Discovering routine behaviour from time series data

Raúl Montoliu

Knowledge-Based Systems
Volume 113, 1 December 2016, Pages 61-74

An incremental algorithm for discovering routine behaviours from smart meter data

Jin Wang, Rachel Cardell-Oliver, Wei Liu

Show more

+ Add to Mendeley Share Cite

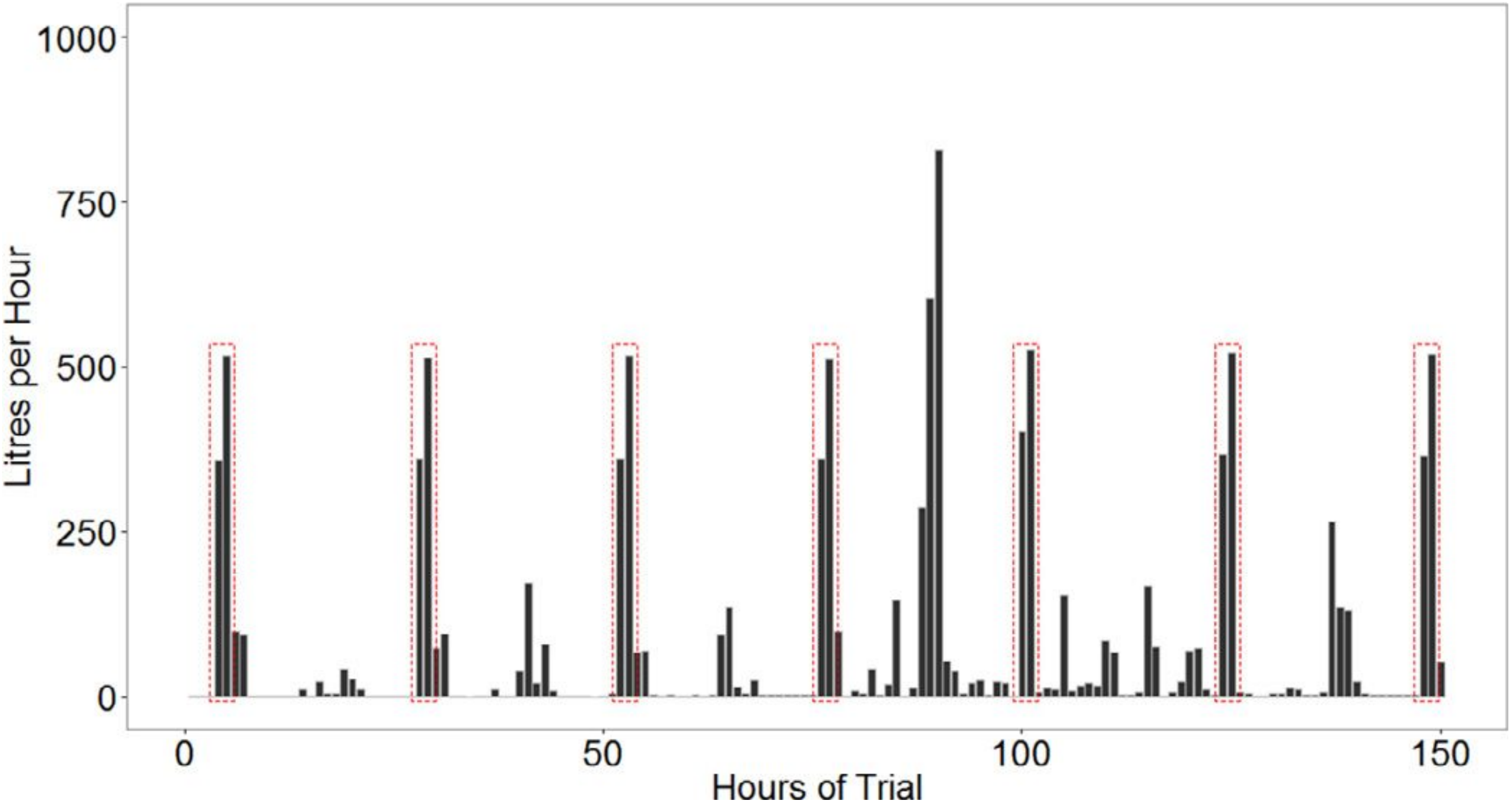
https://doi.org/10.1016/j.knosys.2016.09.016

Get rights and content

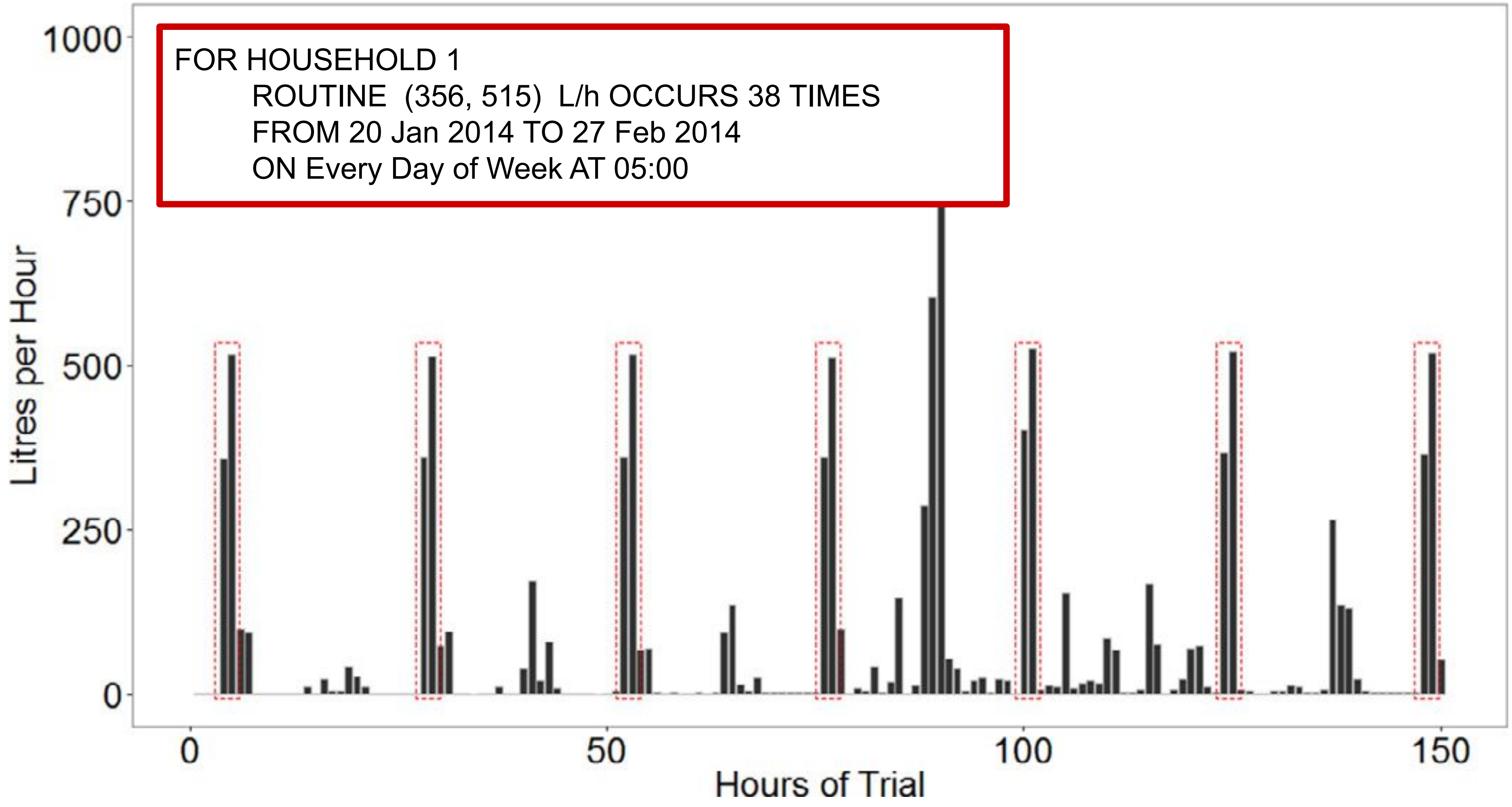
Q1

https://doi.org/10.1016/j.knosys.2016.09.016

A research paper **excellent** written and presented



Litres per Hour
1000
750
500
250
0
0
50
100
150
Hours of Trial


FOR HOUSEHOLD 1
ROUTINE (356, 515) L/h OCCURS 38 TIMES
FROM 20 Jan 2014 TO 27 Feb 2014
ON Every Day of Week AT 05:00
Litres per Hour
1000
750
500
250
0
0
50
100
150
Hours of Trial

The Problem

The problem of discovering routines is **to find all frequently occurring subsequences of variable lengths** in a smart meter time series.

The Problem

The problem of discovering routines is **to find all frequently occurring subsequences of variable lengths** in a smart meter time series.

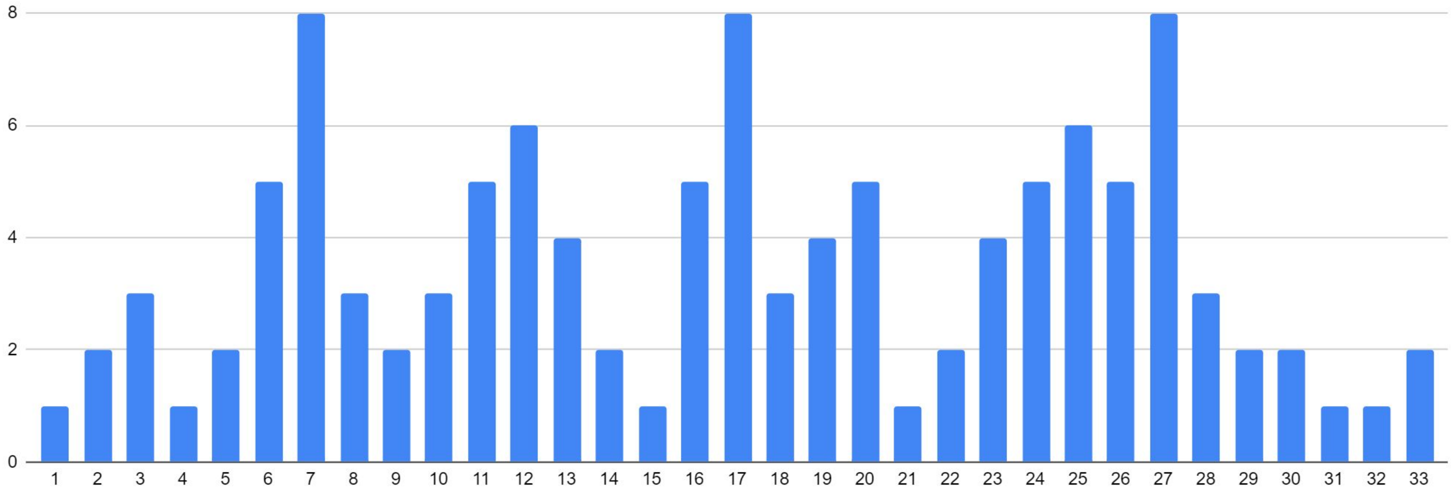
Challenges

1. We are interested in the **shape** and in the **values** of subsequences.
2. There is no prior knowledge about the **length** of the subsequences.
3. The subsequences usually consist of only a **few** elements.

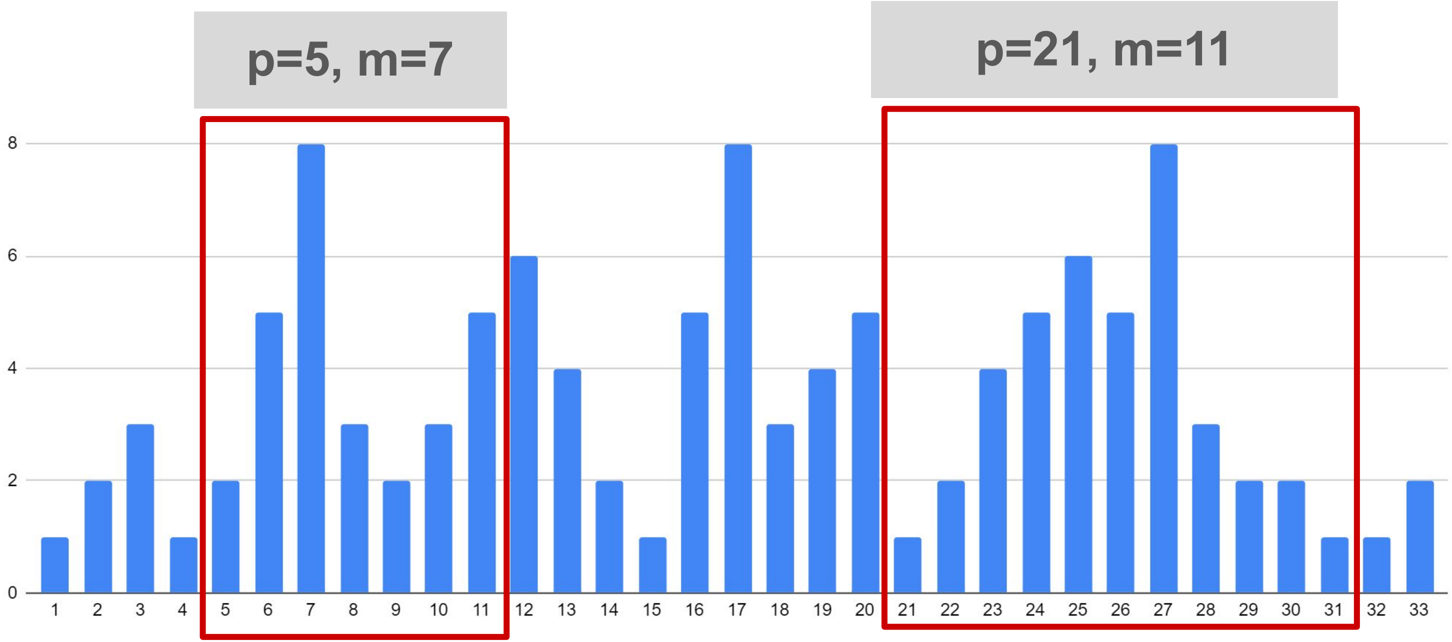
The paper presents

1. **Brute force** algorithm to detect routines
2. **Novel** algorithm to efficiently discover all routines of variable length

Definition 1. *Smart Meter Time Series*: A time series $X = (x_1, x_2, \ldots, x_n)$ is a sequence of n real valued numbers ordered in time.

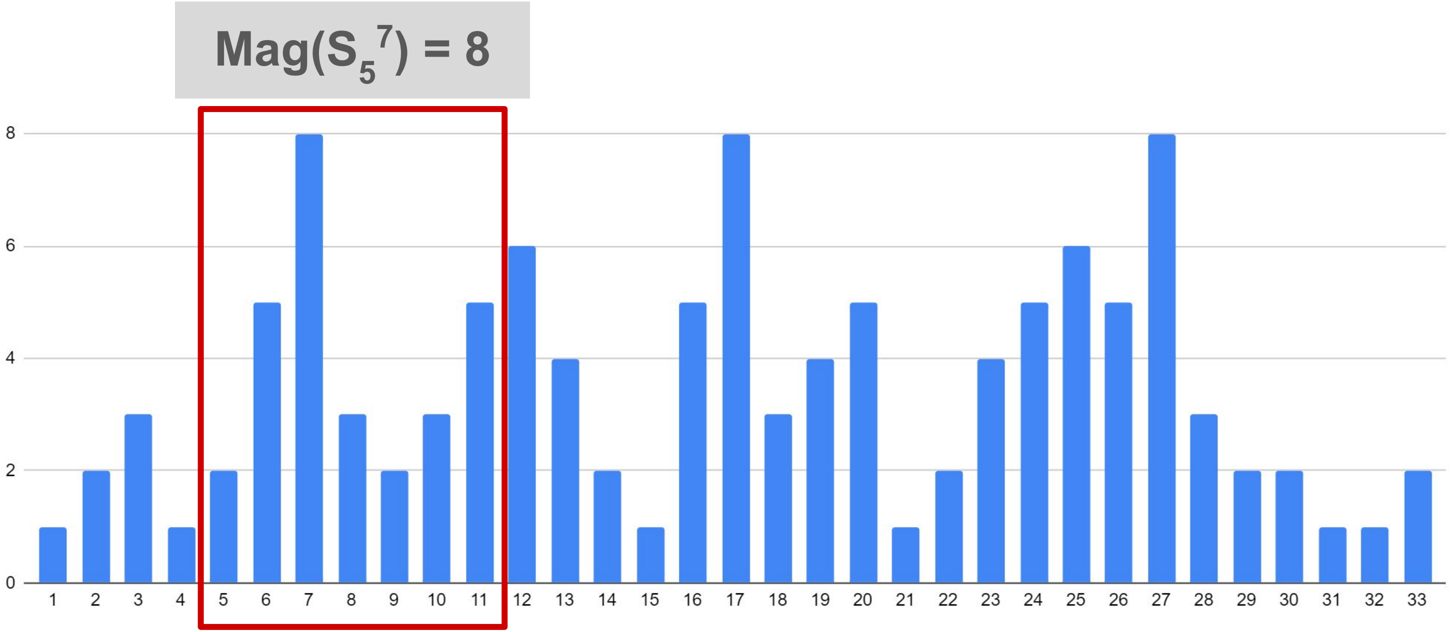


Definition 2. *Subsequence*: Given a time series X of length n, a subsequence S is a subset of m consecutive observations from X, i.e., $S_p^m = (x_p, \ldots, x_{p+m-1})$, where $1 \leq p \leq n - m + 1$, and $m < n$.



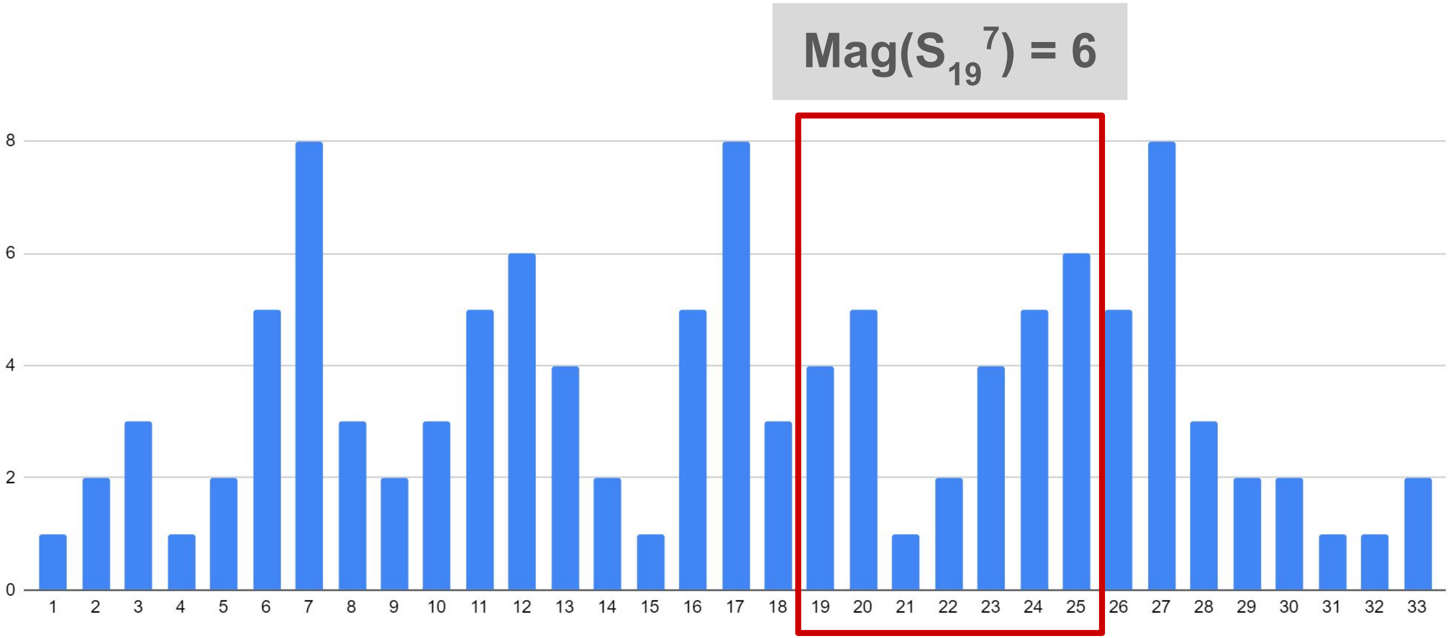
Definition 3. *Magnitude*: Given a subsequence S_p^m of length m, the magnitude of S_p^m is the maximum of all the elements in the subsequence, i.e.,

$$Mag(S_p^m) = \max_{1 \le t \le m} (x_t),$$

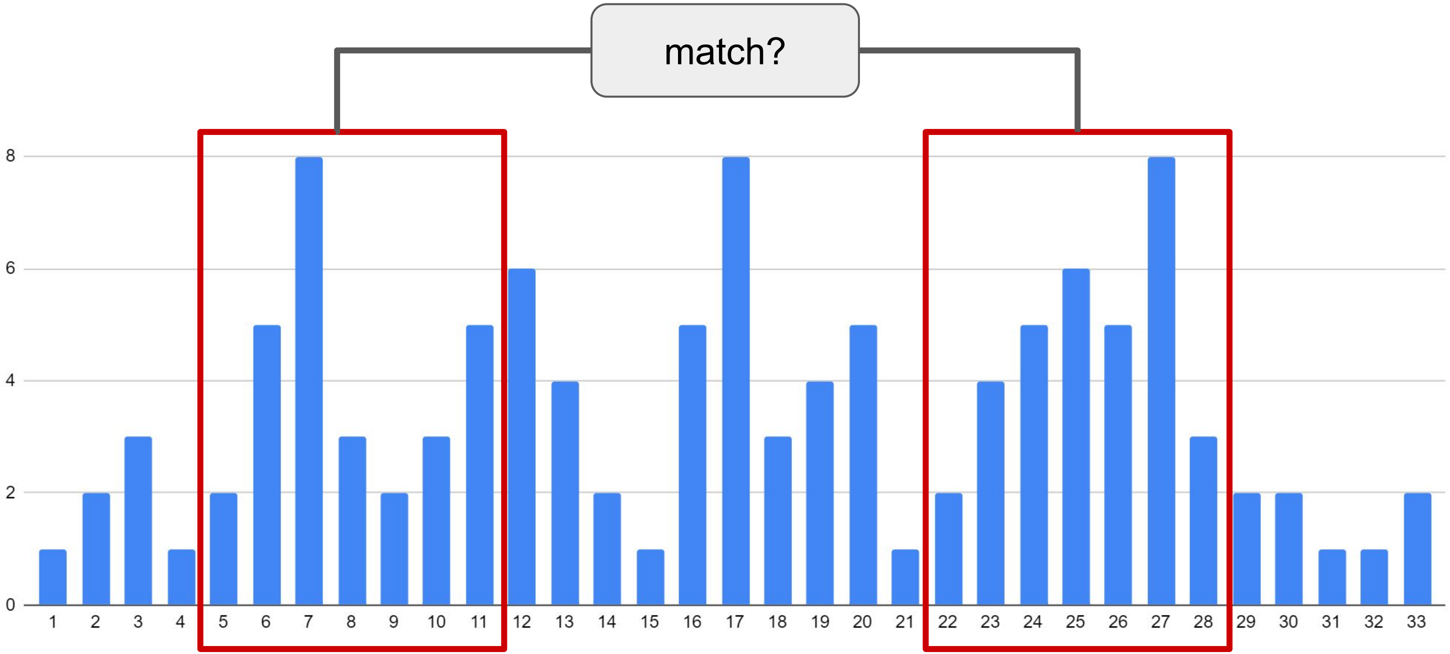


Definition 3. *Magnitude*: Given a subsequence S_p^m of length m, the magnitude of S_p^m is the maximum of all the elements in the subsequence, i.e.,

$$Mag(S_p^m) = \max_{1 \le t \le m} (x_t),$$

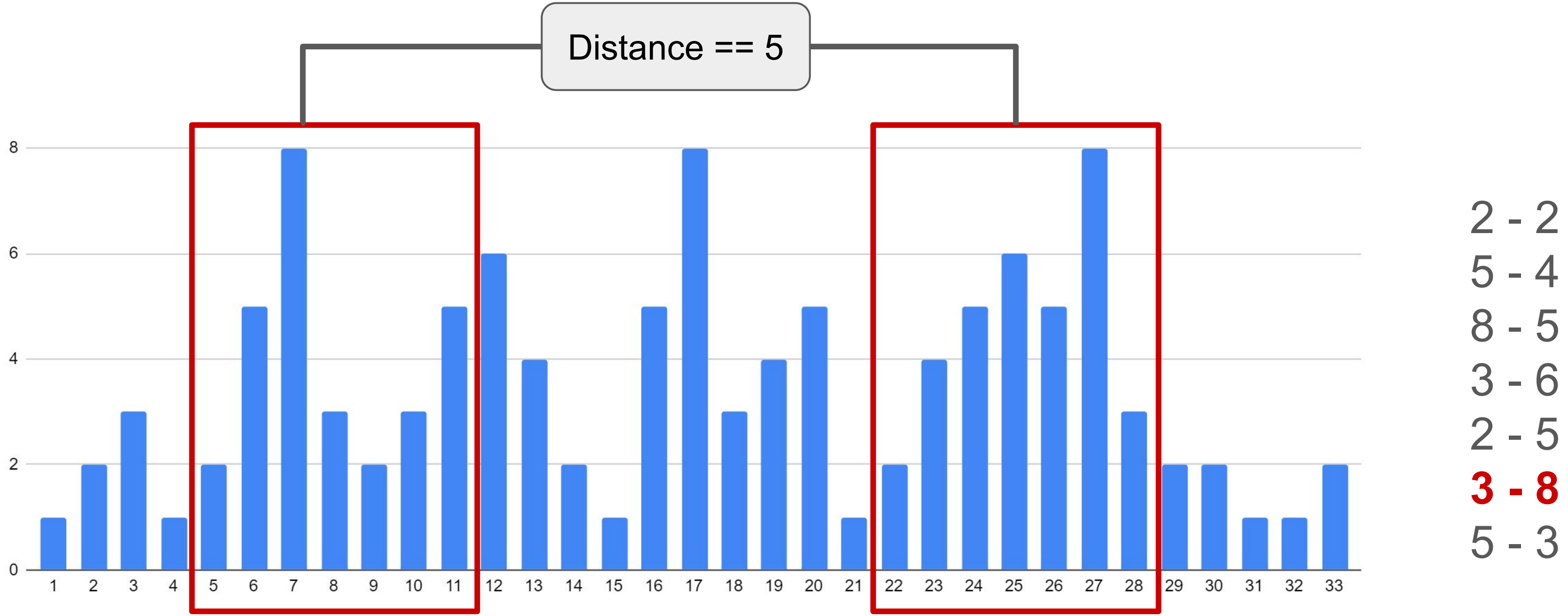


Definition 4. *Match*: Given two subsequences, S_i^m and S_j^m, with the same length of m from X, if the distance between the two subsequences is no greater than a threshold R, i.e., $Dist(S_i^m, S_j^m) \le R$, then the two subsequences are *matched*.



Definition 6. *Distance*: Given two subsequences, S_i^m and S_j^m, of the same length m, the distance between S_i^m and S_j^m is the maximum element-wise difference between S_i^m and S_j^m, i.e.,

$$Dist(S_i^m, S_j^m) = \max_{0 \le t \le m-1} (|x_{i+t} - x_{j+t}|)$$

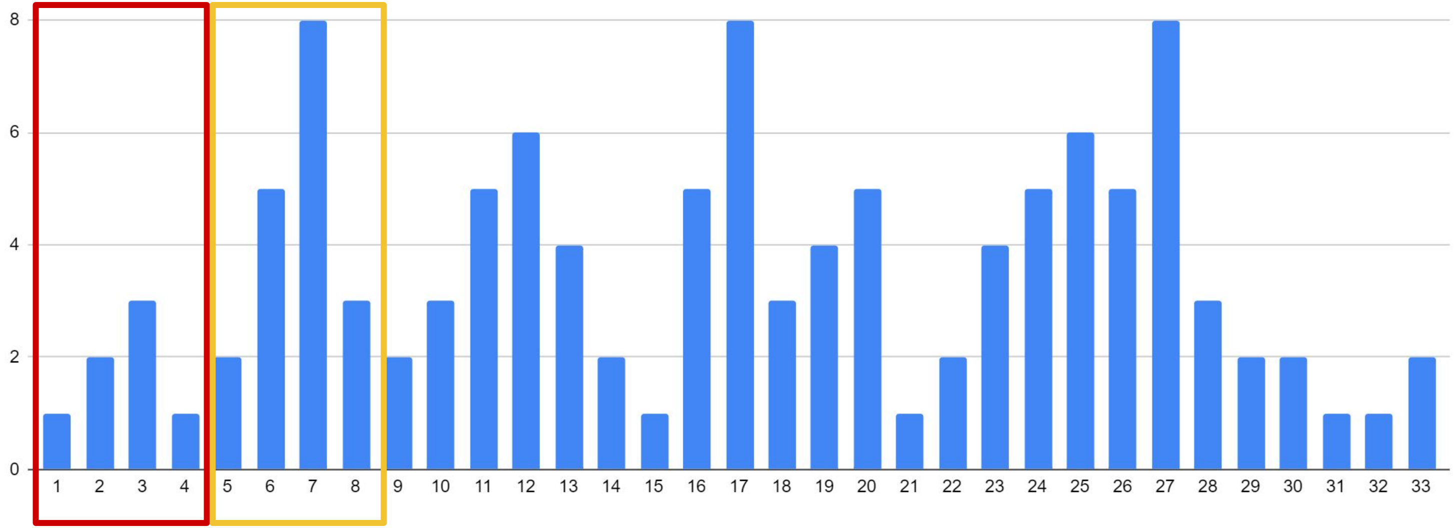


Definition 7. *Motif*: Given a time series X, a subsequence length m, and a distance threshold R, the most significant motif is the subsequence of length m that has most number of matched occurrences under the distance threshold, i.e., $\forall i, j : Dist(S_i^m, S_j^m) \leq R$.

Definition 7. *Motif*: Given a time series X, a subsequence length m, and a distance threshold R, the most significant motif is the subsequence of length m that has most number of matched occurrences under the distance threshold, i.e., $\forall i, j : Dist(S_i^m, S_j^m) \leq R$.

m=4, R=3

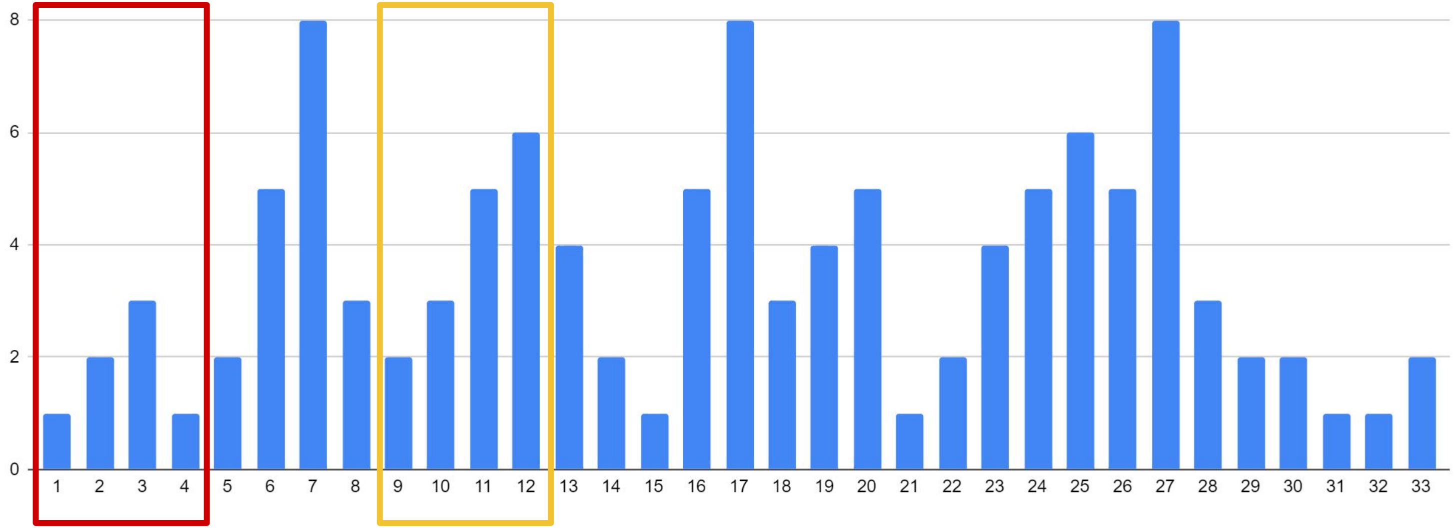
d == 5



Definition 7. *Motif*: Given a time series X, a subsequence length m, and a distance threshold R, the most significant motif is the subsequence of length m that has most number of matched occurrences under the distance threshold, i.e., $\forall i, j : Dist(S_i^m, S_j^m) \leq R$.

m=4, R=3

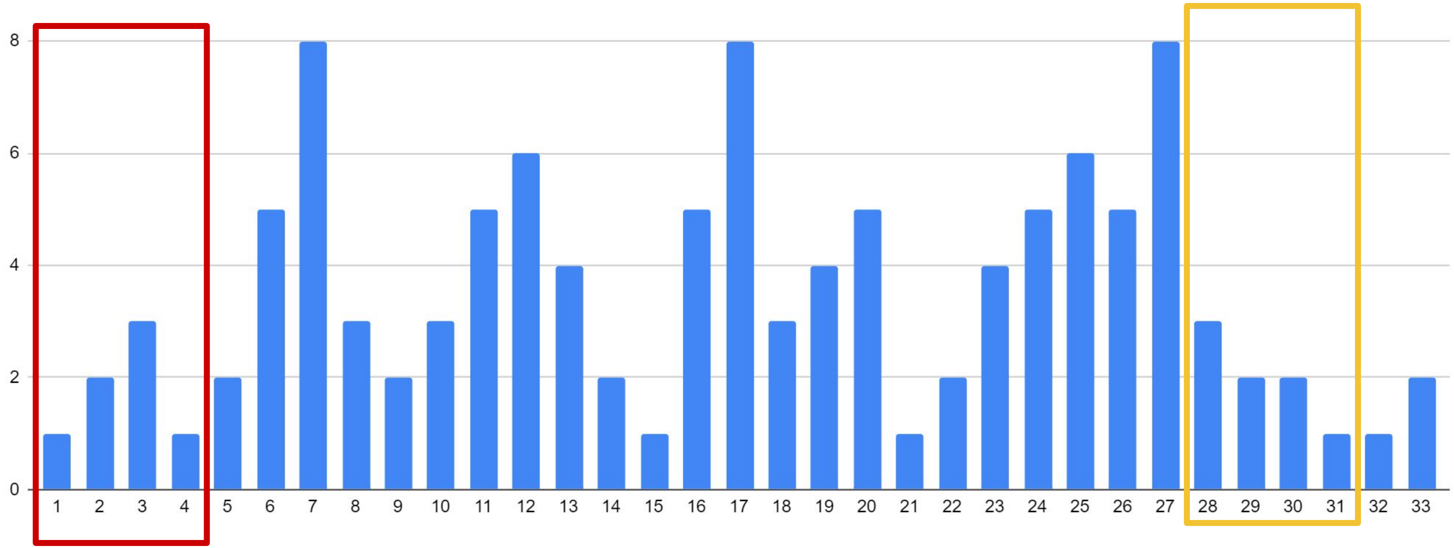
d == 5



Definition 7. *Motif*: Given a time series X, a subsequence length m, and a distance threshold R, the most significant motif is the subsequence of length m that has most number of matched occurrences under the distance threshold, i.e., $\forall i, j : Dist(S_i^m, S_j^m) \leq R$.

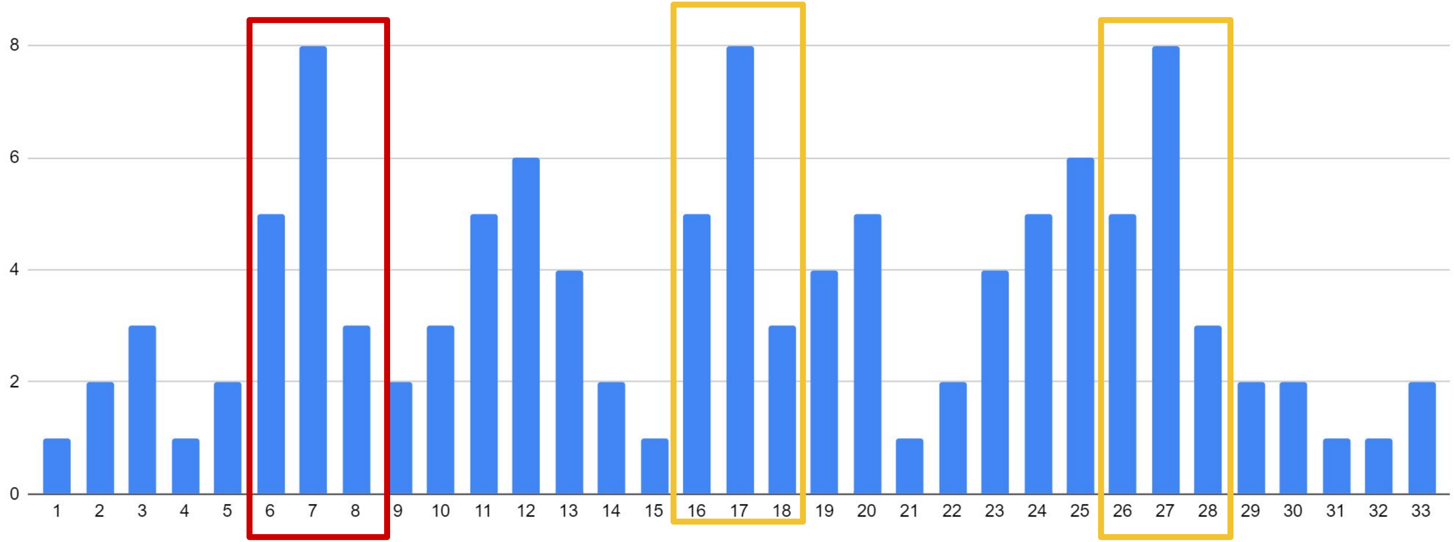
m=4, R=3

d == 2



Definition 7. *Motif*: Given a time series X, a subsequence length m, and a distance threshold R, the most significant motif is the subsequence of length m that has most number of matched occurrences under the distance threshold, i.e., $\forall i, j : Dist(S_i^m, S_j^m) \leq R$.

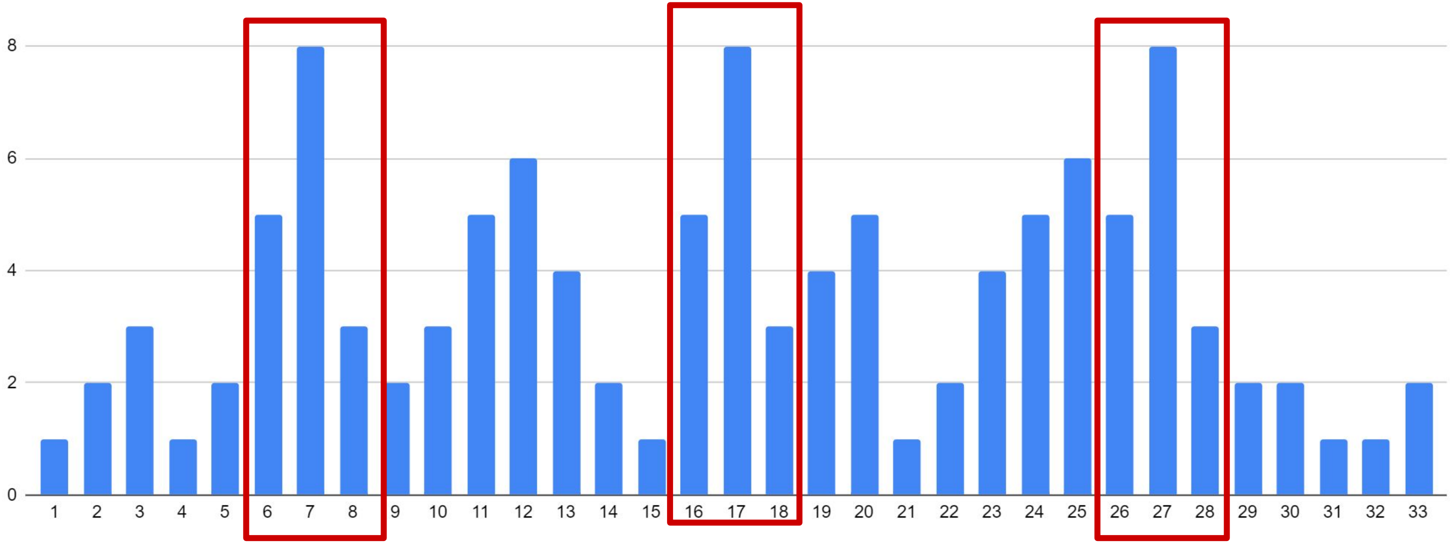
m=3, R=1



Definition 8. *Routine*: Given a frequency threshold C and a magnitude threshold G, a routine is a motif that has at least C matched occurrences in the time series, each of which has at least G magnitude.

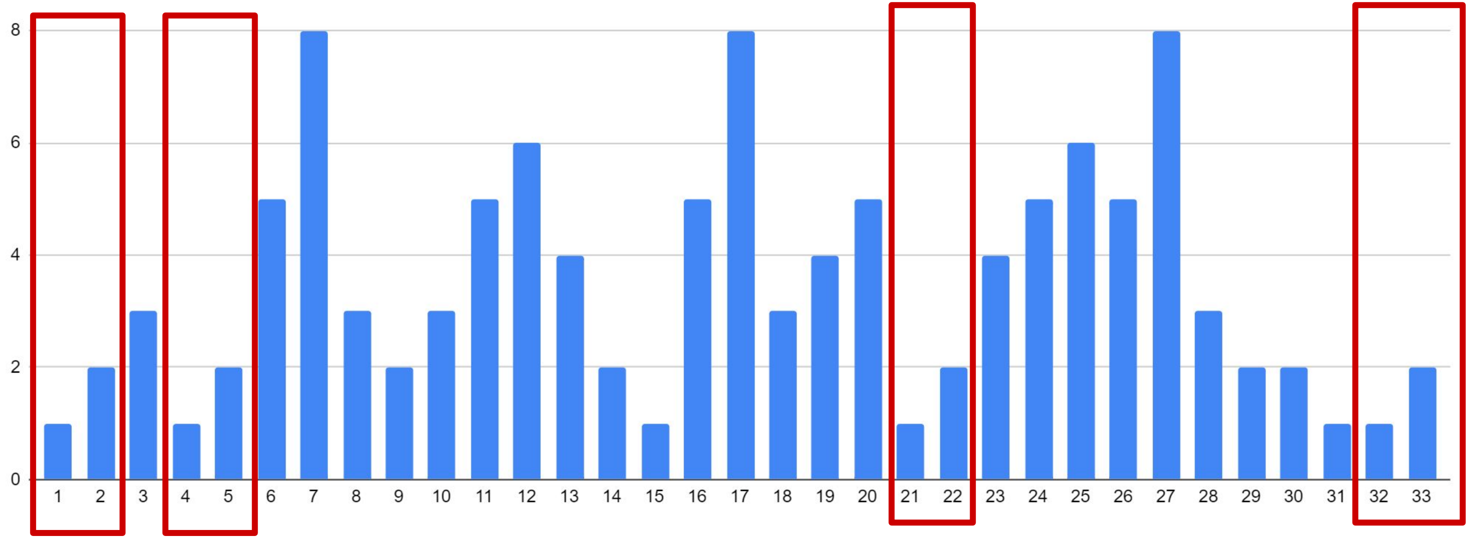
Definition 8. *Routine*: Given a frequency threshold C and a magnitude threshold G, a routine is a motif that has at least C matched occurrences in the time series, each of which has at least G magnitude.

It is a routine with C = 3 and G = 7



Definition 8. *Routine*: Given a frequency threshold C and a magnitude threshold G, a routine is a motif that has at least C matched occurrences in the time series, each of which has at least G magnitude.

It is **not** a routine with C = 3 and G = 7



The routine discovery problem is to find all routines with lengths 1 to m that occur in a smart meter time series X given a magnitude threshold G and frequency threshold C.

m	G	C	R

DRVL

for m in [m_{min}, m_{max}]

DRFL(m)

Algorithm 1: Discovering Routine of Fixed Length (DRFL).

Input : a time series of length n
Parameters: routine length m, distance threshold R; frequency threshold C, magnitude threshold G, overlap parameter ϵ
Output : m-length routines B^m

for $i = 1$ **to** $n - m - 1$ **do**
  extract subsequence S_i^m;
$B^m \leftarrow$ `SubGroup(`S^m, R, C, G`)` ; // See Algorithm 2
for $i = 1$ **to** $|B^m| - 1$ **do**
  for $j = i$ **to** $|B^m|$ **do**
    // See Algorithm 3
    $K_i, K_j \leftarrow$ `OLTest(Inst(`B_i^m`), Inst(`B_j^m`),` ϵ`)` ;
for $i = 1$ **to** $|B^m|$ **do**
  if K_i==*FALSE* **then** remove B_i^m;

Algorithm 1: Discovering Routine of Fixed Length (DRFL).

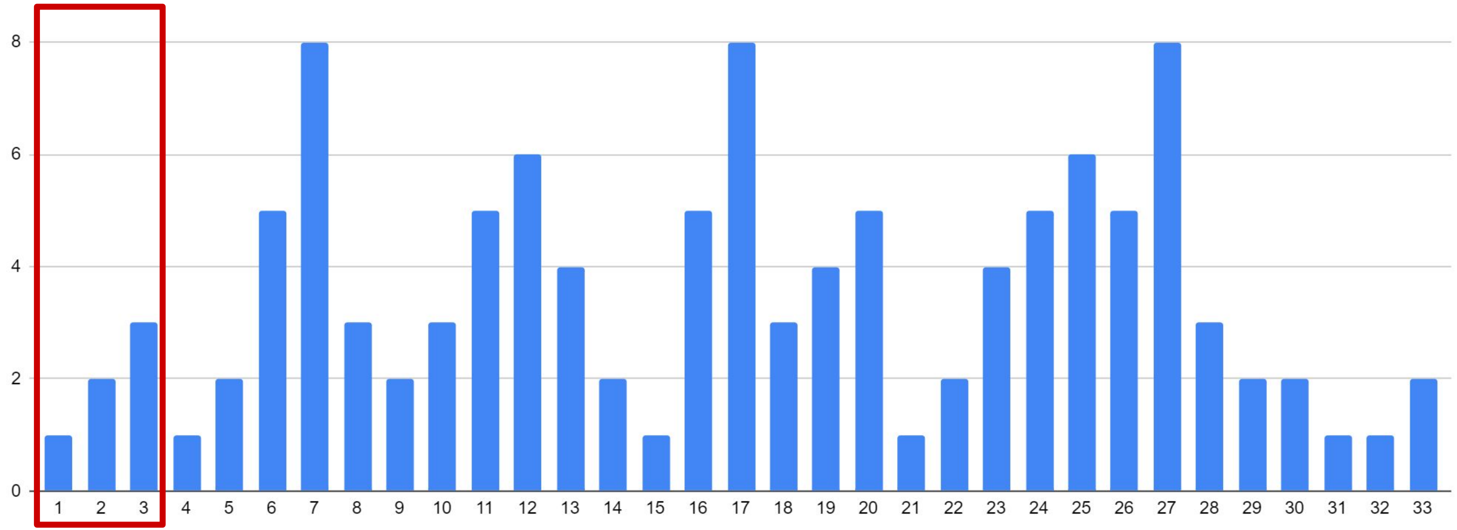
Input : a time series of length n
Parameters: routine length m, distance threshold R; frequency threshold C, magnitude threshold G, overlap parameter ϵ
Output : m-length routines B^m

for $i = 1$ **to** $n - m - 1$ **do**
    extract subsequence S_i^m;
$B^m \leftarrow$ `SubGroup`(S^m, R, C, G) ; // See Algorithm 2
for $i = 1$ **to** $|B^m| - 1$ **do**
    for $j = i$ **to** $|B^m|$ **do**
        // See Algorithm 3
        $K_i, K_j \leftarrow$ `OLTest(Inst(`B_i^m`), Inst(`B_j^m`),` ϵ`)` ;
for $i = 1$ **to** $|B^m|$ **do**
    if K_i==*FALSE* **then** remove B_i^m;

for $i = 1$ **to** $n - m - 1$ **do**

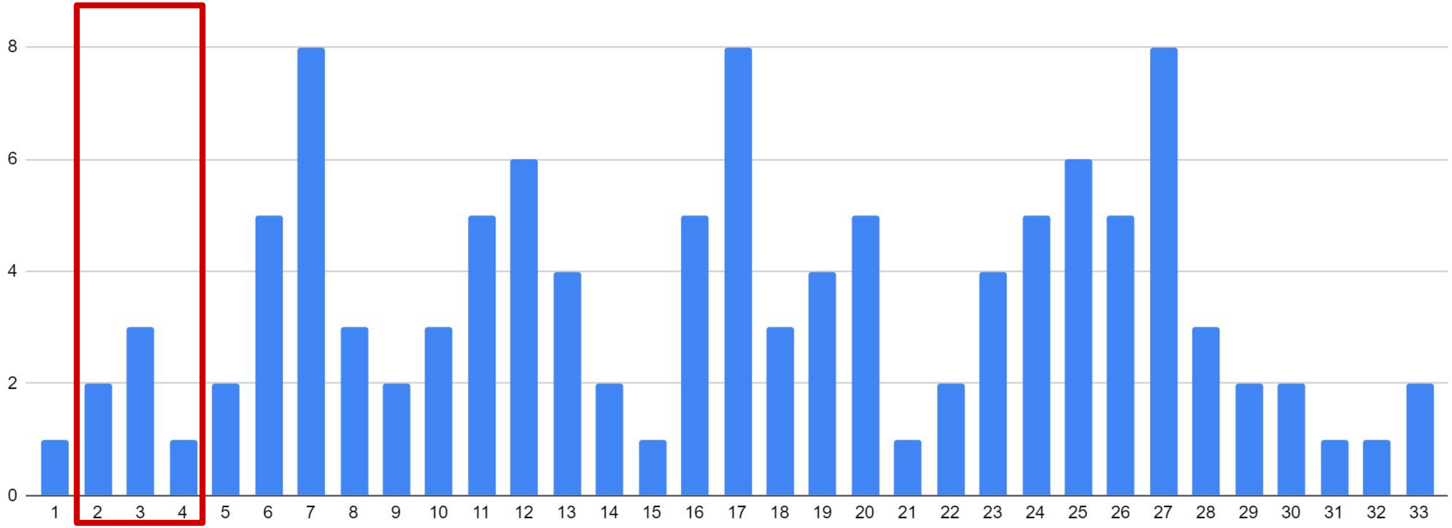
    extract subsequence S_i^m;

m = 3



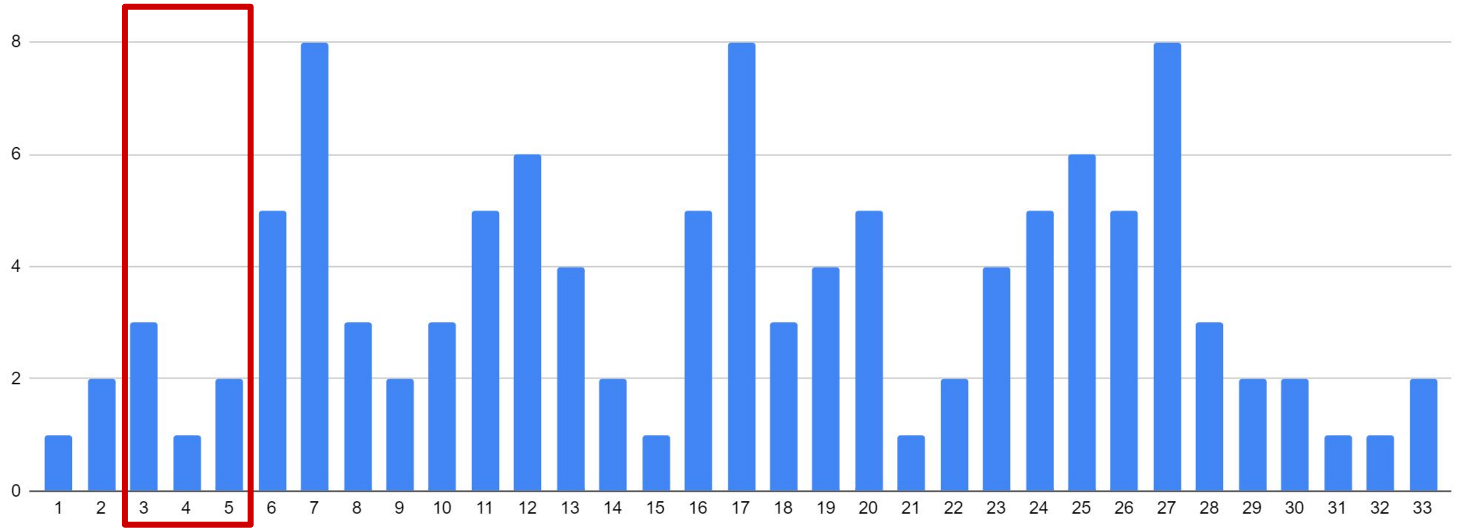
for $i = 1$ **to** $n - m - 1$ **do**
  extract subsequence S_i^m;

m = 3



for $i = 1$ **to** $n - m - 1$ **do**
    extract subsequence S_i^m;

m = 3



Algorithm 1: Discovering Routine of Fixed Length (DRFL).

Input : a time series of length n

Parameters: routine length m, distance threshold R; frequency threshold C, magnitude threshold G, overlap parameter ϵ

Output : m-length routines B^m

for $i = 1$ **to** $n - m - 1$ **do**
    extract subsequence S_i^m;
$B^m \leftarrow$ `SubGroup`(S^m, R, C, G) ;
for $i = 1$ **to** $|B^m| - 1$ **do**
    for $j = i$ **to** $|B^m|$ **do**
        `// See Algorithm 3`
        $K_i, K_j \leftarrow$ `OLTest(Inst(`B_i^m`), Inst(`B_j^m`),` ϵ`)` ;
for $i = 1$ **to** $|B^m|$ **do**
    if K_i==*FALSE* **then** remove B_i^m;

clustering

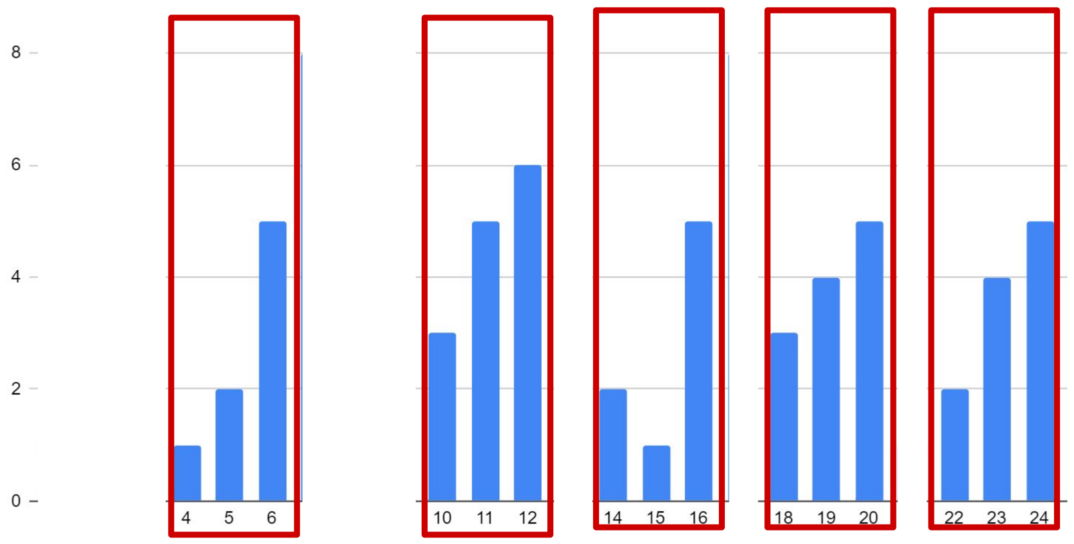
Algorithm 1: Discovering Routine of Fixed Length (DRFL).

Input : a time series of length n
Parameters: routine length m, distance threshold R; frequency threshold C, magnitude threshold G, overlap parameter ϵ
Output : m-length routines B^m

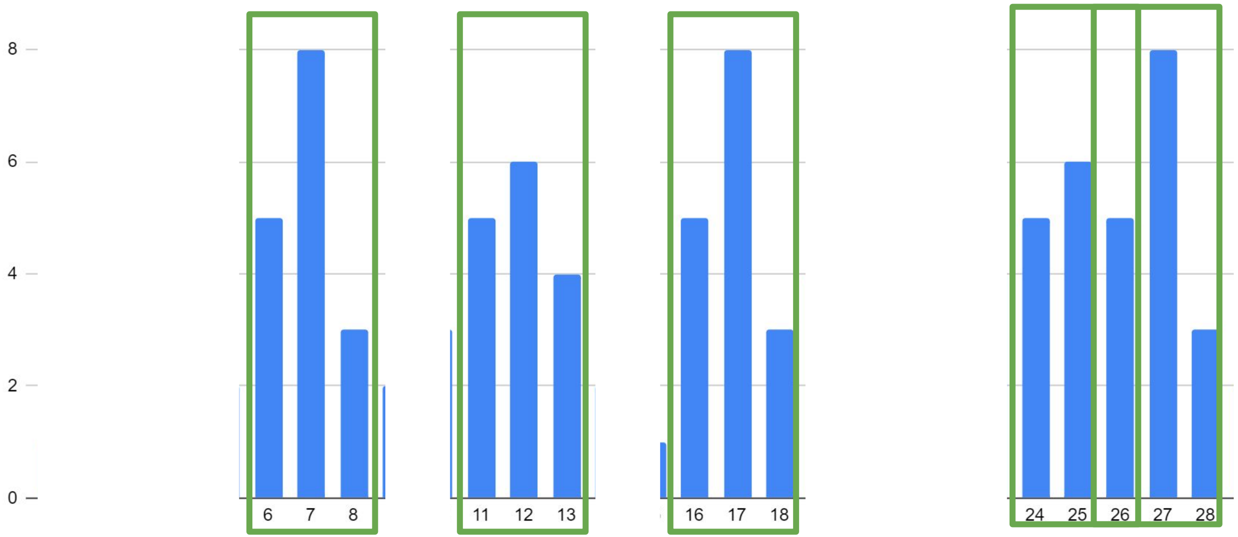
for $i = 1$ **to** $n - m - 1$ **do**
    extract subsequence S_i^m;
$B^m \leftarrow$ `SubGroup`(S^m, R, C, G) ;   `// See Algorithm 2`
for $i = 1$ **to** $|B^m| - 1$ **do**
    for $j = i$ **to** $|B^m|$ **do**
        `// See Algorithm 3`
        $K_i, K_j \leftarrow$ `OLTest(Inst(`B_i^m`), Inst(`B_j^m`),` ϵ`)` ;
for $i = 1$ **to** $|B^m|$ **do**
    if K_i==*FALSE* **then** remove B_i^m;

Overlapping clusters

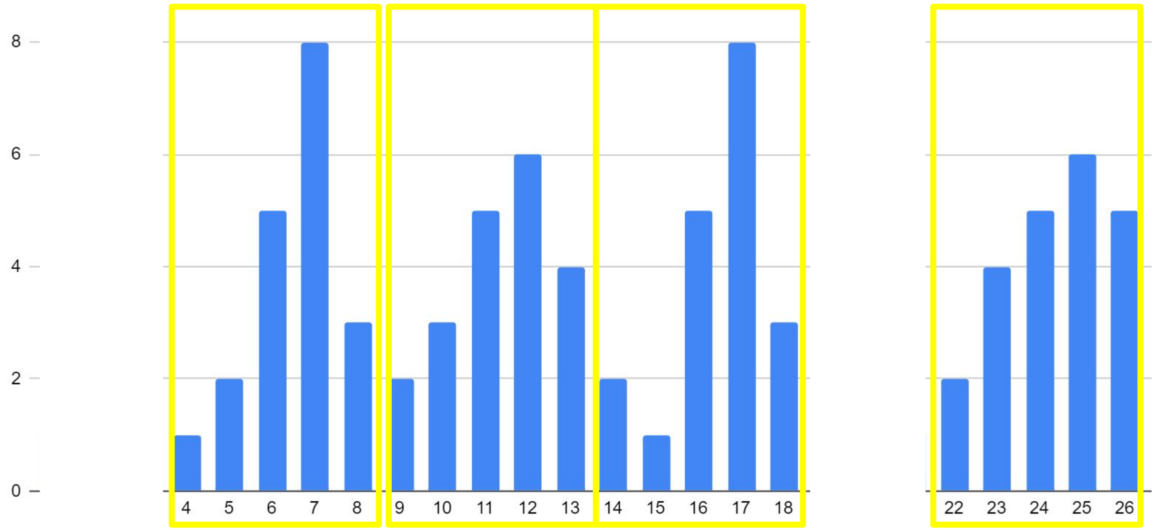
m = 3, R = 2, G = 5, C = 4

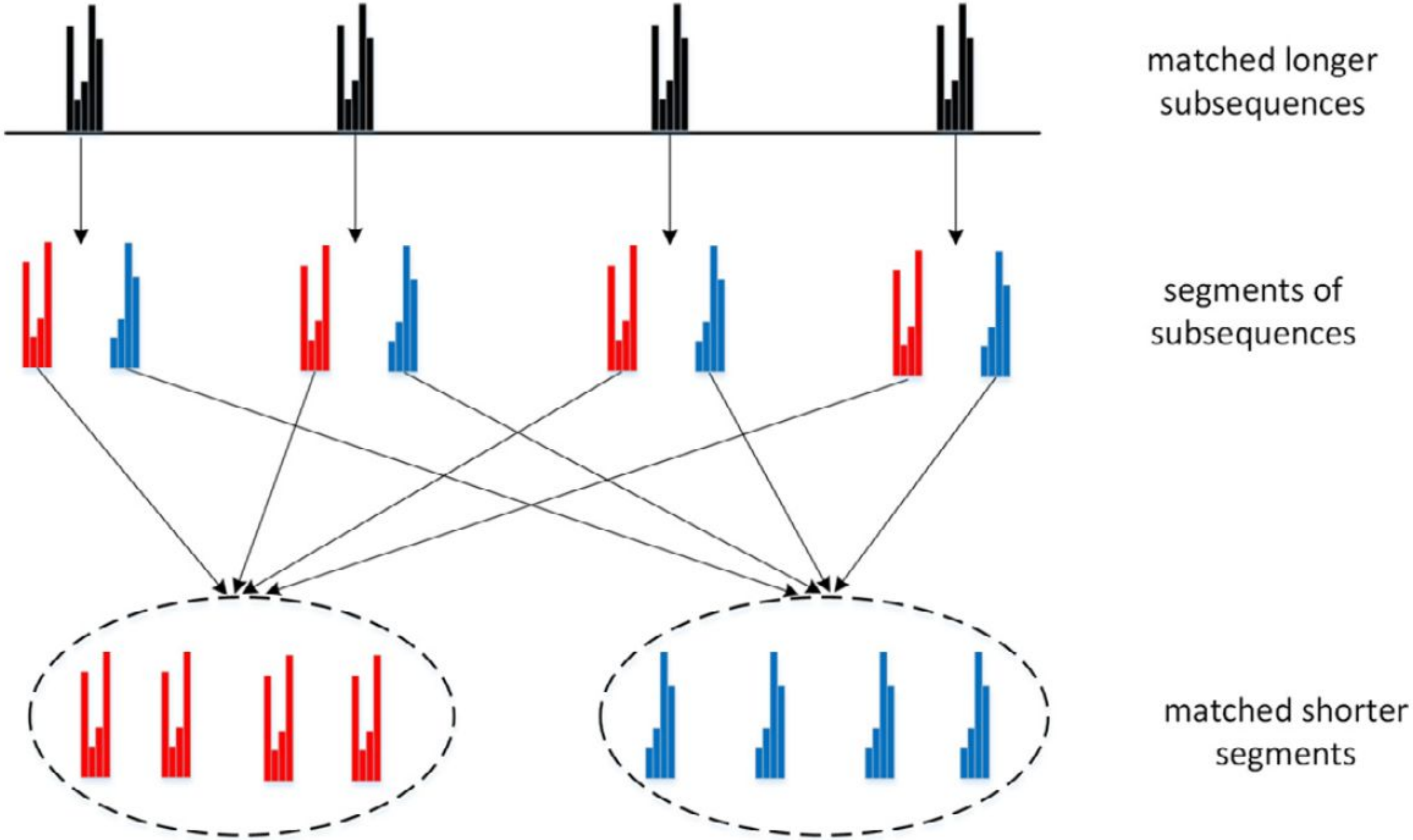


m = 3, R = 2, G = 5, C = 4

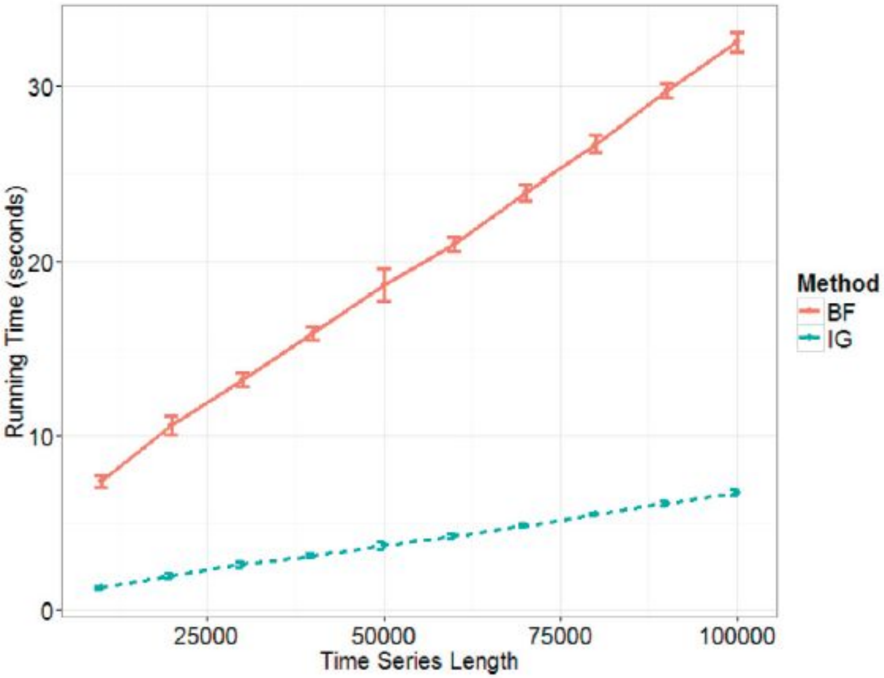


m = 5, R = 2, G = 5, C = 3

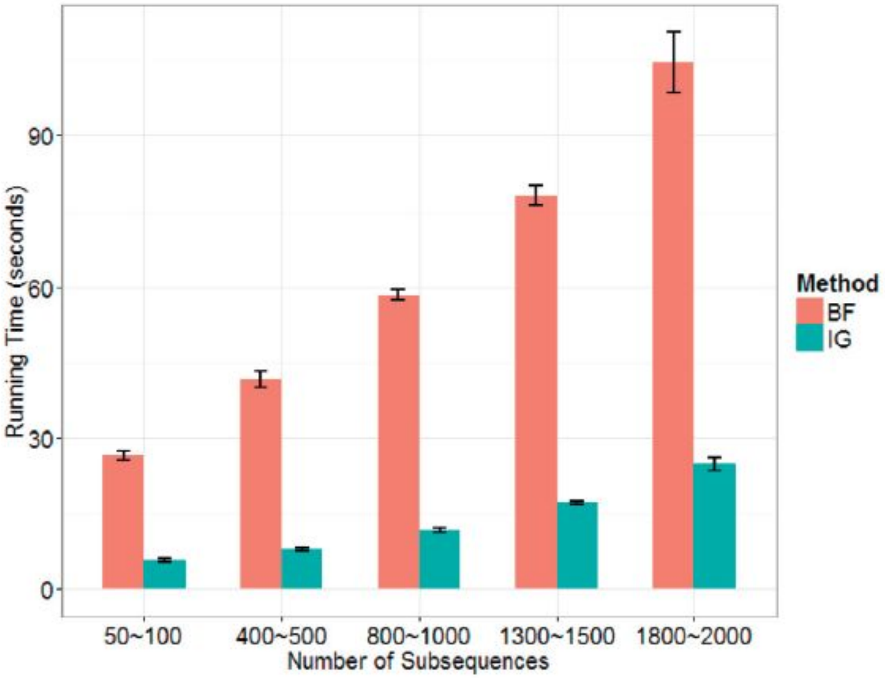



matched longer subsequences
segments of subsequences
matched shorter segments

Experiments on synthetic database



(a) Different sequence lengths.

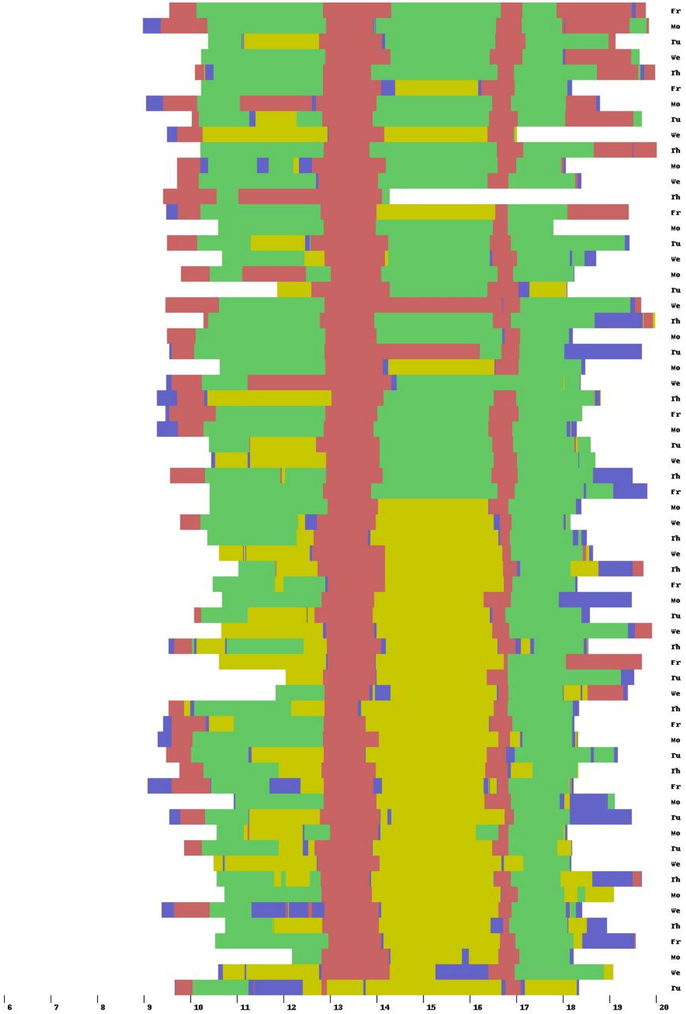


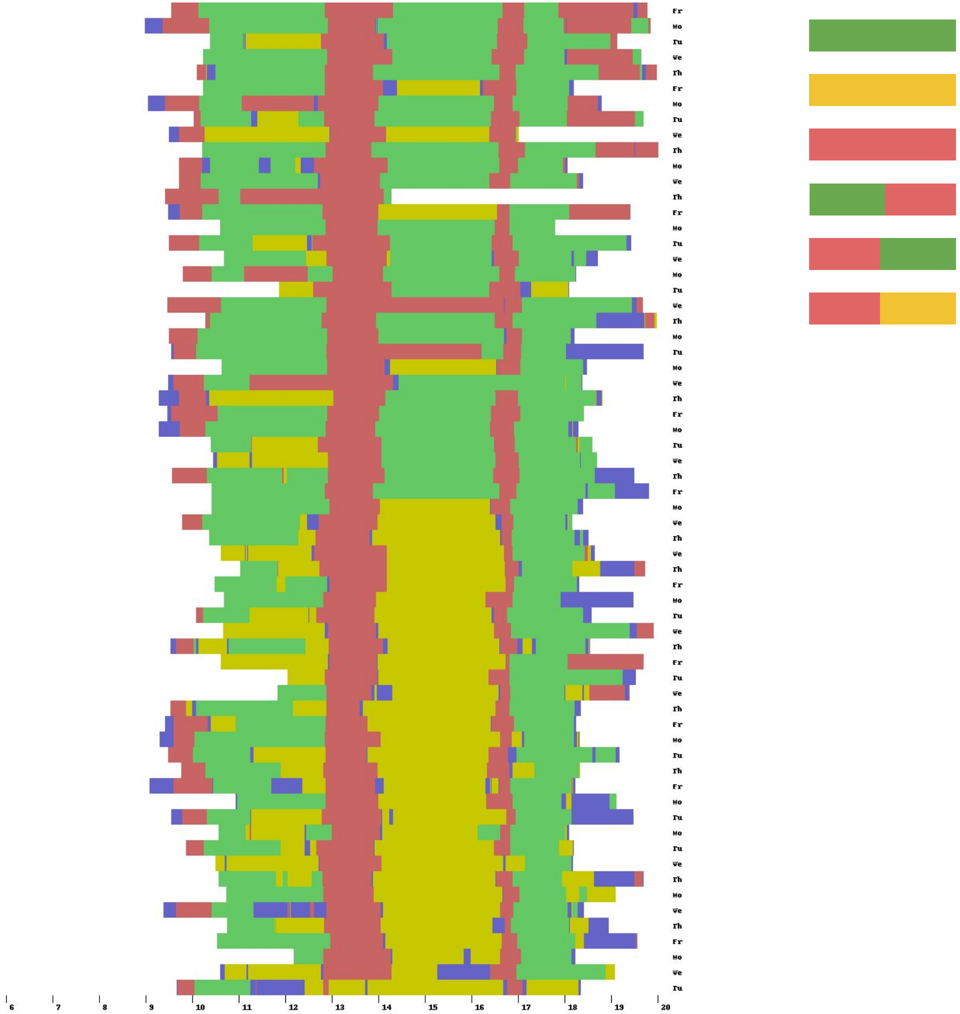
(b) Different numbers of subsequence instances

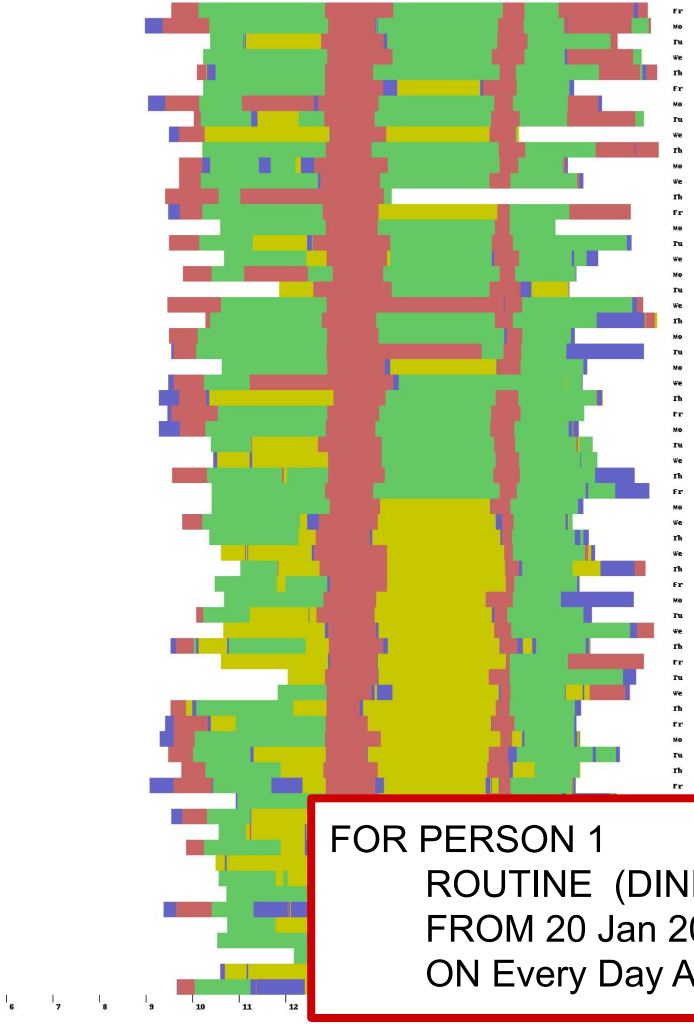
Experiments on real datasets

1. More than one dataset
2. A state of the art method is used to comparisons

This paper is interesting for **us** because...


Fr
Mo
Tu
We
Th
Fr
Mo
Tu
We
Th
Mo
We
Th
Fr
Mo
Tu
We
Mo
Tu
We
Th
Mo
Tu
Mo
We
Th
Fr
Mo
Tu
We
Th
Fr
Mo
We
Th
We
Th
Fr
Mo
Tu
We
Th
Fr
Tu
We
Th
Fr
Mo
Tu
Th
Fr
Mo
Tu
Mo
Tu
We
Th
Mo
We
Th
Fr
Mo
We
Tu
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20


6 7 8 9 10 11 12 13 14 15 16 17 18 19 20



FOR PERSON 1

ROUTINE (DINING ROOM, TV ROOM) OCCURS 23 TIMES

FROM 20 Jan 2014 TO 27 Feb 2014

ON Every Day AT 15:00