Spatio-temporal Mining of Solar-Terrestrial Satellite Observational Data for Distributed Database System

Rie Honda
Kochi University, Japan
honda@is.kochi-u.ac.jp
CODATA2006, Oct 24, 2006

Backgrounds

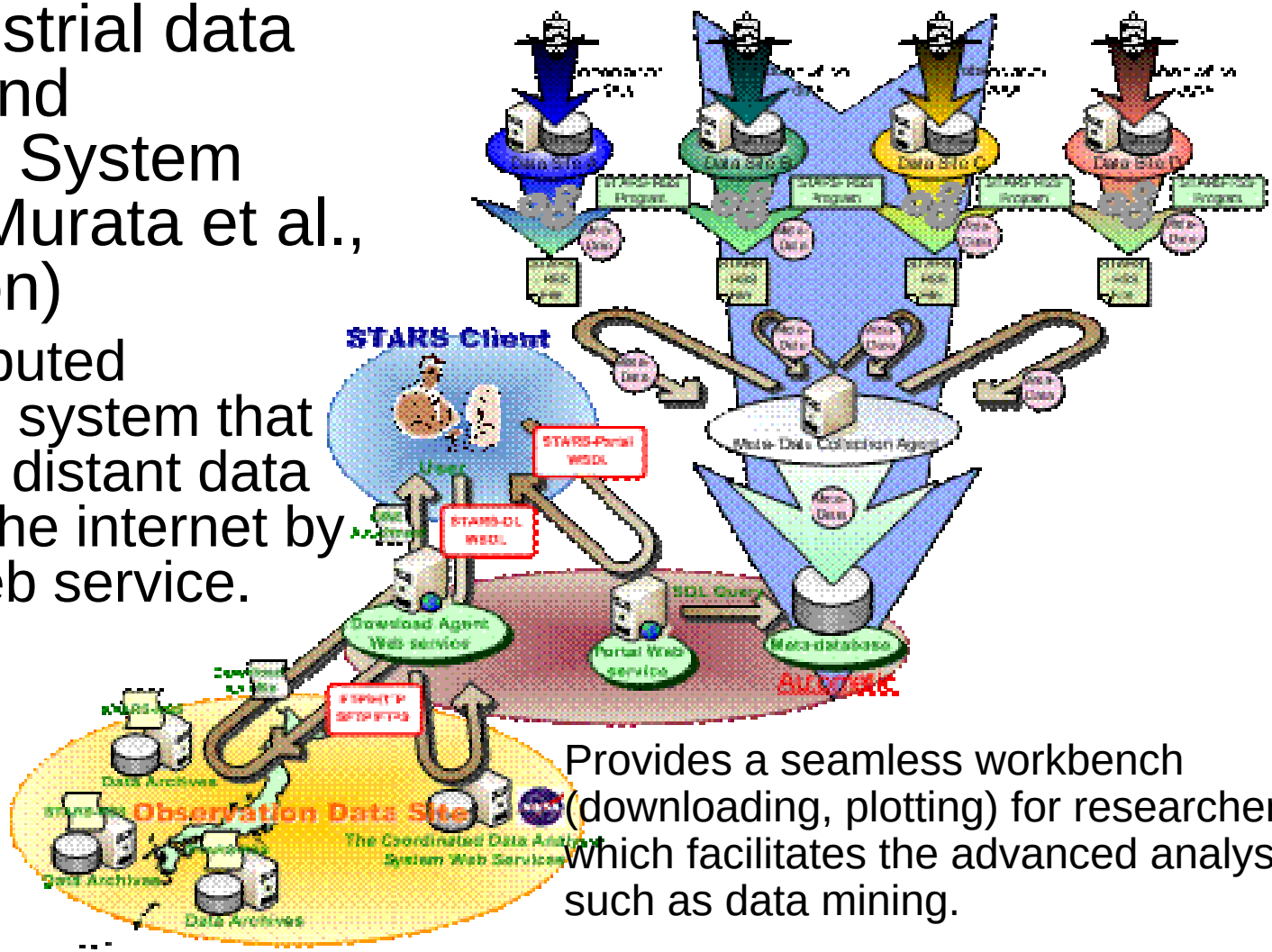
- A large amount of the Solar-terrestrial data obtained by spacecrafts has been accumulated on the database.
  - a various type of spatio-temporal data sampled at different time intervals, and are stored at geographically distant sites.

  →Researchers had difficulties in conducting integrated analysis by using multiple attributes.

•How we can facilitate the knowledge discovery from such a large, inhomogeneous, distributed data system?

Current status of workbench development

- Solar-Terrestrial data Analysis and Reference System (STARS, Murata et al., this session)
  - the distributed database system that connects distant data sites on the internet by using Web service.



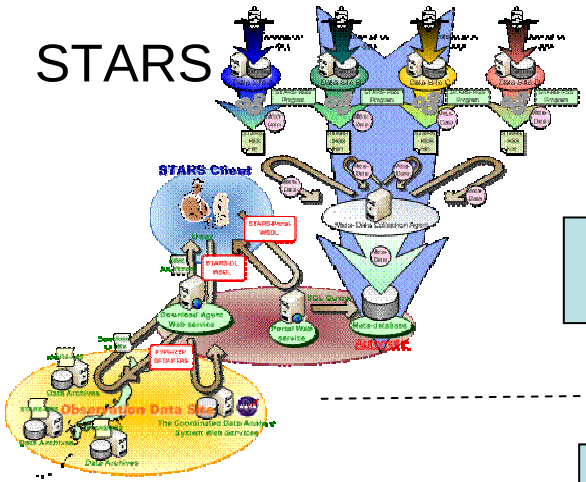
Provides a seamless workbench (downloading, plotting) for researchers which facilitates the advanced analysis such as data mining.

Objectives of this study

- Conducting overall process of data mining on the distributed data system : STARS.
  - Data collection, re-sampling, pattern discovery, examination
- Examination of a practical example of data mining
  - Automatic detection of epochs from Geotail PWI/SFA power spectrum

Overall process of DM

STARS



Current status

Local site

Data collection

Resampling

Mining (clustering)

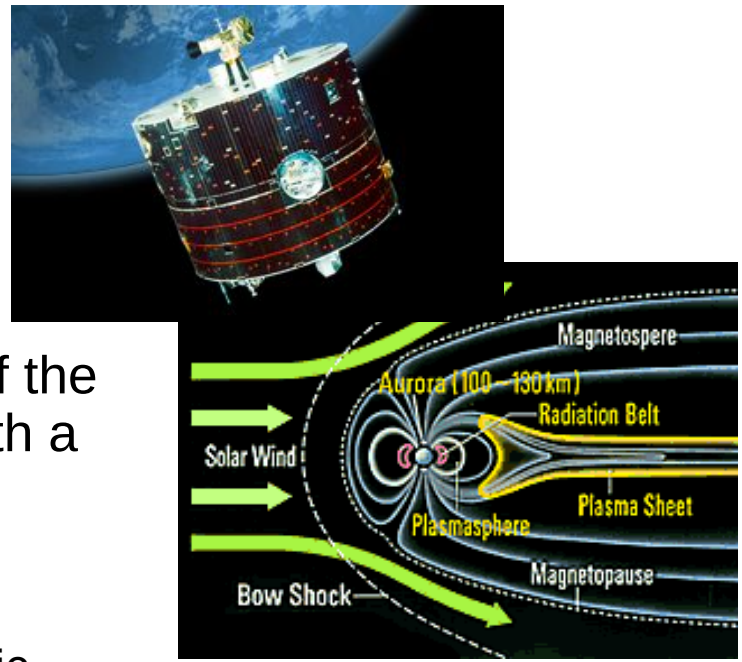
Evaluation

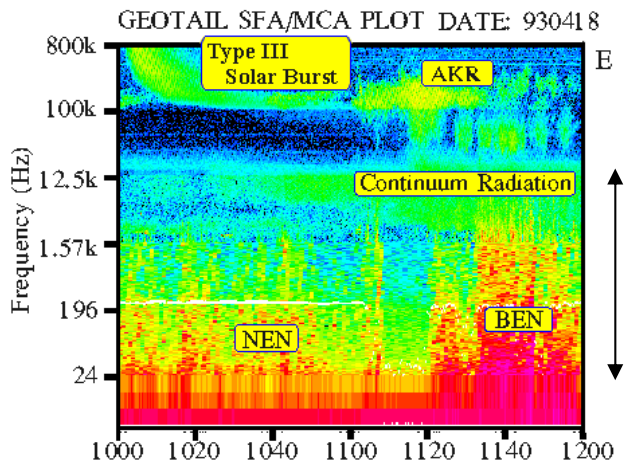
Knowledge discovery

Migration of these processes Into STARS In future

Target dataset

- Geotail dataset
  - Launched on July 24, 1992.
  - Orbit: 8 Re to 210 Re
  - Primary purpose of this mission
    - Study the structure and dynamics of the tail region of the magnetosphere with a comprehensive set of scientific instruments
      - magnetic field (MGF) ,electric field (EFD) ,Plasma (LEP, CPI),Energetic Particle (HEP, EPIC),Plasma Wave (EPIC, PWI)
  - PWI/SFA
    - Spectral information on plasma wave amplitudes



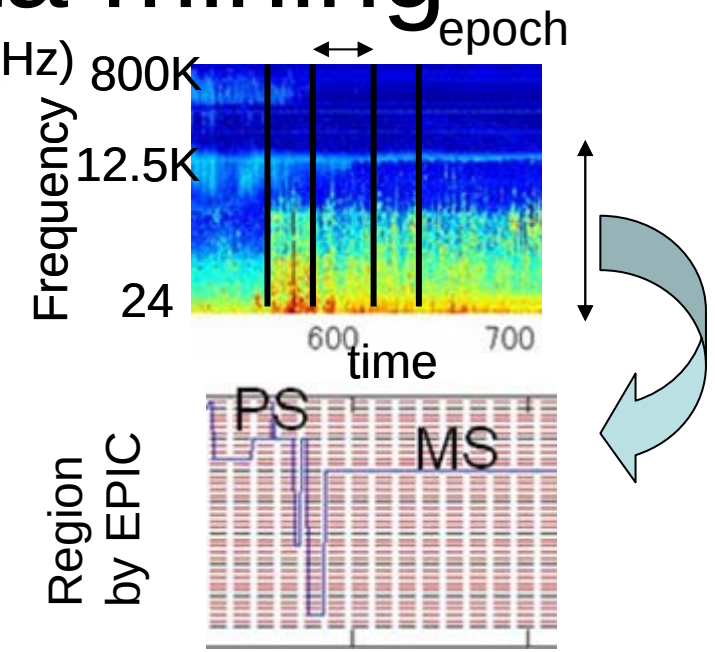


http://www.isas.jaxa.jp/e/enterp/missions/geotail/achiev/typical.shtml

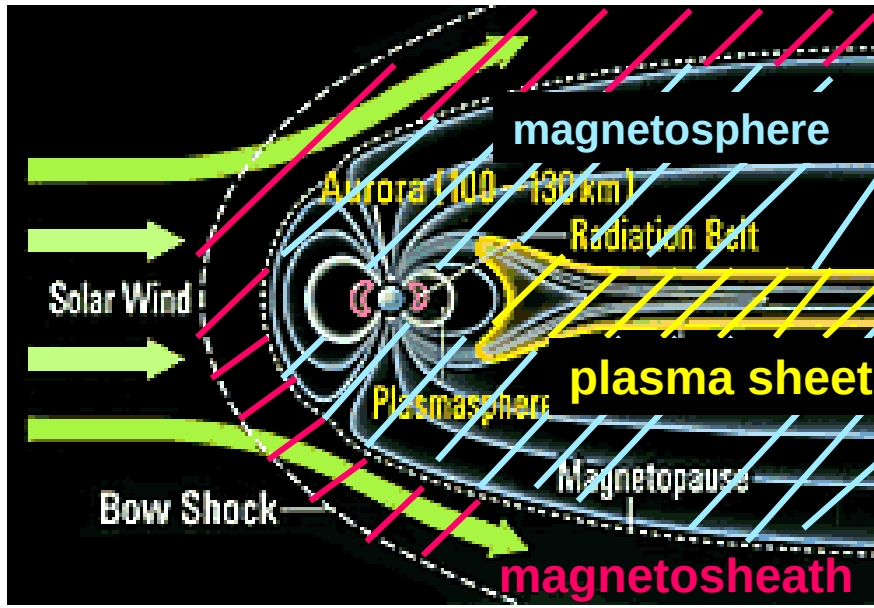
Target of data mining

- Recognizing the epoch from SFA power spectrum at the level of human visual recognition (at least)
  - 1993.1.1 – 1993.9.30
  - Sampling interval: 96sec
  - f :24Hz-12.5KHz

clustering



- Finding the relationship between the epochs and the region where the spacecraft exists.
  - Ground truth: region labeled by EPIC group
    - MS Magnetosheath
    - MG Magnetosphere
    - PS Plasma Sheet
    - BL Boundary Layer:



Clustering Method

- Kohonen's self organizing map (SOM)
  - Unsupervised learning for multi-dimensional vectors

For all x_j,

$$c = \arg\min \left| x_i - m_c(t) \right|$$

- c : winner

For $i \in N_c$

$$m_i(t+1) = m_i(t) + \alpha(t)\left(x_i - m_c(t)\right)$$

Nc : c's neighborhood

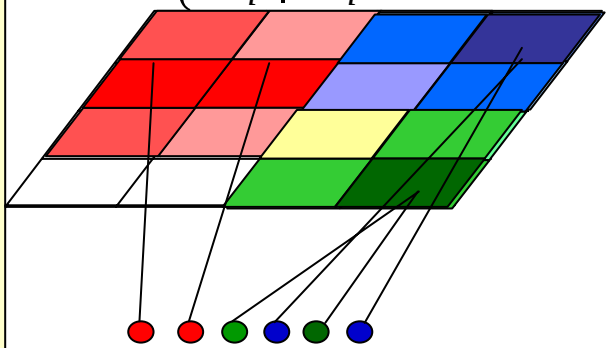
α : the learning rate(0.3)

Iterate the above process for T cycles.

Competition layer(feature map)

Weight vector

$$M = \left\{ m_i \mid m_i \in R^n, i = 1,2,3,\ldots,k \right\}$$

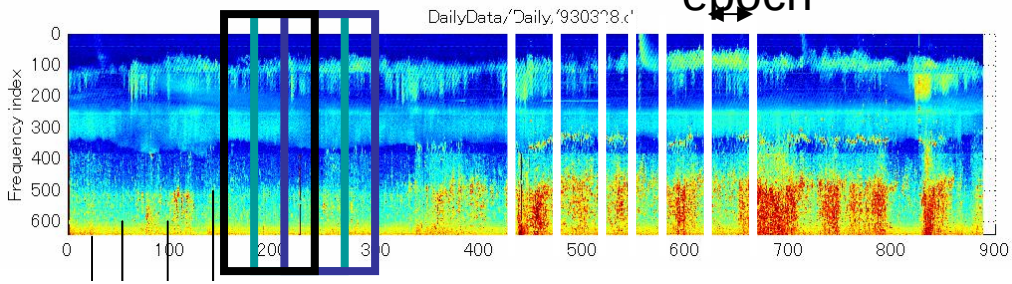


Input layer

input vectors

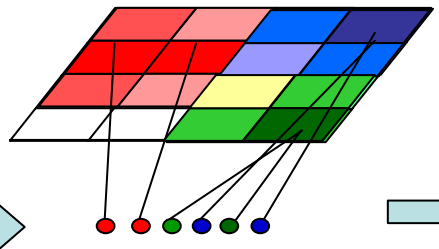
$$X = \left\{ x_i \mid x_i \in R^n, i = 1,2,3,\ldots,d \right\}$$

Two stage SOM for time series of Power spectrum of SFA

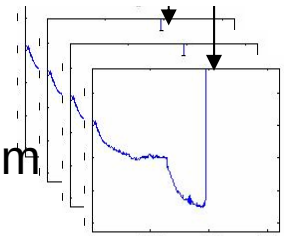


time window: 20 pt/40pt (33min/66min)

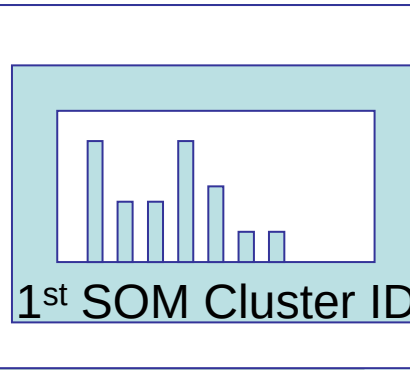
Clustering by 1st SOM



4x4



Power spectrum

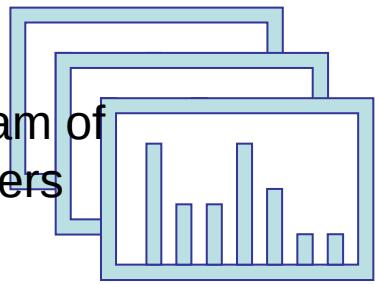


Create the histogram of clusters in each time window

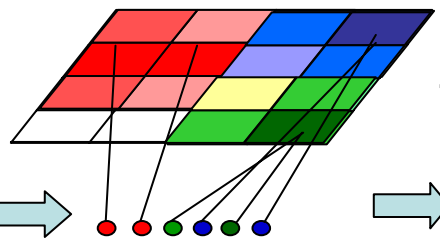
12334567866432312462 ・・・

time series of 1st cluster ID

Clustering by 2nd SOM



Histogram of 1st clusters In time window



4x4

epoch

1 1 1 2 2 2 2 1 1 3 3 ・・・

time series of 2nd cluster ID

Example of result

1993.1.10 (24hour)

SFA power spectrum

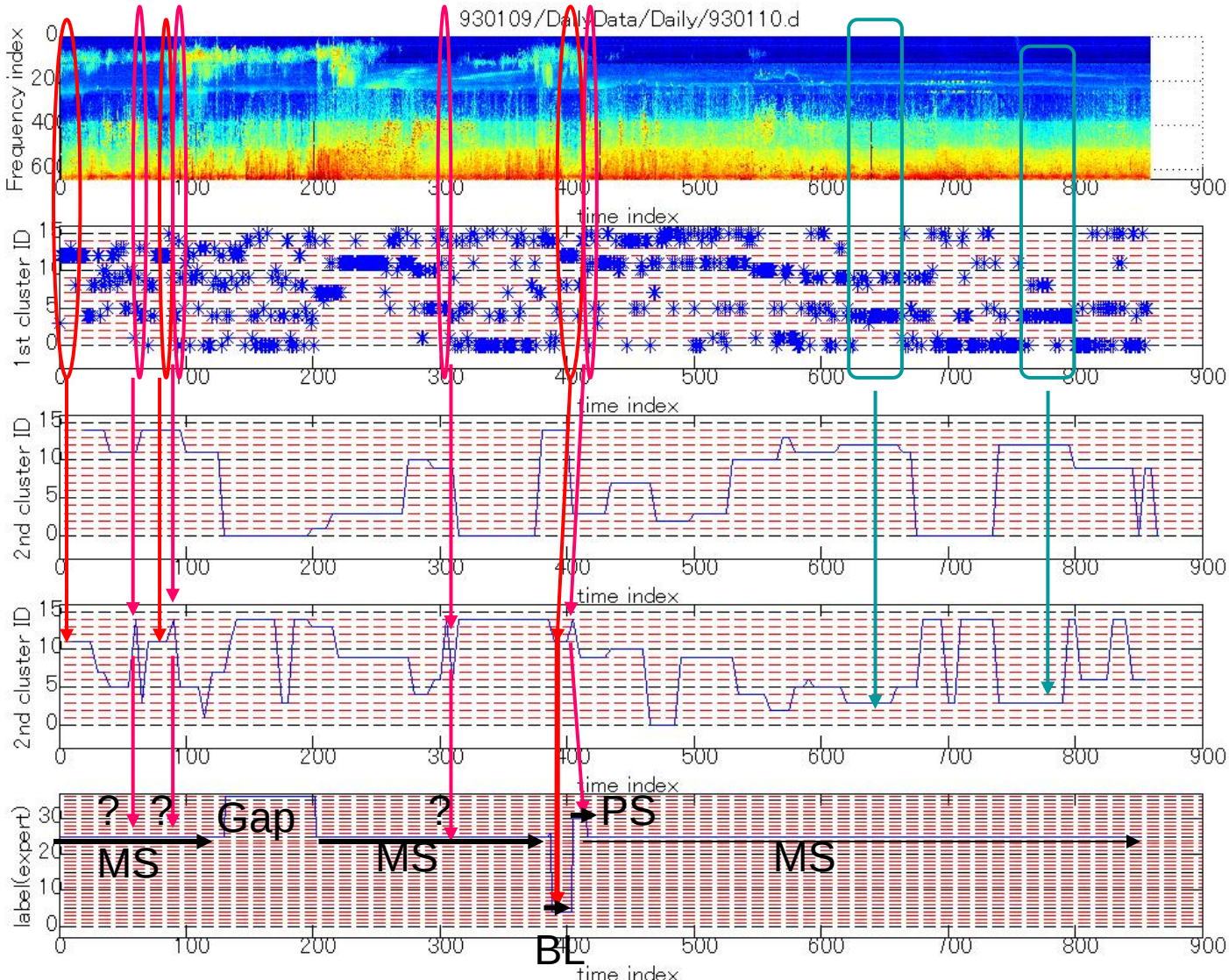
1st Cluster ID

2nd Cluster ID

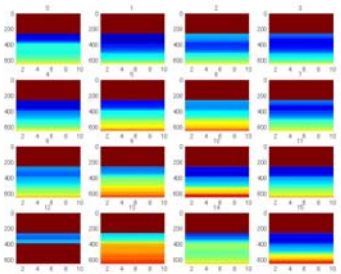
time window 66sec →

time window 33 sec →

Region by EPIC group

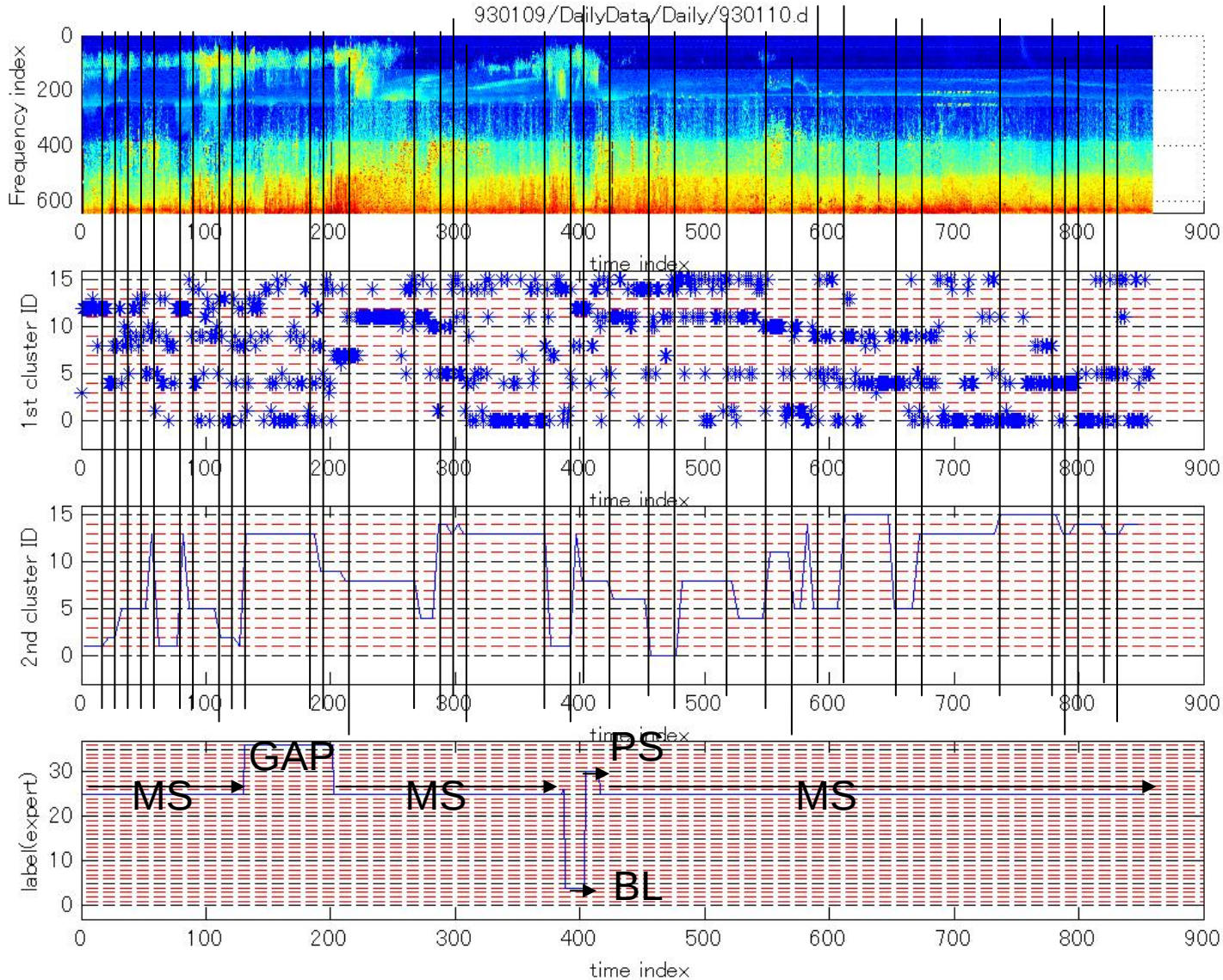


Result: epoch creation



time window
33 min
Offset 480 sec

Higher level Knowledge Discovery
e.g.
Type A epoch occurs within 30 min. after Type B epoch.



Conclusion

- Overall process of the practical data mining case was conducted on STARS
- Two stage SOM successfully found epochs from time series of SFA power spectrum.
- Epochs found by SOM just partly coincide with the labeling of the region by EPIC group.

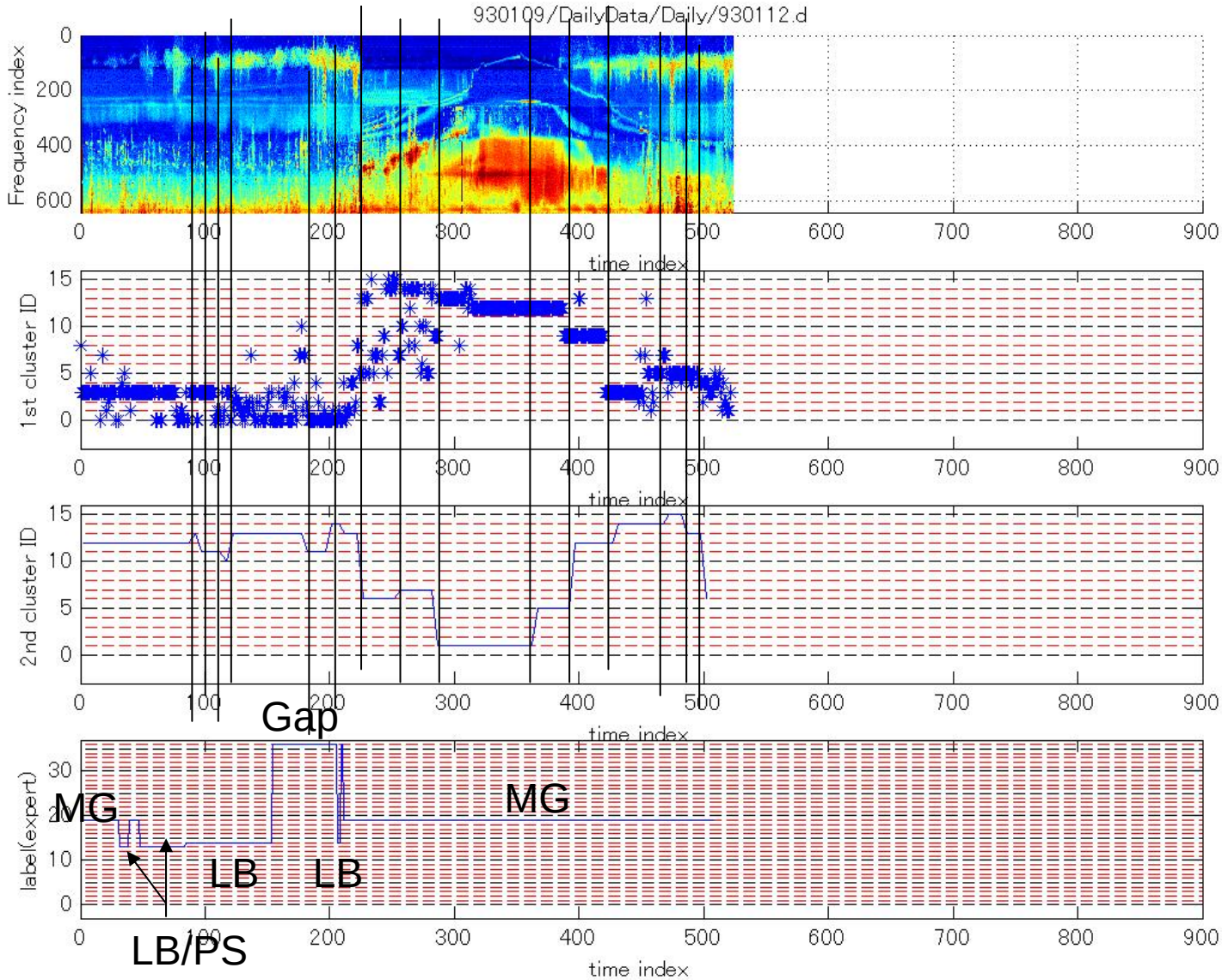
Future directions

- Optimization of clustering by referring to experts knowledge (e.g., number of clusters)
- Extension to multiple attribute data, combination of the characteristics of multiple bands.
- Migration of re-sampling and mining processes into STARS .

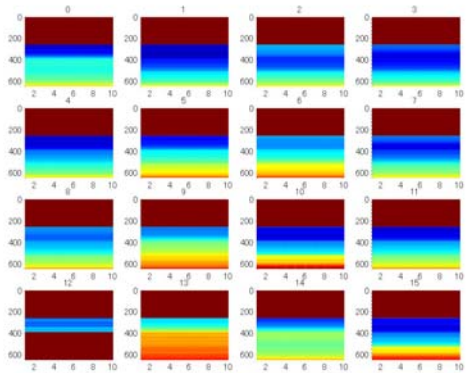
Acknowledgement

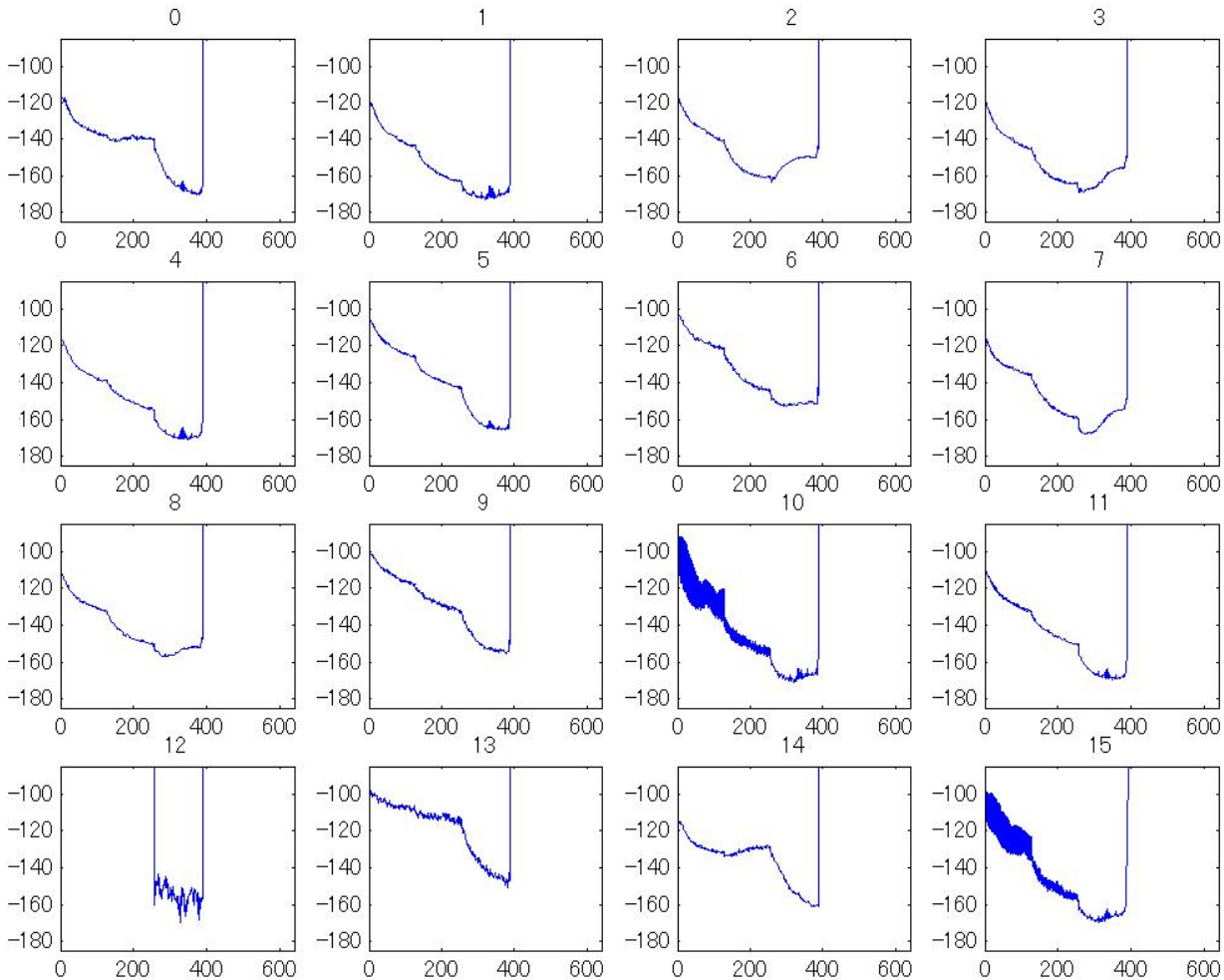
- STARS developing team:
  - K. T. Murata, Kazunori Yamamoto, Takuya, Kumo, Satoshi Ishikura, Eizen Kimura(Ehime University)
  - All the organization involved in Stars PROJECT（JAXA/ISAS, Kyoto Univ., etc)

Result: epoch creation 2

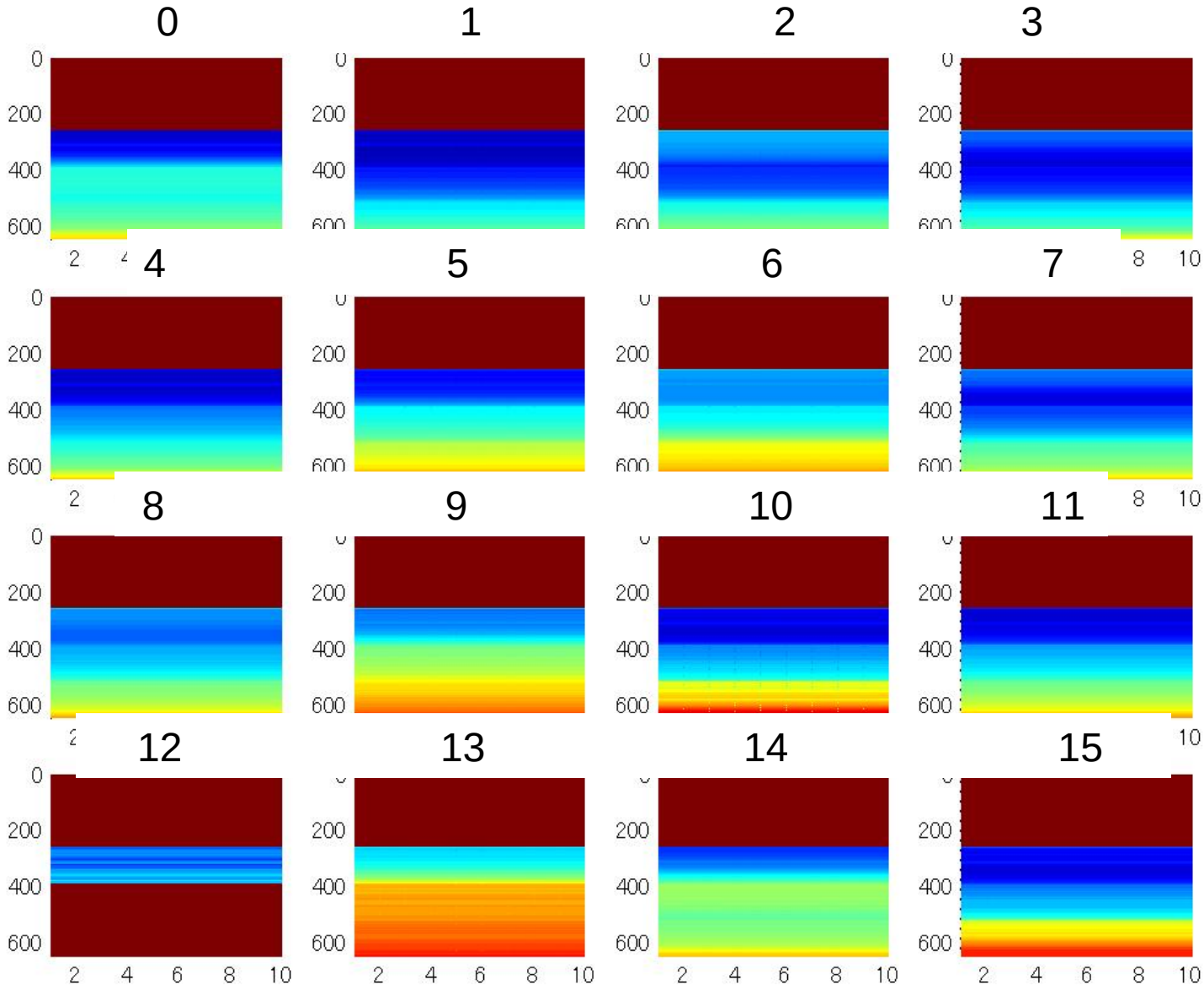


Result of 1st SOM Clustering

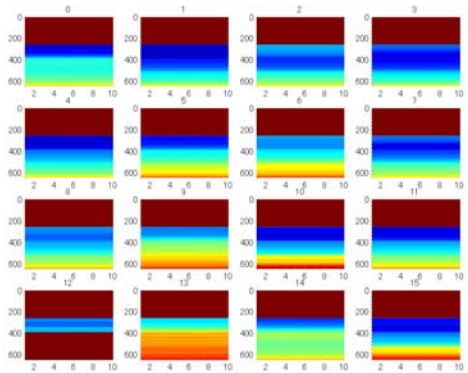


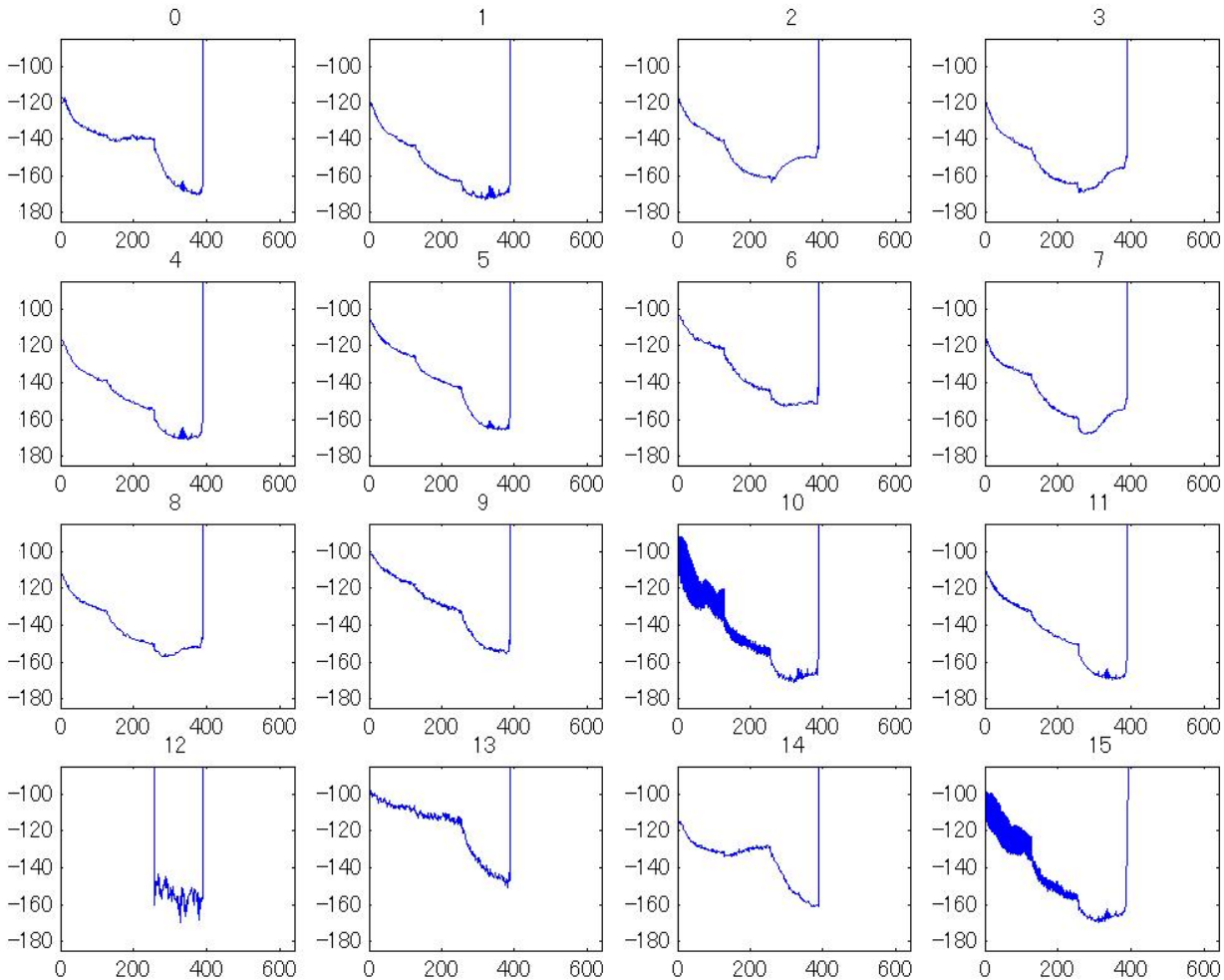


Result of 1st SOM Clustering



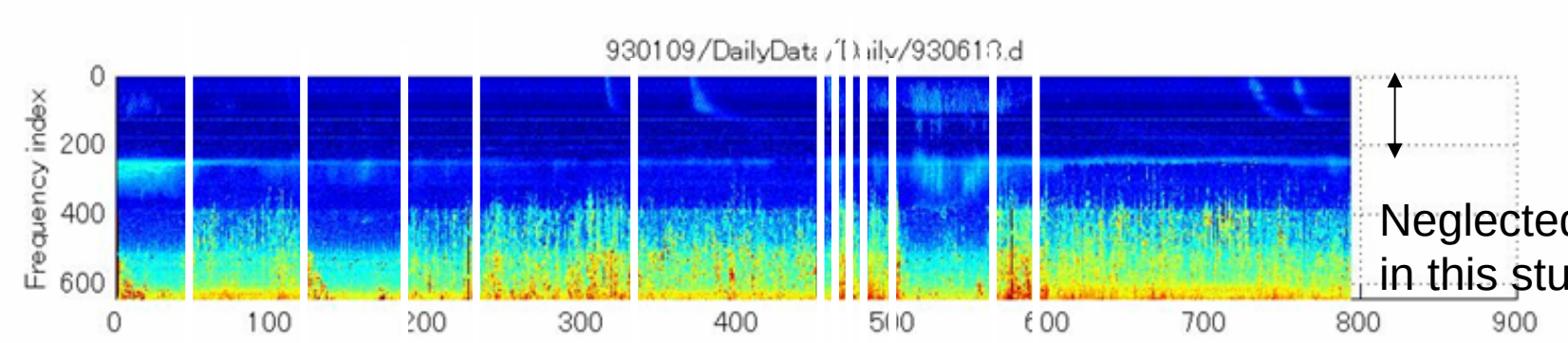
Result of 1st SOM Clustering

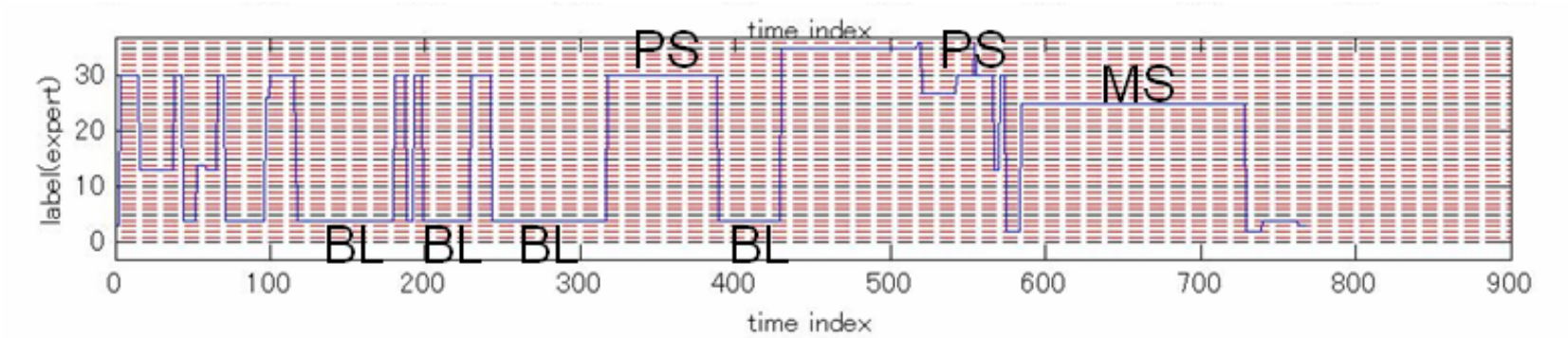




Goal of current study.

Labellng of region by EPIC group

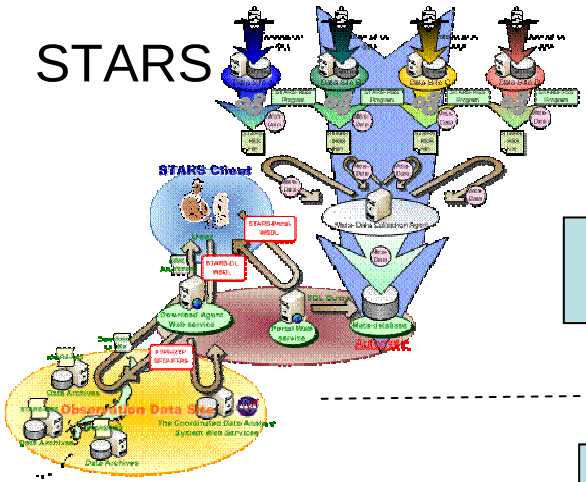




Devide the time series of SFA power spectrum automatically and find the relation between the epoch and the labeling of region by EPIC group.

Overall process

STARS



Current status

Local site

Data collection

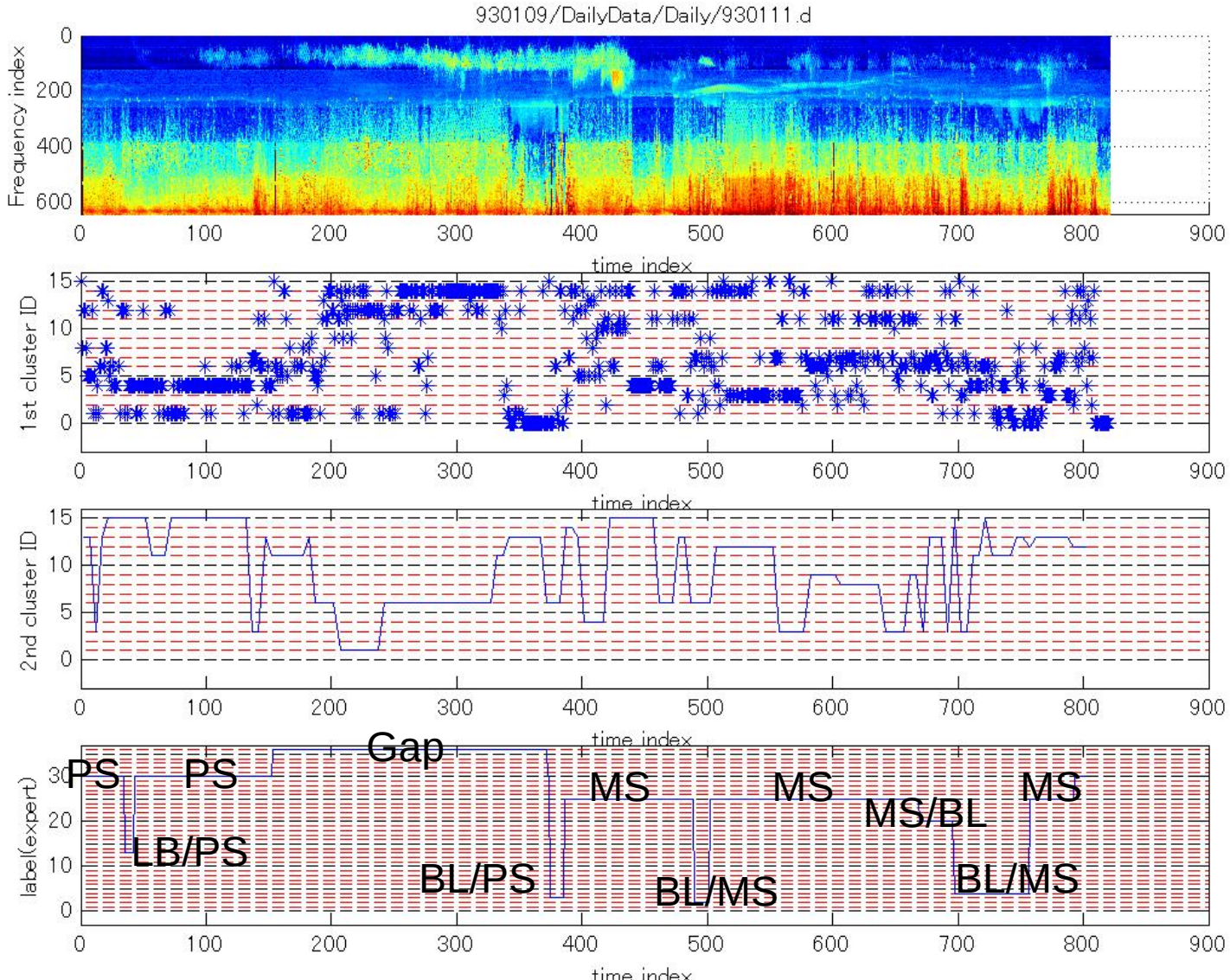
Re-sampling

Mining (clustering)

Evaluation

Knowledge discovery

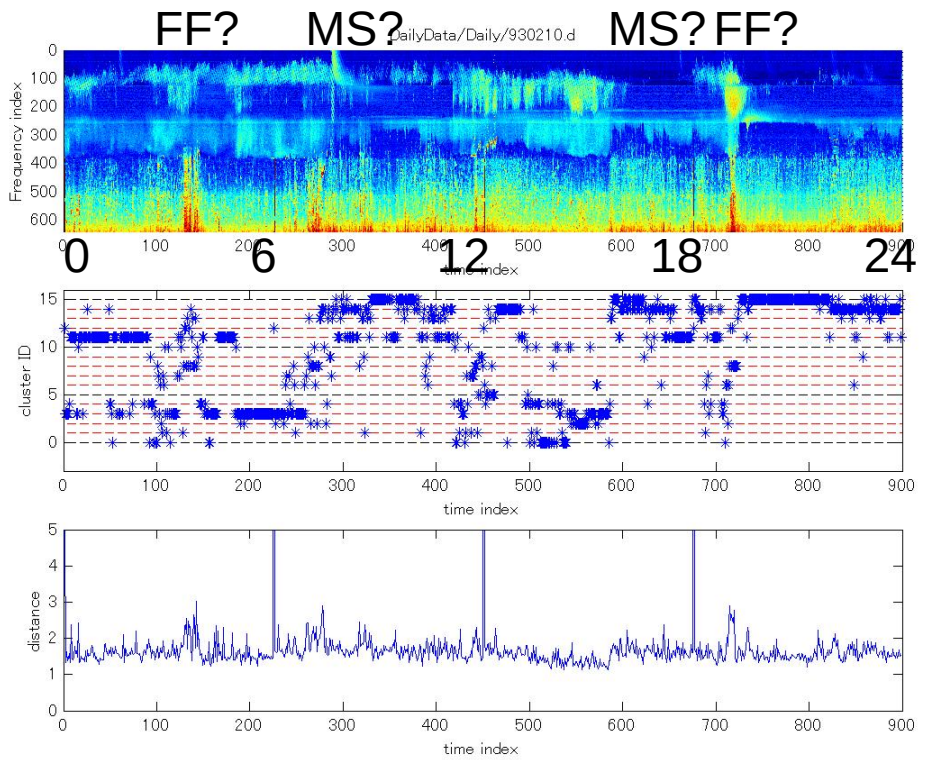
Result 2

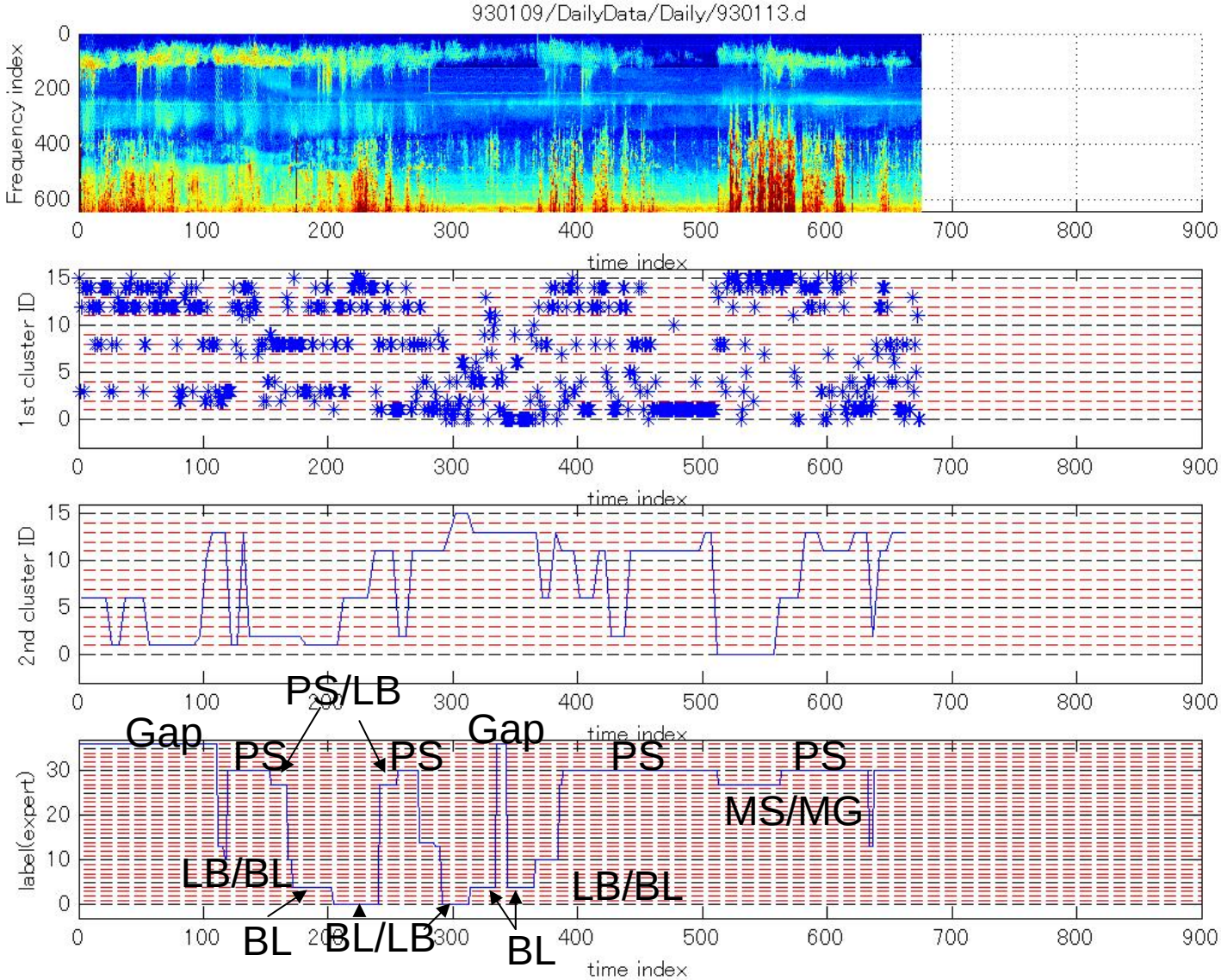

930109/DailyData/Daily/930111.d
Frequency index
time index
1st cluster ID
2nd cluster ID
label(expert)
PS
PS
LB/PS
Gap
BL/PS
MS
BL/MS
MS
MS/BL
MS
BL/MS

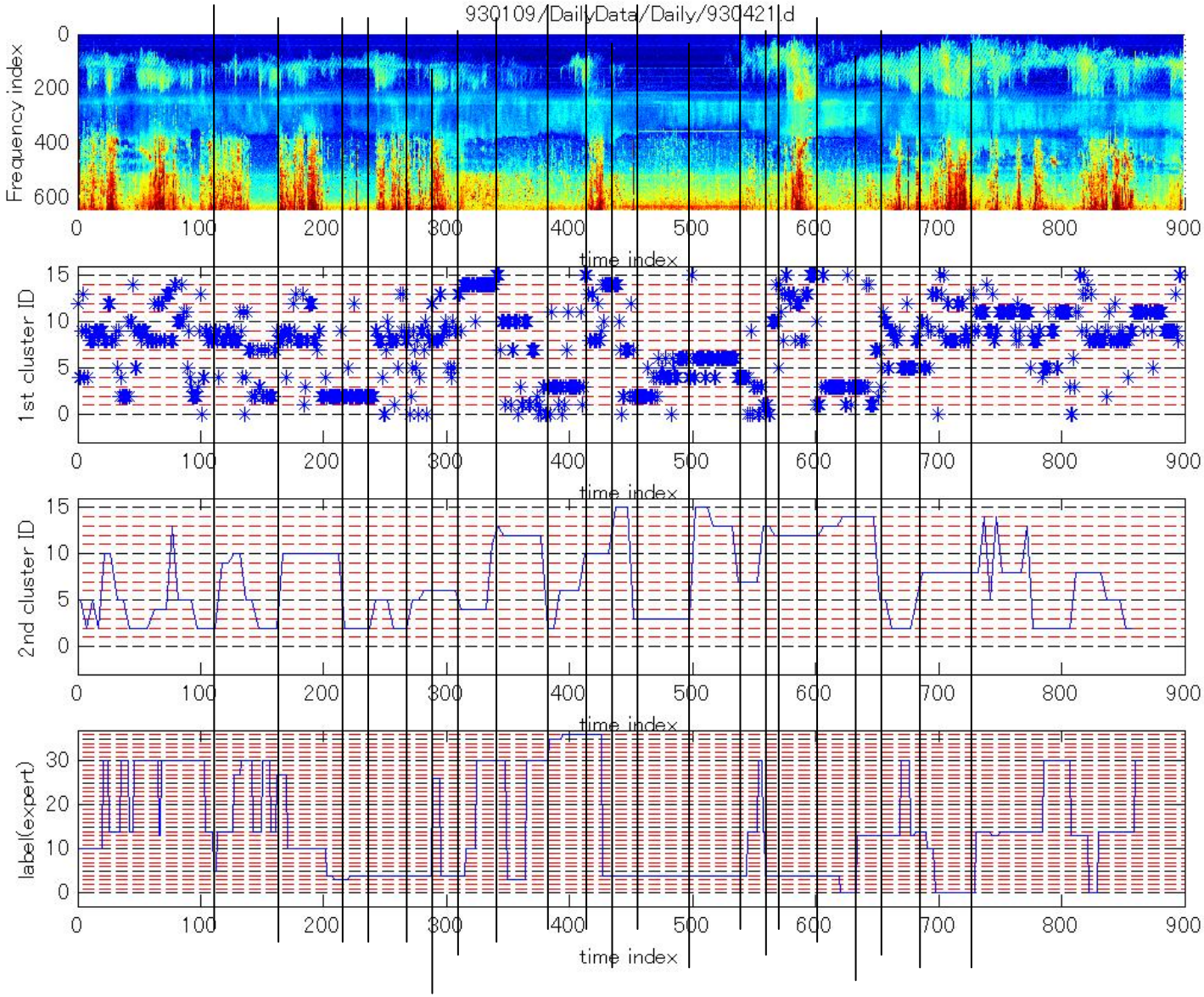
Labeling by EPIC group

93-040 2303 - 93-041 0158 BL = 33
（93-041 0159 - 93-041 PS = 11）
93-041 0210 - 93-041 BL = 33
93-041 0233 - 93-041 PS = 11 Tailward
93-041 0249 - 93-041 BL = 33
93-041 0326 - 93-041 0336 FF = 11 Tailward
93-041 0337 - 93-041 0340 BL = 33
93-041 0341 - 93-041 0348 PS/BL = 13 PSBL
93-041 0349 - 93-041 0652 BL = 33
93-041 0655 - 93-041 0717 PS = 11 Tailward
93-041 0717 - 93-041 0726 LB/BL = 23
93-041 0726 - 93-041 0845 BL/MS = 34
93-041 0845 - 93-041 1010 MS = 44
93-041 1012 - 93-041 1535 BL = 33
93-041 1535 - 93-041 1558 BL/MS = 34
93-041 1558 - 93-041 1720 MS = 44
93-041 1720 - 93-041 1726 BL/MS = 34
93-041 1726 - 93-041 1737 PS = 11
93-041 1739 - 93-041 1820 BL = 33
93-041 1822 - 93-041 1828 PS = 11
93-041 1830 - 93-041 1849 BL = 33
93-041 1850 - 93-041 1857 PS/BL = 13
93-041 1858 - 93-041 1901 BL = 33
93-041 1901 - 93-041 1908 FF = 11 Tailward
93-041 1910 - 93-041 1915 BL = 33
93-041 1915 - 93-041 2135 MS = 44
93-041 2135 - 93-041 2252 BL/MS = 34
93-041 2255 - 93-042 0020 MS = 44

BL Boundary Layer: mantle here; high B(>5g), +Bx(>4
BS Bow Shock [similar to SW] |
Cal Calibration period | d(cs) distance to current sheet
FF Fast Flow event [similar to PS] |
FL Flare, used when there is some flare background |
FR Flow Reversal |
LB Lobe: mostly < 1 keV ions |
MG Magnetosphere: ions > 1-5 keV, 1-10 keV electron
MS Magnetosheath: low B (<5g), -Bx, high dB, EM wa
PS Plasma Sheet: 100 eV+ ions, < ~1 keV electrons (




930109/DailyData/Daily/930113.d
Frequency index
time index
1st cluster ID
time index
2nd cluster ID
time index
label(expert)
time index
Gap
PS
PS/LB
PS
Gap
PS
PS
MS/MG
LB/BL
LB/BL
BL
BL/LB
BL


930109/DailyData/Daily/930421.d
Frequency index
0
200
400
600
0
100
200
300
400
500
600
700
800
900
time index
1st cluster ID
15
10
5
0
0
100
200
300
400
500
600
700
800
900
time index
2nd cluster ID
15
10
5
0
0
100
200
300
400
500
600
700
800
900
time index
label(expert)
30
20
10
0
0
100
200
300
400
500
600
700
800
900
time index

PWI-SFA

- **SFA(Spectral Frequenct a)**
  - **Spectral information on plasma wave amplitudes**

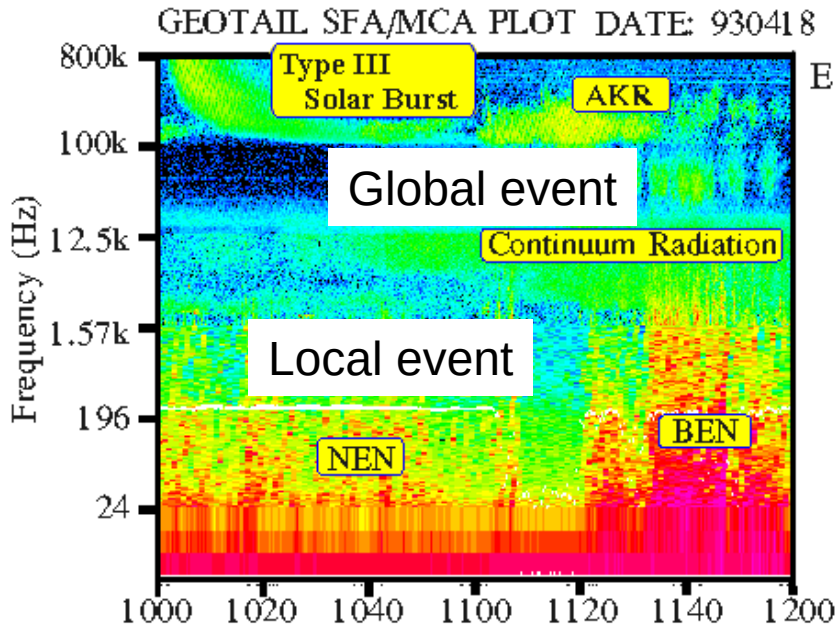


Table 1. Specification of SFA

Band	Frequency Range	Freq. Step	Bandwidth	Source	Sweep
1	24 Hz - 200 Hz	1.3 Hz	2.6 Hz	*B* and *E*	64sec
2	200 Hz - 1600Hz	10.7 Hz	10 Hz	*B*and *E*	64sec
3	1.6 kHz - 12.5 kHz	85.4 Hz	85 Hz	*B* and *E*	8 sec
4	12.5 kHz - 100 kHz	683 Hz	680 Hz	*E* only	8 sec
5	100 kHz - 800 kHz	5.47kHz	5.4 kHz	*E* only	8 sec

References

- 分散強調学習に基づくトピック構造マイニング, 松村, 森永, 山西, **第8回 情報論的学習理論ワークショップ (IBIS 2005)**、2005
- 最新！データマイニング手法:5. 統計的異常検出3手法, 情報処理学会誌, 山西／竹内／丸山, 2005年1月　Vol. 46 No. 1
- 最新！データマイニング手法:2. データスカッシング, 情報処理学会誌, 鈴木, 2005年1月　Vol. 46 No. 1
- **Data Mining and Machine Learning of Time Series Data.** The 14th European Conference on Machine Learning (ECML) and the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD). Dubrovnik, Croatia, 2003 [*http://www.cs.ucr.edu/~eamonn/*]
- Keogh, E. & Folias, T. (2002). The UCR Time Series Data Mining Archive [*http://www.cs.ucr.edu/~eamonn/TSDMA/index.html*].
- Jiawei Han and Micheline Kamber , Data Mining: Concepts and Techniques, 2ed., Morgan Kaufmann Publishers, 2006.
  - [*http://www-faculty.cs.uiuc.edu/~hanj/bk2/index.html*]
- EPICカタログ
  - http://sd-www.jhuapl.edu/Geotail/regime_id.html