Spatio-Temporal Data Mining (Part II)

Donato Malerba and Annalisa Appice

MAESTRA
LEARNING FROM MASSIVE, INCOMPLETELY
ANNOTATED, AND STRUCTURED DATA

UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

KD DE

Dipartimento di
Informatica

Goals

- We will
  - Formalize the problem of learning with spatio-temporal data acquired in real-time through a number of (wireless) remote sensors
  - Clarify the role of various form of correlation
    - Spatial-, temporal, multivariate- correlation
  - Establish links with descriptive and predictive tasks
    - Summarization, interpolation, forecasting, anomaly detection)

Goals

- We will:
  - Point to important results
    - Algorithmic/experimental
    - Applications
  - Describe algorithms and techniques
  - Present open problems

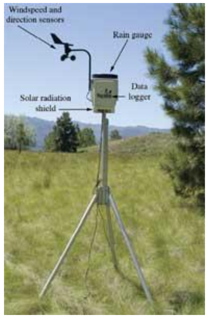
Assumed Background

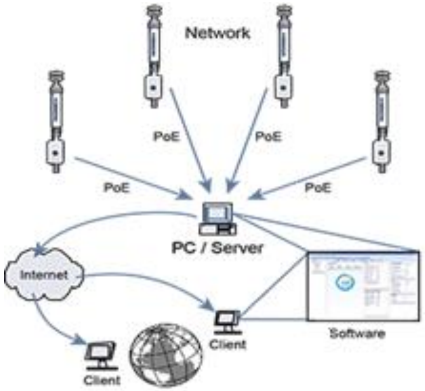
- We assume a basic knowledge of machine learning methods for clustering, classification and regression tasks
  - For background, please see Mitchell (1997)
- Basic knowledge of spatial statistics, time series analysis

Outline

- Data and tasks
- Issues and challenges
- Univariate learning
  - Summarization, interpolation, anomaly/change detection, forecasting
- Beyond the univariate case – multivariate case
  - Summarization vs interpolation
- Open challenges

Sensor networks

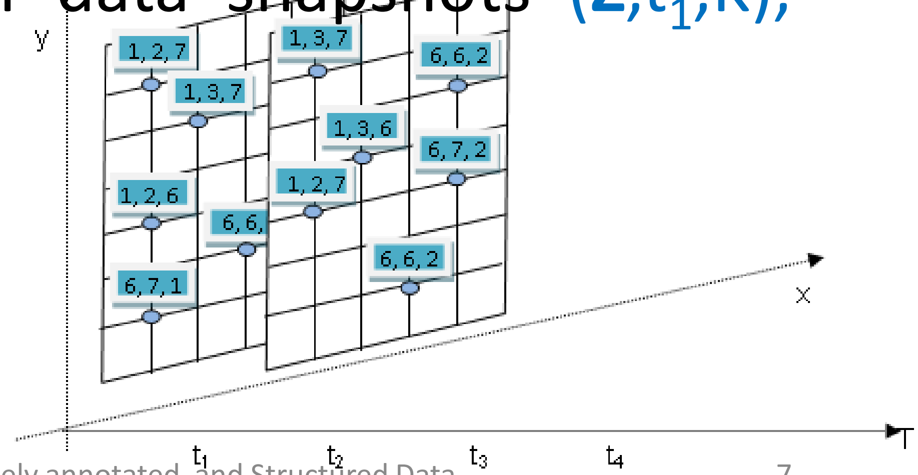




- A set of (wireless) sensor stations K which monitor an environment by collecting geophysical data (temperature, humidity, light,…)
- Each node in a sensor network can be imagined as a small computer, equipped with the basic capacity to sense, process, and act
- Sensors act in dynamic environments, often under adverse conditions

From sensor networks to geophysical time series

- A sensor network is scattered in a (possibly large) region where it is meant to collect data through its sensor nodes
- Every sensor $k \in \mathbf{K}$ measures a space of geophysical fields $\mathbf{Z}=(Z_1,\ldots,Z_m)$ repeatedly at the time points of T
- It feeds a time series of data snapshots $(\mathbf{Z},t_1,K)$, $(\mathbf{Z},t_2,K)$, …



Applications

- Typical applications of sensor networks include monitoring, tracking, and controlling
- Some of the specific applications are photovoltaic plant controlling, habitat monitoring, traffic monitoring, and ecological surveillance

Online data

- Sensor networks for climate data
  - National Oceanic and Atmospheric Administration (NOAA) data (http://www.ncdc.noaa.gov/)
- Sensor networks for ecology
  - Nature serve data (http://www.natureserve.org/biodiversity-science/conservation-topics/data)
- Sensor networks for energy markets
  - National Renewable Energy Lab data (http://www.nrel.gov/)
  - PVGIS (http://photovoltaic-software.com/pvgis.php)

Data scenario

- Time series of data that are measured repeatedly over a set of sensor stations.
  - The spatial location of a sensor station is modeled by means of point coordinates (e.g., latitude and longitude).
  - The spatial locations of the sensors are known, distinct and invariant, while the number of recording sensors may change in time: a sensor may be inactive and transmit no data for a time interval.
  - Active sensors transmit measurements for a number of numeric variables (multi-variate data) and they are synchronized in the transmission time.

Tasks

- Training on (incomplete) past data snapshots ($\mathbf{Z}$,t_1,K), ($\mathbf{Z}$,t_2,K), …, ($\mathbf{Z}$,t_n,K), in order to
  - Predict on missing data in some data snapshot ($\mathbf{Z}$,t_i,K) with $1 \leq i \leq n$ -- INTERPOLATION
  - Forecast some next data snapshot ($\mathbf{Z}$,t_i,K) with $i > n$ -- FORECASTING
  - Perform anomaly/change detection in the last data snapshot ($\mathbf{Z}$,t_n,K) -- ANOMALY/CHANGE DETECTION

Issues & challenges

- Spatial dimension:

  → Inferences on spatial correlation: how data taken at a relatively close location behave similarly to each other (proximity relation, network structure, local & global indexes, non stationariety)

- Temporal dimension:

  → Inferences on the temporal correlation: how many future observations can be predicted from past behavior (stationariety vs variation, concept drift, anomalies)

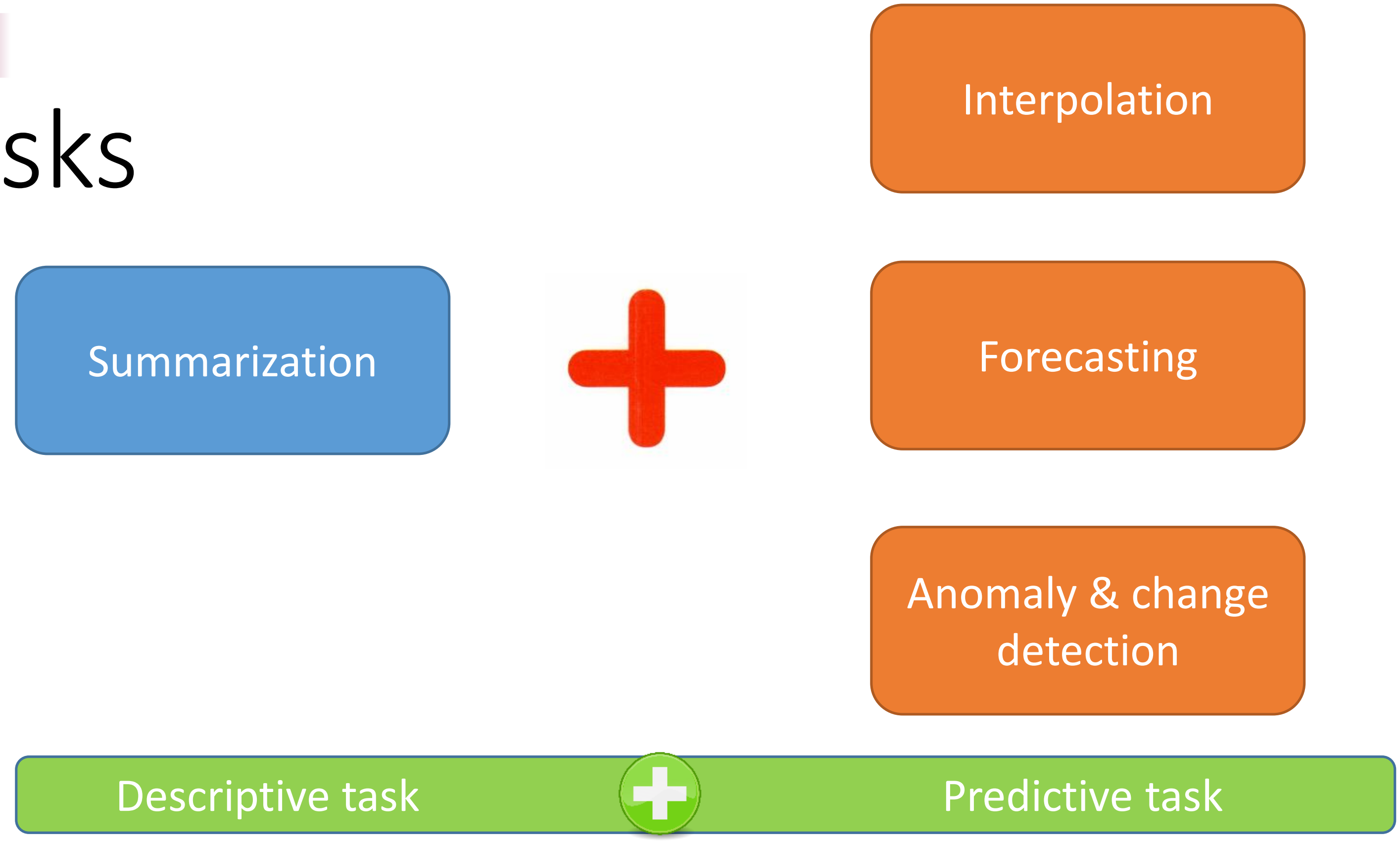
- Multivariate data:

  → Inferences on cross-correlation of variables measured at the same site, as well as at close sites

Additional issues & challenges

- Huge volume of data which cannot be entirely recorded for future analysis.
  - Computing data aggregates, discarding real data, using data aggregates in future analysis.
- Sensed data must be processed on-line (patterns are computed in (near) real time).
- Computation can be distributed in-network

Tasks



Interpolation

Summarization

Forecasting

Anomaly & change detection

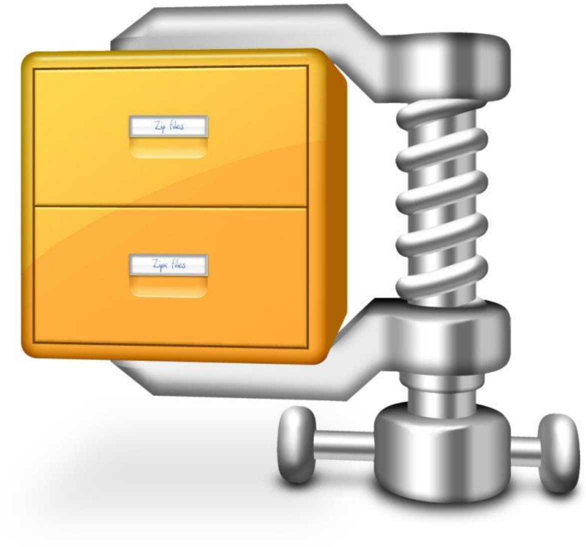
Descriptive task

Predictive task

- Associative classification (B. Liu et al., KDD 1998; M. Ceci & A. Appice, J. Intell. Inf. Syst, 2006, J. Yuan et al, Intelligent Data Analysis 2015)
- Predictive clustering (H. Blockeel et al., ICML 1998; D. Stojanova et al., Ecologial Informatics 2013; S. Dzeroski et al., KDID 2006)

Summarization

“Derive a compact representation of data for storage”



Summarization

- Sampling (uniform vs stratified): (S. Acharya et al., SIGMOD 2000)
- Discrete Fourier Transform: signal processing technique (Y. Zhu & D. Shasha, VLDB 2002)
- Histograms (optimal, equal-width, end-biased): summary structures used to capture the distribution of values (A.C. Gilbert, STOC 2002; M. Greenwald & S. Khanna, ACMSIGMOD Rec 2001; F. Furfaro et al., Knowl Inf Syst. 2008)

Summarization

- Sketches: approximation algorithms which allow the estimation of frequency moments and aggregates over joins (N. Alon et al., STOC 1996; J. Hershberger et al., Algoritmica 2006)
- Wavelets: projection of a sequence of data onto an orthogonal set of basis vectors (N. Alon, STOV 1996, Y. Matias, VLDB 2000)
- SAX: reduction of a numeric time series to a string of arbitrary length (J. Lin et al., Data Min Knowl Discov, 2007)
- Cluster analysis (S. Nassar & J. Sander, SSDBM 2007, M. Kontaki et al, DAWAK 2008)

Summarization in sensor network anaysis

- Centric summarization: deployed on the server station of networks by aggregating (spatial and/or temporal) correlated data.
  - Spatial cluster analysis snapshot by snapshot (X. Ma et al., APWeb/WAIM 2007)
  - Temporal cluster analysis sensor by sensor (P.P. Rodrigues et al., ECMLPKDD 2008)

Summarization in sensor network anaysis

- In-network summarization: sensor on-board summarization of data, only the summary is transferred to the centralized station → data communication and energy usage can be minimized
  - sampling, k-means or wavelet computed on the sensor (R. Chiky and G. Hébrail, DaWaK 2008)
  - simple aggregates (sum, count, histogram), computed along a tree-coordinating schema (Z. Chen, WTS 2010)
  - Spatio-temporal clustering (data which are autocorrelated both in space and time computed along a tree-coordinating schema (S. Yoon and C. Shahabi, ACM Trans Sens Netw 2007)

Trend cluster (spatial+temporal)

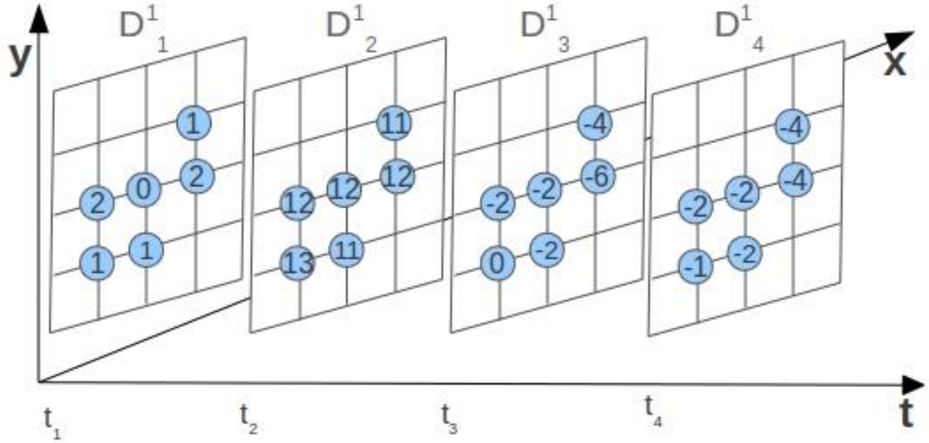
A trend cluster (Ciampi et al, KES 2010)is a triple:

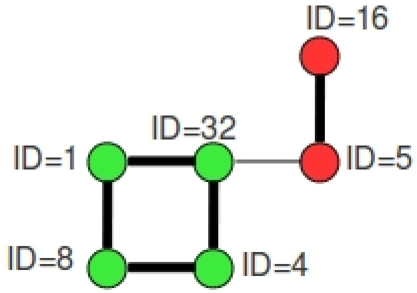
[W;C; T]

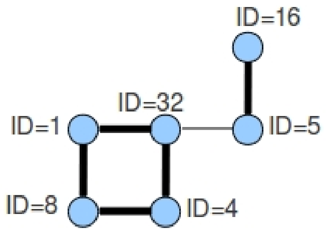
where:

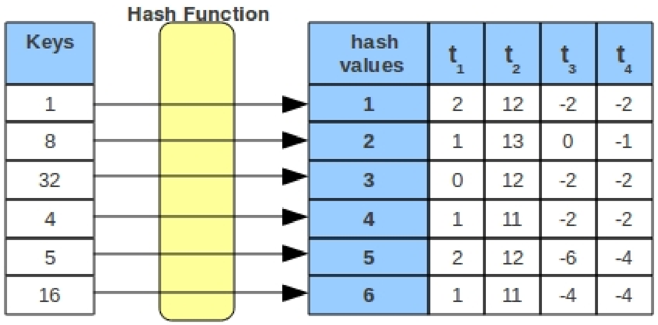
1. W is a time horizon along which field data were collected;
2. C is a cluster of spatially close sensors which transmitted values whose temporal variation was similar along the time horizon of the window;
3. T is the time series which represents the trend of the clustered measures as they were collected at the transmission time points comprised in W.

Trend cluster discovery



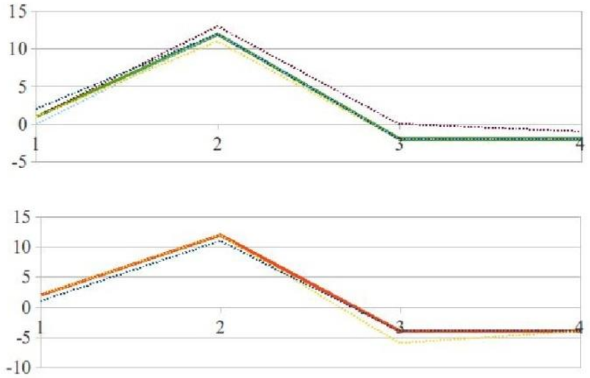






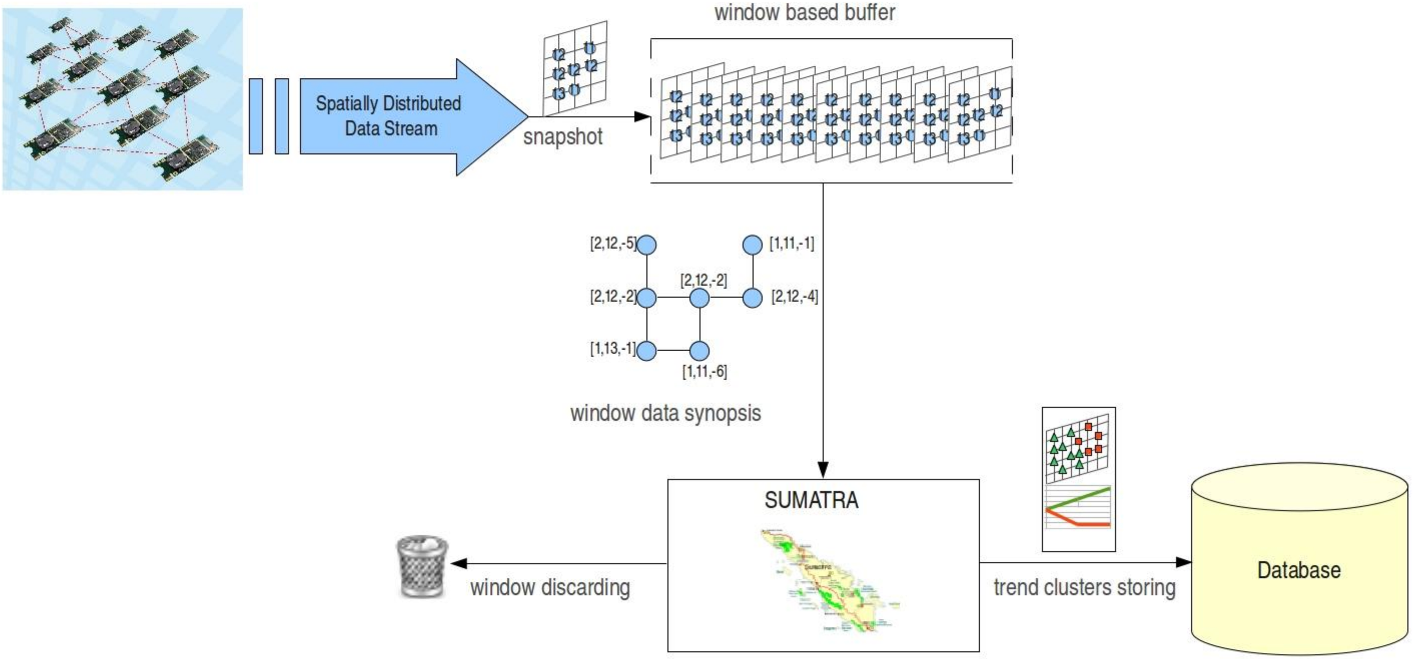
Keys
1
8
32
4
5
16

hash values	t_1	t_2	t_3	t_4
1	2	12	-2	-2
2	1	13	0	-1
3	0	12	-2	-2
4	1	11	-2	-2
5	2	12	-6	-4
6	1	11	-4	-4



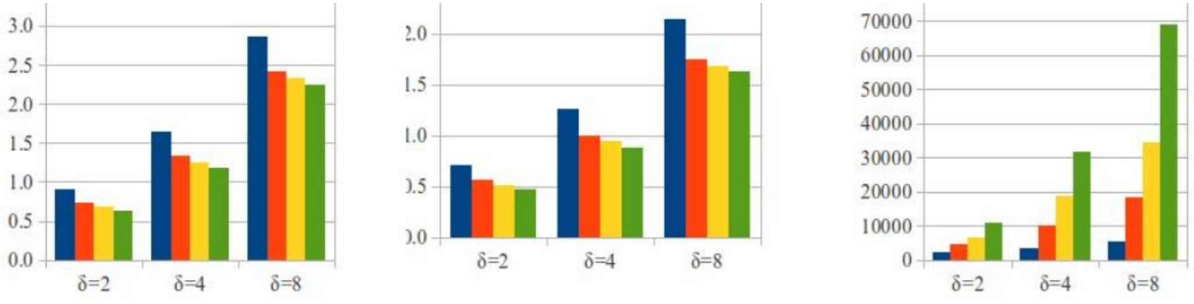
Trend cluster discovery

- Count-based window model - SUMATRA (Appice et al., DAMI 2015)



http://www.di.uniba.it/~appice/software/SUMATRATRECI/index.htm

South-America air climate

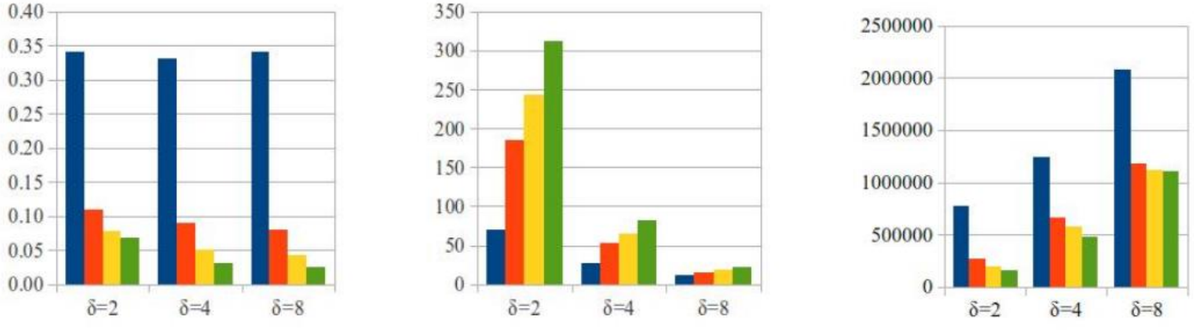


(a) RMSE

(b) MAE

(c) Average Computation Time per Window

- Monthly mean temperature: 6477 sensors

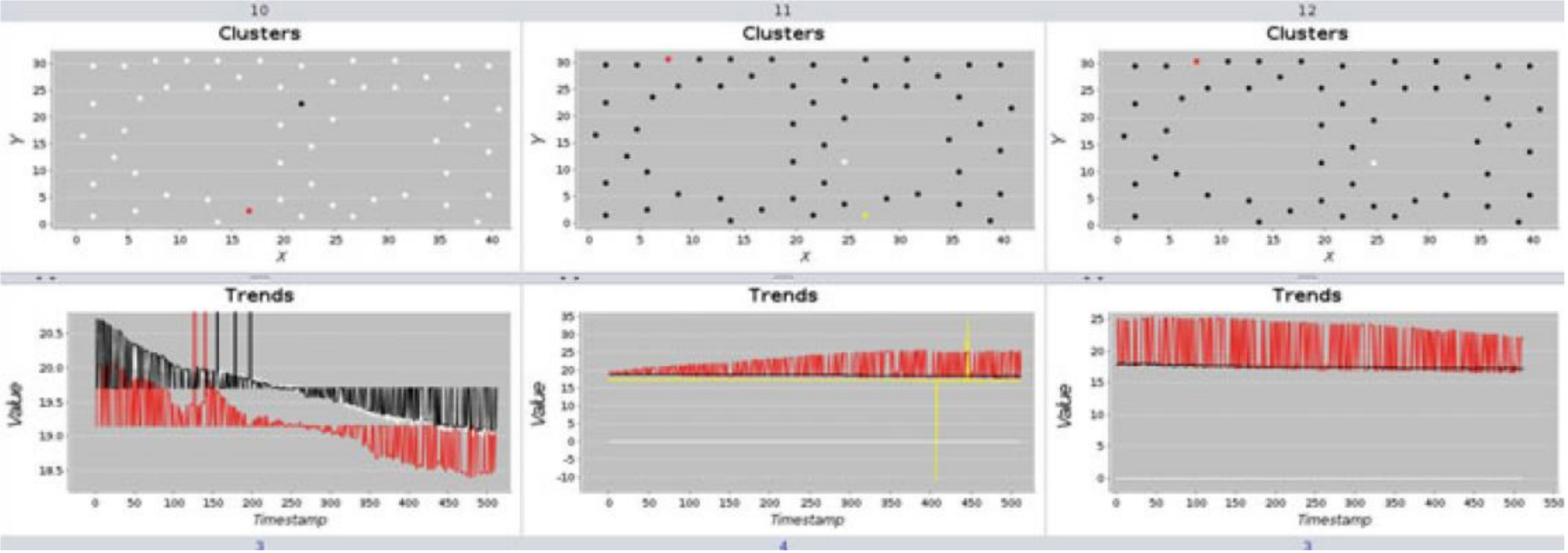


(d) Compression size

(e) Average Number Of Clusters per Window

(f) Total Computation Time

Intel Berkeley Lab

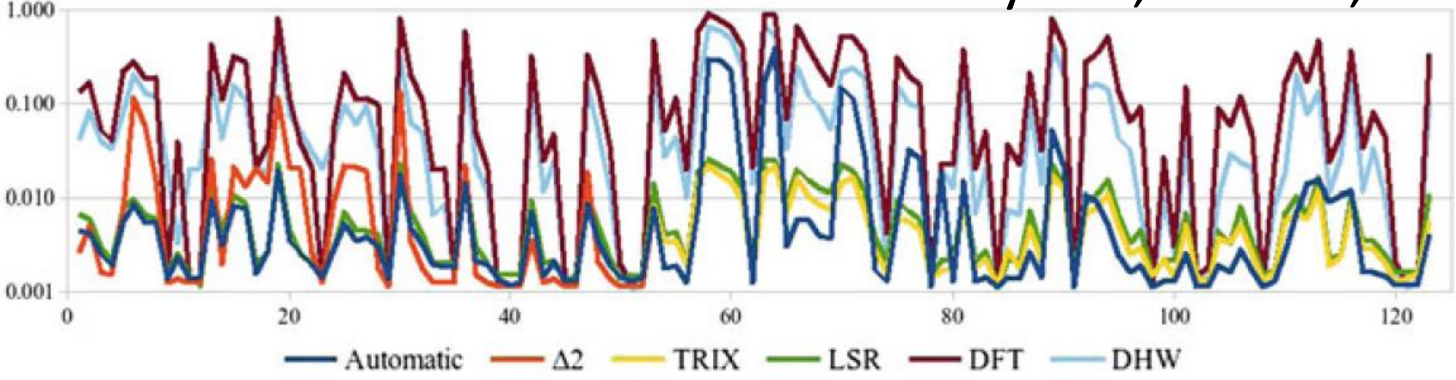


- Temperature every 31 secs: 54 sensors
- W=256, δ=2.5

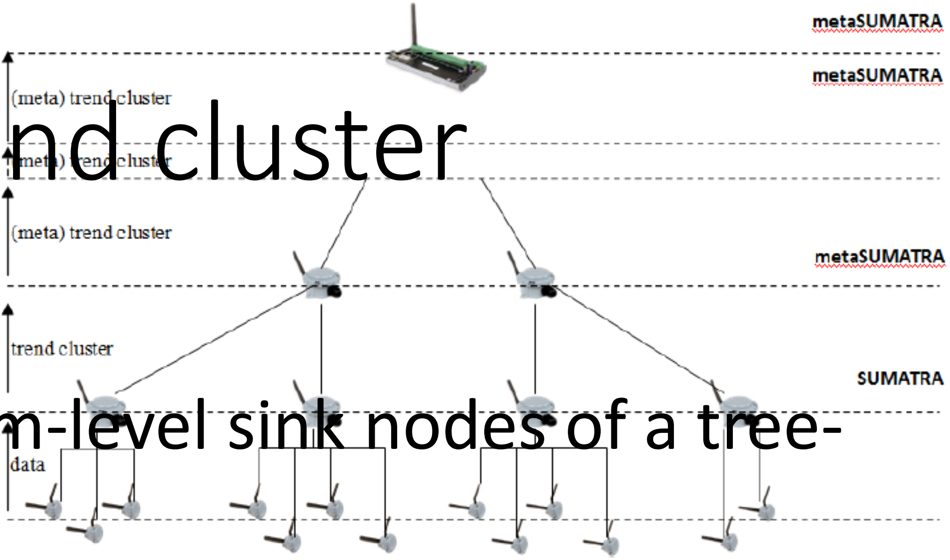
Further work

- Trend compression (A. Ciampi et al., CIDM 2011) by:
  - Discrete Fourier Transform
  - Haar Wavelets
  - Sampling
  - Least Square regression
- On-line selection of a trend compression technique (A. Appice et al., DAMI 2015)

Intel Berkeley Lab, W=256, δ=2.5



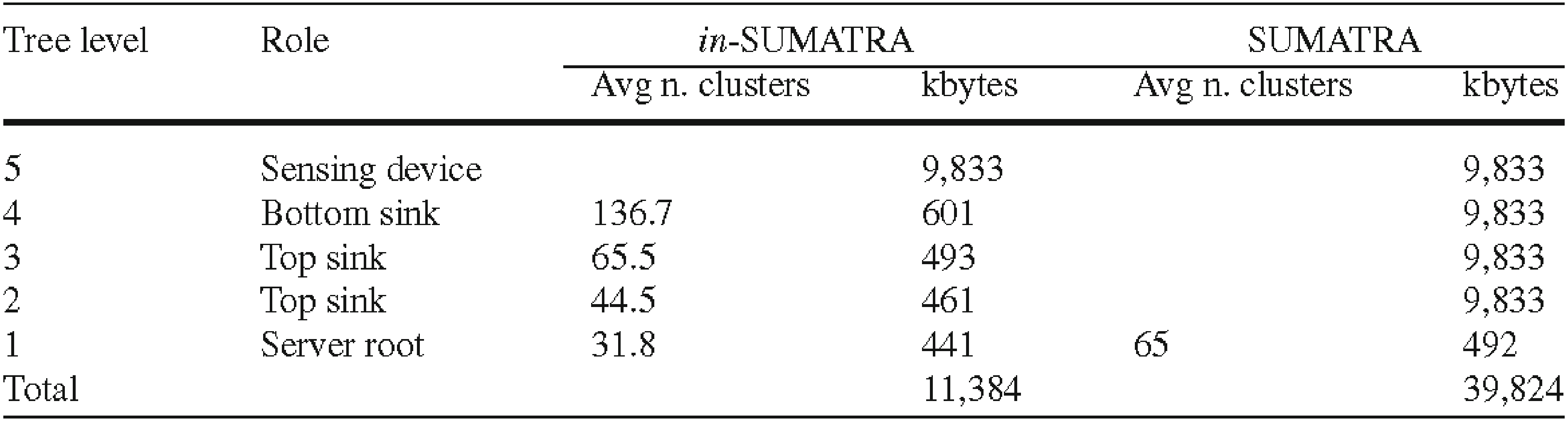
In-network trend cluster discovery



- SUMATRA on the bottom-level sink nodes of a tree-based WSN
- metaSUMATRA on the top-level sink nodes
- A top-level sink:
    1. receives trend clusters from its child sinks
    2. gathers together received trend clusters
    3. propagates (meta) trend clusters to the parent sink

(A. Appice et al., DAMI 2015)

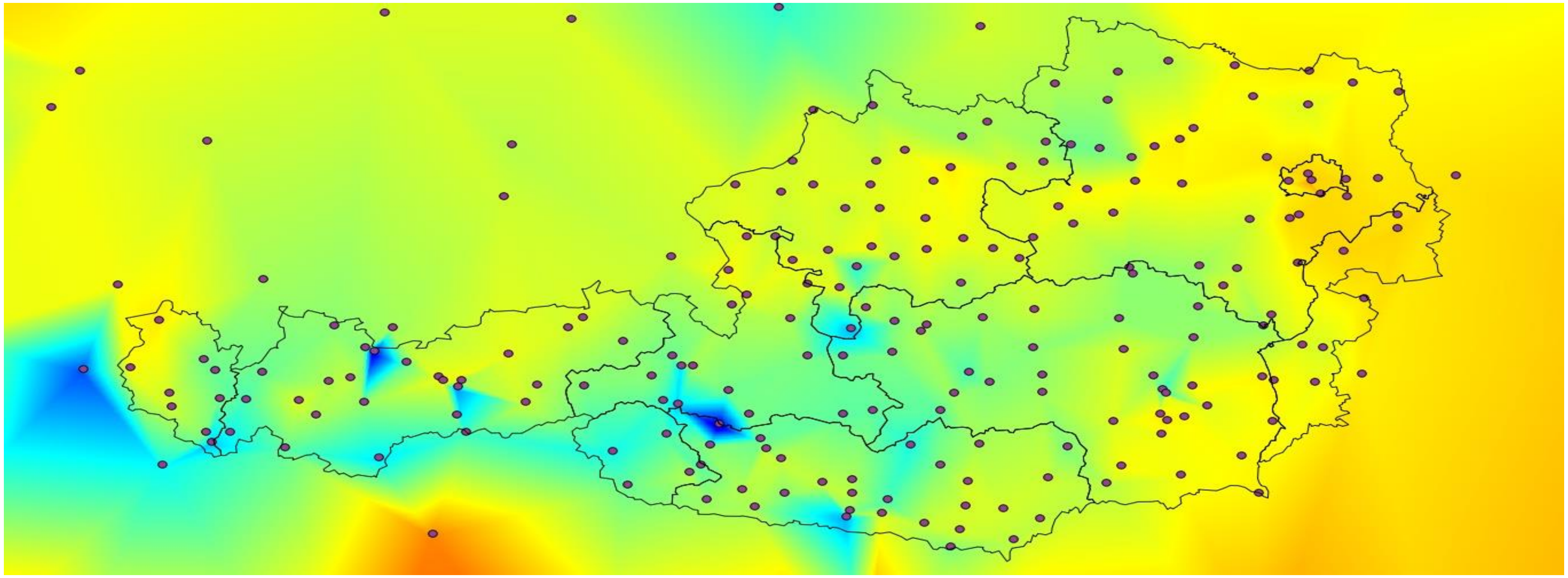
South-America air climate



Tree level	Role	*in*-SUMATRA		SUMATRA	
		Avg n. clusters	kbytes	Avg n. clusters	kbytes
5	Sensing device		9,833		9,833
4	Bottom sink	136.7	601		9,833
3	Top sink	65.5	493		9,833
2	Top sink	44.5	461		9,833
1	Server root	31.8	441	65	492
Total			11,384		39,824

Interpolation

“Supplement, smooth and standardize observational data”



Interpolation - spatial

- Inverse distance weighting: to calculate an unknown field value in a geographic location based on the degree of similarity in a neighbourhood (D. Shepard, ACM 1968)
- Radial basis functions: to calculate an unknown field value in a geographic location based on the degree of smoothing in a neighbourhood (G.F. Lin & L.H. Chen, Journal of Hydrology, 2004)
- Kriging (N. Cressie, 1993): to calculate an unknown field value in a geographic location based on a linear combination of data in a neighbourhoods. Weights are based on the computation of a variogram. The variogram represents an approximate measure of the spatial dissimilarity of the observed data

Interpolation - spatial

- Kriging is more complex than IDW,
  - The variogram computation cost scales as the cube of the number of observed data
  - Kriging is highly dependent on a reliable estimation of the variogram
- but it has the undeniable advantage of computing the best linear unbiased estimator of the correlation model
- However, the accuracy of an IDW interpolator often approaches the accuracy of a Kriging interpolator, especially for smooth fields (G.Y. Lu & D.W. Wong, Journal of Computers and Geosciences, 2008)

Interpolation – spatio-temporal

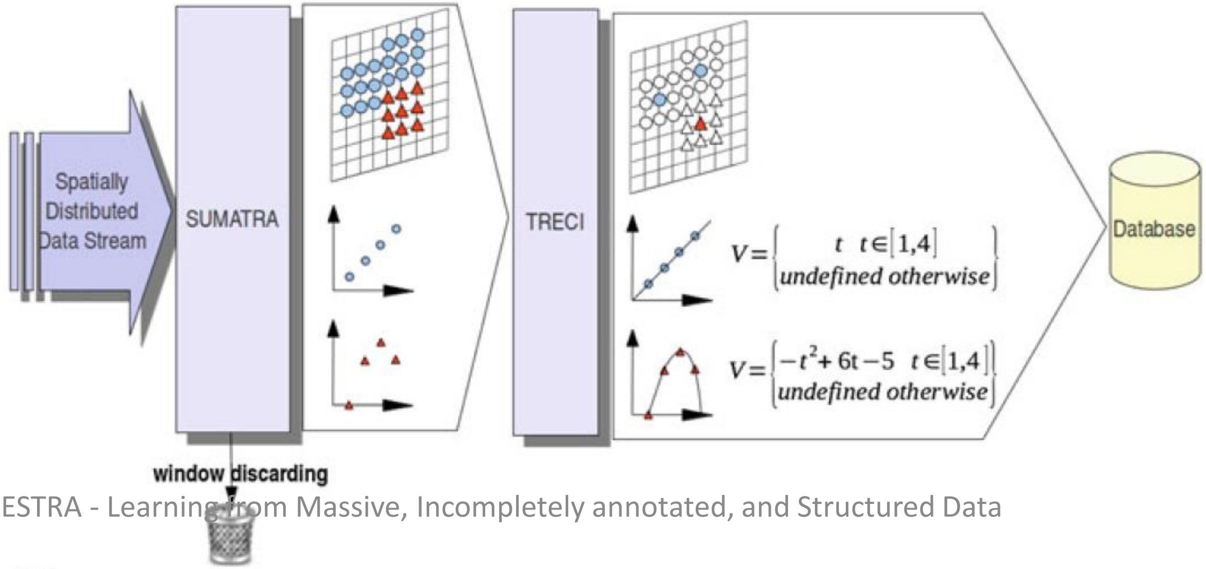
- First performing spatial interpolation and then reducing temporal interpolation to the application of simple methods (such as linear or spline interpolation) to the sequence of snapshots of spatially interpolated data (L. Li et al., SARA 2011)
- First interpolating time series of data for each relevant location and then using them as sampled observations for the application of a traditional spatial interpolator (L. Li, GIS: Exploring Data for Decision Making, 2009)

Interpolation – spatio-temporal

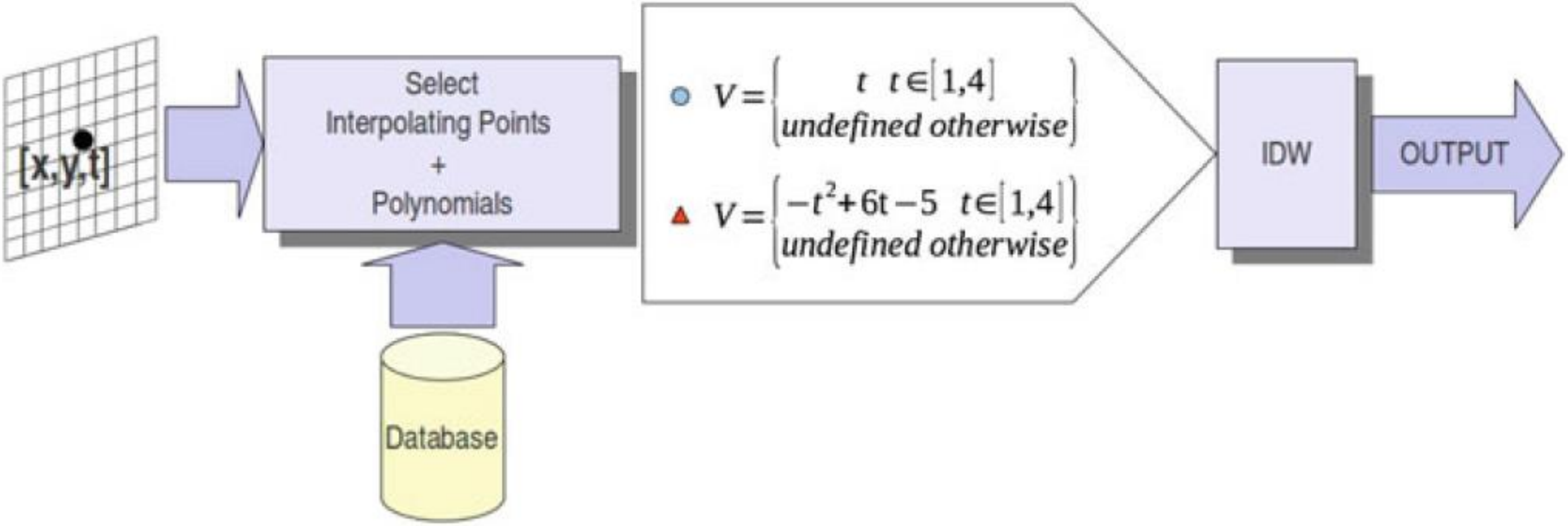
- The true integration of the spatial and temporal data component essentially based on the application of a dynamic model, like the Kalman filter or the Markov Random field, to consecutive snapshots of data
  - e.g. Kriging + Kalman Filter (W.S. Kerwin & J.L. Prince, *IEEE Transactions on Signal Processing, 1999)*
- Non-stationary time series analysis (trend and armonic component + spatiotemporal model of log-transformed data are computed. The model consists of trend and noise and represents the spatiotemporal variations (R. Romanowicz et al., Environmental Modeling Software , 2006)

Interpolation – trend cluster

- For each trend cluster (A. Appice et al., J. Spatial Information Science 2013),
  - Extract a shape-dependent (quadtree-based) sample of clustered sensors (key sensors)
  - Determine a (polynomial) regression model of the time law underlying the trend time series
- Key sensors and regression coefficients (trend) are stored


Spatially Distributed Data Stream
SUMATRA
TRECI
$V=\begin{cases} t & t\in[1,4] \\ \text{undefined otherwise} \end{cases}$
$V=\begin{cases} -t^2+6t-5 & t\in[1,4] \\ \text{undefined otherwise} \end{cases}$
Database
window discarding

Interpolation – trend cluster


[x,y,t]
Select Interpolating Points + Polynomials
Database
V = { t t∈[1,4] ; undefined otherwise }
V = { −t²+6t−5 t∈[1,4] ; undefined otherwise }
IDW
OUTPUT

- IDW applied to regression polyline-based predictions determined for the key sensors, at the time t

http://www.di.uniba.it/~appice/software/SUMATRATRECI/index.htm

Interpolation – trend cluster

- Trend cluster + Kriging (Guccione et al., MSM/MUSE 2011)
  - Trend cluster discovery to reduce the amount of data to mine for the variogram estimation
  - Trend cluster discovery + transfer learning to adapt the variogram learned at a time along the trend time series

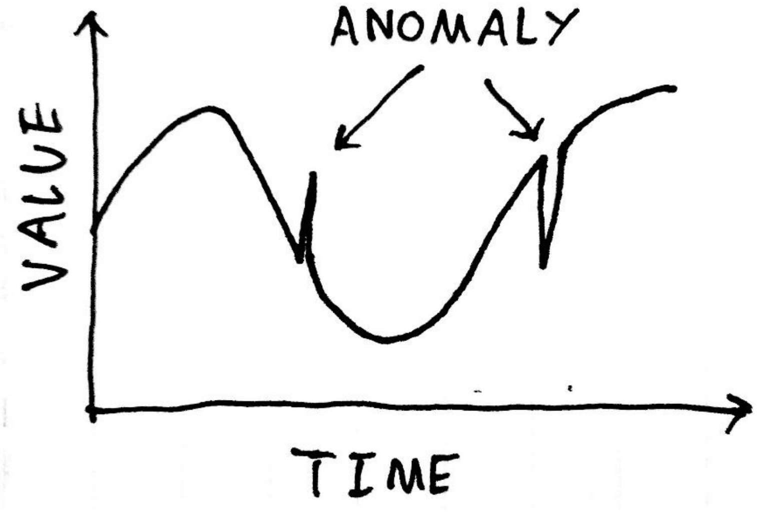
South America air climate

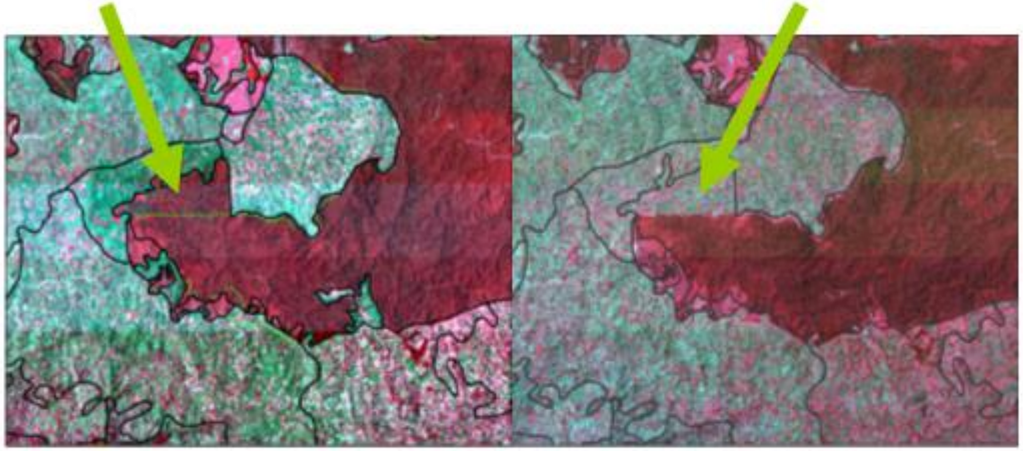
	SUMATRA	Sumatra + Sampling	TRECI
average D			5
size (Kbytes)	548.6	200.1	168.8
rmse	1.25	1.86	1.97

	Baseline	50%Sensors switching-off	50% Time points jumping-on	50 %Sensors switching-off and 50% time points jumping-on
TC+IDW	1.97	2.48	2.72	2.90
TC+Kriging	1.94	2.08	-	-

Anomaly/Change detection

"Detect exceptional (anomaly) or stable (change) abnormal behaviour in data"





Anomaly detection

- Time series analysis (M. Gupta et al., IEEE Trans. Knowl. Data Eng., 2013): semi-supervised, supervised, unsupervised learning
- Spatio-temporal data mining: spatial neighbourhood+ time window (S. Subramaniam, VLDB 2006, Franke et al., ACM 2009, Appice et al., Springer Briefs in Computer Science 2014)

Change detection

- Gradual changes (drift) vs and abrupt changes (shift) in the data distribution
- Incremental learning strategy + Gradual forgetting mechanisms (E. Lughofer et al., Appl. Soft Comput. 2011); (adaptive) window methods (R. Klinkenberg, IDA 2004, Gama et al., SBIA 2004); Page–Hinkley test (R. Sebastiao et al., SensorKDD 2008)

Anomaly and change detection

- A local polynomial fitting method + forward and backward prediction errors, (Z. Li et al., PAKDD 2007)
- Model fitting + outlier detection + quarantene to identify changes (*M. Pechenizkiy et al., SIGKDD Explor. 2009)*
- A change is alerted in the presence of outliers detected simultaneously in a snapshot *(Bakker et al., KDD 2009)*

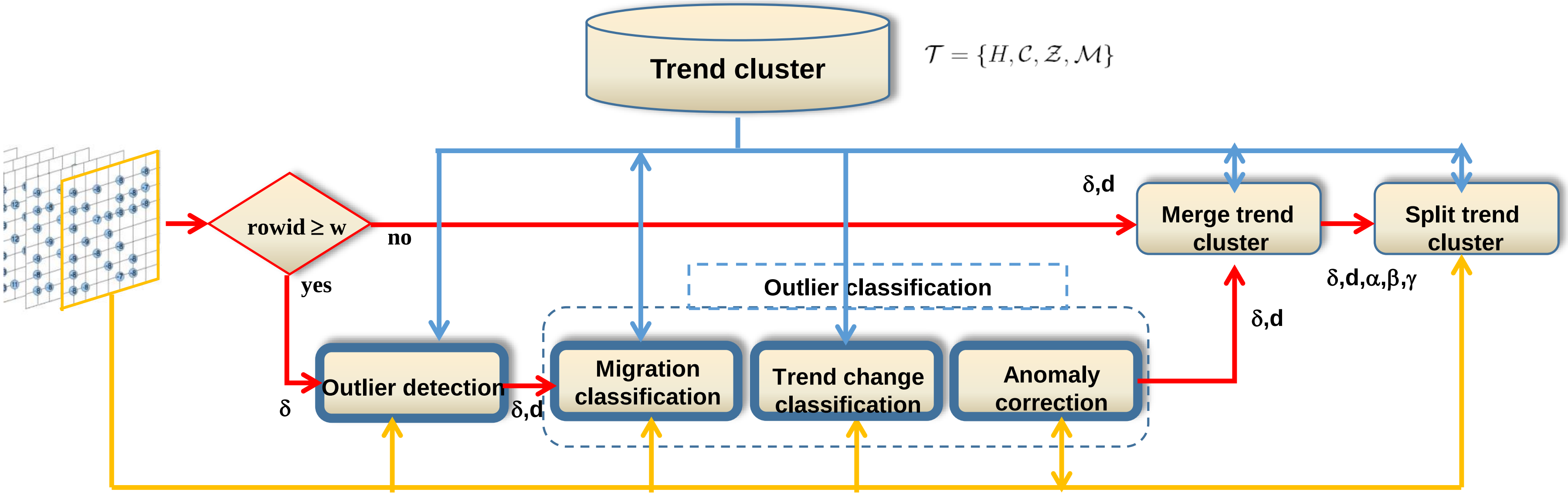
Anomaly and change detection - trend cluster

- Performing an incremental modeling phase of a geophysical data stream by accounting for spatial and temporal autocorrelation of data
- Detecting outliers (data which do not conform the model)
- Classifying outliers in anomalies and change points
  1. Correcting anomalies
  2. Changing the data model when change points are met

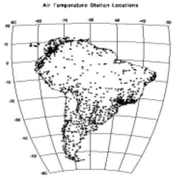
(Appice et al., Information Science 2014)

http://www.di.uniba.it/~appice/software/SWOD/index.htm

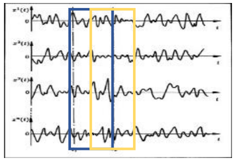
Anomaly and change detection - trend cluster



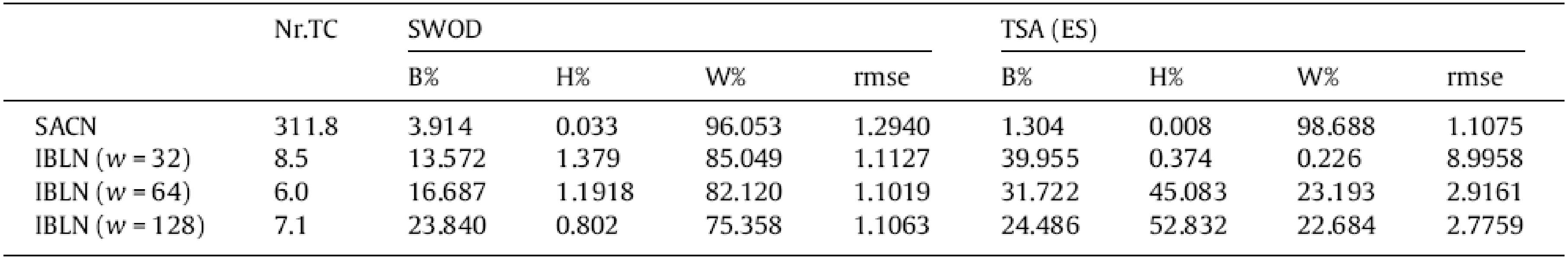
Anomaly and change detecion - trend cluster



South American Climate network (SACN): monthly-mean temperature ([-7.6 to 32.9]°C) form 6477 sensors between 1960 and 1990 .



Intel Berkeley Lab network (IBLN): temperature ([9.75–34.6]°C) every 31s from 54 sensors irregularly deployed in the Intel Berkeley Research lab between February 28th and April 5th 2004.



	Nr.TC	SWOD				TSA (ES)			
		B%	H%	W%	rmse	B%	H%	W%	rmse
SACN	311.8	3.914	0.033	96.053	1.2940	1.304	0.008	98.688	1.1075
IBLN ($w = 32$)	8.5	13.572	1.379	85.049	1.1127	39.955	0.374	0.226	8.9958
IBLN ($w = 64$)	6.0	16.687	1.1918	82.120	1.1019	31.722	45.083	23.193	2.9161
IBLN ($w = 128$)	7.1	23.840	0.802	75.358	1.1063	24.486	52.832	22.684	2.7759

Forecasting

“Predict the future”

Time series analysis

- Exponential smoothing model: averages a time series up to the current sample by Brown (more weight to recent data), Holt (correcting a linear tendency in the trend part), Winters (assuming the seasonality) models (A.C. Harvey, *1989*)
- ARIMA – family model: linear combinations are determined through the estimation of the autocorrelation function of the time series (ar, arma, arima, auto.arima) (G. E. P. Box and G. M. Jenkins, 1994)
- Multi-variate AR model: linear model of multiple time series used to forecast the target (var) (Lutkerpohl, 2005)

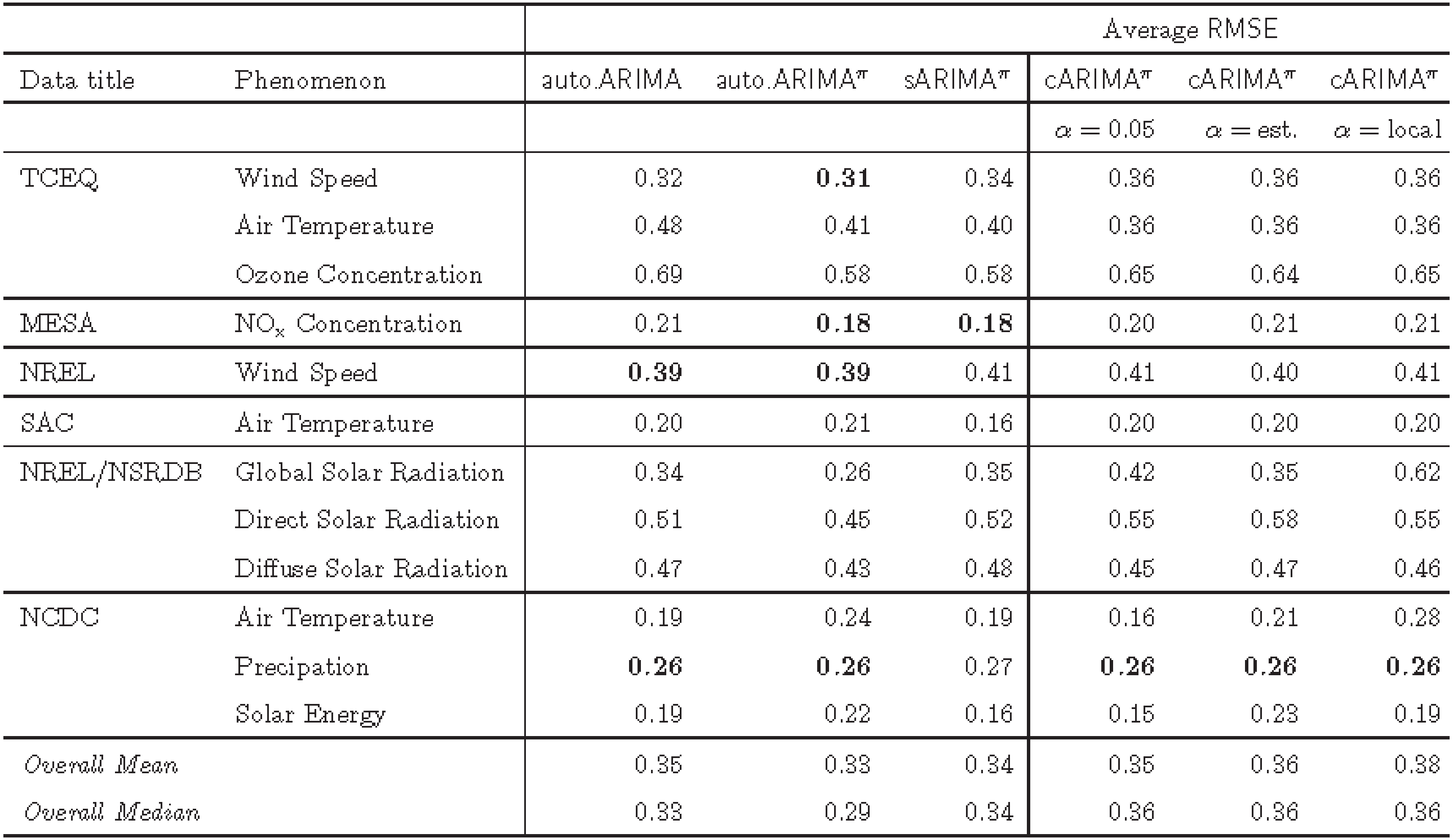
Spatial-aware forecasting

- Pokrajac and Obradovic (2001) have extended ARIMA by adding a term of auto-regressive disturbance, in order to model spatial-temporal correlation of residuals over defined neighborhood structures.
- Kamarianakis et al. (2005) have extended ARIMA, in order to account for the property of spatial correlation by expressing each data point at the time point t and the location (x;y) as a linearly weighted combination of data lagged both in space and time.
- Saengseedam and Kantanantha (2014) have used linear mixed models (LMMs) with spatial effects under a Bayesian framework.

Spatial-aware forecasting

- Luna and Genton (2005), as well as Barbosa et al. (2006) have explored the possibility of learning a VAR model from a vector of variables composed of the same variable observed at neighboring sites.
  - Sites are grouped in neighborhoods according to user-defined specifications.
- Pravilovic et al (2013, 2014) addressed the forecasting task with ARIMA-family models by dealing with spatial and temporal correlation when choosing parameters p,d and q and determining coefficients ϕ and σ
- Pravilovic et al (2014) integrated the spatio-temporal clustering analysis and forecasting in the same learning process

Spatial-aware ARIMA

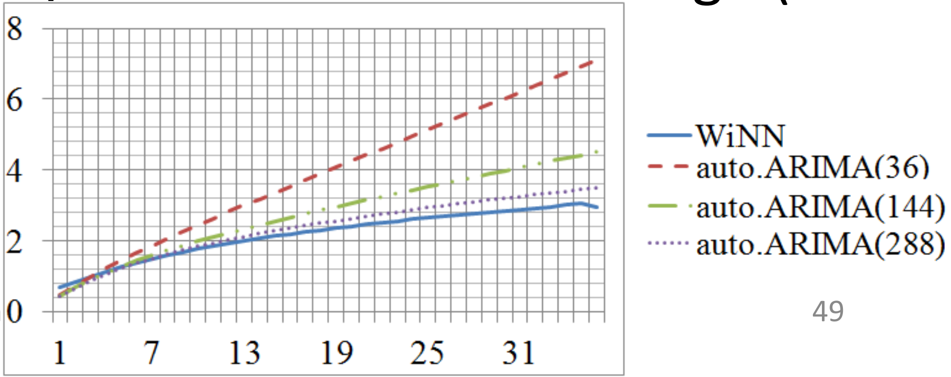


		Average RMSE					
Data title	Phenomenon	auto.ARIMA	auto.ARIMA$^\pi$	sARIMA$^\pi$	cARIMA$^\pi$	cARIMA$^\pi$	cARIMA$^\pi$
					$\alpha = 0.05$	$\alpha = \text{est.}$	$\alpha = \text{local}$
TCEQ	Wind Speed	0.32	**0.31**	0.34	0.36	0.36	0.36
	Air Temperature	0.48	0.41	0.40	0.36	0.36	0.36
	Ozone Concentration	0.69	0.58	0.58	0.65	0.64	0.65
MESA	NO_x Concentration	0.21	**0.18**	**0.18**	0.20	0.21	0.21
NREL	Wind Speed	**0.39**	**0.39**	0.41	0.41	0.40	0.41
SAC	Air Temperature	0.20	0.21	0.16	0.20	0.20	0.20
NREL/NSRDB	Global Solar Radiation	0.34	0.26	0.35	0.42	0.35	0.62
	Direct Solar Radiation	0.51	0.45	0.52	0.55	0.58	0.55
	Diffuse Solar Radiation	0.47	0.43	0.48	0.45	0.47	0.46
NCDC	Air Temperature	0.19	0.24	0.19	0.16	0.21	0.28
	Precipation	**0.26**	**0.26**	0.27	**0.26**	**0.26**	**0.26**
	Solar Energy	0.19	0.22	0.16	0.15	0.23	0.19
Overall Mean		0.35	0.33	0.34	0.35	0.36	0.38
Overall Median		0.33	0.29	0.34	0.36	0.36	0.36

Spatial-aware forecasting

- Time-series clustering + Predictive clustering
  - Temporal-based power wind forecasting (S. Pravilovic et al., DS 2014)
- Spatial neighborhood + ARIMA
  - Spatio-temporal power wind forecasting (V. Almeida & J. Gama, ICAIS 2014)
- Spatio-temporal adaptive neighborhood+ Knn
  - Spatio-temporal based power wind forecasting (A. Appice et al., DS 2015)

Rmse/NREL Data

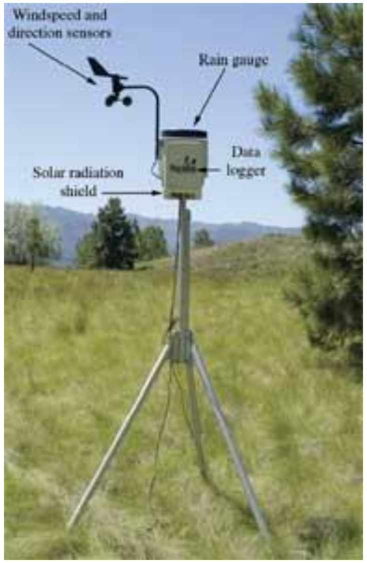


Spatial-aware forecasting

- A distributed system for storing huge amounts of data, gathered from energy production plants and weather prediction services
  - HBase over Hadoop framework
- One-day ahead forecast of PV energy production based on Artificial Neural Networks (with both structured and non-structured output prediction) (M. Ceci et al., IDEAS 2015)

Multivariate case

“Dealing with cross-correlation of various geophisical fields”



State of the art - spatial

- Stojanova et al. (2013) propose computing the mean of global measures (Moran I and global Getis C), computed for distinct variables of a vector, as a global indicator of spatial autocorrelation of the vector by blurring cross-correlations between separate variables
- Dray et al. (2006) explore the theory of the principal coordinates of neighbour matrices and develop the framework of Moran's eigenvector maps
  - They demonstrate that their framework can be linked to spatial autocorrelation structure functions also in multivariate domains

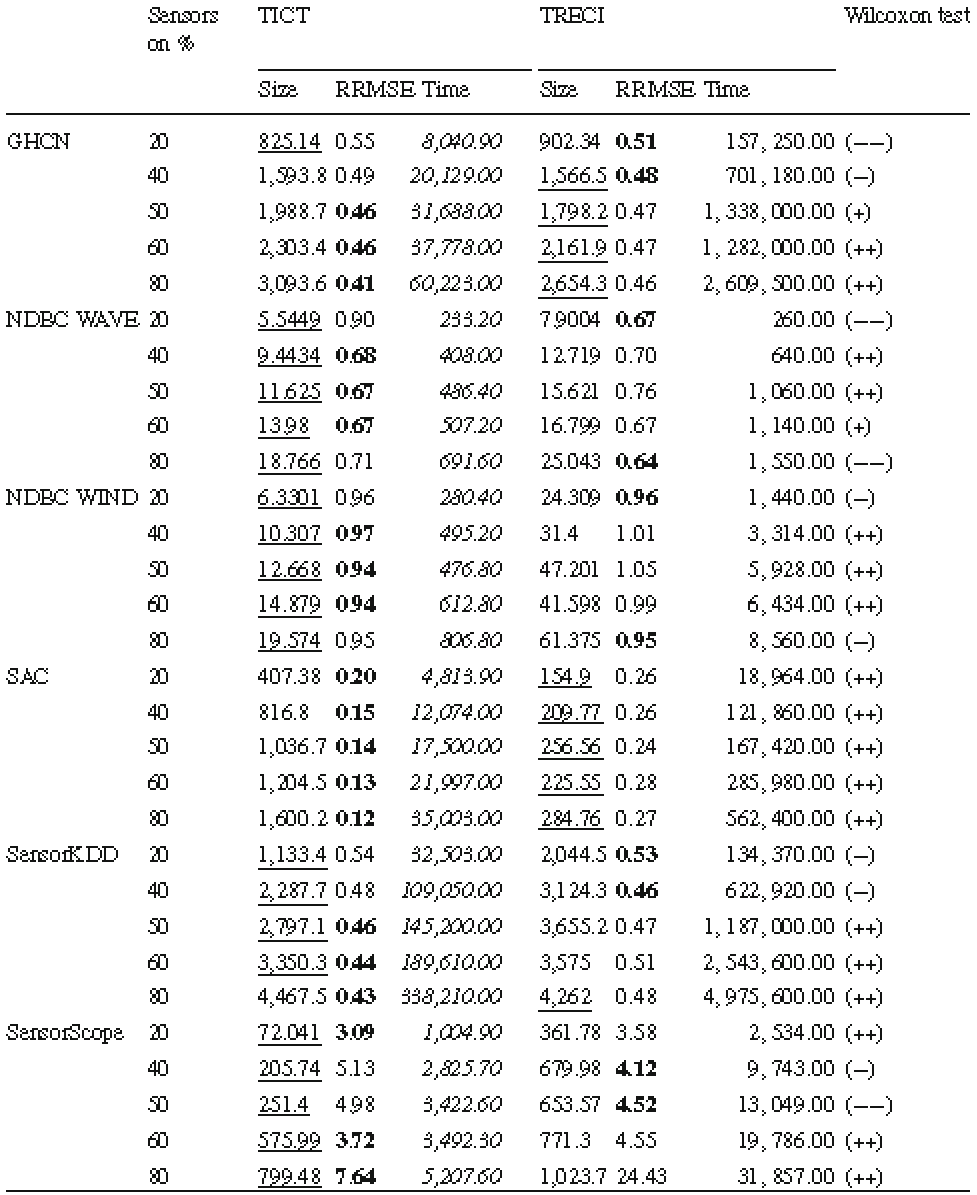
State of the art -spatial

- Dray and Jombart (2011) propose:
  - a two-step procedure, where data are first summarized with PCA. In a second step, any univariate (either global or local) spatial measure can be applied to PCA scores for each axis separately
  - An approach that finds coefficients to obtain a linear combination of variables, which maximizes the product between the variance and the global Moran measure of the scores
- Appice & Malerba (2014): interpolative clustering
  - Model the spatial autocorrelation when collecting the data records of multiple geophysical variables in a sensor network
  - Use this model to compute compact tree-based summaries of actual data that are discarded
  - Inject computed summaries into predictive (IDW-based) inferences to yield accurate estimations of geophysical data at any space location

State of the art – spatio-temporal

- Time-evolving interpolative clustering (A. Appice & D. Malerba, DAMI 2014)
  - We look for the interpolative clusters, which manifest change in the property (mean and variance) of the spatial autocorrelation of the clustered data
  - We build again only sub-tree of the existence tree, which do not cluster the new records appropriately (i.e. leaf conditions are not satisfied on the new data snapshot)

Time-evolving interpolative clustering



	Sensors on %	TICT			TRECI			Wilcoxon test
		Size	RRMSE	Time	Size	RRMSE	Time	
GHCN	20	825.14	0.55	*8,040.90*	902.34	**0.51**	157, 250.00	(−−)
	40	1,593.8	0.49	*20,129.00*	1,566.5	**0.48**	701, 180.00	(−)
	50	1,988.7	**0.46**	*31,688.00*	1,798.2	0.47	1, 338, 000.00	(+)
	60	2,303.4	**0.46**	*37,778.00*	2,161.9	0.47	1, 282, 000.00	(++)
	80	3,093.6	**0.41**	*60,223.00*	2,654.3	0.46	2, 609, 500.00	(++)
NDBC WAVE	20	5.5449	0.90	*233.20*	7.9004	**0.67**	260.00	(−−)
	40	9.4434	**0.68**	*408.00*	12.719	0.70	640.00	(++)
	50	11.625	**0.67**	*486.40*	15.621	0.76	1, 060.00	(++)
	60	13.98	**0.67**	*507.20*	16.799	0.67	1, 140.00	(+)
	80	18.766	0.71	*691.60*	25.043	**0.64**	1, 550.00	(−−)
NDBC WIND	20	6.3301	0.96	*280.40*	24.309	**0.96**	1, 440.00	(−)
	40	10.307	**0.97**	*495.20*	31.4	1.01	3, 314.00	(++)
	50	12.668	**0.94**	*476.80*	47.201	1.05	5, 928.00	(++)
	60	14.879	**0.94**	*612.80*	41.598	0.99	6, 434.00	(++)
	80	19.574	0.95	*806.80*	61.375	**0.95**	8, 560.00	(−)
SAC	20	407.38	**0.20**	*4,813.90*	154.9	0.26	18, 964.00	(++)
	40	816.8	**0.15**	*12,074.00*	209.77	0.26	121, 860.00	(++)
	50	1,036.7	**0.14**	*17,500.00*	256.56	0.24	167, 420.00	(++)
	60	1,204.5	**0.13**	*21,997.00*	225.55	0.28	285, 980.00	(++)
	80	1,600.2	**0.12**	*35,003.00*	284.76	0.27	562, 400.00	(++)
SensorKDD	20	1,133.4	0.54	*32,503.00*	2,044.5	**0.53**	134, 370.00	(−)
	40	2,287.7	0.48	*109,050.00*	3,124.3	**0.46**	622, 920.00	(−)
	50	2,797.1	**0.46**	*145,200.00*	3,655.2	0.47	1, 187, 000.00	(++)
	60	3,350.3	**0.44**	*189,610.00*	3,575	0.51	2, 543, 600.00	(++)
	80	4,467.5	**0.43**	*338,210.00*	4,262	0.48	4, 975, 600.00	(++)
SensorScope	20	72.041	**3.09**	*1,004.90*	361.78	3.58	2, 534.00	(++)
	40	205.74	5.13	*2,825.70*	679.98	**4.12**	9, 743.00	(−)
	50	251.4	4.98	*3,422.60*	653.57	**4.52**	13, 049.00	(−−)
	60	575.99	**3.72**	*3,492.30*	771.3	4.55	19, 786.00	(++)
	80	799.48	**7.64**	*5,207.60*	1,023.7	24.43	31, 857.00	(++)

Open challenges

- Designing in-network spatio-temporal algorithms
- Integrating big data technologies
- Dealing with the covariance in autocorrelation measures of several variables
- Dealing with multivariate case in spatio-temporal forecasting, as well as in anomaly/change detection task
- Completing the bridge between time series analysis and stream data mining

Thank you for the attention

References

- E. Biagioni and K. Bridges, The application of remote sensor technology to assist the recovery of rare and endangered species. International Journal of High Performance Computing Applications , vol. 16, pp. 315–324, 2002

- WA J. Johnson, M. Ruiz, J. Lees, and M. Welsh, Monitoring volcanic eruptions with a wireless sensor network. in Proceedings of the Second European Workshop on Wireless Sensor Networks, EWSN 2005

- J. Burrell, T. Brooke, and R. Beckwith, Vineyard computing: Sensor networks in agricultural production. IEEE Pervasive Computing, vol. 3, no. 1, pp. 38–45, 2004

- V. Weber, Smart sensor networks: technologies and applications for green growth, OECD Conference on ICTs, the environment and climate change, 2009

- F. Fumarola, A. Appice, D. Malerba, A Business Intelligence Solution for Monitoring Efficiency of Photovoltaic Power Plants, ISMIS 2014, 518-523, 2014

- A. Appice, S. Pravilovic, A. Lanza, D. Malerba, Very Short-Term Wind Speed Forecasting Using Spatio-Temporal Lazy Learning, Discovery Science 2015, 9-16, 2015

- W. Liu, W. Hsu, and Y. Ma, CBA: Integrating Classification and Association Rule Mining. In KDD'98, New York, NY, Aug. 1998.

- M. Ceci, A. Appice, Spatial associative classification: propositional vs structural approach. J. Intell. Inf. Syst. 27(3): 191-213, 2006

- J. Yuan, Z. Wang, M. Han, Y. Sun, A lazy associative classifier for time series, Intelligent Data Analysis, vol. 19, no. 5, pp. 983-1002, Ios Press 2015

- H. Blockeel, L. De Raedt, J. Ramon, Top-Down Induction of Clustering Trees. ICML 1998: 55-63, 1998

- D. Stojanova, M. Ceci, A. Appice, A. Malerba, A. Dzeroski, Dealing with spatial autocorrelation when learning predictive clustering trees. Ecological Informatics 13: 22-39, 2013

- S. Dzeroski, V. Gjorgjioski, I. Slavkov, J. Struyf, Analysis of Time Series Data with Predictive Clustering Trees. KDID 2006: 63-80, 2006

- S. Acharya, PB Gibbons, V. Poosala, Congressional samples for approximate answering of group-by queries. In: Proceedings of the international conference on management of data, SIGMOD 2000. ACM, New York, pp 487–498, 2000

- Y. Zhu, D. Shasha, Statstream: statistical monitoring of thousands of data streams in real time. In: Proceedings of the 28th international conference on very large data bases, VLDB 2002. VLDB Endowment, pp 358–369, 2002

- AC Gilbert, S. Guha, P. Indyk, Y. Kotidis, S. Muthukrishnan, MJ Strauss, Fast, small-space algorithms for approximate histogram maintenance. In: Proceedings of the 24th annual ACM symposium on theory of computing, STOC 2002. ACM, New York, pp 389–398, 2002

- M. Greenwald , S. Khanna, Space-efficient online computation of quantile summaries. ACMSIGMOD Rec 30(2):58–66, 2001

- F. Furfaro, GM Mazzeo, D. Saccà, C Sirangelo, Compressed hierarchical binary histograms for summarizing multi-dimensional data. Knowl Inf Syst 15(3):335–380, 2008

References

- N. Alon, Y.Matias, M. Szegedy,The space complexity of approximating the frequency moments. In: Proceedings of the 28th Annual ACM symposium on theory of computing, STOC 1996. ACM, New York, pp 20–29, 1996

- Y. Matias, JS Vitter, M. Wang, Dynamic maintenance of wavelet-based histograms. In: Proceedings of the 26th international conference on very large data bases, VLDB 2000. Morgan Kaufmann, San Francisco, pp 101–110, 2000

- J. Lin,EJ Keogh,L. Wei, S. Lonardi, Experiencing sax: a novel symbolic representation of time series. Data Min Knowl Discov 15(2):107–144, 2007

- S. Nassar, J. Sander, Effective summarization of multi-dimensional data streams for historical stream mining. In: Proceedings of the 19th international conference on scientific and statistical database management, SSDBM 2007. IEEE Computer Society, p 30, 2007

- M. Kontaki, AN Papadopoulos, Y. Manolopoulos, Continuous trend-based clustering in data streams. In: Proceedings of the 10th international conference on data warehousing and knowledge discovery, DaWaK 2008. Lecture notes in computer science, vol 5182. Springer, Berlin, pp 251–262, 2008

- X. Ma, S. Li, Q. Luo, D. Yang, S. Tang, Distributed, hierarchical clustering and summarization in sensor networks. In: Proceedings of the Joint 9th Asia-PacificWeb and 8th international conference on web-age information management and advances in data and web management, APWeb/WAIM 2007, Springer, Berlin, pp 168–175, 2007

- PP Rodrigues, J. Gama, LMB Lopes, Clustering distributed sensor data streams. In: Proceedings of the European Conference on machine learning and Knowledge discovery in databases. Lecture notes in computer science, vol 5212. Springer, Berlin, p 282–297, 2008

- R. Chiky, G. Hébrail, Summarizing distributed data streams for storage in data warehouses. In: Proceedings of the 10th international conference on datawarehousing and knowledge discovery, DaWaK 2008. Lecture notes in computer science, vol 5182. Springer, Berlin, pp 65–74, 2008

- Z. Chen, S. Yang, L. Li, Z. Xie, A clustering approximationmechanism based on data spatial correlation in wireless sensor networks. In: Proceedings of the 9th conference on wireless telecommunications symposium, WTS 2010. IEEE Press, Piscataway, pp 208–214, 2010

- S. Yoon, C. Shahabi, The clustered aggregation (cag) technique leveraging spatial and temporal correlations in wireless sensor networks. ACM Trans Sens Netw 3(1), 2007

- A. Ciampi, A. Appice, D. Malerba, Summarization for Geographically Distributed Data Streams. KES (3) 2010: 339-348, 2010

- A. Appice, A. Ciampi, D. Malerba, Summarizing numeric spatial data streams by trend cluster discovery. Data Min. Knowl. Discov. 29(1): 84-136, 2015

- A. Ciampi, A. Appice, D. Malerba, P. Guccione, Trend cluster based compression of geographically distributed data streams. CIDM 2011: 168-175, 2011

- D. Shepard, A two-dimensional interpolation function for irregularly-spaced data. In: Proceedings of the 1968 23rd ACM National Conference. ACM '68. ACM, 1968, pp. 517–524, 1968

- G.F. Lin and L.H. Chen, A spatial interpolation method based on radial basis function networks incorporating a semivariogram model. In Journal of Hydrology, Volume 288, Issues 3–4, pp. 288–298, 2004

References

- N. Cressie, Statistics for spatial data. Wiley, New York, 1993
- Lu, G. Y., and Wong, D. W. An adaptive inverse-distance weighting spatial interpolation technique. Journal of Computers and Geosciences 34, 1044–1055, 2008
- L. Li, X. Zhang, J. Holt, J. Tian and R. Piltner, Spatiotemporal interpolation methods for air pollution exposure. In Proc. Ninth Symposium on Abstraction, Reformulation, and Approximation, SARA (2011), M. R. Genesereth and P. Z. Revesz, Eds.,AAAI, 2011
- L. Li, Spatiotemporal Interpolation Methods, in GIS: Exploring Data for Decision Making. ETD collection for University of Nebraska-Lincoln, 2009.
- WS Kerwin and JL Prince, The Kriging update model and recursive spacetime function estimation. IEEE Transactions on Signal Processing 47, 11 (1999), 2942– 2952, 1999.
- R. Romanowicz, P. Young, P. Brown and P. Diggle, A recursive estimation approach to the spatio-temporal analysis and modelling of air quality data. Environmental Modeling Software 21, 6, 759–769, 2006
- A. Appice, A. Ciampi, D. Malerba, P. Guccione, Using trend clusters for spatiotemporal interpolation of missing data in a sensor network. J. Spatial Information Science 6(1): 119-153, 2013
- P. Guccione, A. Appice, A. Ciampi, D. Malerba, Trend Cluster Based Kriging Interpolation in Sensor Data Networks. MSM/MUSE 2011: 118-137, 2001
- M. Gupta, J. Gao, A. Charu, C.J. Han, Outlier detection for temporal data: a survey, IEEE Trans. Knowl. Data Eng. 25, 1–20, 2013
- S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, D. Gunopulos, Online outlier detection in sensor data using non-parametric models, in: VLDB, pp. 187–198, 2006
- M. Franke, O. Gertz, Outlier region detection and exploration in sensor networks, in: ACM SIGMOD International Conference on Management of Data, ACM, 2009, pp. 1075–1078, 2009
- A. Appice, A. Ciampi, F. Fumarola, D. Malerba, Data Mining Techniques in Sensor Networks: Summarization, Interpolation and Surveillance, Springer Briefs in Computer Science, Springer, 2014.
- E. Lughofer, P. Angelov, Handling drifts and shifts in on-line data streams with evolving fuzzy systems, Appl. Soft Comput. 11 , 2057–2068, 2011
- J. Gama, P. Medas, G. Castillo, P. Rodrigues, Learning with drift detection, in: A. Bazzan, S. Labidi (Eds.), Advances in Artificial Intelligence SBIA 2004, LNCS, vol. 3171, Springer, 2004, pp. 286–295, 2004
- R. Klinkenberg, Learning drifting concepts: example selection vs. example weighting, Intell. Data Anal. 8, 281–300, 2004

References

- R. Sebastiao, J. Gama, P.P. Rodrigues, J. Bernardes, Monitoring incremental histogram distribution for change detection in data streams, in: SensorKDD 2008, LNCS, vol. 5840, Springer, 2008, pp. 25–42,2008
- Z. Li, H. Ma, Y. Mei, A unifying method for outlier and change detection from data streams based on local polynomial fitting, in: PAKDD 2007, LNCS, vol. 4426, Springer, pp. 150–161, 2007
- M. Pechenizkiy, J. Bakker, I. Zliobaite, A. Ivannikov, T. Karkkainen, Online mass flow prediction in cfb boilers with explicit detection of sudden concept drift, SIGKDD Explor. 11, 109–116
- J. Bakker, M. Pechenizkiy, I.Zˇ liobaite˙ , A. Ivannikov, T. Karkkainen, Handling outliers and concept drift in online mass flow prediction in cfb boilers, in: Sensor KDD 2009, ACM, pp. 13–22, 2009
- A. Appice, P. Guccione, D. Malerba, A. Ciampi, Dealing with temporal and spatial correlations to classify outliers in geophysical data streams. Inf. Sci. 285: 162-180, 2014
- A.C. Harvey, Forecasting, Structural Time Series Models and the Kalman Filter, Cambridge University Press, 1989
- G. E. P. Box and G. M. Jenkins. Time Series Analysis: Forecasting and Control. Prentice Hall PTR, 3rd edition, 1994.
- H. Lutkepohl, New Introduction to Multiple Time Series Analysis, Springer-Verlag Berlin Heidelberg, 2005.
- Pokrajac, Z. Obradovic, Improved spatial-temporal forecasting through modelling of spatial residuals in recent history, in: Proceedings of the First SIAM International Conference on Data Mining, SDM 2001, Chicago, IL, USA, April 5-7, 2001, pp. 1–17, 2001
- Y. Kamarianakis, P. Prastacos, Space–time modeling of traffic flow, Computers & Geosciences 31 (2), 119–133, 2005
- P. Saengseedam, N. Kantanantha, Spatial time series forecasts based on Bayesian linear mixed models for rice yields in Thailand, in: Proceedings of the International MultiConference of Engineers and Computer Scientists, IMECS 2014, vol. II, Newswood Limited, pp. 1007–1012, 2014
- S. M. Barbosa, M. E. Silva, M. J. Fernandes, Multivariate autoregressive modelling of sea level time series from TOPEX/Poseidon satellite altimetry, Nonlinear Processes in Geophysics 13 (2), 177–184, 2006
- S. Pravilovic, A. Appice, D. Malerba, An Intelligent Technique for Forecasting Spatially Correlated Time Series. AI*IA 2013: 457-468, 2013
- S. Pravilovic, A. Appice, D. Malerba, Integrating Cluster Analysis to the ARIMA Model for Forecasting Geosensor Data. ISMIS 2014: 234-243, 2014
- S. Pravilovic, A. Appice, A. Lanza, D. Malerba. Wind Power Forecasting Using Time Series Cluster Analysis. Discovery Science 2014 276-287, 2014

References

- V. Almeida, J.Gama. Collaborative Wind Power Forecast. ICAIS 2014: 162-171, 2014
- M. Ceci, R. Corizzo, F. Fumarola, M. Ianni, D. Malerba, GM, E. Masciari, M. Oliverio, A. Rashkovska, Big Data Techniques For Supporting Accurate Predictions of Energy Production From Renewable Sources. IDEAS 2015: 62-71, 2015 S. Dray, P. Legendre and PR Peres-Neto, Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (pcnm). Ecol Model 196(34):483–493, 2006
- S. Dray and T. Jombart, Revisiting guery’s data: introducing spatial constraints in multivariate analysis. Ann Appl Stat 5(4):2278–2299, 2011
- A. Appice, D. Malerba, Leveraging the power of local spatial autocorrelation in geophysical interpolative clustering. Data Min. Knowl. Discov. 28(5-6): 1266-1313, 2014