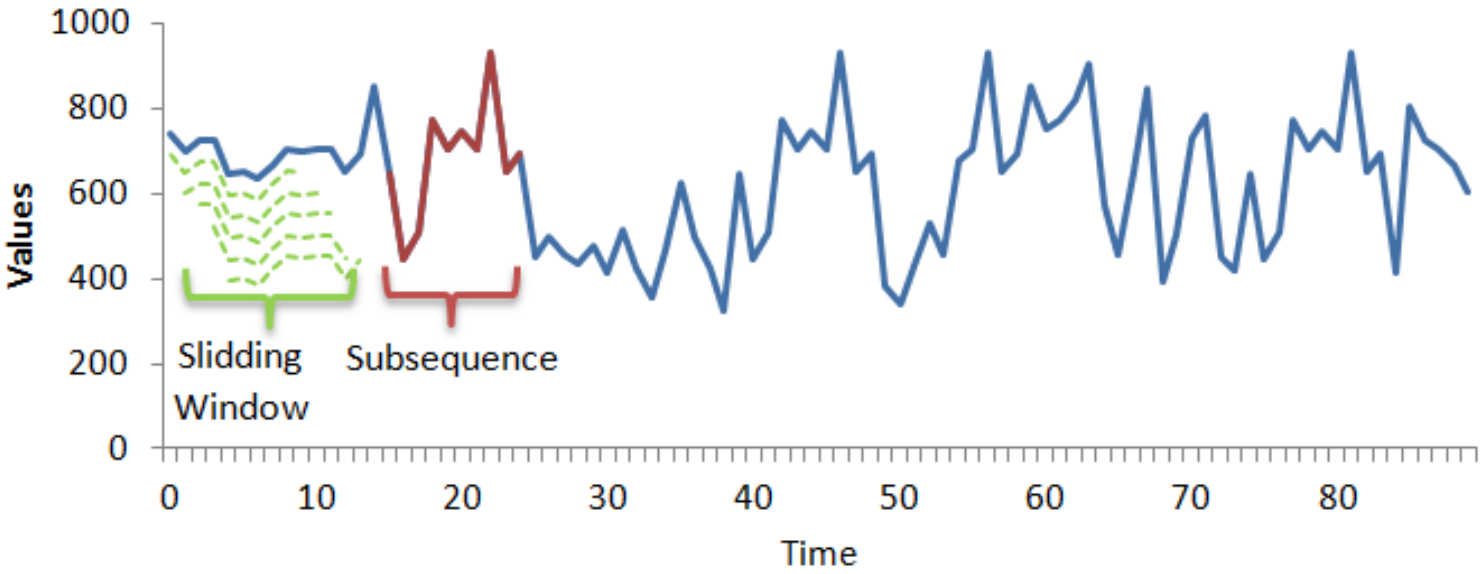
[1] P. Patel, E. Keogh, J. Lin, and S. Lonardi, "Mining motifs in massive time series databases," in Proceedings - IEEE International Conference on Data Mining, ICDM, 2002, pp. 370–377
[2] A. Mueen, "Time series motif discovery: Dimensions and applications," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 4, no. 2, pp. 152–159, 2014
[3] S. Torkamani and V. Lohweg, "Survey on time series motif discovery," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 7, no. 2, 2017.

Discovering motifs in spatial-time series

- Various time-series phenomena present different behaviors when observed at points of space
  - Series collected by sensors and IoT
- Spatial-time series: each time series is associated to a position in space
  - Fixed position: Analysis of points/regions
  - Variable position: trajectory data
- Motifs might not be discovered when analyzing each time series
  - They may be frequent if we consider different spatial-time series at some time interval or some spatial range
  - Finding patterns that are frequent in a constrained space and time, *i.e.*, "finding spatial-time motifs", may enable us to comprehend how a phenomenon occurs concerning space and time
- Formalize spatial-time motifs
  - Spatial-time motif needs to occur both in time and space to become interesting
- Present an approach to discover them
- How to rank spatial-time motifs
  - There can be many discovered motifs

Background: time series

- A **time series** t is an ordered sequence of values in time: $< t_1, \cdots, t_m >, t_i \in \mathbb{R}$
- A **subsequence** is a continuous sample of a time series with a defined length
  - $seq_{n,p}(t)$: p-th **subsequence** of size n in a time series t is a sequence $< t_p, \ldots, t_{p+n-1} >$, where $|seq_{n,p}(t)| = n$ and $1 \leq p \leq |t| - n$
- **Sliding windows** consist in exploring all possible subsequences of a time series
  - function $sw_n(t)$ produces a matrix W of size $(|t| - n + 1)$ by n
    - each line w_i in W is the i-th subsequence of size n from t
  - Given $W = sw_n(t)$, $\forall w_i \in W, w_i = seq_{n,i}(t)$



A spatial-time series st is a pair (t, p), such that a time series t with an associated position p

Formalizing motifs discovery in time series

- Given a sequence q and time series t, q is a **motif** in t with support σ, if and only if q is included in t at least σ times
- The length of a motif q ($|q|$) is also known as word size
- Given a sequence q and a time series t where $W = sw_{|q|}(t)$, $motif(q, t, \sigma) \leftrightarrow \exists\ R \subseteq W$, $(|R| \geq \sigma)$, such that $\forall w_i \in R$, $w_i = q$
  - Motifs are not previously known and are discovered when scanning the entire data [1]
- Many approaches were proposed in the literature to discover motifs in time series
  - They require some data preprocessing such as normalization and indexing before running the motif discovery algorithms to increase the performance and precision of results

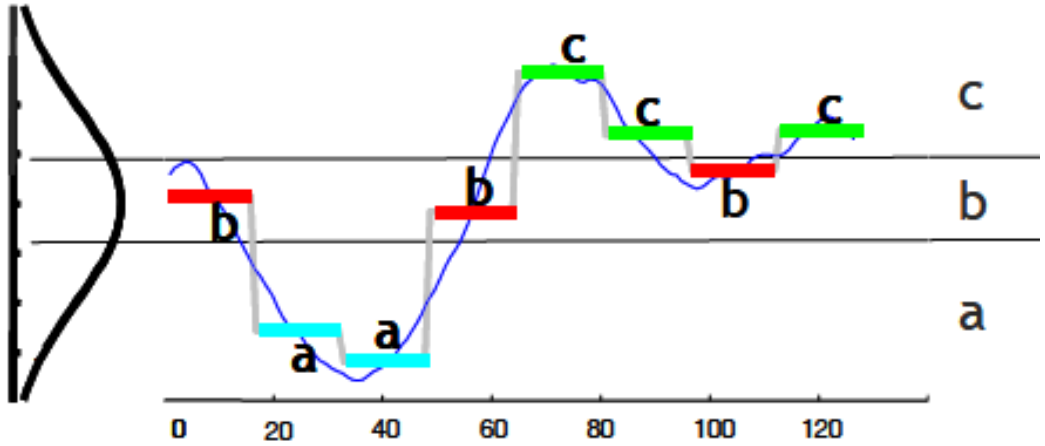
[1] A. Mueen, "Time series motif discovery: Dimensions and applications," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 4, no. 2, pp. 152–159, 2014

Background: Normalization

- Normalization is commonly used to adjust scale of data
  - Z-score apply a linear transformation where t_i is an observation of the time series t, μ_t is the average, σ_t is the standard deviation of the time series, and t' is the transformed time series with mean equals to zero and one as standard deviation
    - $t_i' = \frac{(t_i - \mu_t)}{\sigma_t}$
  - Min-max apply a linear transformation to the original data, where the minimum value ($min(t)$) and the maximum value ($max(t)$) are used to transform each value t_i of to another value t_i' in a range varying from $[0,1]$
    - $t_i' = \frac{t_i - min(t)}{max(t) - min(t)}$

Background: SAX

- SAX is an indexing technique that partitions the domain of a variable into ranges such that each range is associated with a particular symbol [1]
- The SAX alphabet size defines the number of partitions for the domain (ex: a, b, c)



[1] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: a novel symbolic representation of time series," Data Mining and Knowledge Discovery, vol. 15, no. 2, pp. 107–144, 2007

Background: Approaches for motifs discovery

- Brute force approach is the simplest method
  - It has a high computational cost [2] and is indicated for discovering sequences of smaller size [3]
  - The coverage and accuracy are complete. It makes all possible comparisons
- The random projections approach was proposed to handle large dataset by reducing dimensionality
  - Optimizes search by randomly selects some of sliding windows columns for search [4]
  - Collision matrix that masks the projected columns (subsequence matrix and candidate search sequence)
- Sort the motifs according to their relevance [5]
  - A standard classification method is k-motif which considers the number of occurrences of the motifs in time series
  - Sorted according to their relevance degree (some motifs can be similar to a straight line, i.e., may not be relevant)
  - Such motifs can be low qualified or discarded to avoid distorting the analysis
  - Assess the relevance of motifs based how expected is the motif to occurs [5]

[2] A. Mueen, E. Keogh, Q. Zhu, S. Cash, and B. Westover, "Exact discovery of time series motifs," in Society for Industrial and Applied Mathematics - 9th SIAM International Conference on Data Mining 2009, Proceedings in Applied Mathematics, 2009, vol. 1, pp. 469–480.
[3] L. Li and S. Nallela, "Probabilistic discovery of motifs in water level," in 2009 IEEE International Conference on Information Reuse and Integration, IRI 2009, 2009, pp. 388–393
[4] J. Buhler and M. Tompa, "Finding motifs using random projections," Journal of computational biology, vol. 9, no. 2, pp. 225–242, 2002
[5] N. C. Castro and P. J. Azevedo, "Significant motifs in time series," Statistical Analysis and Data Mining, vol. 5, no. 1, pp. 35–53, 2012.

Related work: motifs in spatial-time

- Oates [1] focused on analyzing repetitive sequences of moving objects
  - For that, they developed a grammar, applied SAX indexing, and searched for motifs over trajectory
  - Difference: we do not have a moving object. Sensors are fixed, and we analyze a phenomenon that occurs at each position throughout the time
- Du [2], space is modeled by discrete attributes that resemble states of an object
  - they refer to the state of companies in the stock market
  - It is state-space model where a trajectory is the registration of state transitions
  - Differences: The modeled phenomenon may not be constrained in space and time

[1] T. Oates, A. P. Boedihardjo, J. Lin, C. Chen, S. Frankenstein, and S. Gandhi, "Motif discovery in spatial trajectories using grammar inference," in Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, 2013, pp. 1465–1468.
[2] X. a Du, R. a Jin, L. b Ding, V. E. a Lee, and J. H. b T. Jr, "Migration motif: A spatial-temporal pattern mining approach for financial markets," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 1135–1143

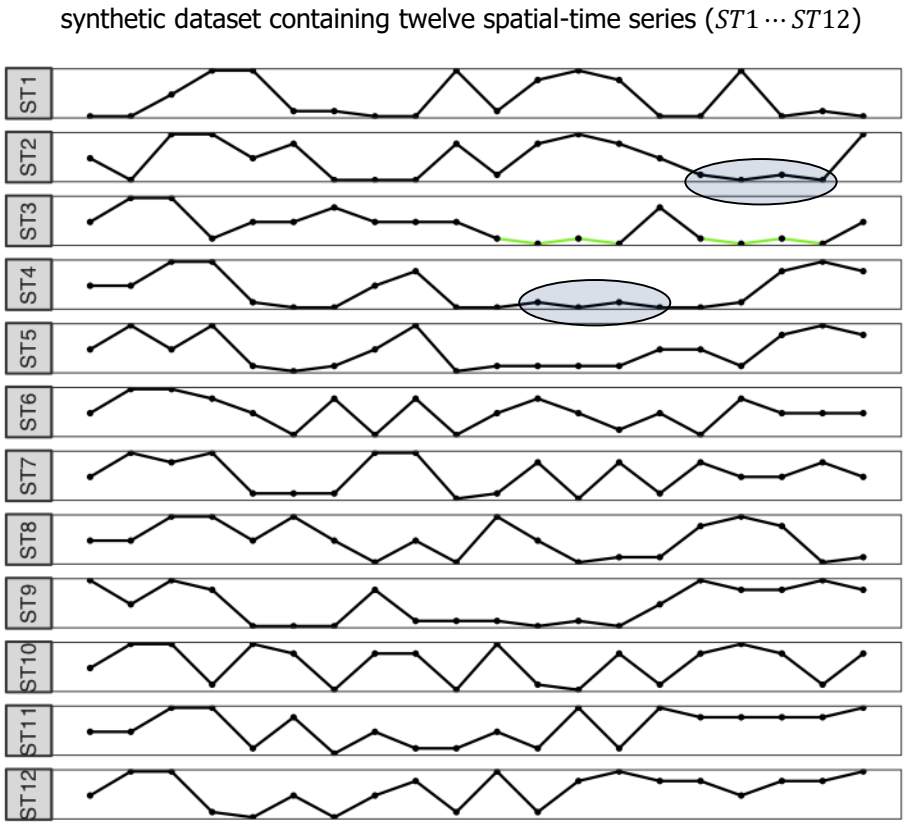
Related work: Multivariate time series

- Tanaka [1] applied the method to the multidimensional time-series transforming into one-dimensional time-series using the Principal Component Analysis
- Son [7] proposed two new algorithms: one based on R-tree and the other is based on dimensionality reduction through Skyline index

[1] . Tanaka, K. Iwamoto, and K. Uehara, "Discovery of time-series motif from multi-dimensional data based on MDL principle," Machine Learning, vol. 58, no. 2–3, pp. 269–300, 2005
[7] N. T. Son and D. T. Anh, "Discovery of time series k-motifs based on multidimensional index," Knowledge and Information Systems, vol. 46, no. 1, pp. 59–86, 2016

Discovering Motifs in Space-Time Series

- Spatial-time series is a more complex scenario
- known motif discovery method on each spatial-time series for a support $\sigma \geq 2$
  - green worm-like found only in $ST3$
  - Other equivalent worm-like shape are not discovered: $(ST2, ST4)$



Discovering Space-Time Motifs

- We are interested in finding motifs that occur in a constrained space and time
- A **block** b is a couple $(\{st\}, i)$ where $\{st\}$ is a set of neighbors spatial-time series and i is a time interval
  - The size of a block b is the product of the number of spatial-time series with interval length: $|b| = |st| \cdot |i|$
- Let B be a partition of S into blocks b.
- Let σ and κ be two support values such that $\sigma \geq \kappa$
  - A subsequence q is a **spatial-time motif** if and only if there exists a block b such that q is included at least σ times in it and q occurs in at least κ different spatial-time series inside b
- From the definition above, the problem can be summarized as *the discovery of spatial-time motifs in a spatial-time series dataset*

Combined Series Approach (CSA)

- $\boldsymbol{CSA(D, w, sb, tb, \sigma, \kappa)}$
  - $DS \leftarrow normSAX(D, a)$
  - $stmotifs \leftarrow discoverSTMotifs(DS, w, sb, tb, \sigma, \kappa)$
  - $rstmotifs \leftarrow rankSTMotifs(stmotifs)$
  - return $rstmotifs$

- $discoverSTMotifs(DS, w, sb, tb, \sigma, \kappa)$
  - $B \leftarrow partition(DS, sb, tb)$
  - $stmotifs \leftarrow \emptyset$
  - $cs \leftarrow combine(b_{i,j})$
  - $motifs \leftarrow discover(cs, w, \sigma)$
  - $stmotifs \leftarrow validate(motifs, \sigma, \kappa) \cup stmotifs$
  - return $stmotifs$

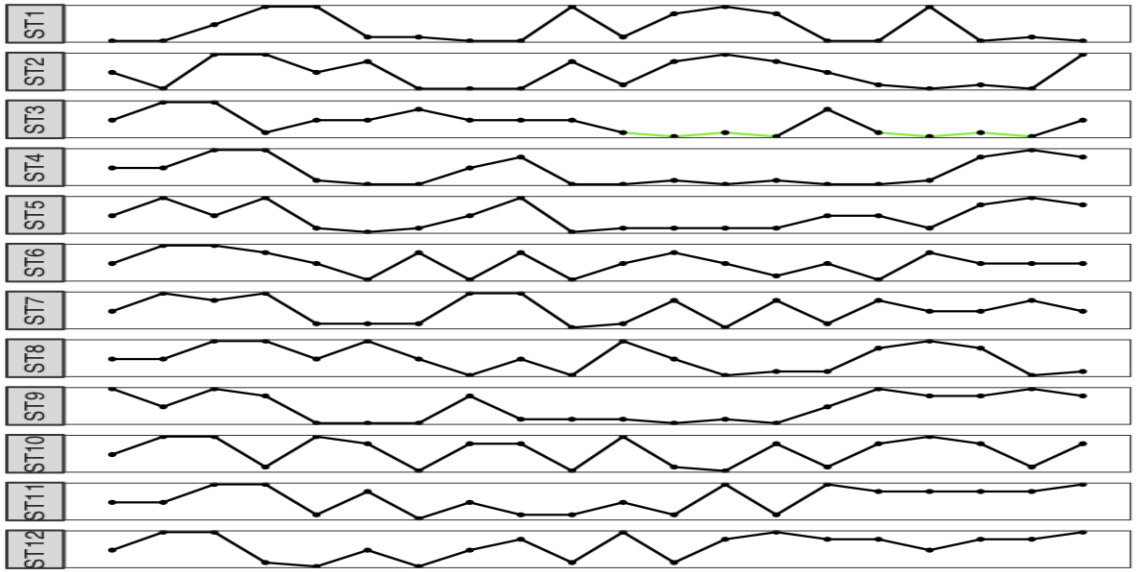
- $normSAX(D, a)$
  - $Dz \leftarrow zscore(D)$
  - $DS \leftarrow SAX(Dz, a)$
  - return DS

- $rankSTMotifs(stmotifs)$
  - $stmotifs \leftarrow group(stmotifs)$
  - $ent_i = \sum_{k=1}^{|ft(m_i)|} \left(\frac{ft(m_i)_k}{n} \cdot log_2 \left(\frac{ft(m_i)_k}{n} \right) \right)$
  - $O_i \leftarrow occurrences(m_i)$
  - $occ_i \leftarrow log_2(O_i)$
  - $dist_i \leftarrow \frac{1}{aw(mst(wam(O_i)))}$
  - $rank = proj(norm(ent, occ, dist))$
  - return $order(stmotifs, rank)$

Analysis Using Synthetic Dataset: Normalization and indexing

(i) alphabet size ($a = 5$)
(ii) spatial block size ($sb = 4$)
(iii) temporal block size ($tb = 10$)
(iv) word size ($w = 4$)
(v) thresholds $\sigma = 2$ and $\kappa = 2$

$normSAX$



tb = 10 tb = 10

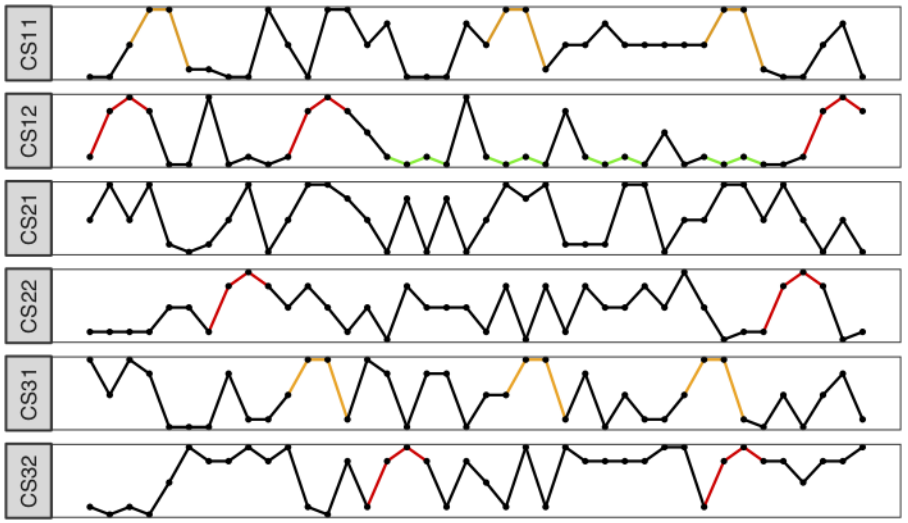
sb = 4 sb = 4 sb = 4

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ST1	a	a	c	e	e	b	b	a	a	e	b	d	e	d	a	a	e	a	b	a
ST2	c	a	e	e	c	d	a	a	a	d	b	d	e	d	c	b	a	b	a	e
ST3	c	e	e	b	c	c	d	c	c	c	b	a	b	a	d	b	a	b	a	c
ST4	c	c	e	e	b	a	a	c	d	a	a	b	a	b	a	a	b	d	e	d
ST5	c	e	c	e	b	a	b	c	e	a	b	b	b	b	c	c	b	d	e	d
ST6	c	e	e	d	c	a	d	a	d	a	c	d	c	b	c	a	d	c	c	c
ST7	c	e	d	e	b	b	b	e	e	a	b	d	a	d	b	d	c	c	d	c
ST8	c	c	e	e	c	e	c	a	c	a	e	c	a	b	b	d	e	d	a	b
ST9	e	c	e	d	a	a	a	d	b	b	b	a	b	a	c	e	d	d	e	d
ST10	c	e	e	b	e	d	a	d	d	a	e	b	a	d	b	d	e	d	b	d
ST11	c	c	e	e	b	d	a	c	b	b	c	b	e	b	e	d	d	d	d	e
ST12	c	e	e	b	a	c	a	c	d	b	e	b	d	e	d	d	c	d	d	e

Analysis Using Synthetic Dataset: Combining series and discovering motifs

- Dataset has 12 spatial-time series and 20 observations
- $combine$
  - The dataset is partitioned into 6 blocks (40 observations)
  - $sb = 4$ and $tb = 10$
- $discover$
  - In each cs, the $discover$ identifies all motif with $\sigma \geq 2$
- $validate$
  - Check κ constraint

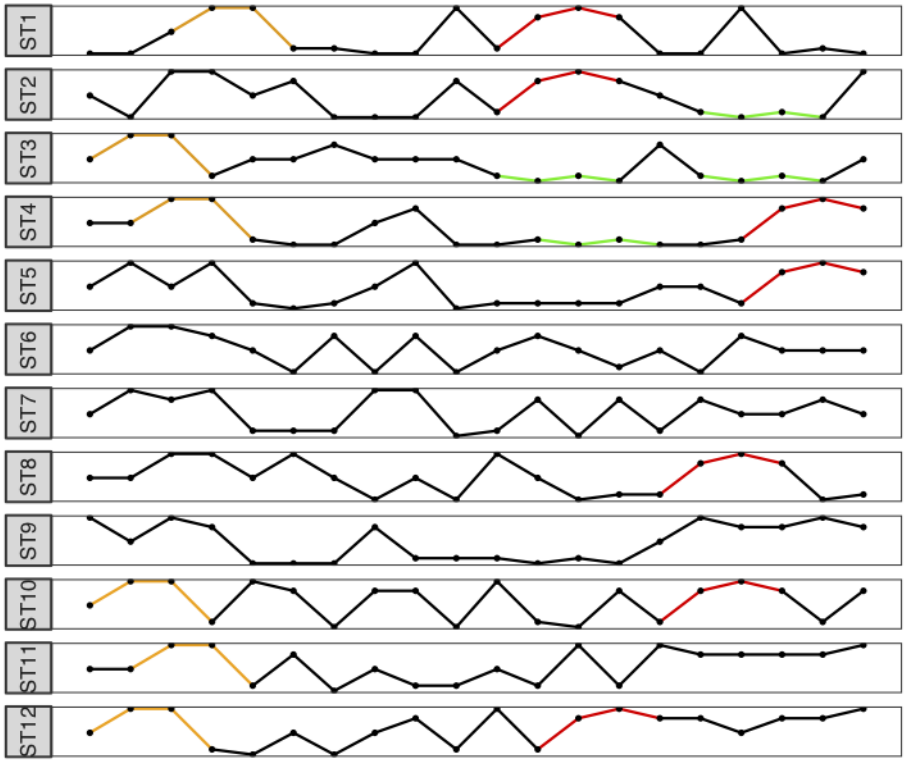
Combined time series



Analysis Using Synthetic Dataset: Combining series and discovering motifs

- Motifs discovered are marked with colors red, green, and orange
- The majority of motifs discovered using the CSA approach had not been found earlier

Spatial-time series



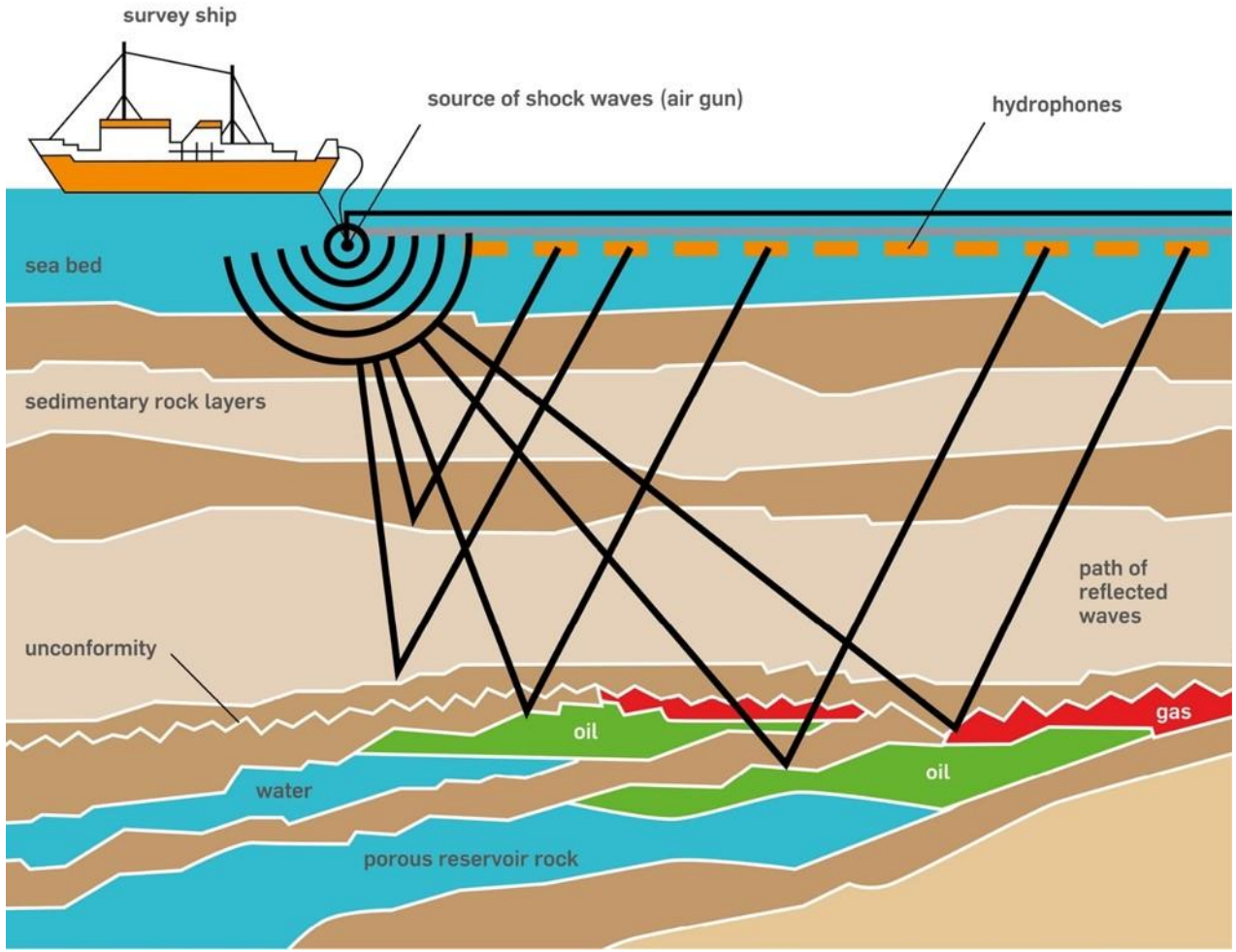
Analysis Using Synthetic Dataset: Ranking

- Motifs are grouped according neighbor blocks
- Dimensions used to rank the identified motifs
  - Entropy (Ent.) for both $bded$ and $ceeb$ was 1.5 (3/4 distinct characters)
  - Distance metric (Dist.) is the reciprocal of the average weight of the minimum spanning tree that connects identified occurrences for each motif (the closer to one, the better it is)
  - Occurrences (Occ.) consider the log_2 of the occurrences
- Ranking (Rank) combines the normalized dimensions (Ent., Dist. and Occ.) projecting it to normalized vector $(\sqrt{\frac{1}{3}}, \sqrt{\frac{1}{3}}, \sqrt{\frac{1}{3}})$

Motif	Trad.	CSA	Ent.	Dist.	Occ.	Rank
bded	-	7	1.5	0.53	2.81	1.52
ceeb(1)	-	3	1.5	0.71	1.58	1.17
baba	2	4	1.0	0.83	2.00	0.95
ceeb(2)	-	3	1.5	0.47	1.58	0.71

Seismic Analysis Example

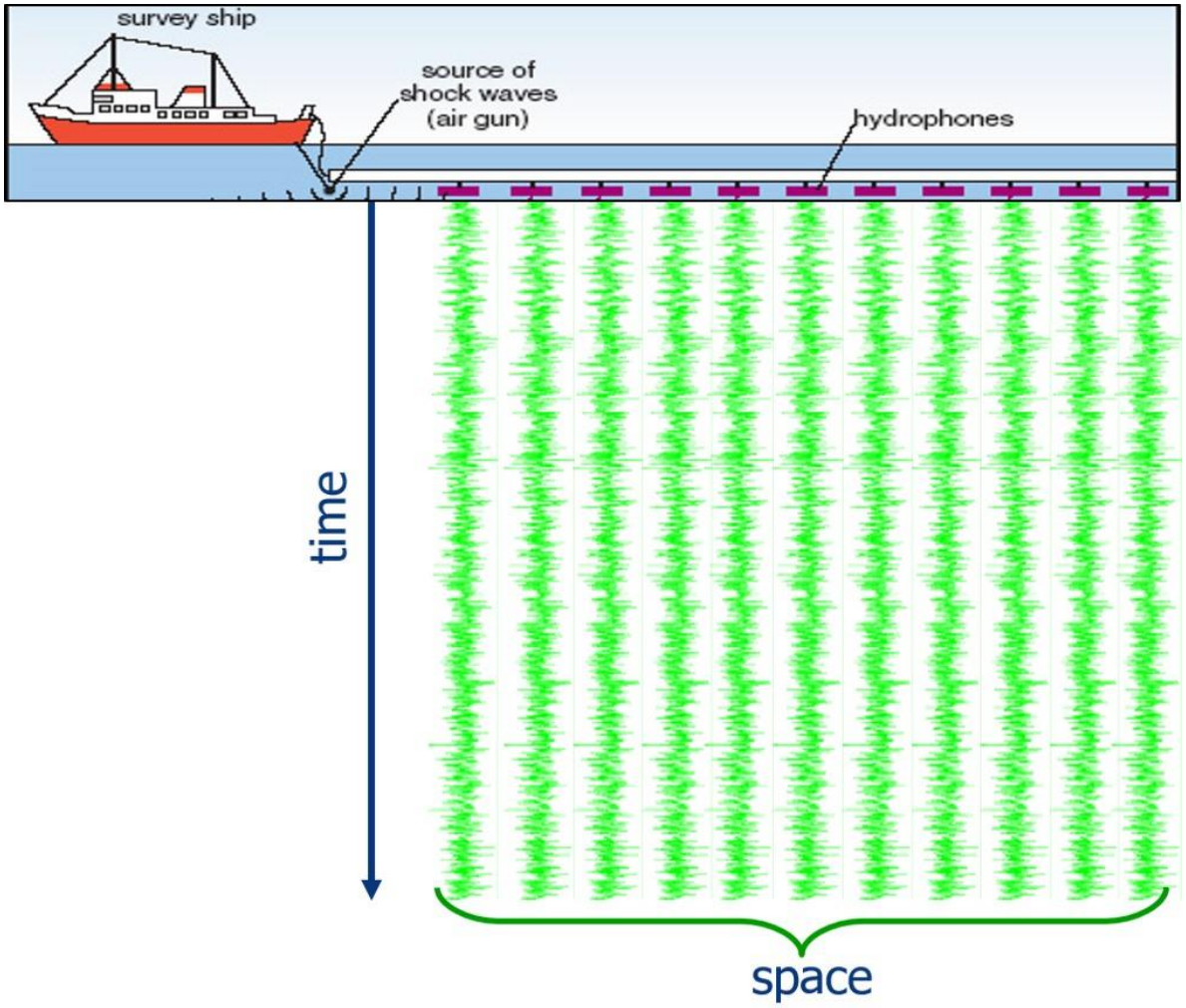
We applied CSA in the Netherlands seismic spatial-time series dataset, named F3 Block
Each spatial-time series has a position in which the geophone or hydrophone is placed



Source: https://krisenergy.com/company/about-oil-and-gas/exploration/

Seismic Analysis Example
Spatial-time series dataset

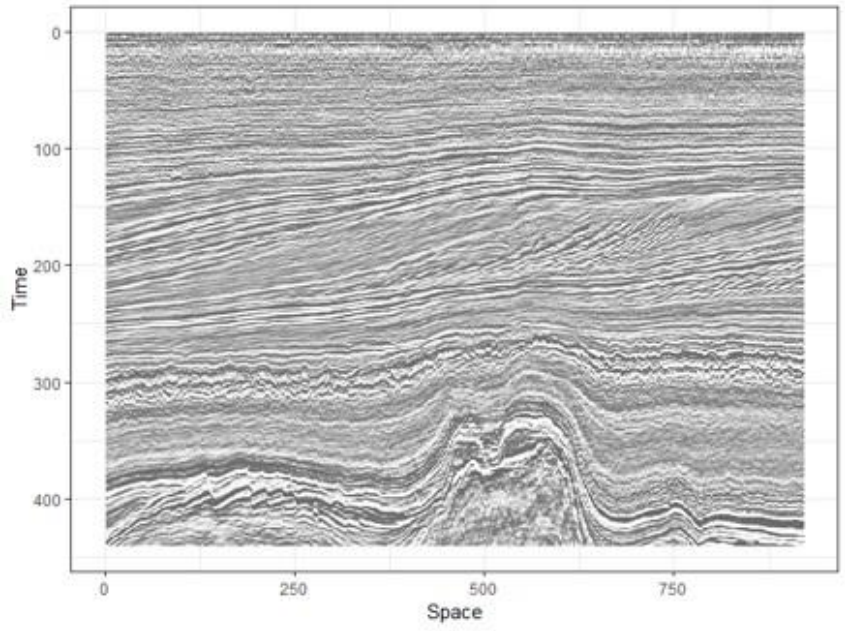
The dataset is organized into inlines (920 spatial-time series with 440 observations in each)
We selected the inline 401 since it has been mapped by seismic specialists

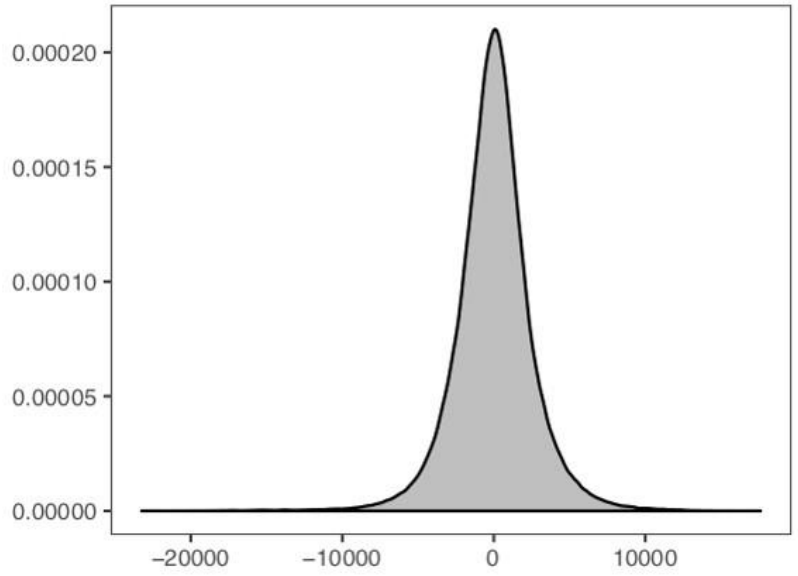


Each receiver produces a **spatial-time series** related to a specific position of the surface

(Scientific)
Seismic Analysis Example

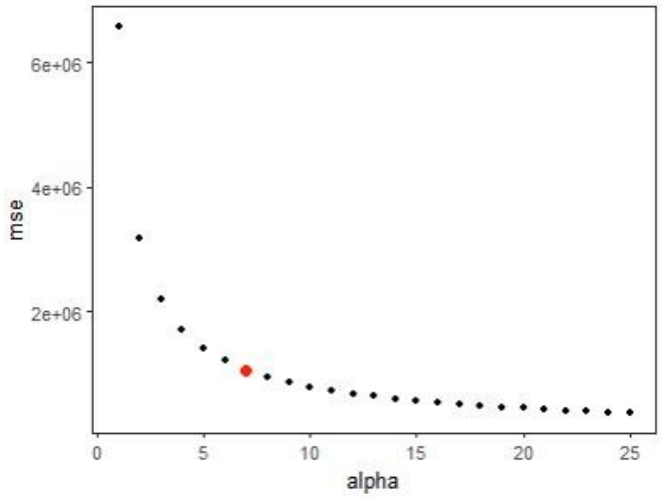
- The values of observations represents the wave amplitude reflected from the subsoil
- The probability density function (PDF): frequency distribution with a high concentration of values close to zero varying from -10000 to 10000





Experimental Setup

- The CSA Algorithm requires parameters $alpha$, $word$, tb, sb, σ, and κ
- The $alpha$ was chosen based on the data adjustment
  - We varied the alphabet size for SAX encoding from 1 to 25 using maximum curvature analysis for the MSE for each alphabet size
- The CSA is available as an R Package (STMotif)
- Experimental evaluation was conducted in a cluster with 24 cores using SparkR
- The experimental evaluation ran at wall-time of 1.3 hours



Parameter	Description (explored values)
alpha	Size of the alphabet for SAX indexing (fixed at 7)
word	Length of motif word (from 3 to 7)
tb x *sb*	Temporal and spatial block size (40x10, 20x20, 10x40)
σ	Minimum number of occurrences inside each block (from 2 to 7)
κ	Minimum number of spatial-time series with occurrences inside each block (from 1 to 5)

Analysis of Spatial-Time Motifs (overall performance according to orientation)

- Discovered motifs and their occurrences and computational time as we vary block size (tb and sb), $word$, σ, and κ
- Three orientations: vertical rectangle (tb = 40; sb = 10), square (tb = 20; sb = 20), and horizontal rectangle (tb = 10; sb = 40) and the traditional approach

Block orientation	motifs	sets of occur.	discovery time (min)	ranking time (min)	elapsed time (min)
Traditional (440x1)	43	449	1.8	2.0	4.7
CSA Vertical (40x10)	85	673	1.6	1.8	4.2
CSA Square (20x20)	114	772	1.4	1.6	3.8
CSA Horizontal (10x40)	105	705	0.9	1.2	2.9

Analysis of Spatial-Time Motifs (performance according to word size and orientation)

- Number of discovered motifs and the sets of occurrences
- As we increase the word size, the number of discovered motifs decrease
- The number identified motifs for CSA when compared to traditional approach becomes more significant as we increase the word size
- The highest number of identified motifs occurred in CSA Square orientation for word size equals to four
- The computation time (in minutes) for all discovered motifs also decreases as we increase the word size
  - It is due to the ranking function overhead

Block orientation	*word*	motifs	sets of occurrences	total time (min)
Traditional	3	139	95862	9.5
	4	65	6809	5.4
	5	7	369	3.0
	6	2	72	2.7
	7	1	17	2.6
CSA Vertical (40x10)	3	168	62278	8.0
	4	163	13980	4.7
	5	60	2988	3.3
	6	23	761	2.7
	7	11	229	2.5
CSA Square (20x20)	3	184	62324	6.7
	4	221	16887	4.5
	5	103	4182	3.1
	6	42	1157	2.4
	7	19	352	2.1
CSA Horizontal (10x40)	3	187	52199	5.5
	4	219	12901	3.7
	5	89	2918	2.3
	6	25	628	1.6
	7	7	149	1.2

Analysis of Spatial-Time Motifs (varying σ and κ)

- Influence of σ and κ in the number of discovered motifs
  - Orientation: square block; word size = 4
- As we increase σ, lower number of occurrences are identified
- As we increase κ constraints, the number of occurrences also decreases

	σ					
κ	2	3	4	5	6	7
1	42725	30052	21349	13559	9621	6959
2	42253	29938	21297	13527	9589	6927
3	-	29640	21191	13461	9530	6895
4	-	-	20073	13184	9368	6758
5	-	-	-	11900	8800	6490

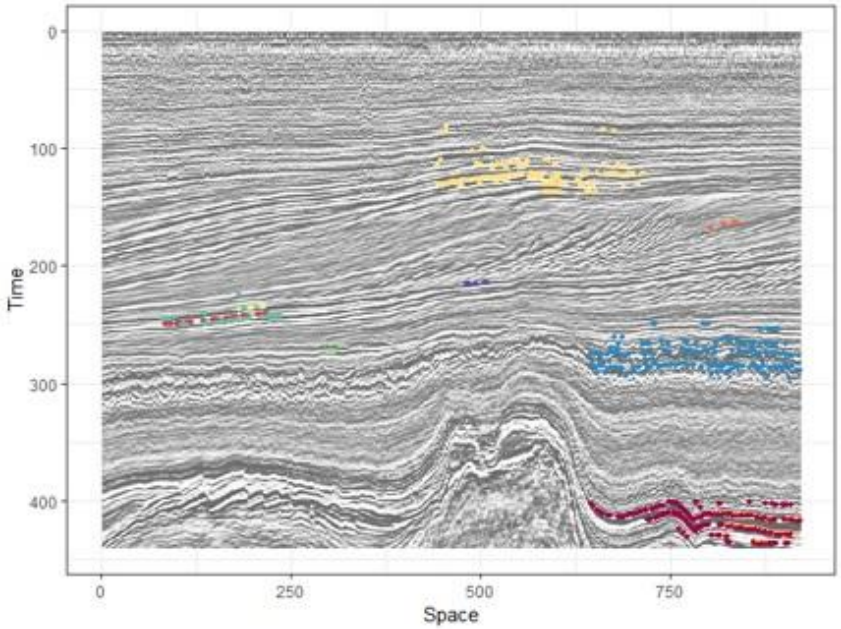
Analysis of Ranking of Spatial-Time Motifs

- Parameters: Orientation: Square block; Word = 4, σ = 3, κ = 3
- The highest ranked motif ($aagg$) presented a good distance value, an average entropy, and a high number of occurrences
- The second place ($dfge$), although having small number of occurrences, it had a good distance value and a good entropy
- The third place ($aaag$) was similar to the first motif, but the smaller number of occurrences
- The fourth place ($ggfa$) compensated for the smaller number of occurrences with an excellent distance metric
- The fifth place ($egfa$) was similar to the second place

motif	distance	entropy	occurrences	rank
$aagg$	0.74	1.0	8.28	1.57
$dfge$	0.83	2.0	3.16	1.46
$aaag$	0.85	0.8	7.06	1.45
$ggfa$	1.00	1.5	3.17	1.40
$egfa$	0.75	2.0	3.17	1.39

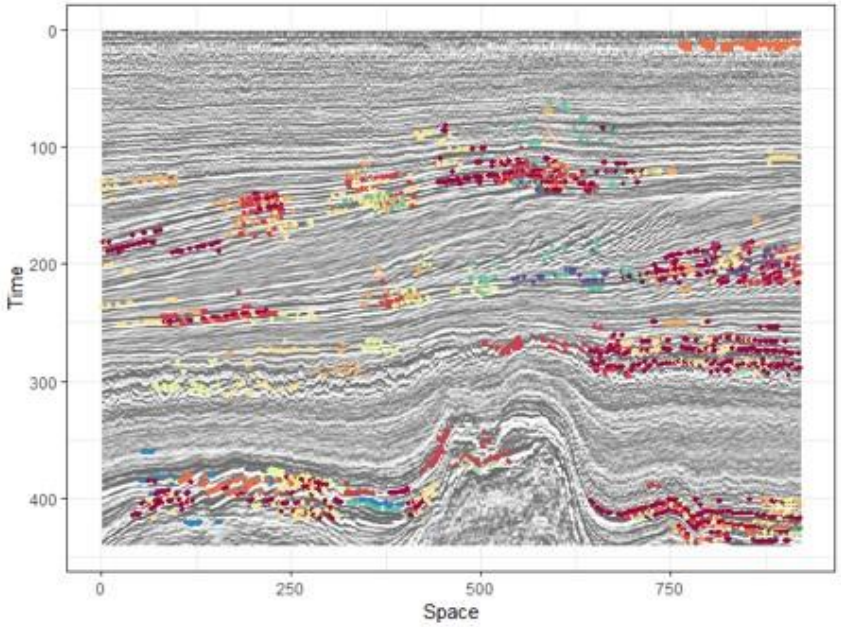
Analysis of Top-k Spatial-Time Motifs (according to ranking)

- top-ten discovered motifs
  - The places where the motifs were plotted are in agreement with annotations from specialists where seismic horizons are located
  - Yellow ones are close to a gas reservoir



Analysis of Top-k Spatial-Time Motifs (using ranking as a filter)

- Top-ten distinct motifs sorted by their occurrences
  - Using ranking value to filter: Values greater than 1.0
- The occurrences of motifs matched more regions where seismic horizons are located
- Ranking function was conceived for general purpose usage and did not focus on any aspect to target seismic horizons



Discovering Tight Space-Time Sequences

Spatial-time sequence mining (motivation example)

- Quantified-self movement, where people wear connected bracelets giving their position and inferring their activities
  - A brand might discover some habits regarding sports and food at coarse grain
  - "people who jog in the morning step by a vegan shop once a week"
- There might be fine-grained behaviors that cannot be extracted
  - They concern a niche (extremely low support)
  - "people in Manhattan who jog at 7 am and have lunch near Time Square, spend 1 to 2 minutes in front of the buildings displays"
- The challenge is to extract both the pattern (*e.g.*, jog, lunch, buildings displays), its occurrence time (*e.g.*, from 7 am to lunchtime), and the location where it occurs (*e.g.*, Time Square)

Discovery challenge

- We are interested in finding tight space-time sequences
  - Sequences that are constrained in space and time
  - Sequences that may not be frequent in the entire dataset
  - Sequences that may be frequent inside a time interval and space range (spatiotemporal blocks)
- The primary challenge is to discover these blocks and the frequent sequences they contain

Related Work: Trajectories datasets

- Collection of events of the same moving object at different spatial locations and times [1,3]
- It includes works that search for routes frequently used
- The position of moving objects may be inaccurate, some approaches address the position uncertainty to better recognize common trajectories [2]
- Difference: We are not interested in moving objects that have the same behavior. Instead, we are interested in regions and time intervals when some events are related (constrained) and relevant (with high local support)

[1] Y. Huang, L. Zhang, e P. Zhang, "A framework for mining sequential patterns from spatio-temporal event data sets", *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, nº 4, p. 433–448, 2008.
[2] Y. Li, J. Bailey, L. Kulik, e J. Pei, "Mining probabilistic frequent spatio-temporal sequential patterns with gap constraints from uncertain databases", in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2013, p. 448–457.
[3] F. Giannotti, M. Nanni, F. Pinelli, e D. Pedreschi, "Trajectory pattern mining", in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007, p. 330–339.

Related Work: Event-based datasets

- Correspond to the majority of related work
- Find sequences constrained by space and time
  - For that, data is partitioned according to spatial or temporal dimensions, and events are related whenever they preserve certain proximity [1,2,3]
- Difference: they differ since all identified sequences are frequent in the entire dataset. In our work, sequences may only be frequent inside spatiotemporal blocks (*i.e.*, a time interval and space range)

[1] H. Alatrista-Salas, J. Azé, S. Bringay, F. Cernesson, N. Selmaoui-Folcher, e M. Teisseire, "A knowledge discovery process for spatiotemporal data: Application to river water quality monitoring", *Ecological Informatics*, vol. 26, nº P2, p. 127–139, 2015.
[2] I. Tsoukatos e D. Gunopulos, "Efficient mining of spatiotemporal patterns", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2121, p. 425–442, 2001.
[3] A. Julea, N. Méger, C. Rigotti, E. Trouvé, P. Bolon, e V. Lăzărescu, "Mining pixel evolutions in satellite image time series for agricultural monitoring", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6870 LNAI, p. 189–203, 2011..

Related Work: Emerging patterns

- Correspond to solve the problem where data are continuously added to the database
  - Previously identified patterns may become irrelevant, and new patterns may emerge [1,3]
  - Some initiatives in emerging spatiotemporal datasets have been developed so far [2]
- Difference: these works are complementary to ours since all identified patterns in both datasets (initial or updated) have high support

[1] C.-Y. Tsai e Y.-C. Shieh, "A change detection method for sequential patterns", *Decision Support Systems*, vol. 46, nº 2, p. 501–511, 2009.
[2] Y.-L. Chen e Y.-H. Hu, "Constraint-based sequential pattern mining: The consideration of recency and compactness", *Decision Support Systems*, vol. 42, nº 2, p. 1203–1215, 2006.
[3] B. Aydin e R. A. Angryk, "Spatiotemporal event sequence mining from evolving regions", in *Proceedings - International Conference on Pattern Recognition*, 2017, p. 4172–4177.

Formalization

- Let **t** be a **time-stamped sequence (TS)** < $v_1, v_2, \ldots, v_n$ >, **sequence** of items, where $|t| = n$ is the number of items in t
- A sequence s = < $w_1, w_2, \ldots, w_k$ > is **included** in TS if there exist integers $i_1 < i_2 < \cdots < i_k$ such that $w_1 = v_{i_1}, w_2 = v_{i_2}, \cdots, w_k = v_{i_k}$
- Let $P = \{p_1, p_2, \ldots, p_m\}$ be a set of positions, a **spatial time-stamped sequence (STS)** d is a couple (p, t) where $p \in P$ and t is TS
- A **STS dataset** D is a set of STS
- An STS $d = (p, t)$ is said to support a sequence s if s is included in t
- The **support** of a sequence s in D is the number of STS in D in which s is included
- The **frequency** of a sequence s in D is the fraction of STS in D that supports s: $freq(s, D) = \frac{sup(s,D)}{|D|}$
- Given a user's minimum threshold $\gamma \in]0..1]$, a sequence s is said to be **frequent** if $freq(s, D) \geq \gamma$

Formalization

- A **spatial range** (or simply **range**) $r = (p_s, p_e)$ is defined by a start position p_s and an end position p_e
  - We define the set of all potential ranges over D as PR
- The set of STS that belong to range r is defined as $Tr(r) = \{d : d \subseteq D, p_s \leq d.p \leq p_e\}$
  - The frequency of s over $Tr(r)$ in STS for r is denoted by $freq(s, r)$
  - The support of s over $Tr(r)$ in STS is denoted by $sup(s, r)$
- A **ranged sequence** sr is a triple (s, r, fr) where s is a *sequence*, r is a *range*, and fr is the *frequency* of the *sequence* s over the *range* r: $fr = freq(s, r)$
- A **time interval** (or simply **interval**) $i = (i_s, i_e)$ is defined by a start time i_s and an end time i_e
  - The length of an interval i is given by: $|i| = i_e - i_s + 1$.
  - We define the set of all possible intervals over D as PI

Formalization

- A **block** b is a couple (r, i) where r is a range ($r \in PR$) and i is an interval ($i \in PI$): $|b| = |b.r| \cdot |b.i|$.
  - We define the set of all possible blocks over a range r as $PB(r)$
- A sequence s ***occur*** in b if the first element of s is inside in b.
  - $occur(s, b)$ is a set of pairs (p, t) that corresponds to the beginning* of s in b
- The **support** of a *sequence* s in a block b, $sup(s, b)$ is the number of occurrences of s in b: $|occur(s, b)|$
  - Given a user's minimum threshold $\delta \in]0..1]$, a sequence s is said to be **frequent** in a block b if $freq(s, b) \geq \delta$.
- A **blocked sequence** sb is a triple (s, b, fr) where s is a *sequence*, b is a *block*, and fr is the *frequency* of s over b: $fr = freq(s, b)$

Problem statement

- Considering an STS dataset D, the problem we address is to find sequences in D that are frequent in constrained spatial range and time interval
  - The goal is to discover these ranges and intervals and the frequent sequences they contain
- In the following definitions, we introduce the notions of solid-ranged sequence and solid-blocked sequence that are fundamental for STSM algorithm
- Their intuition is to respectively support the identification of spatial range and time-space blocks where a pattern is frequent

Spatial-Temporal Sequence Miner - STSM

- STSM is the algorithm to address the problem definition
- We introduce the notions of solid-ranged sequence and solid-blocked sequence
- Their intuition is to respectively support the identification of spatial range and time-space blocks where a pattern is frequent

Solid-Ranged Sequence

- Let sr be a ranged sequence of range r, sequence s, and frequency fr. Then, sr is called a **solid-ranged sequence** iff the following conditions hold:
  1) $fr \geq \gamma$
  2) $\forall r_2 \in PR \mid r \subseteq r_2$, we have either a) or b) or both:
     a) $freq(s, r_2) < \gamma$
     b) $sup(s, r_2) = sup(s, r)$
  3) $\forall r_2 \in PR$ such that $r_2 \subset r$, $sup(s, r_2) < sup(s, r)$
- The first condition ensures that sr represents a sequence that is frequent over its associated range
- The second condition ensures that the size of r is maximal
- The third condition ensures that the size of r is minimal

Solid-Ranged Block

- Let sb be a blocked sequence with a block b, sequence s, and frequency ifr. Then, sb is called a **solid-blocked sequence** iff the following conditions hold:
  1) $\exists\ sr \in \mathbf{SR}_{|s|} \mid b.r \subseteq sr.r\ and\ s = sr.s$
  2) $fr \geq \delta$
  3) $\forall b_2 \in PB(r) \mid b \subseteq b_2$, we have either a) or b) or both:
     a) $freq(s, b_2) < \delta$
     b) $sup(s, b_2) = sup(s, b)$
  4) $\forall b_2 \in PB(r) \mid b_2 \subset b, sup(s, b_2) < sup(s, b)$
  5) $sup(s, b) > 1$
- The first condition ensures that the range of sb is within the range of a solid-ranged sequence sr
- The second condition ensures that s corresponds to a sequence that is frequent in b
- The third condition ensures that the size of b is maximal
- The fourth condition ensures that the size of b is minimal
- The fifth condition avoids trivial solid-blocked sequences that contain only a single occurrence of s in b

General Principle

- $STSM(D, \gamma, \delta)$
  - $C_1 \leftarrow generateCandidates(D, nil)$
  - $k \leftarrow 0$
  - *Repeat*
    - $k \leftarrow k + 1$
    - $SR_k \leftarrow solidRangedSequences(D, C_k, \gamma)$
    - $SB_k \leftarrow solidBlockedSequences(D, SR_k, \delta)$
    - $C_{k+1} \leftarrow generateCandidates(D, SB_k)$
  - Until $C_{k+1} \neq \emptyset$
  - return $\{SB_1, \cdots, SB_k\}$

Toy example

- 10 Spatial-TSS $(d_1, d_2, \ldots, d_{10})$
- Each Spatial-TSS has 6 observations (v_1 to v_6)
- Consider a frequency threshold $\gamma = \frac{1}{2}$ for D
  - <a> (frequent in 5 of the 10 TSS)

D / t	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
v_1	**a**	b	c	d	t	q	**i**	g	**a**	h
v_2	k	l	m	n	p	q	**u**	s	t	v
v_3	w	**e**	**e**	x	y	m	**a**	r	d	a
v_4	h	**o**	**o**	g	**e**	i	e	**i**	c	b
v_5	**i**	j	k	l	**o**	z	n	**u**	z	p
v_6	**u**	**a**	r	s	t	$\propto$	c	d	f	**a**

Example of Spatial-Temporal Series

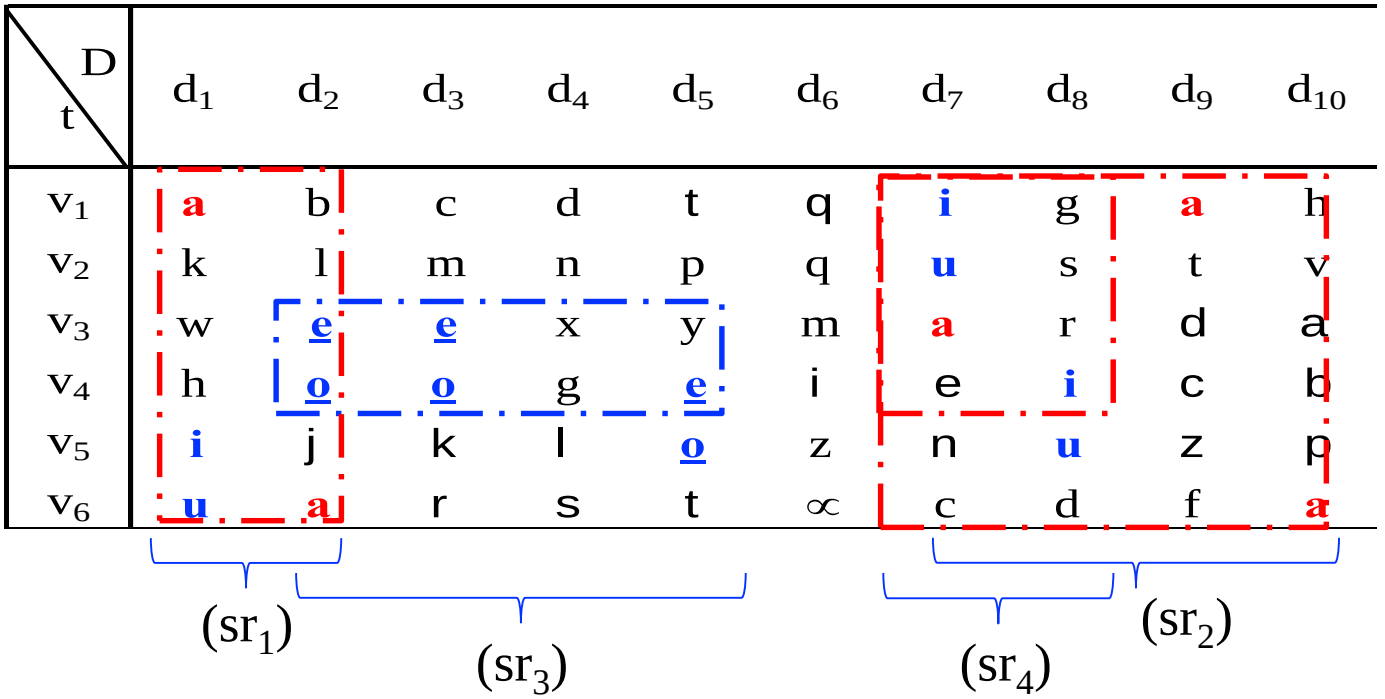
- Consider a frequency threshold $\gamma = \frac{1}{2}$ for SR
  - <a> (100% sr_1), <a> (75% sr_2),
  - <e>,<o>,<e,o> (75% sr_3), <i>,<u>,<i,u> (100% sr_4)

D / t	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
v_1	**a**	b	c	d	t	q	**i**	g	**a**	h
v_2	k	l	m	n	p	q	**u**	s	t	v
v_3	w	**e**	**e**	x	y	m	**a**	r	d	a
v_4	h	**o**	**o**	g	**e**	i	e	**i**	c	b
v_5	**i**	j	k	l	**o**	z	n	**u**	z	p
v_6	**u**	**a**	r	s	t	$\propto$	c	d	f	**a**

(sr_1) (sr_3) (sr_4) (sr_2)

Example of Spatial-Temporal Series

- Consider a frequency threshold $\delta = \frac{3}{4}$ for SB
  - <e>,<e,o> (75% blue block from sr_3)

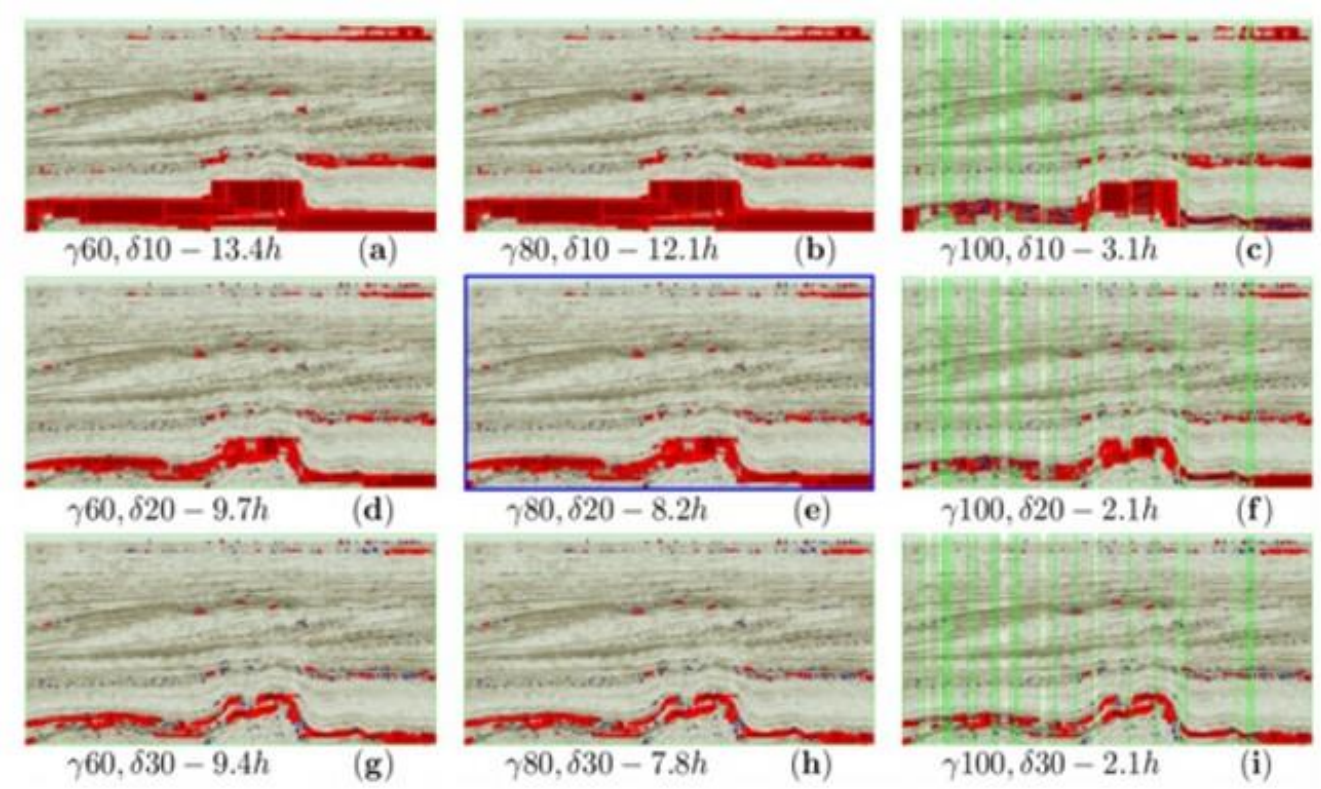


Experimental Analysis
Seismic Dataset

- The algorithm proposed in this work allows for users to set solid range threshold γ and solid block threshold δ constraints.
- Finding adequate values for these thresholds depends on the characteristics of input dataset/application.
  - Lower values for such constraints can lead to the identification of a large number of non-useful frequent sequences
  - Higher values for these thresholds can result in the detection of a small number of frequent sequences that may become too small to be interesting
- We explored the combination of solid range threshold γ (60%, 70%, 80%, 90%, and 100%) with solid block threshold δ (10%, 20%, 30%, 40%, and 50%)

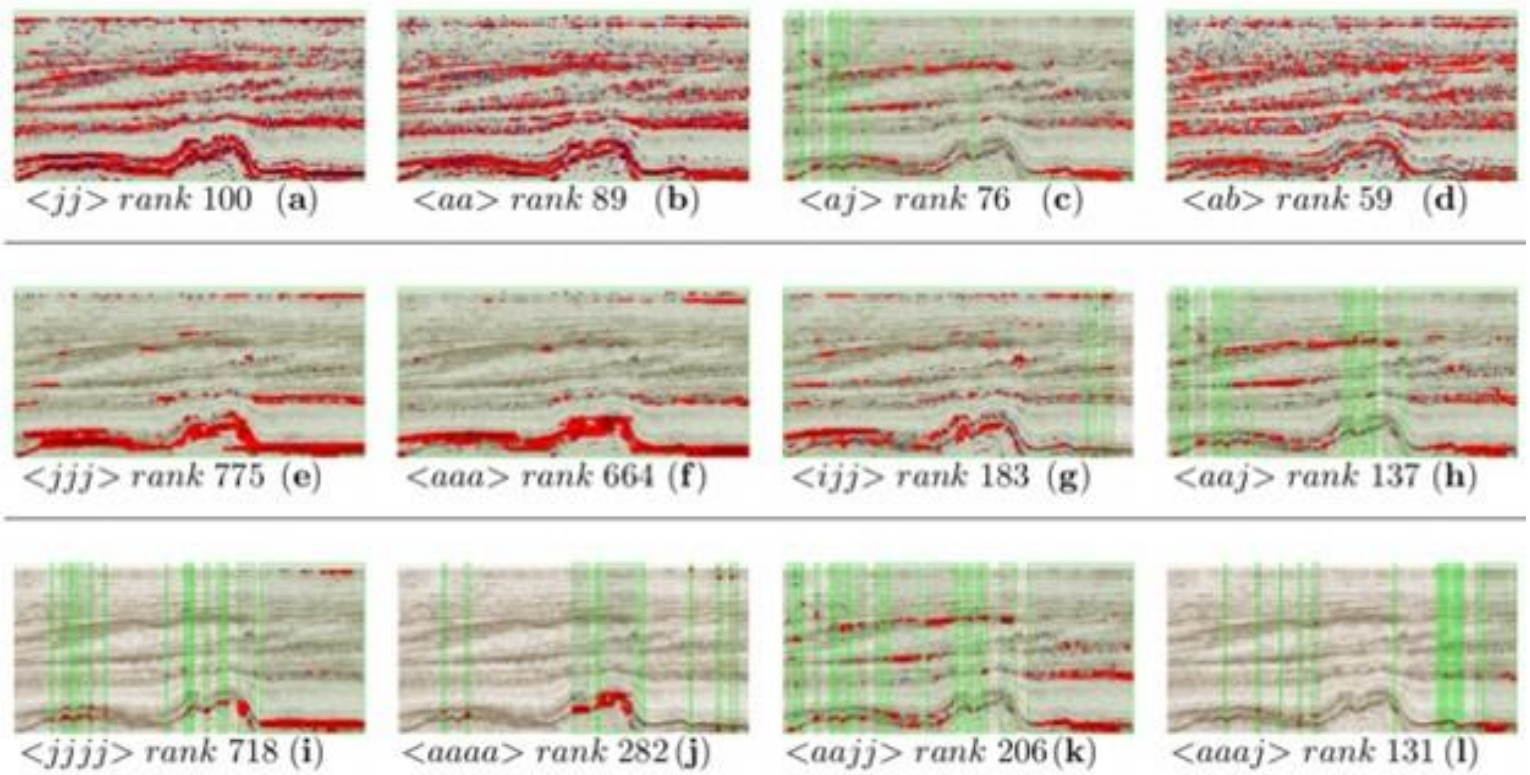
Experimental Evaluation: STSM Parameters

- Calibration of thresholds γ = 80% and δ = 20% for sequence <a, a, a> produce finer grained blocks with a more complete overlap of the horizon



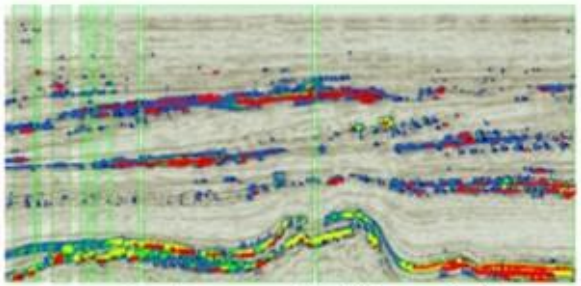
Experimental Evaluation: Best ranked patterns

- High number of patterns detected
  - Ranked to prioritize best results using a simple density criteria (mean block size of all solid-blocked sequences for s)
  - The best patterns follow potential horizons

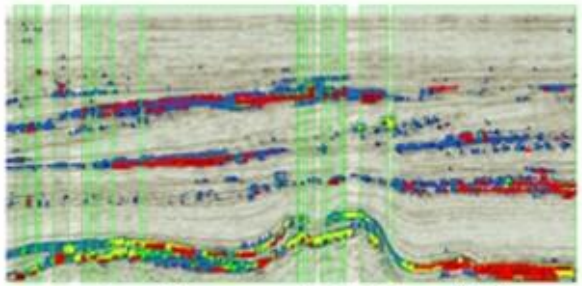


Experimental Evaluation: bright-spots

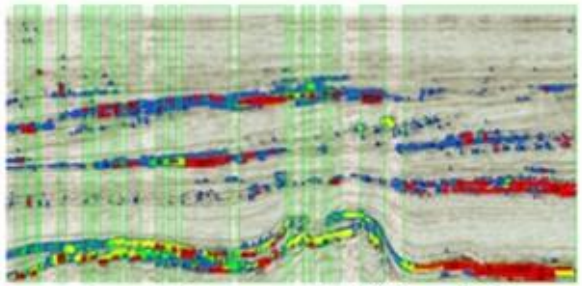
- Some of the high ranked discovered patterns follow previously knowns potential bright-spots for this dataset
  - *Bright spots* are rare patterns that occur when there is an inversion of the wave phase



<aj> rank 76 (c)



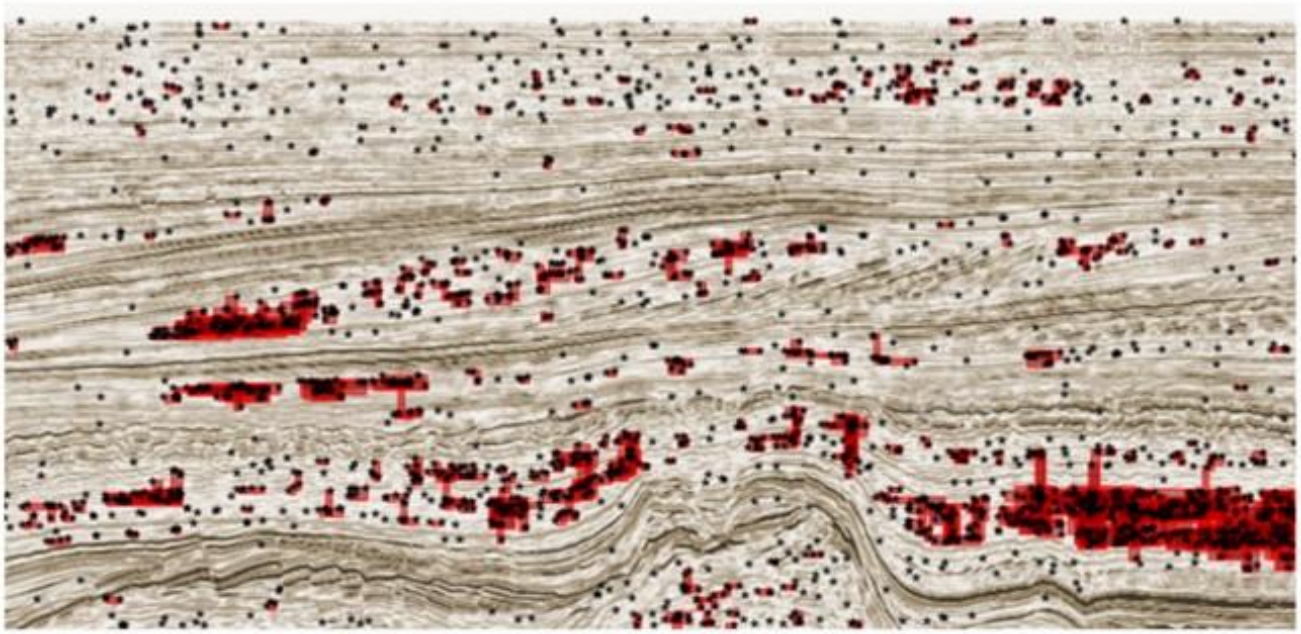
<aaj> rank 137 (h)



<aajj> rank 206 (k)

Experimental Evaluation: STSM and GSTSM

- Identified occurrences by STSM (marked as red) correspond to seismic horizons
- Many occurrences from GSTSM (marked as black) correspond to noise



Comparison

Comparisons between Motifs and Sequence Mining (according to the observations)

- Motif identification
  - Can work directly with time series (exact match motif)
  - Univariate/ Multi-variate time series
  - No item-set support
- Sequence-pattern mining
  - Indexed time-series only
  - Item-set support
  - May find multi-variate patterns among different dimensions

Comparisons between Motifs and Sequence Mining (according to space-time dimensions)

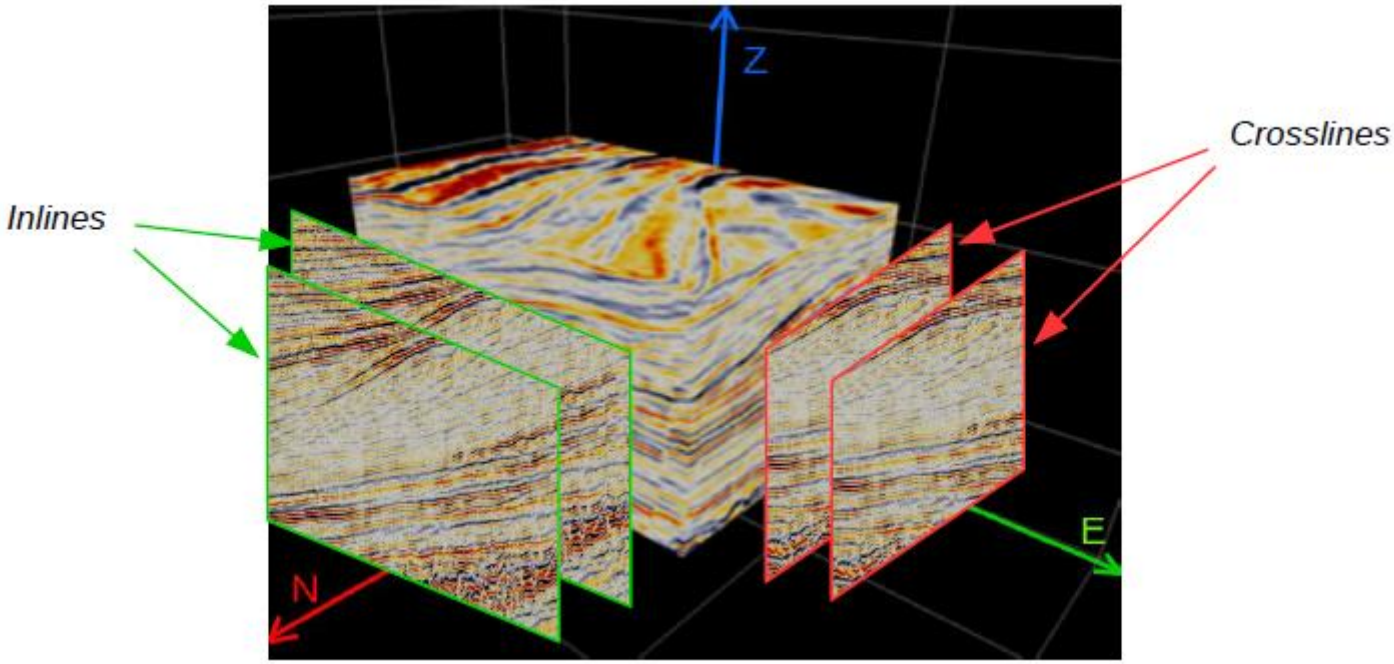
- Motif identification
  - $S \cdot T$
  - $S \cdot S \cdot T$
  - $S \cdot S \cdot T \cdot T$ (challenge)
- Sequence-pattern mining
  - $S \cdot T$
  - $S \cdot S \cdot T$ (challenge)
  - $S \cdot S \cdot T \cdot T$ (challenge)

Comparisons between Motifs and Sequence Mining (according to performance)

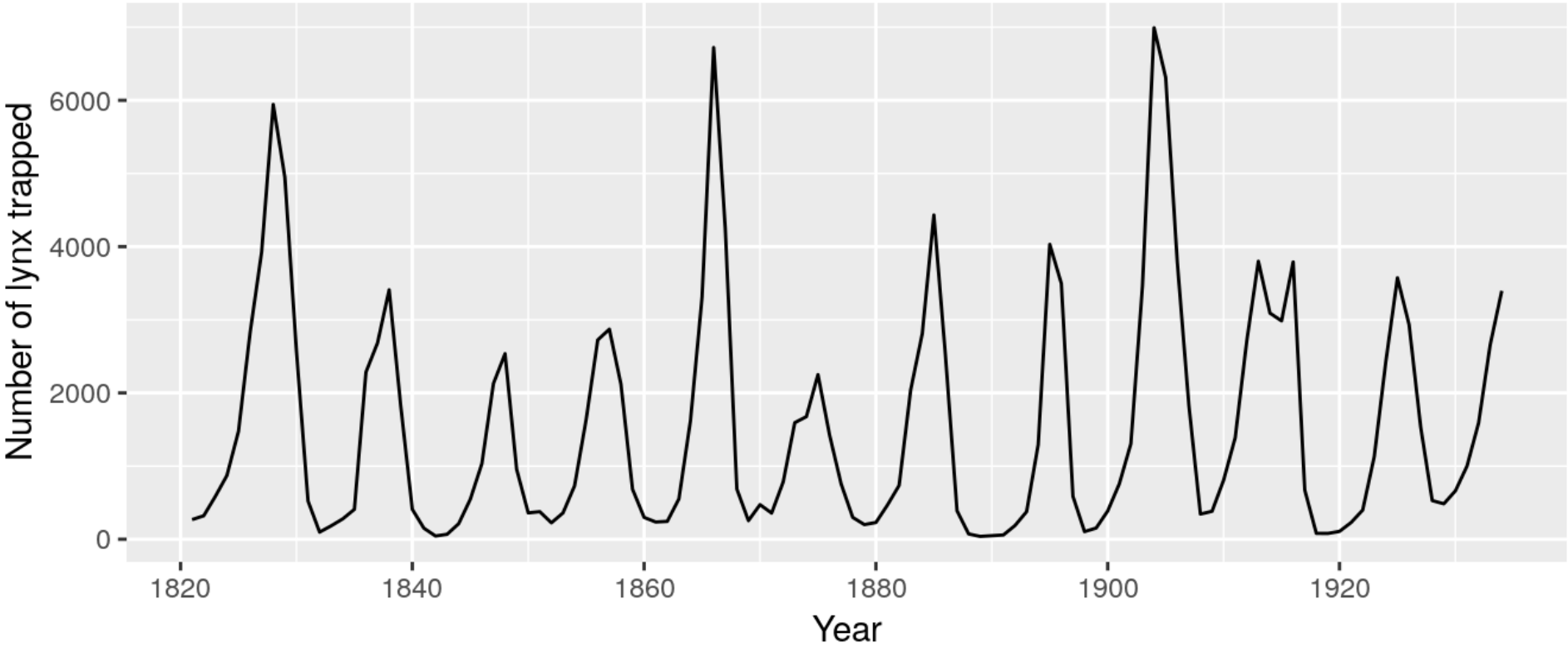
- Motif identification
  - Limited for small-sequences (without random-projection)
- Sequence-pattern mining
  - Seems to scale up better
    - Our approach needs to guarantee antimonotonicity property of Aprori

Challenges

Spatial order in higher dimensions (2D Space and 1D time)

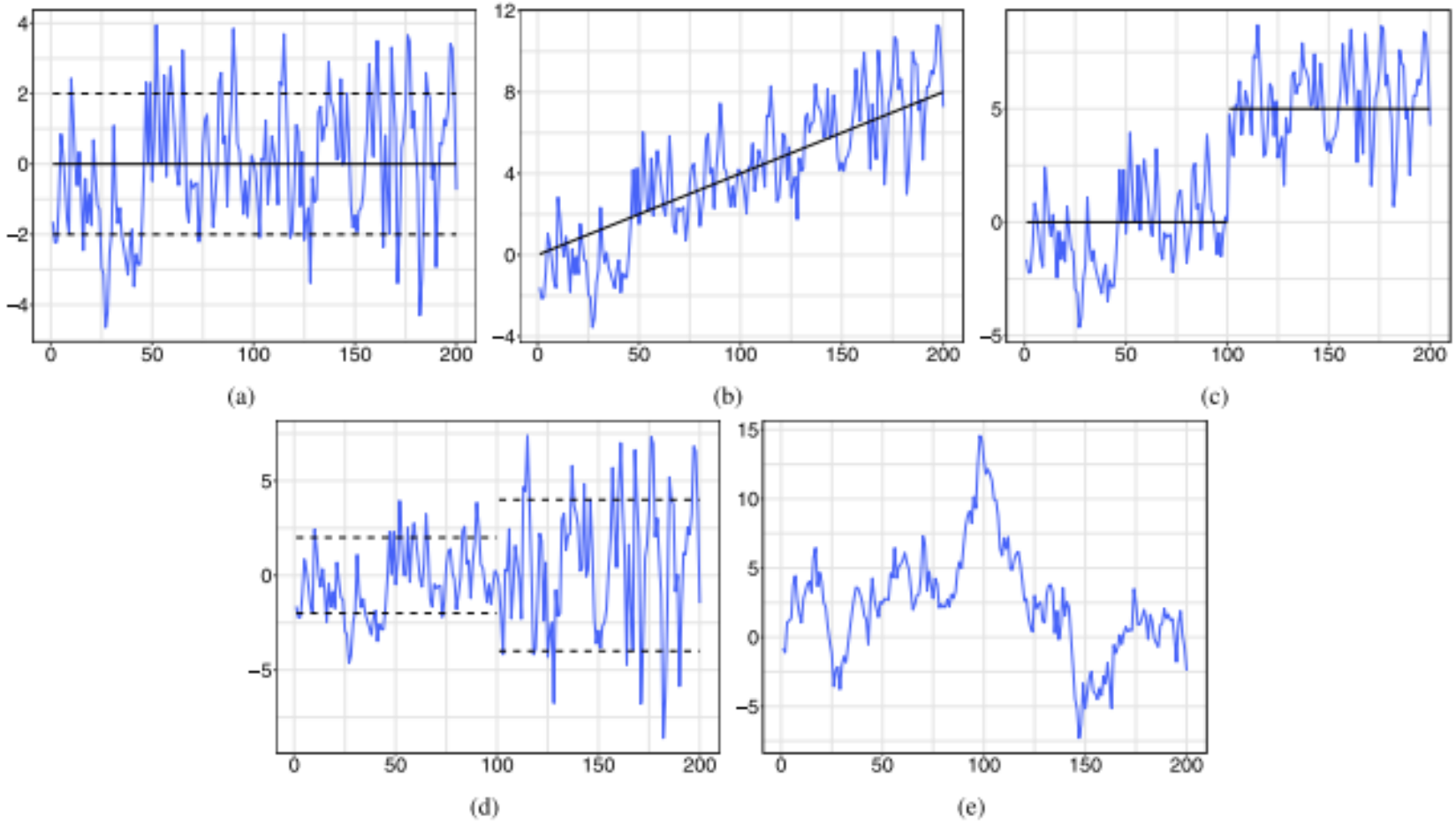


Seasonal patterns



Hyndman, 2011 - https://robjhyndman.com/hyndsight/cyclicts/

Non-stationarity

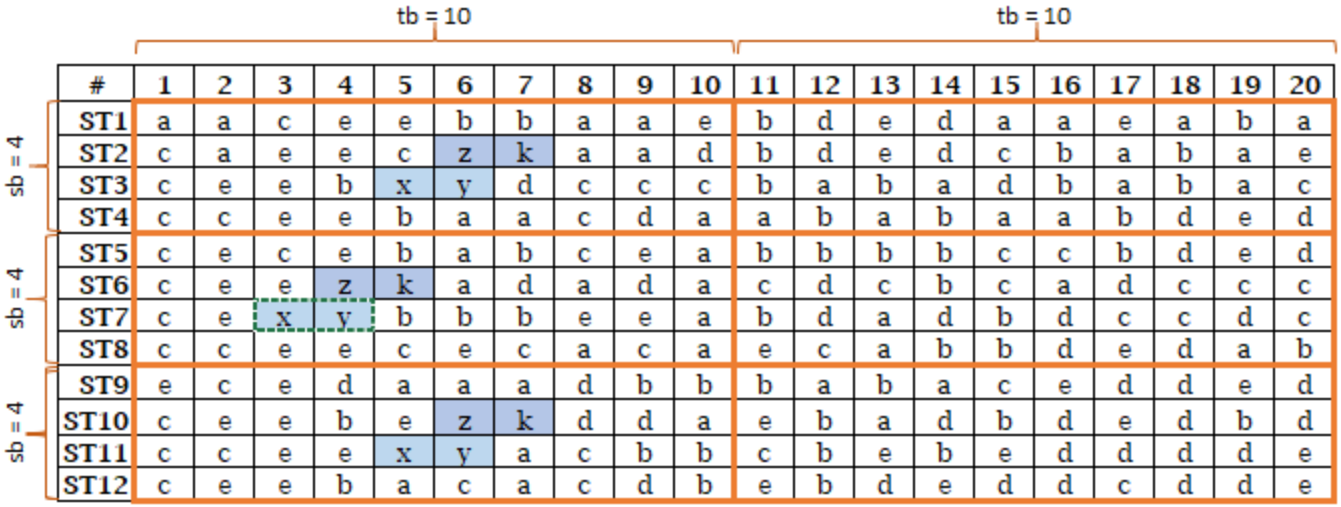


R. Salles, K. Belloze, F. Porto, P. H. Gonzalez, e E. Ogasawara, "Nonstationary time series transformation methods: An experimental review", *Knowledge-Based Systems*, nov. 2018.

Parameter-free methods

- General
  - SAX indexing alphabet
- Motifs
  - Word size
  - Block size (time x space)
  - Temporal and Spatial thresholds
- Spatial-temporal sequence-mining
  - Temporal and Spatial thresholds

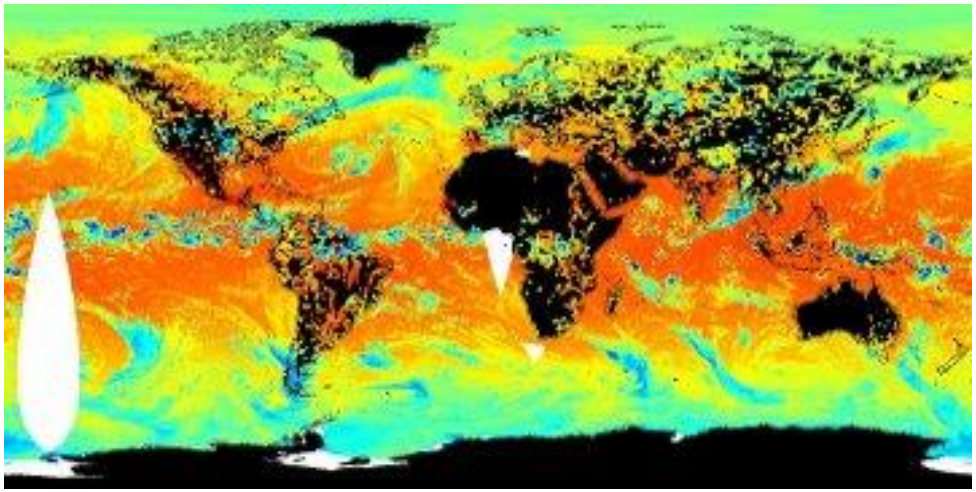
Complex patterns



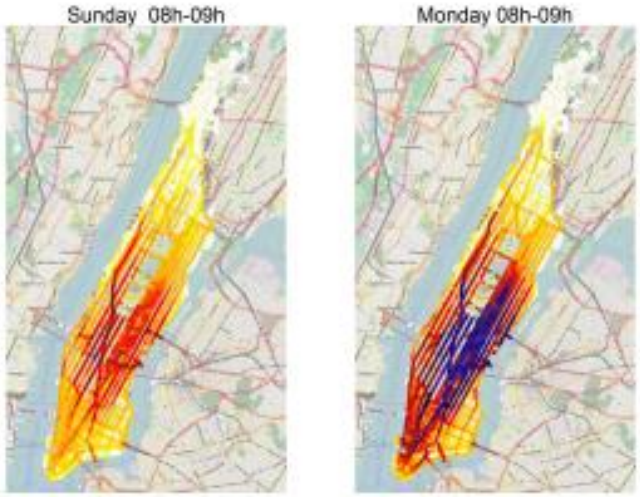
#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ST1	a	a	c	e	e	b	b	a	a	e	b	d	e	d	a	a	e	a	b	a
ST2	c	a	e	e	c	z	k	a	a	d	b	d	e	d	c	b	a	b	a	e
ST3	c	e	e	b	x	y	d	c	c	c	b	a	b	a	d	b	a	b	a	c
ST4	c	c	e	e	b	a	a	c	d	a	a	b	a	b	a	a	b	d	e	d
ST5	c	e	c	e	b	a	b	c	e	a	b	b	b	b	c	c	b	d	e	d
ST6	c	e	e	z	k	a	d	a	d	a	c	d	c	b	c	a	d	c	c	c
ST7	c	e	x	y	b	b	b	e	e	a	b	d	a	d	b	d	c	c	d	c
ST8	c	c	e	e	c	e	c	a	c	a	e	c	a	b	b	d	e	d	a	b
ST9	e	c	e	d	a	a	a	d	b	b	b	a	b	a	c	e	d	d	e	d
ST10	c	e	e	b	e	z	k	d	d	a	e	b	a	d	b	d	e	d	b	d
ST11	c	c	e	e	x	y	a	c	b	b	c	b	e	b	e	d	d	d	d	e
ST12	c	e	e	b	a	c	a	c	d	b	e	b	d	e	d	d	c	d	d	e

New domains

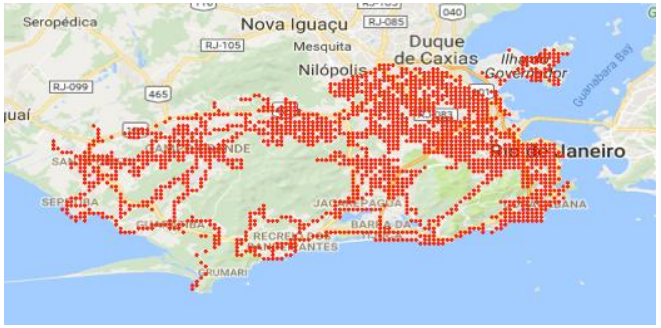
Climate Data (sea surface temperature, wind) [1]



New York Taxis [2]



Urban mobility [3]



[1] Nasa, 2018 - https://podaac.jpl.nasa.gov/dataset/VIIRS_NPP-OSPO-L2P-v2.4
[2] J. A. Deri, F. Franchetti, e J. M. F. Moura, "Big data computation of taxi movement in New York City", in *2016 IEEE International Conference on Big Data (Big Data)*, 2016, p. 2616–2625.
[3] A. B. Cruz *et al.*, "Detecção de anomalias frequentes no transporte rodoviário urbano", in *Proceedings of the 33nd Brazilian Symposium on Databases (SBBD)*, 2018.

Results

- Papers
  - [1]R. Campisano, F. Porto, E. Pacitti, F. Masseglia, e E. Ogasawara, “Spatial Sequential Pattern Mining for Seismic Data”, in XXXI Brazilian Symposium on Databases, Salvador, BA, 2016
  - [2]R. Campisano et al., “Discovering Tight Space-Time Sequences”, in DaWak 2018, 2018
- R Packages
  - [3] H. Borges, A. Bazaz, e E. Ogasawara, “STMotif: Discovery of Motifs in Spatial-Time Series”, *The Comprehensive R Archive Network*, 2018. https://cran.r-project.org/web/packages/STMotif/index.html
- Master degree defense
  - M. Dutra, Discovering Motifs in Spatial-Time Series Seismic Datasets, 2016, co-advised with Fabio Porto
  - R. Campisano, Sequence Mining In Spatial-Time Series, 2017 co-advised with Florent Masseglia
- Ph.D. on going
  - H. Borges, Researching on Motifs and Spatial-Temporal Sequence Mining, co-advised with Esther Pacitti

Data Science Research Group

Comparing Motif Discovery Techniques with Sequence Mining in the Context of Space-Time Series

CEFET/RJ

Eduardo Ogasawara
eogasawara@ieee.org
http://eic.cefet-rj.br/~eogasawara