UNIL | Université de Lausanne

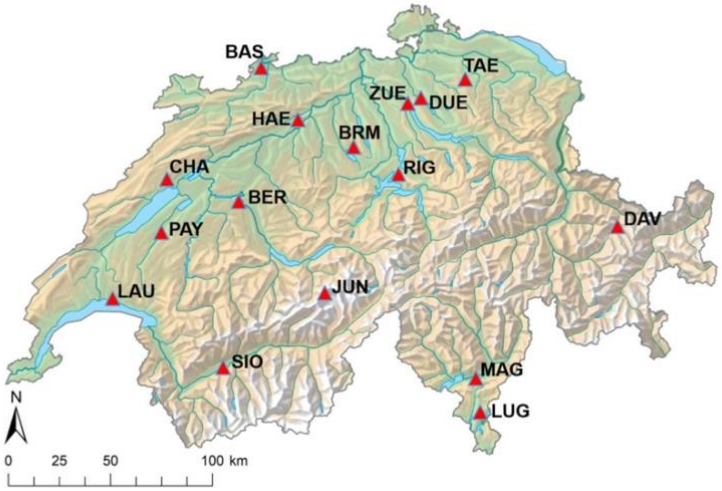
Advanced Exploratory Analysis of Air Pollution Multivariate Spatio-Temporal Data

M. Kanevski, F. Amato, F. Guignard

Institute of Earth Surface Dynamics, University of Lausanne

Mikhail.Kanevski@unil.ch

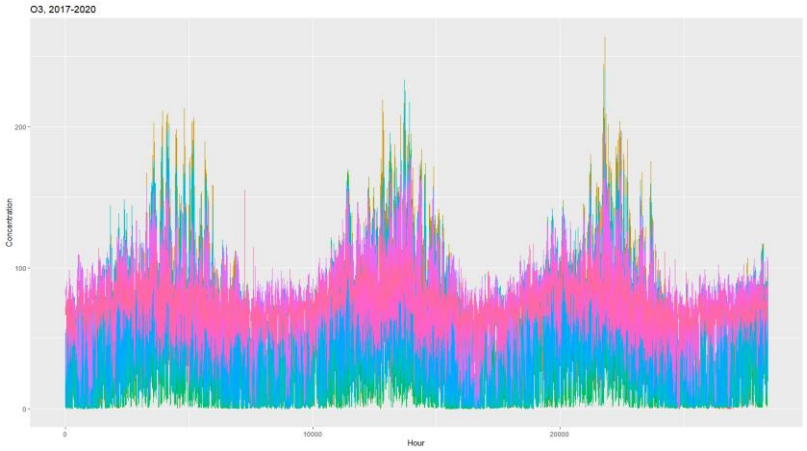
Objectives

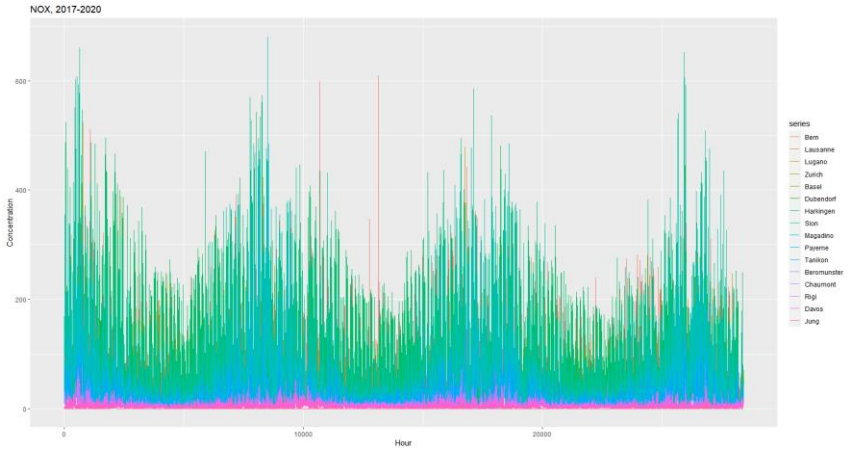


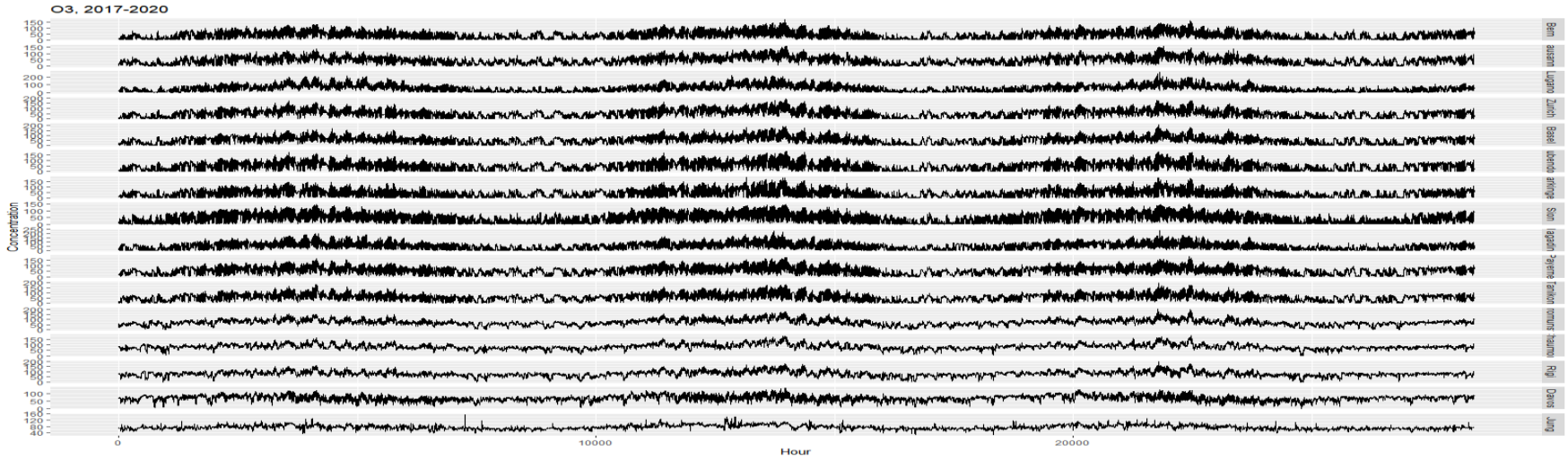
The main objective of the study is to perform an advanced exploratory analysis of air pollution multivariate data. Monitoring stations represent different regions and land use zones (urban, rural, etc.)

The main tools applied include: analysis of statistical and fractal properties of the series using Morisita index , data representation via features and time series clustering.

Examples of time series for all stations (hourly data from 2017, NABEL Swiss monitoring network- 16 stations): O3 (left, below), NOX (right). Typical properties: complex dynamics, periodicities, outliers, missing values..







Morisita index

- There are many measures used to quantify the complexity of time series: entropy, (multi)fractals, time series features, etc. In the present research the Morisita index (MI) is introduced to time series. Originally, index was proposed to study clustering of spatial data. Later, it was generalised to the multiple point index and it was shown that MI can be efficiently used to estimate intrinsic dimension [1] of data and to study feature selection problems in machine learning [2].

1. A new estimator of intrinsic dimension based on the multipoint Morisita index Pattern Recognition, v.48, 2015, p. 4070, J. Golay, M. Kanevski

2. Feature selection for regression problems based on the Morisita estimator of intrinsic dimension Pattern Recognition, v. 70, p. 126, J. Golay, M. Leuenberger, M. Kanevski

Intrinsic dimension

- If we consider measurements of air pollution as points embedded into 16 dimensional Euclidean space, we can estimate an intrinsic dimension (ID) of the corresponding manifold, on which they evolve. It helps to quantify the complexity of the complete pollution pattern and «redundancy» in data.

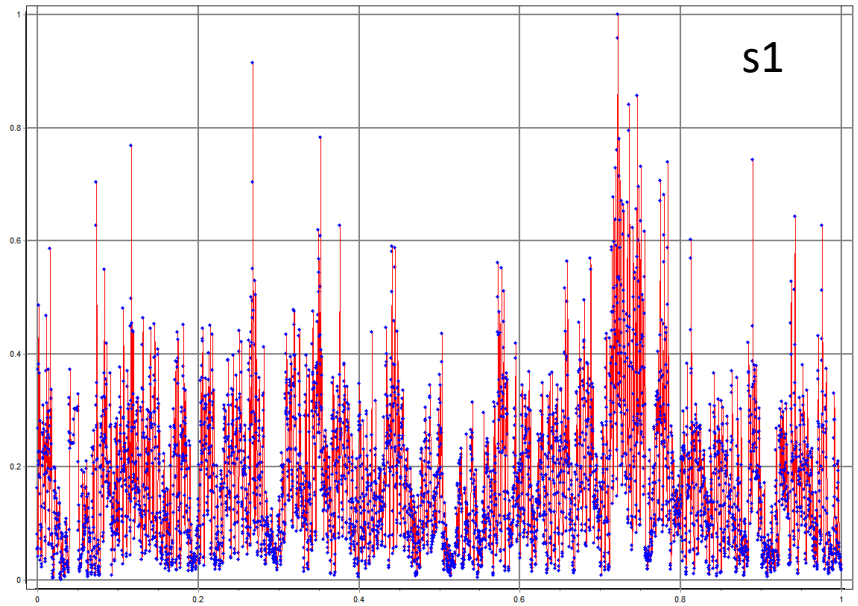
- There are many methods to estimate ID [1], including fractal approach, nearest neighbour methods, entropies, etc. In the present research we use a method based on Morisita index. The corresponding R package «IDmining» can be found on CRAN.

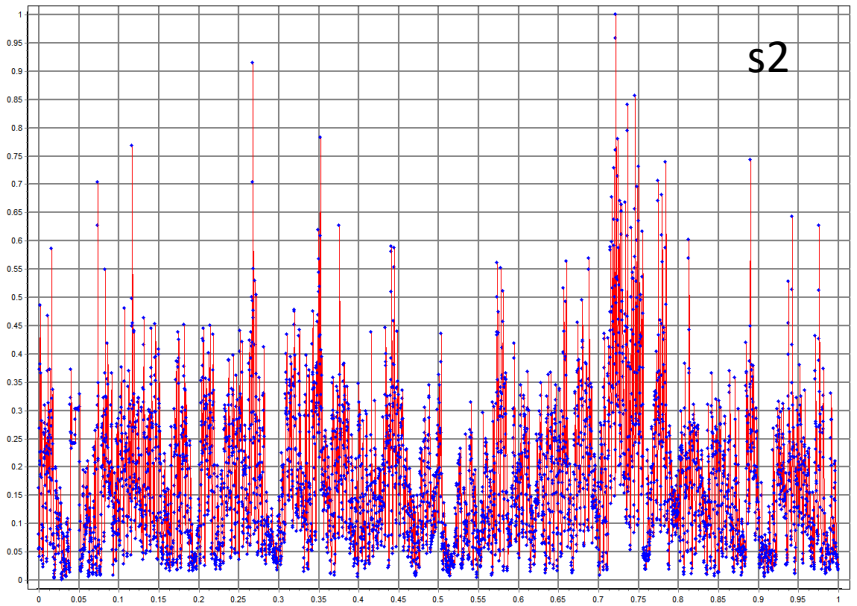
- The method is quite similar to box counting approaches in fractal theory.

Multipoint Morisita index (m-Morisita):

$$I_{m,\Delta} = Q^{(m-1)} \frac{\sum_{i=1}^{Q} n_i(n_i-1)(n_i-2)\ldots(n_i-m+1)}{N(N-1)(N-2)\ldots(N-m+1)}$$

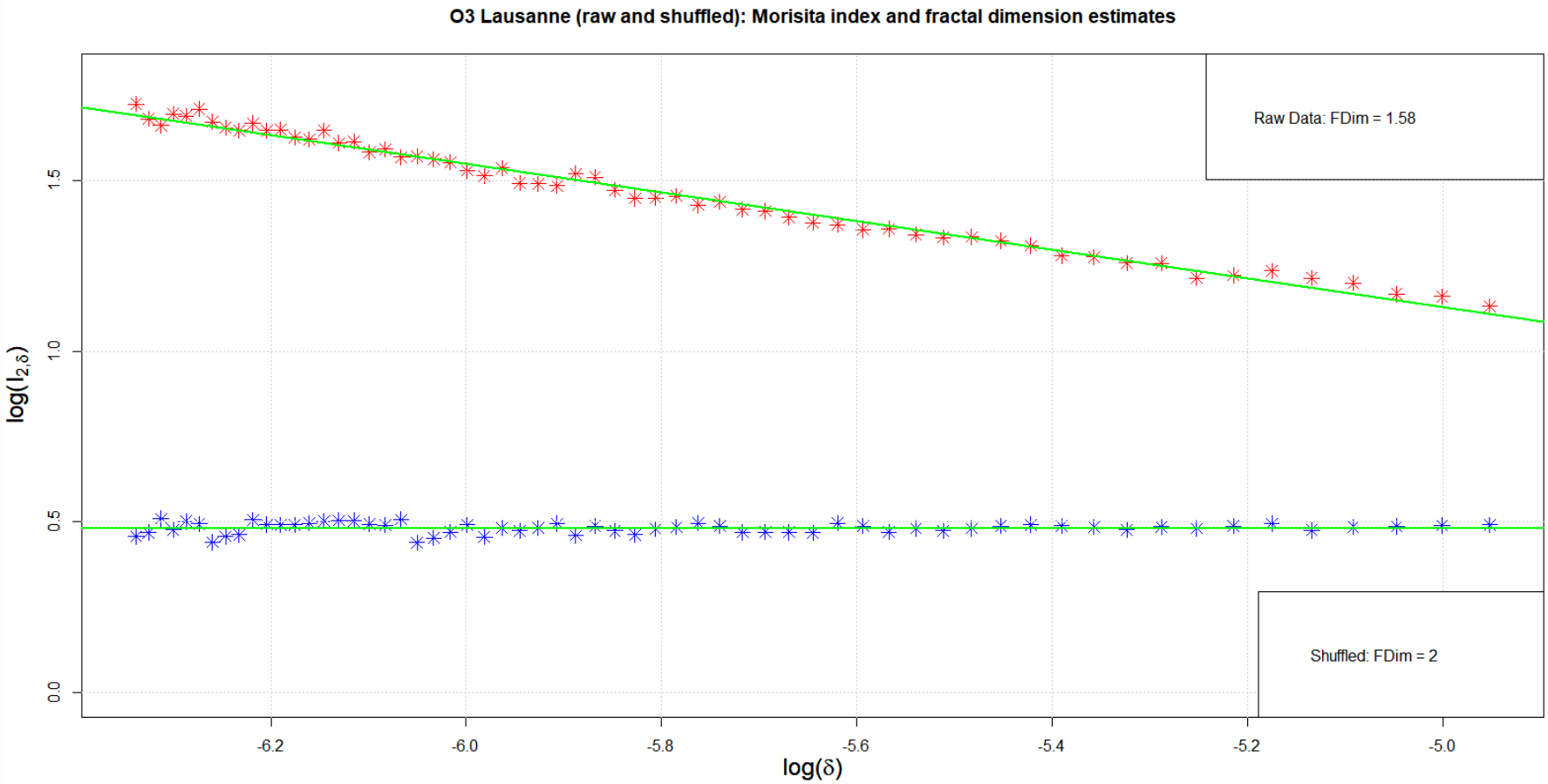
where n_i (i=1,2...,Q) is the number of points in the i-th cell, and N is the total number of points. *Interpretation: Normalized probability to find m-points in a cell at a given scale. Classical Morisita index: m=2*





Morisita index is calculated by covering a data pattern by a changing grid (scale = s1,s2,…) and counting the number of points in each cell, see formula above. In the present case 2d space is the following [t,f(t)].

An example of ID and fractal dimension estimates for Lausanne O3 raw and shuffled data using MI. Figures: log(MI) vs log(scale)



ID estimates of the monitoring

Euclidean dimension (ED) equals to the number of stations considered

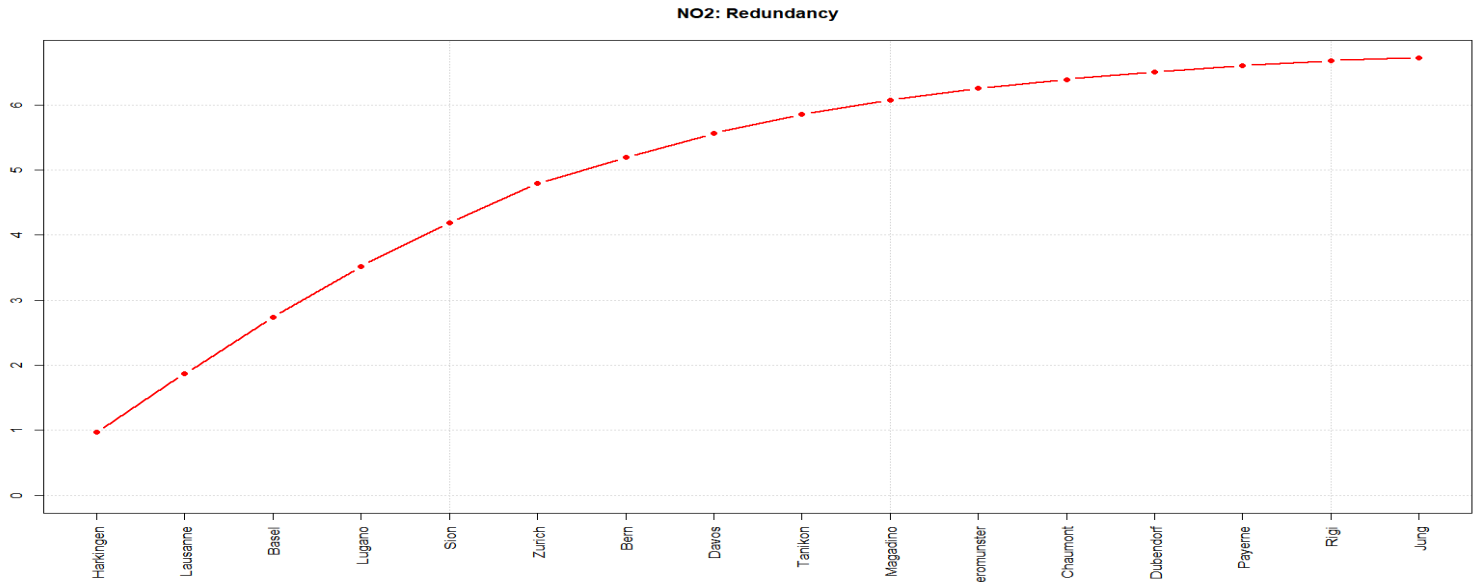
- "The ID estimate for NO2 equals to 6.54 (ED=16)"
- "The ID estimate for O3 equals to 5.37 (ED=16)"
- "The ID estimate for PM2.5 equals to 3.66 (ED=9)"
- "The ID estimate for NOX equals to 5.53 (ED=16)“
- "The ID estimate for SO2 equals to 3.15 (ED=9)“
- Randomly generated pattern in 9d with the same number of points, gives a value of ID ≈ 9.
- If we add a randomly generated time series to PM2.5 data (ED=9+1), the ID becomes equal to 4.67. The difference is ~1.

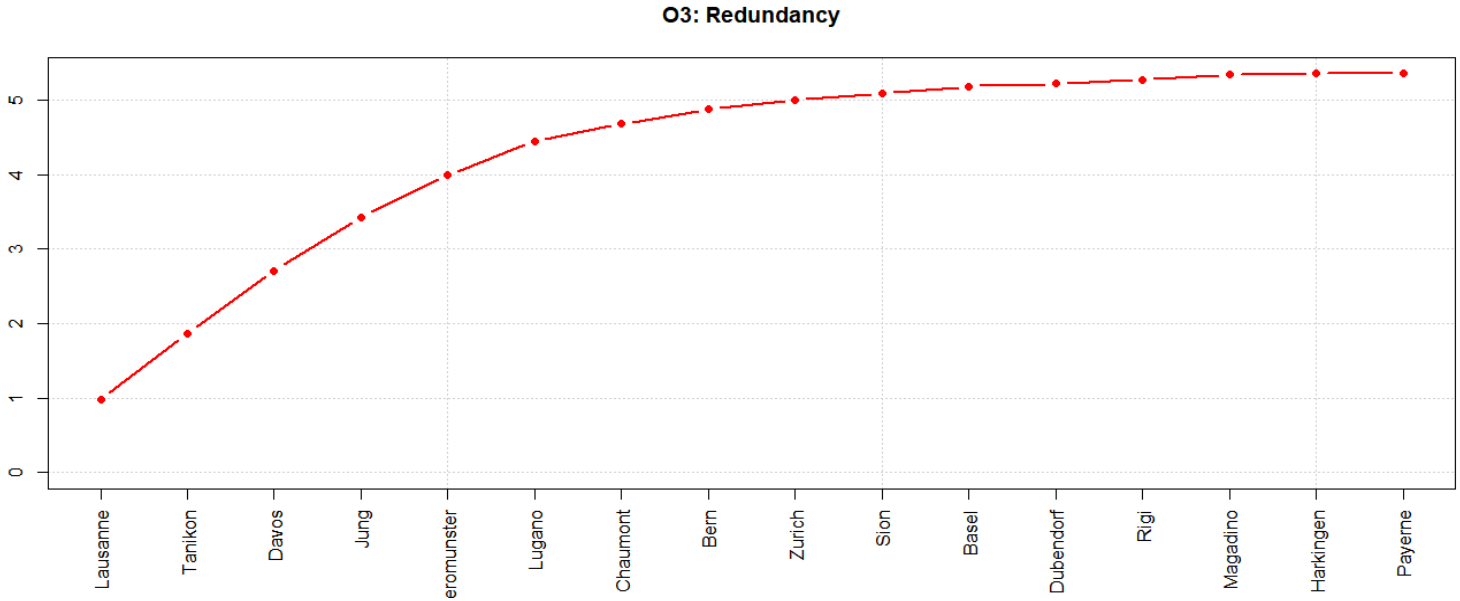
It is evident, that all ID dimensions are smaller than the Euclidean dimensions.

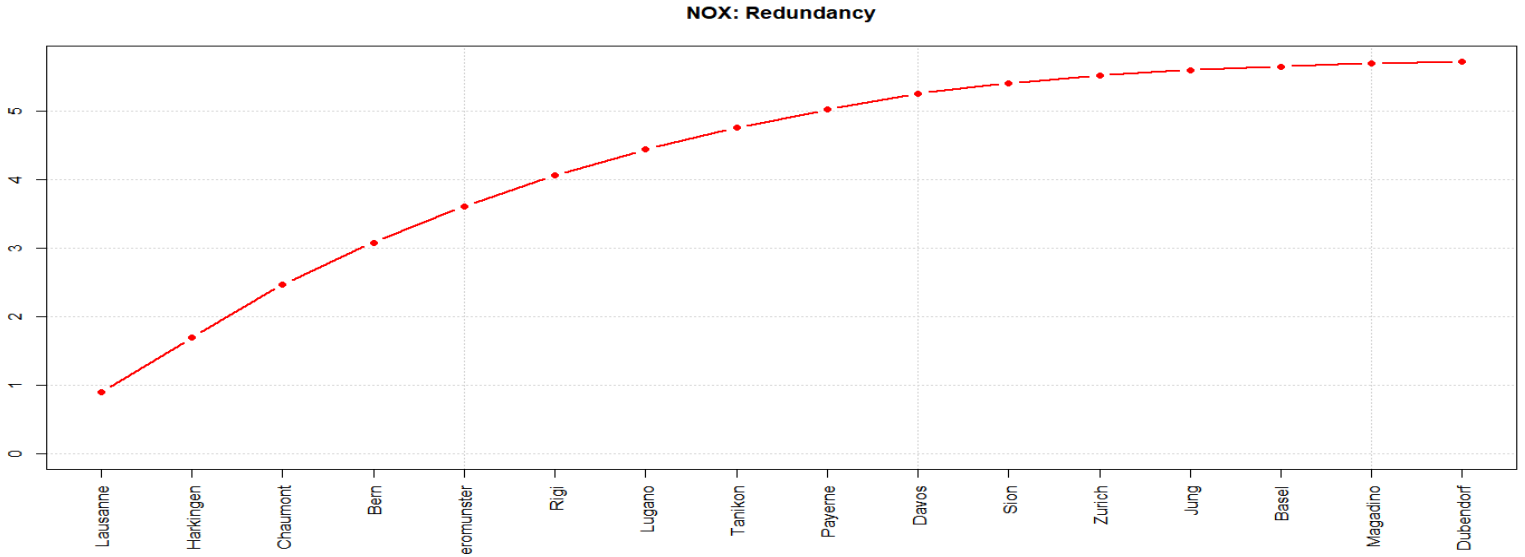
Redundancy in data

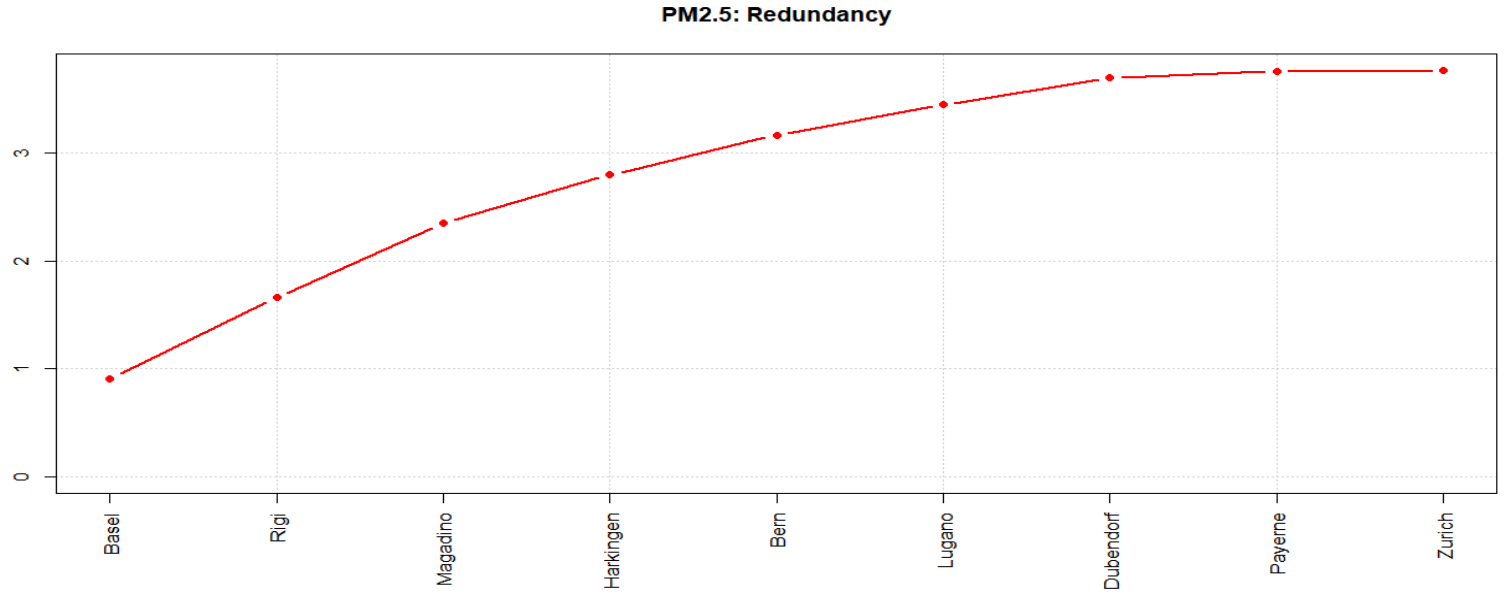
- Redundancy study deals, in part, with the analysis of nonlinear dependencies in data. Redundancy is an important concept in feature selection tasks in machine learning. Morisita based ID estimate was proposed to study the redundancy in high dimensional data (J. Golay, M. Kanevski “Unsupervised feature selection based on the Morisita estimator of intrinsic dimension” Knowledge-Based Systems, 2017, pp. 125-134) .

The results are presented (next slide) via ranking of the variables according to their redundancy: from the less redundant (left) to the most redundant (right). Redundant variable is, basically, the variable that can be represented as a linear/nonlinear function of non redundant variables. Such analysis can be interesting, for example, in constructing nonlinear multivariate models of time.


NO2: Redundancy
0
1
2
3
4
5
6
Harkingen
Lausanne
Basel
Lugano
Sion
Zurich
Bern
Davos
Tanikon
Magadino
eromunster
Chaumont
Dubendorf
Payerne
Rigi
Jung


O3: Redundancy
0
1
2
3
4
5
Lausanne
Tanikon
Davos
Jung
eromunster
Lugano
Chaumont
Bern
Zurich
Sion
Basel
Dubendorf
Rigi
Magadino
Harkingen
Payerne


NOX: Redundancy
0
1
2
3
4
5
Lausanne
Harkingen
Chaumont
Bern
eromunster
Rigi
Lugano
Tanikon
Payerne
Davos
Sion
Zurich
Jung
Basel
Magadino
Dubendorf


PM2.5: Redundancy
0
1
2
3
Basel
Rigi
Magadino
Harkingen
Bern
Lugano
Dubendorf
Payerne
Zurich

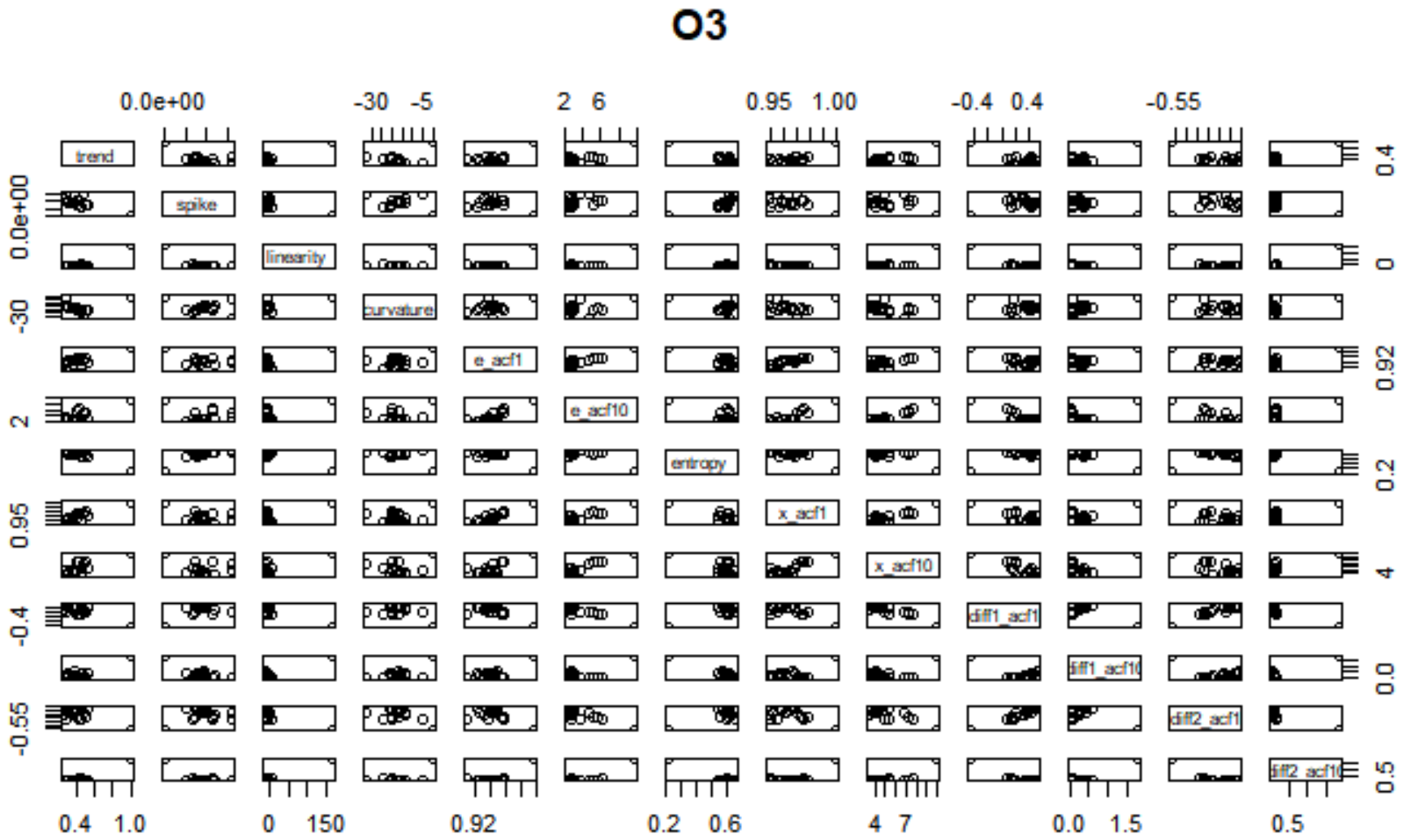
Time series representation via features

- Extraction of features from complex times series is a popular approach in clustering and classifications. Hundreds, even thousands, of features were proposed [1]. Some features can well discriminate time series, which considerably improves clustering tasks [2].

1. B.D. Fulcher, M.A. Little, N.S. Jones Highly comparative time-series analysis: the empirical structure of time series and their methods. J. Roy. Soc. Interface 10, 83 (2013).

2. F. Amato, M. Laib, F. Guignard, M. Kanevski. Analysis of air pollution time series using complexity-invariant distance and information measures. Physica A: Statistical Mechanics and its Applications, v.5471, 2020

Scatterplot matrix of the features extracted from O3 time series (trend, spike, linearity, etc.) using R package «tsfeatures»



Conclusions

- Nowadays, there is a huge variety of time series (TS) methods and tools efficiently applied in intelligent exploratory data analysis (IEDA)
- IEDA is an important (*in many cases the most important*) phase in TS analysis, modelling and forecasting.
- In the present paper real data on air pollution are considered with a major contribution in the introduction of the methods based on Morisita index to TS. This approach has an important potential in quantifying the complexity and dependencies both in univariate and multivariate time series

FNSNF
Swiss National Science Foundation

75 NRP
Big Data
National Research Programme