



**UNIVERSITÀ
DEGLI STUDI
DI TRIESTE**

**Dipartimento di Scienze
Chimiche e Farmaceutiche**

SOMEnv: un free software tool con Graphical User Interface per l'identificazione di profili ricorrenti di odori tramite elaborazione con algoritmo Self-Organizing Map di dati registrati da IOMS

S. Licen (slicen@units.it)^a, M. Franzon^b, T. Rodani^c, S. Cozzutto^d, P. Barbieri^a

^a Dept. of Chemical and Pharmaceutical Sciences, University of Trieste, Via L. Giorgieri 1, 34127 Trieste, Italy;

^b eXact lab s.r.l., via Beirut, 2 - 34151 Trieste (Italy);

^c AREA Science Park - Padriciano, 99 34149 Trieste, Italy

^d ARCO SolutionS s.r.l., spin-off company of the Dept. of Chemical and Pharmaceutical Sciences, University of Trieste, Via L. Giorgieri 1, 34127 Trieste, Italy



**UNIVERSITÀ
DEGLI STUDI
DI TRIESTE**

Dipartimento di Scienze Chimiche e Farmaceutiche

UNITA' DI RICERCA IN CHIMICA ANALITICA AMBIENTALE

UNIVERSITA' DEGLI STUDI DI TRIESTE



Prof. Gianpiero Adami



Prof. Pierluigi Barbieri



Prof. Matteo Crosera



Dr. Sabina Licen



UNITA' DI RICERCA IN CHIMICA ANALITICA AMBIENTALE

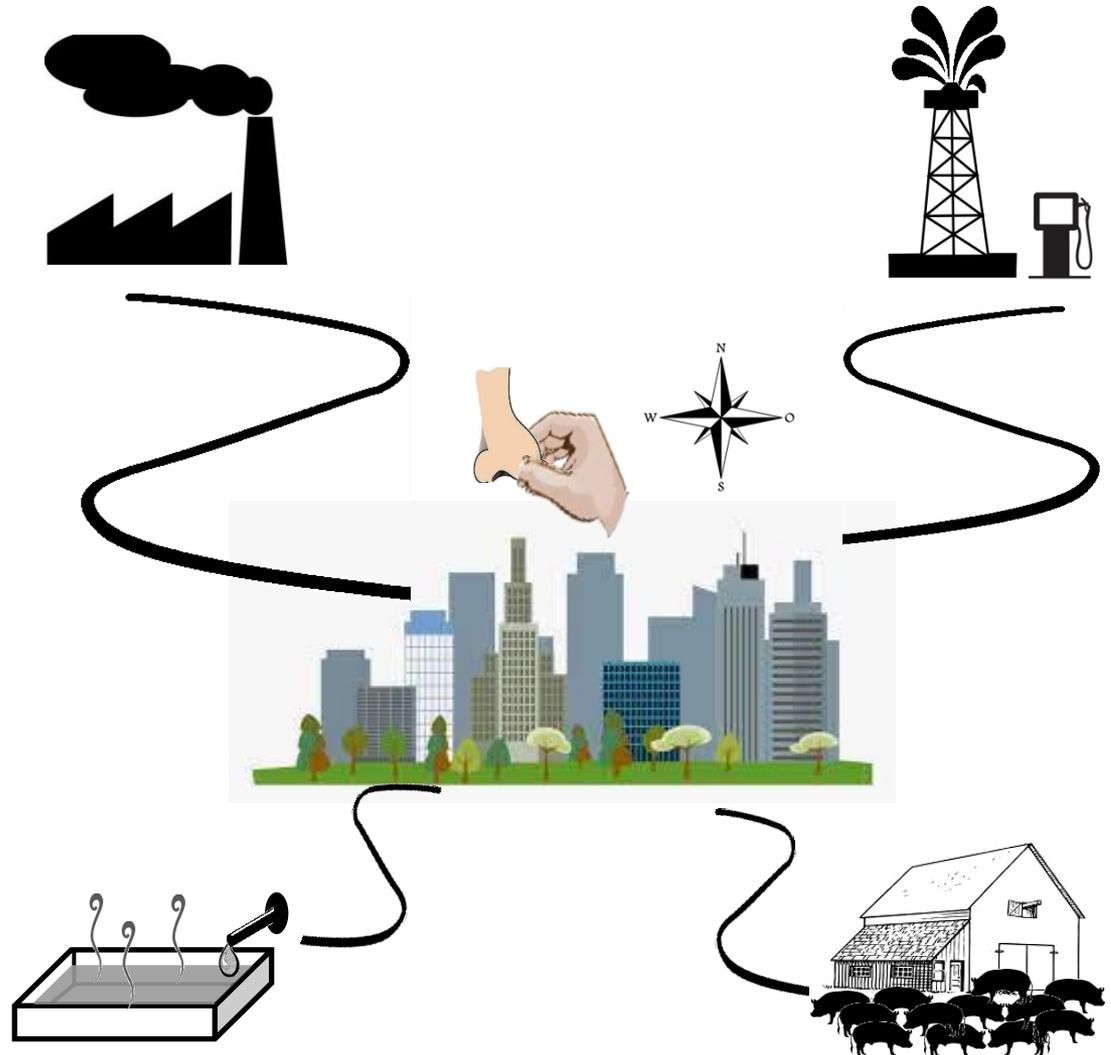
UNIVERSITA' DEGLI STUDI DI TRIESTE

Attività di ricerca principali:

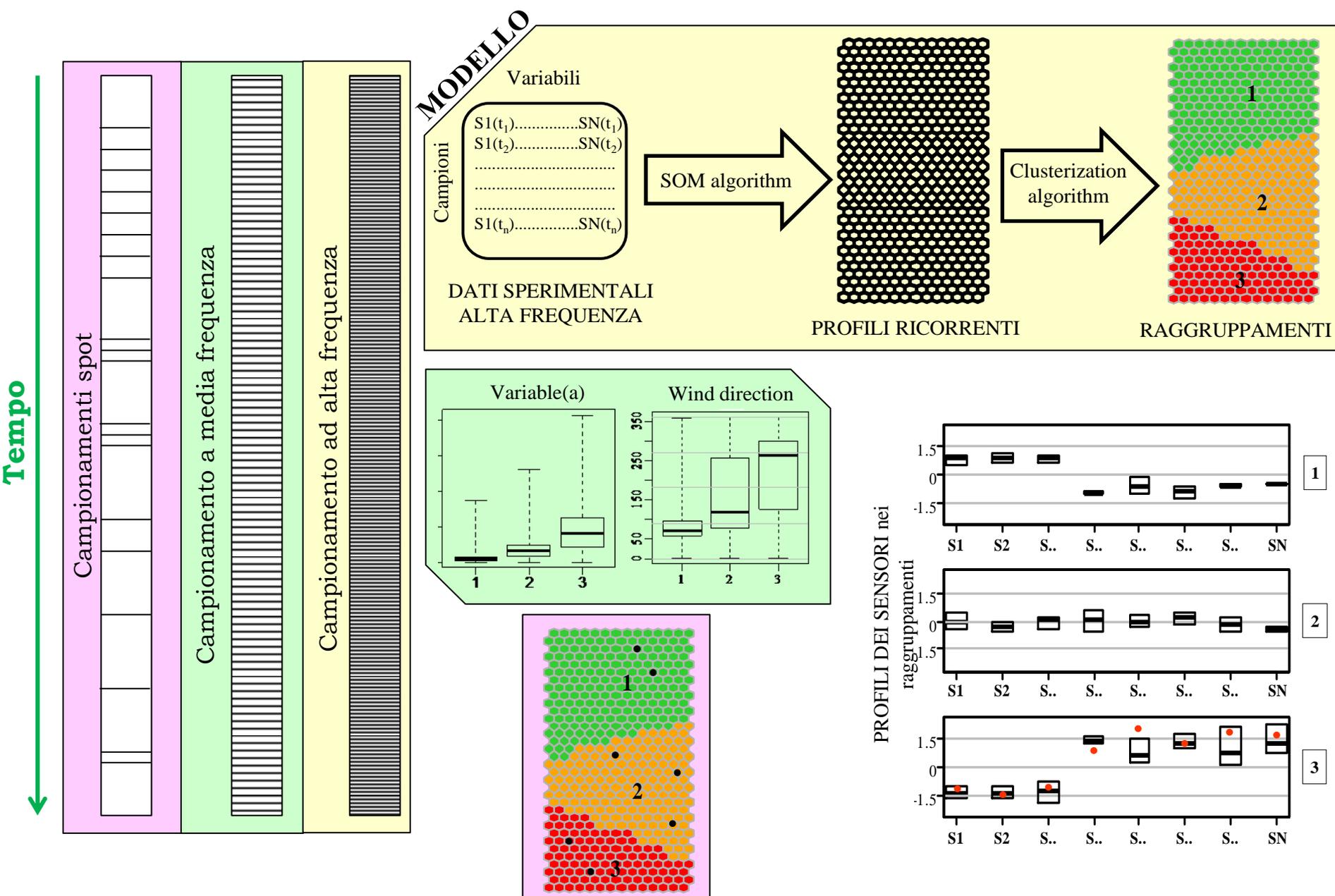
- Caratterizzazione di Composti Organici Volatili nelle valutazioni ambientali, per diagnosi cliniche e qualità della vita;
- Caratterizzazione di aerosol e bioaerosol atmosferici e indoor;
- Sviluppo di metodi innovativi per l'analisi multivariata di dati da monitoraggio ambientale;
- Caratterizzazione di nanoparticelle nell'ambiente e nanotossicologia;
- Caratterizzazione di profili elementali di acque e sedimenti marini.

Individuazione della/e sorgente/i di odore

- Quale/i di questi?
- Quando?
- Per quanto tempo?



Integrazione dei dati disponibili



Algoritmo Self-Organizing Map (SOM)

- Famiglia delle reti neurali artificiali;
- **Tecnica di analisi multivariata NON supervisionata;**
- Basato sulla comparazione di distanze multidimensionali tra vettori di dati (es. distanza euclidea);
- **Il *noise* viene automaticamente ridotto durante il training del modello;**
- Si possono elaborare «**big data**» (es. monitoraggio annuale dati al minuto = 525'600 campioni);
- **I risultati sono visualizzabili ed esplorabili in mappe bidimensionali**

Il modello «**impara**» dai dati sperimentali

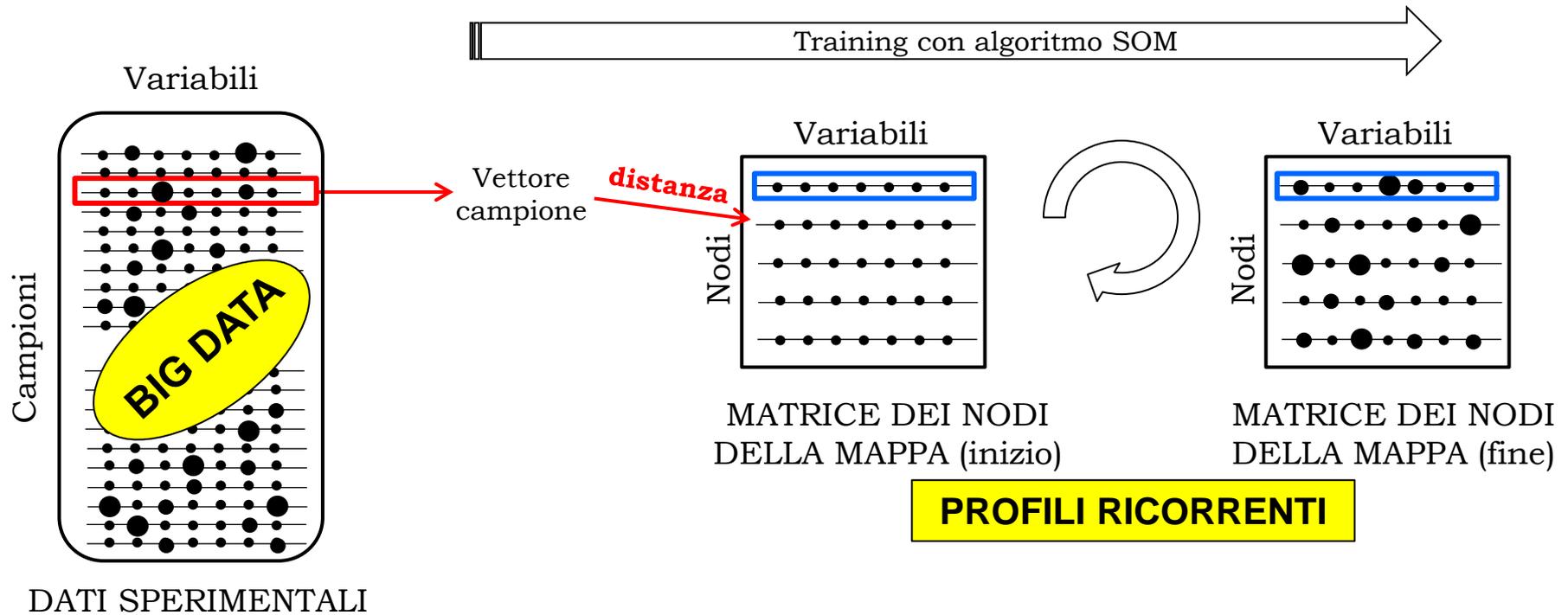
Non necessaria conoscenza a priori sulla **classificazione** dei dati

Può rappresentare anche relazioni **non lineari** tra variabili

Non è necessario pre-processamento per ridurre il rumore

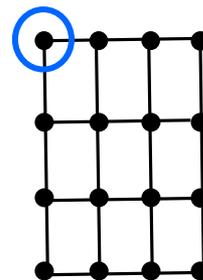


Come funziona l'algoritmo SOM

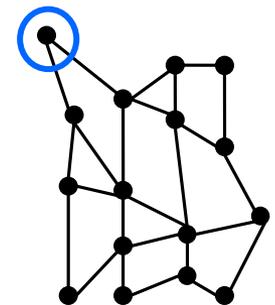
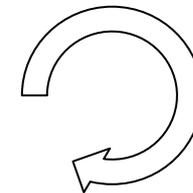


MANTIENE:

- a) Tutte le variabili sperimentali;
- b) La variabilità dei dati sperimentali (eccetto il *noise*).

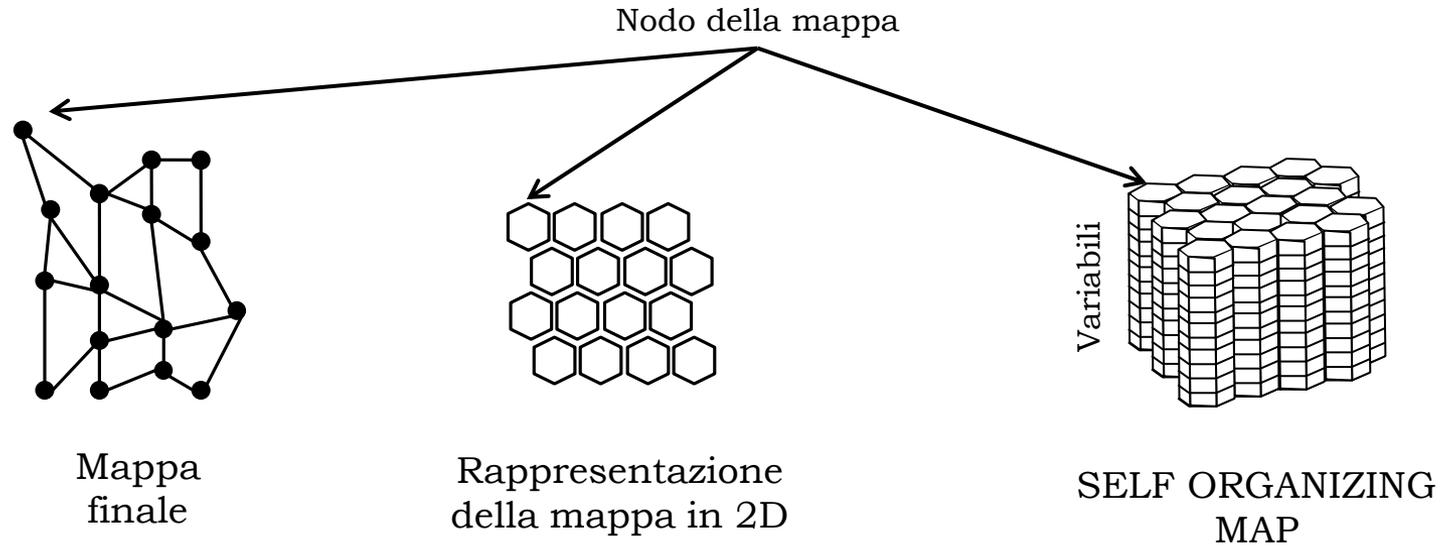


Mappa
inizializzata



Mappa
finale

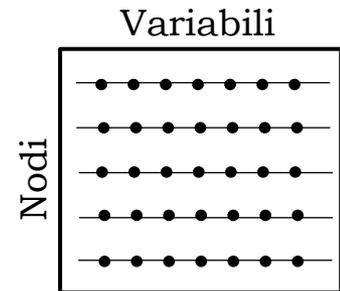
Come funziona l'algoritmo SOM (2)



Inizializzazione della mappa

L'utilizzatore deve scegliere:

- il **numero di nodi** della mappa;
- il **rapporto tra le dimensioni** della mappa (base x altezza);
- i **valori** delle variabili con cui inizializzare la matrice prima del training.



Soluzione:

- regole euristiche di Vesanto, basate su autovalori e autovettori della matrice di dati.



SOMEnv package

- Basato sull'algoritmo SOM presente nel **kohonen package** di Wehrens, Buydens & Kruisselbrink (2007-2018)
- Regole euristiche di **Vesanto** calcolate automaticamente (fonte: SOMtoolbox per MATLAB)
- Interfaccia Grafica (**GUI**) che si apre nel **browser** predefinito;
- Risultati esplorabili attraverso diversi tipi di **grafici**;
- Facilità di **download** di grafici e risultati.
- ***<https://cran.r-project.org/web/packages/SOMEnv/index.html>***

Microchemical Journal 165 (2021) 106181

Contents lists available at [ScienceDirect](#)

 **Microchemical Journal** 

journal homepage: www.elsevier.com/locate/microc

SOMEnv: An R package for mining environmental monitoring datasets by Self-Organizing Map and k-means algorithms with a graphical user interface



Sabina Licen^{a,*}, Marco Franzon^b, Tommaso Rodani^c, Pierluigi Barbieri^a

Special Issue:
“*Chemometrics in
Environmental Analytical
Chemistry*”

SOMEnv package loading

<https://cran.r-project.org/>



> install.packages("SOMEnv")

```
RGui (64-bit)
File Edit View Misc Packages Windows Help

R Console
> library(SOMEnv)
#####
Welcome to 'SOMEnv' package
Find citation details with citation("SOMEnv")
Use SomEnvGUI() to start the Graphical User Interface in the default browser
For further information see the scrolling Help in the Graphical User Interface
#####
> SomEnvGUI() |
```

SOMEnv Graphical User Interface (v 1.1.1)



An R package for mining information from multivariate environmental high frequency data by Self-Organizing Map and k-means clustering algorithms

Authors: S. Licen, M. Franzon, T. Rodani, P. Barbieri

Licence: GPL-3.0

Large datasets up to 100 MB are allowed!

Help

Tab 1: Load Data

Use Browse button to select the data file. Data are imported in the GUI using the `import` function from `openair` package. The data must be in table format in a txt file with date variable (or datetime variable) in the first column (the header must be 'date') and numeric variables in the following columns. The user has to write the correct datetime format according to the instructions for `strptime` function used in import function. Column and decimal separator have to be selected as well. Press Load button to load the data. When loaded, the first six rows (header) and the last six rows (tail) appear in the lower part of the screen thus the user can check if the data have been uploaded correctly. Under the above mentioned tables the number of total uploaded rows as well as the number of deleted rows (containing NA values) are presented. If the date format is not correct and/or one or more columns contain non numeric values some warnings appear. When the data are correctly loaded the user can move to the next tab.

Tab 2: SOM training

The SOM training is based on the use of `som` function from `kohonen` package. The parameter selection for training the SOM

Load data | SOM training | SOM Map | Kmeans clustering | Projection | Daily profiles

Load experimental data:

Browse... No file selected

Date format: %Y-%m-%d %H:%M:%S

Separator:

Decimal:

Load

Experimental data header

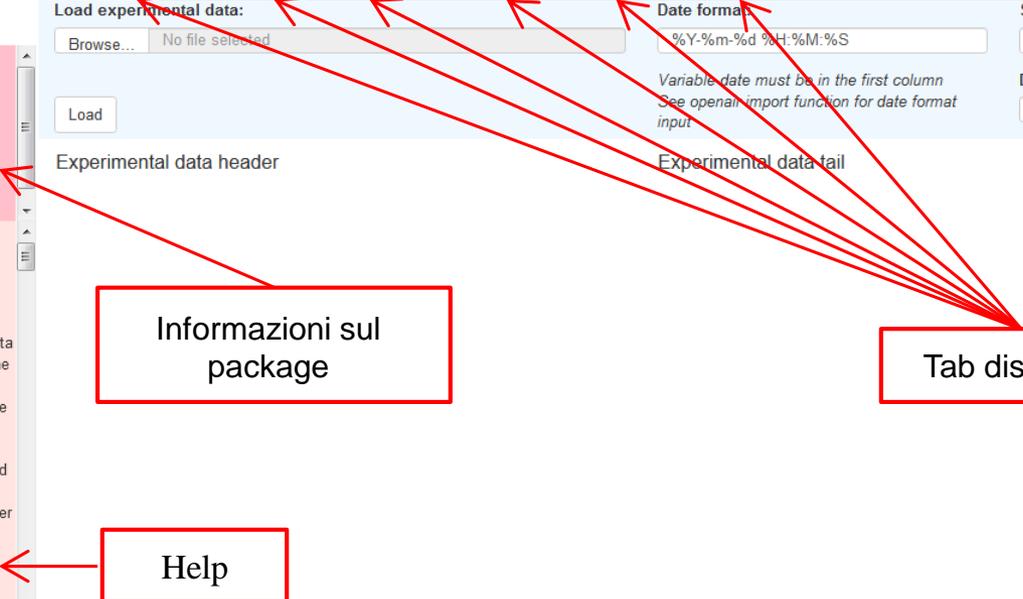
Experimental data tail

Variable date must be in the first column
See `openair::import` function for date format input.

Informazioni sul package

Tab disponibili

Help



SOMEnv package: Data loading Tab



An R package dedicated to the analysis of multivariate environmental high frequency data by Self-Organizing Map and k-means clustering algorithms

Large datasets up to 100 MB are allowed!

Authors: S. Licen, M. Franzon, T. Rodani, P. Barbieri

Licence: GPL-2.0

Help

Tab 1: Load data Use Browse button to select the data file. Data are imported in the GUI using the `import` function from `openair` package. The data must be in table format in a txt file with date variable (or datetime variable) in the first column (the header must be 'date') and numeric variables in the following columns. The user has to write the correct datetime format according to the instructions for `strptime` function used in `import` function. Column and decimal separator have to be selected as well. Press Load button to load the data. When loaded, the first six rows (header) and the last six rows (tail) appear in the lower part of the screen thus the user can check if the data have been uploaded correctly. Under the above mentioned tables the number of total uploaded rows as well as the number of deleted rows (containing NA values) are presented. If the date format is not correct and/or one or more columns contain non numeric values some warnings appear. When the data are correctly loaded the user can move to the next tab.

Tab 2: SOM training The SOM training is based on the use of `som` function from `kohonen` package. The parameter selection for training the SOM has not default values but some heuristic rules by Vesanto et al. (2000) are widely used, these rules have

[Load data](#) [SOM training](#) [SOM Map](#) [Kmeans clustering](#) [Projection](#) [Daily profiles](#)

Load experimental data:

Browse... OPC_site_A_selection.txt

Upload complete

Load

Date format: %Y-%m-%d %H:%M:%S

Separator: .

Decimal: .

*Variable date must be in the first column
See openair import function for date format input*

Experimental data header

date	PM03	PM05	PM07	PM1	PM2	PM3	PM5	PM
2015-07-01 00:00:00	34087	1319	354	184	90	21	0	
2015-07-01 00:01:00	34478	1357	347	196	85	21	3	
2015-07-01 00:02:00	35346	1406	362	199	100	25	5	
2015-07-01 00:03:00	36581	1560	401	232	117	20	3	
2015-07-01 00:04:00	34928	1380	421	266	137	33	9	
2015-07-01 00:05:00	34974	1434	423	265	131	25	5	

Experimental data tail

date	PM03	PM05	PM07	PM1	PM2	PM3	PM5	PM
2015-07-16 23:54:00	65395	3421	1285	825	418	98	19	
2015-07-16 23:55:00	64951	3420	1283	789	415	90	21	
2015-07-16 23:56:00	66134	3562	1246	759	387	97	23	
2015-07-16 23:57:00	66021	3541	1292	795	414	76	22	
2015-07-16 23:58:00	64728	3376	1164	758	401	88	14	
2015-07-16 23:59:00	62954	3122	1034	639	324	62	12	

Number of uploaded rows = 22958

Number of deleted NA rows = 0

SOMEnv package: SOM training Tab



An R package dedicated to the analysis of multivariate environmental high frequency data by Self-Organizing Map and k-means clustering algorithms

Large datasets up to 100 MB are allowed!

Authors: S. Licen, M. Franzon, T. Rodani, P. Barbieri

Licence: GPL-2.0

Help

Tab 1: Load data Use Browse button to select the data file. Data are imported in the GUI using the `import` function from `openair` package. The data must be in table format in a txt file with date variable (or datetime variable) in the first column (the header must be 'date') and numeric variables in the following columns. The user has to write the correct datetime format according to the instructions for `strptime` function used in import function. Column and decimal separator have to be selected as well. Press Load button to load the data. When loaded, the first six rows (header) and the last six rows (tail) appear in the lower part of the screen thus the user can check if the data have been uploaded correctly. Under the above mentioned tables the number of total uploaded rows as well as the number of deleted rows (containing NA values) are presented. If the date format is not correct and/or one or more columns contain non numeric values some warnings appear. When the data are correctly loaded the user can move to the next tab.

Tab 2: SOM training The SOM training is based on the use of `som` function from `kohonen` package. The parameter selection for training the SOM has not default values but some heuristic rules by Vesanto et al. (2000) are widely used, these rules have

Load data

SOM training

SOM Map

Kmeans clustering

Projection

Daily profiles

Train SOM

Number of SOM map rows:

24

Map size:

small

Number of epochs:

2

Neighborhood:

gaussian

Number of SOM map cols:

8

Number of units = 192

Neigh.radius:

3,1

Codebook header

Unit	PM03	PM05	PM07	PM1	PM2	PM3	F
1	91694.60	8839.90	3379.05	2152.53	1065.25	220.85	46
2	88336.39	7987.58	3039.04	1935.15	958.15	199.58	42
3	83114.48	6677.72	2508.60	1592.94	787.79	164.79	36
4	76982.80	5256.82	1927.63	1216.01	598.77	125.39	27
5	71774.67	4233.34	1505.63	941.10	460.06	95.93	20
6	67912.13	3657.40	1266.00	784.61	380.81	78.80	17

Summary

SOM of size 24x8
Number of samples used for training: 22958
Number of prototypes(units): 192
Topology: hexagonal
Neighbourhood function: gaussian
Distance measure used: Euclidean
Learning algorithm: pbatch
Number of training epochs: 2

Download SOM output

Map quality parameters

Evaluate QE

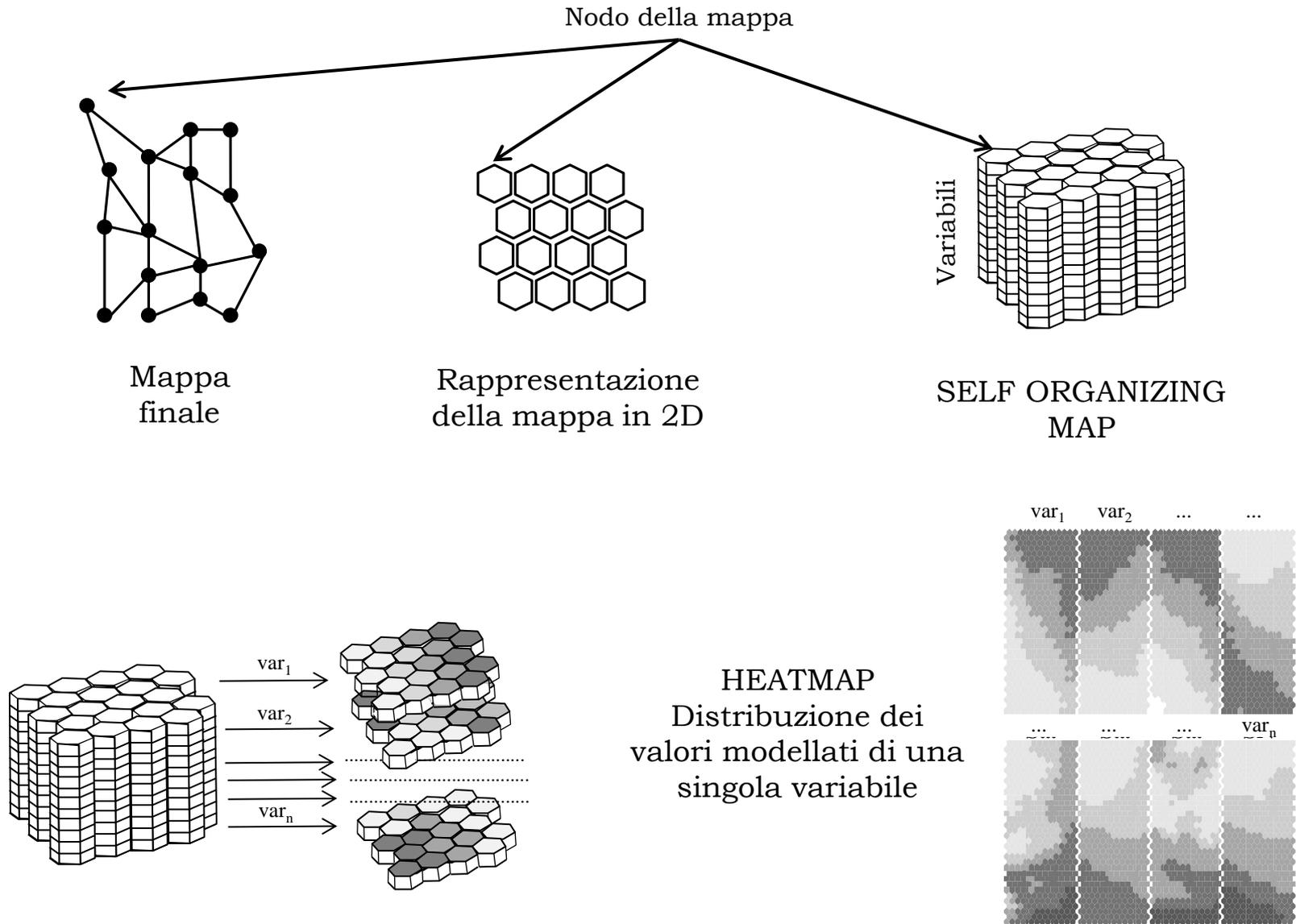
0.691985

Evaluate TE

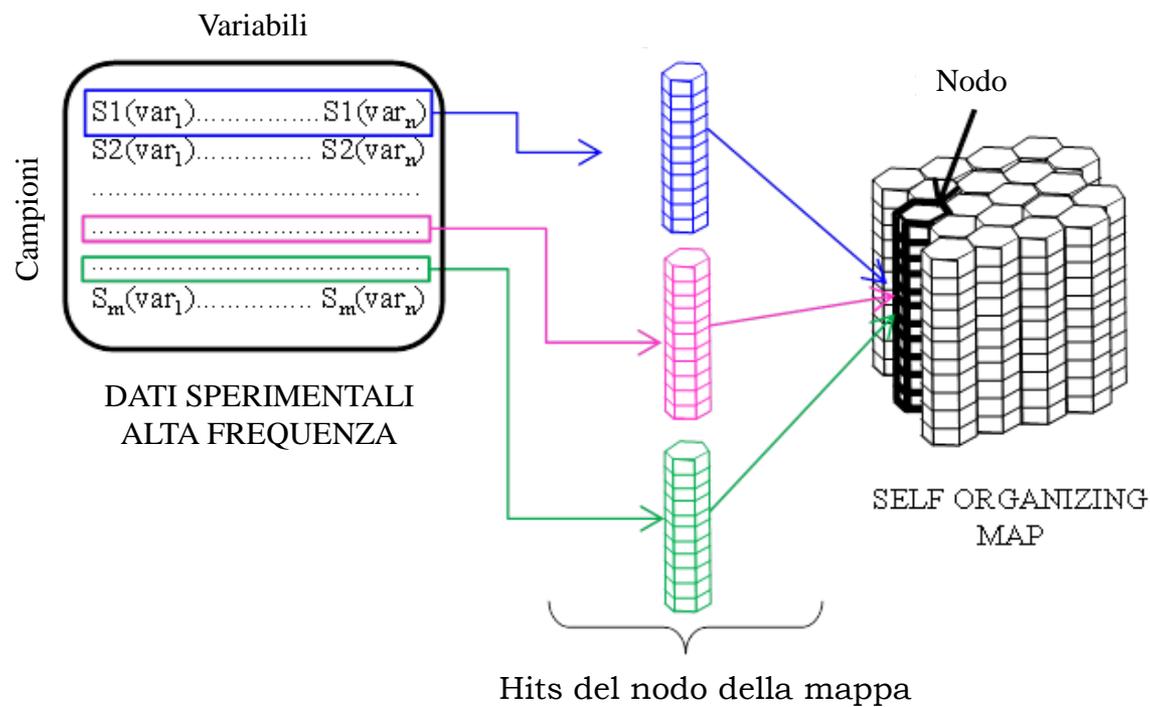
0.0453

(The TE calculation can take several minutes)

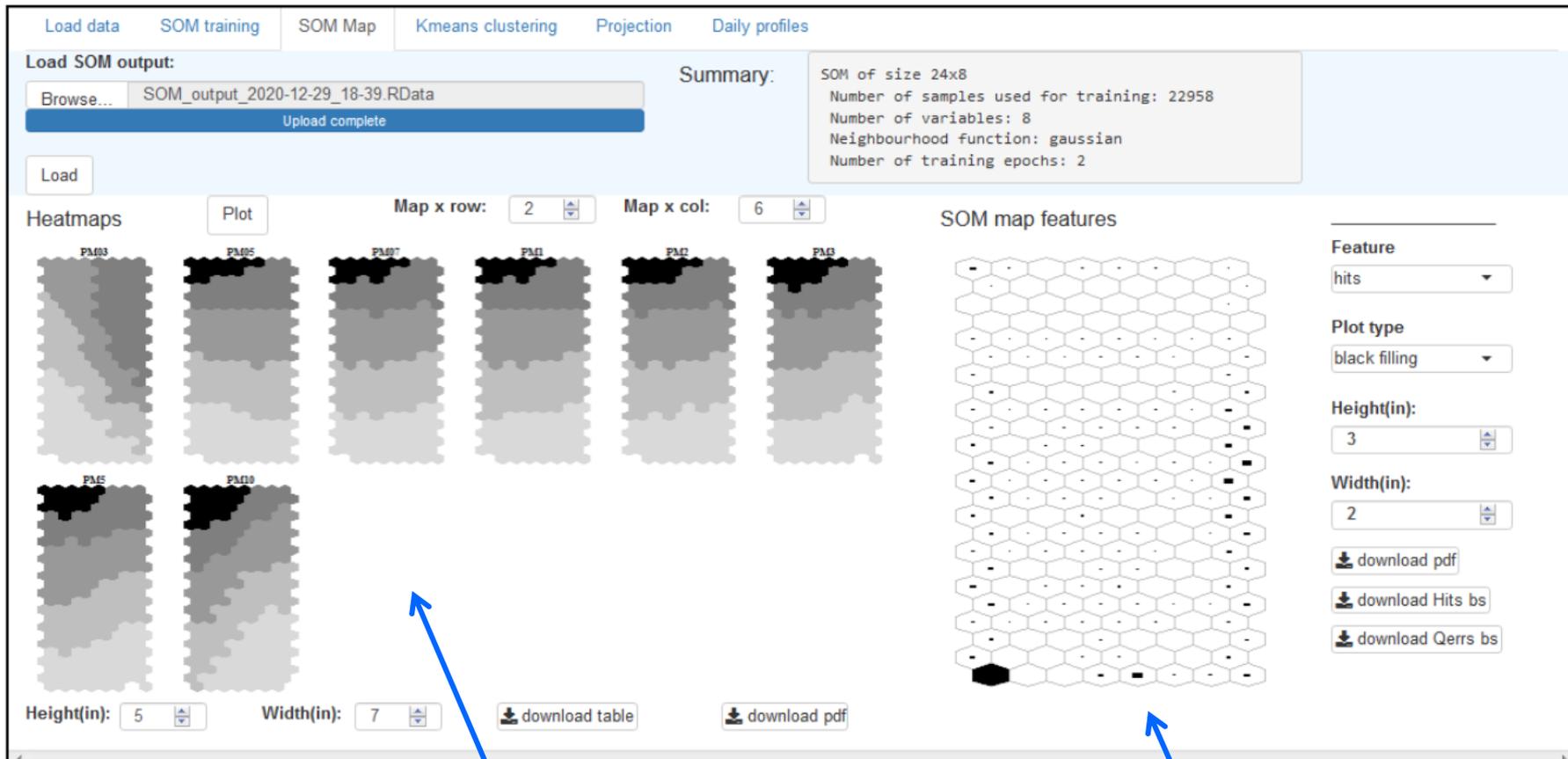
Relazioni tra variabili: esplorare le heatmaps



Quanti dati sperimentali sono rappresentati da ogni profilo ricorrente? Hits



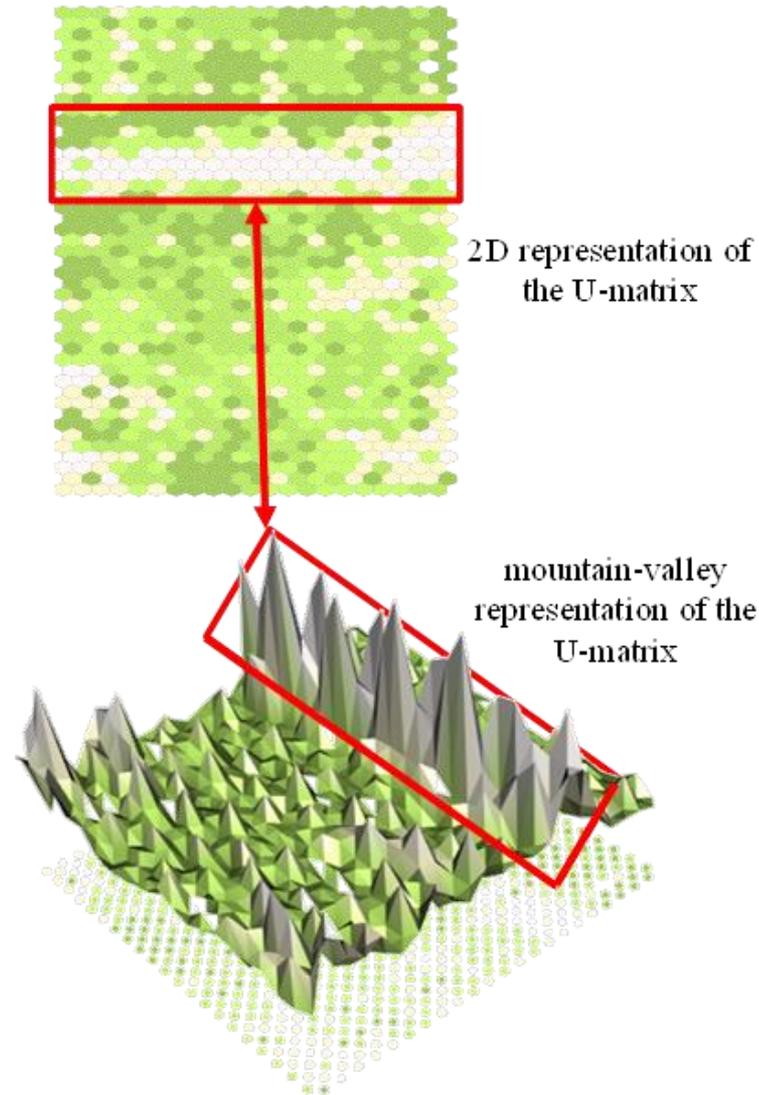
SOMEnv package: SOM map (visualization) Tab



Heatmaps
(scala di grigi da bianco a nero
secondo i quartili)

Hits

Quanto sono diversi tra loro i profili ricorrenti? Unified Distance Matrix (U-matrix)



SOMEnv package: SOM map (visualization) Tab

Unified distance matrix

Load data SOM training **SOM Map** Kmeans clustering Projection Daily profiles

Load SOM output:
Browse...
Upload complete

Load

Summary:
SOM of size 24x8
Number of samples used for training: 22958
Number of variables: 8
Neighbourhood function: gaussian
Number of training epochs: 2

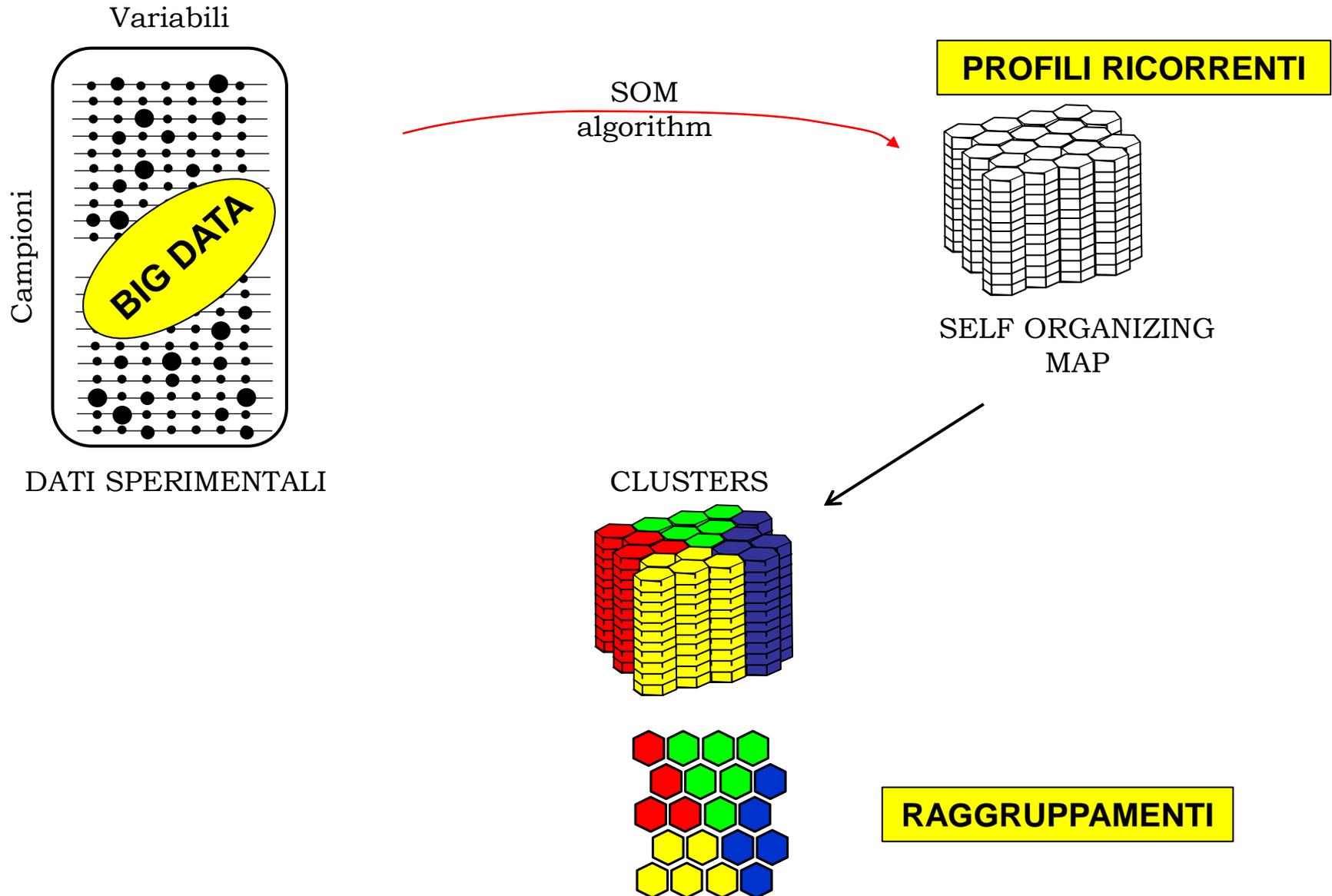
Heatmaps Plot Map x row: 2 Map x col: 6

SOM map features

Feature: umat
Plot type: gs
Height(in): 3
Width(in): 2
download pdf
download Hits bs
download Qerrs bs

Height(in): 5 Width(in): 7 download table download pdf

Raggruppamento di profili ricorrenti: k-means clustering



SOMEnv package: K-means clustering Tab

K-means clustering training

Profili delle variabili modellate nei raggruppamenti



Davis Bouldin index: indice di migliore raggruppamento

SOMEnv package: K-means clustering Tab

Load data SOM training SOM Map Kmeans clustering Projection Daily profiles

Max clusters:

8

Load file?

Cluster numbers:

3,4,2,1

Iterating times (x 100 epochs):

3

(Remember to load SOM output in SOM map tab!)

N.clusters:

4

Cl 1



Cl 2



Cl 3



Cl 4



Plot

Run

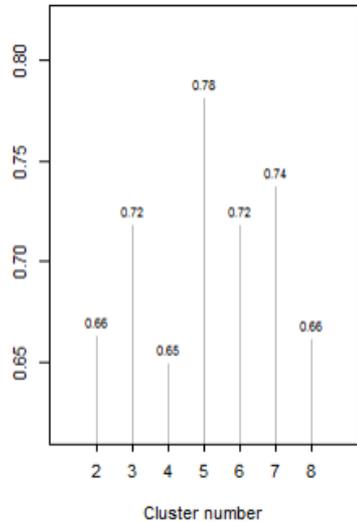
Download Kmeans output

Set seed?

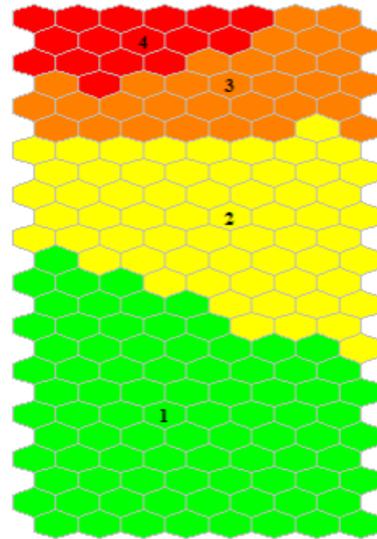
Yes

No

DB-index plot

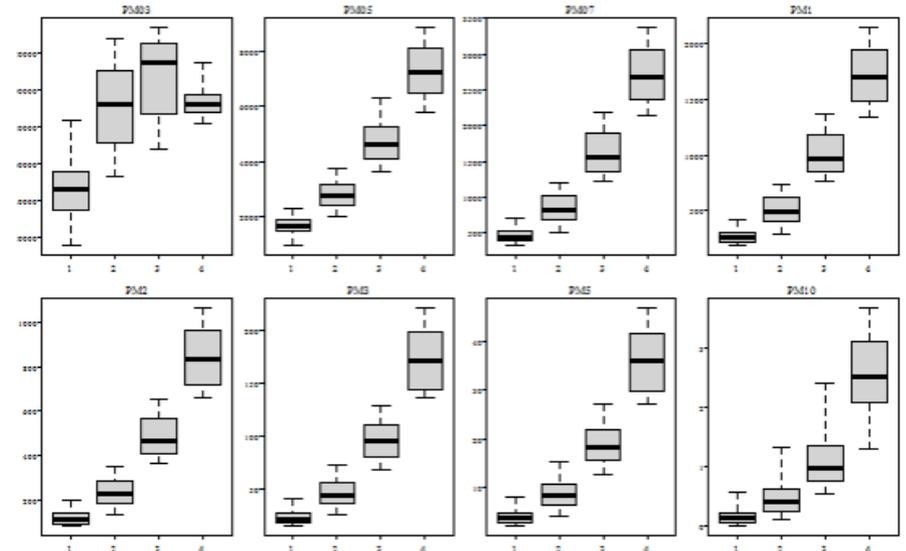


SOM map



Cluster profiles

by variable



H(in):

5

W(in):

7

pdf

pdf

Map x row:

2

Map x col:

4

pdf

SOMEnv package: daily profiles Tab

Load data SOM training SOM Map Kmeans clustering Projection Daily profiles

Dataset

training

Starting date:

01/07/2015

Select sample:

2015-07-07 00:01:00

Units cluster assignment

Data cluster assignment

Obs/Day

1440

Ending date:

16/07/2015

BMU = 1

Cluster = 4

Qe = 8.751

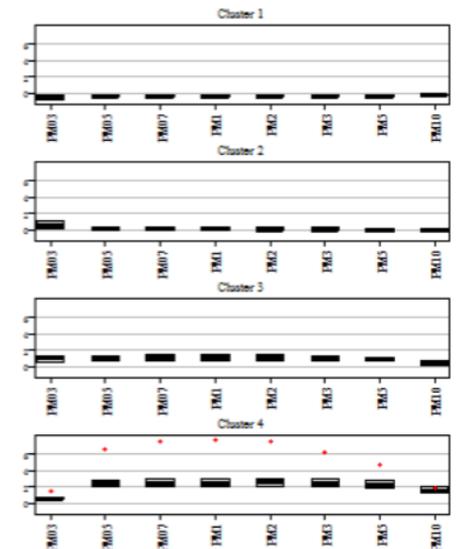
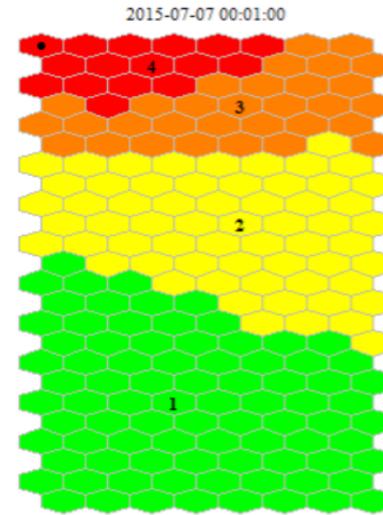
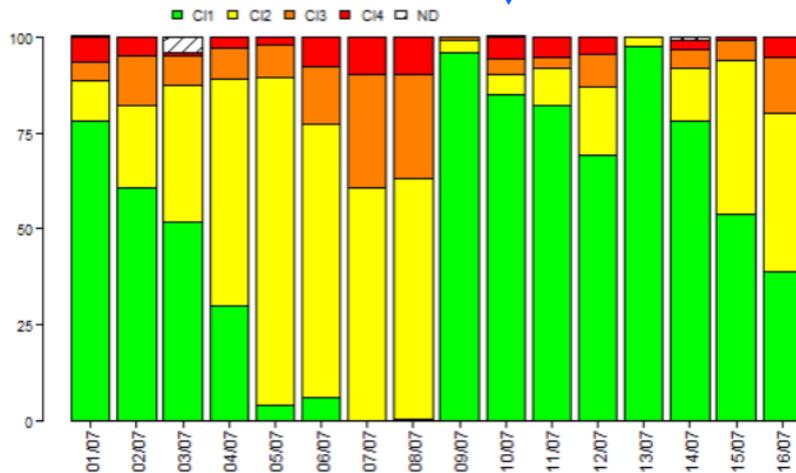
% giornaliera di presenza di ogni raggruppamento

Daily graph

Xlab dim:

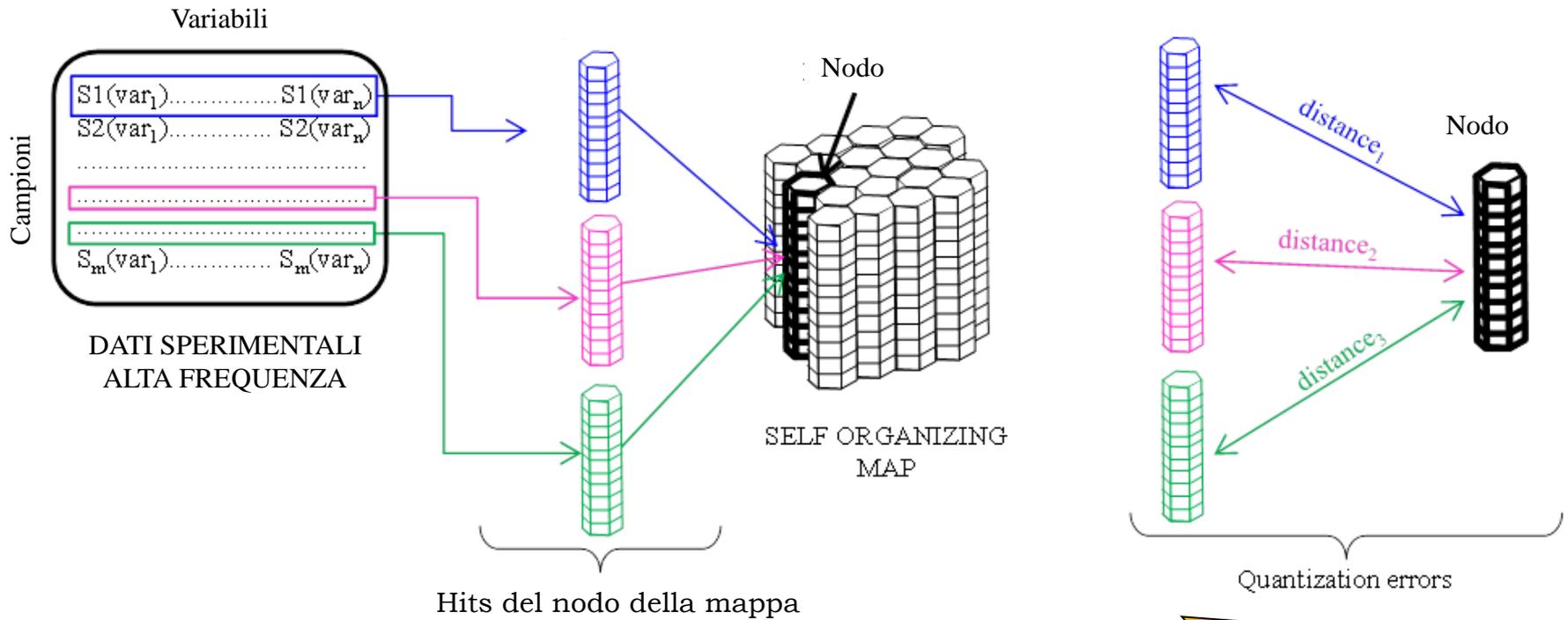
1

Sample profile



H(in): 5 W(in): 7 pdf table overall tab monthly tab pdf

Individuare outliers: Quantization Errors



SOMEnv package: daily profiles Tab

esplorazione del singolo dato
(vettore campione) registrato



SOMEnv package: projection Tab

Proiettare dati esterni nel modello:

- dati registrati in tempi successivi;
- dati rilevati alle sorgenti

Load data SOM training SOM Map Kmeans clustering **Projection** Daily profiles

Load dataset:

Browse... OPC_site_B_selection.txt
Upload complete

Load *(Remember to load SOM output in SOM map tab!)*

Number of uploaded rows = 23034
Number of deleted NA rows = 0

Date format: %Y-%m-%d %H:%M:%S

Variable date must be in the first column
See openair import function for date format input

Separator: ,

Decimal: .

Dataset header

date	PM03	PM05	PM07
2015-07-01 00:00:00	55284	2653	841
2015-07-01 00:01:00	54863	2672	862
2015-07-01 00:02:00	54486	2743	913
2015-07-01 00:03:00	54531	2878	1032

Dataset tail

date	PM03	PM05	PM07
2015-07-16 23:54:00	88856	5771	1993
2015-07-16 23:55:00	88956	5638	1954
2015-07-16 23:56:00	90578	5338	1569
2015-07-16 23:57:00	98076	6214	2058

Projection

Summary

Projection of 23034 samples on SOM of size 24x8
Topology: hexagonal
Neighbourhood function: gaussian
Distance measure used: Euclidean
Number of training epochs: 2

Download Projection

Load projection?

SOM (projection)



Feature: hits

Plot type: grayscale

Height(in): 3

Width(in): 2

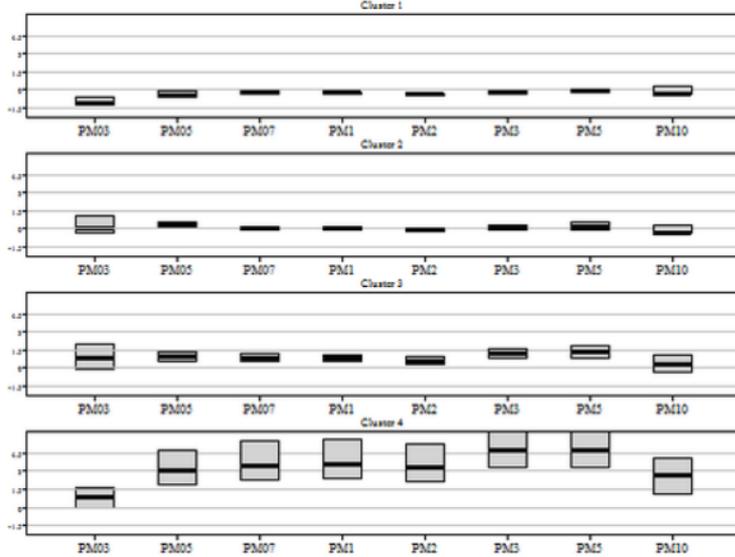
download pdf

download Hits bs

download Qerrs bs

Clusters (projection)

Proj by cluster



Cluster 1

Cluster 2

Cluster 3

Cluster 4

download pdf

Conclusioni

SOMEnv package per R software è un pacchetto **free** che consente di individuare:

- (i) profili ricorrenti di odori;
- (ii) profilo dell'aria di background;
- (iii) frequenze della presenza dei diversi profili di odori;
- (iv) possibili outlier o sensor fault;

e di comparare:

- (i) i profili ricorrenti con quelli delle sorgenti;
- (ii) i raggruppamenti con dati esterni, es. segnalazioni, campionamenti puntuali, dati registrati da altri strumenti.

**GRAZIE PER
L'ATTENZIONE**

slicen@units.it

<https://www.researchgate.net/profile/Sabina-Licen-3>