Literature Survey:

# Challenges in Big data analytics

(Assunção *et al.*, 2015) challenges:

Better data management – efficient methods to store filter, transform and retrieve

Need for better visualisation, report generation and storytelling data presentation with narrative visualisation

user interaction

# Challenges for data analytics applied to climate science and construction

The use of big data analytics is underutilised in the domain of weather data (Jain and Jain, 2017), better big data weather forecasting could benefit the construction industry by helping build more efficient buildings and provide an “*improved quality of life”.*

The latter benefit could be assumed to include buildings that provide better occupant comfort and are better suited to purpose.

The construction industry is responsible for massive amounts of greenhouse gases (Dimoudi and Tompa, 2008). Through their lifecycle buildings consume 50% of all total energy demand and contribute 50% of all CO2 emissions. Using big data analytics to design buildings could reduce these impacts.

# Environmental design strategies

Primary goal for environmental design is to reduce the energy consumed by buildings, estimated to be around 40% the total global energy consumption (Omer, 2008). Much of this energy is expended on heating lighting and cooling.

Energy efficiency can be defined as the minimising the amount of energy consumed to achieve thermal comfort for occupants. Currently the energy required to maintain thermal comfort accounts for 60-70% of energy consumed in non-industrial buildings (Omer, 2008).

Better understanding of human response to climatic context (bioclimatic) can result in buildings that require less energy for heating and cooling (Olgyay and Olgyay, 2015, p11)

Bioclimatic chart (Olgyay and Olgyay, 2015, p23) based on the conditions the chart recommends wind for ventilation, radiation from sun, shading and or evaporative cooling

Typical approach divides into 5 broad climate zones as in (Hausladen, Liedl and De Saldanha, 2011) based on conditions averaged over the course of a year. The complexity of the problem is clear in the dimensionality of 20 cities 17 variables

### Thermal comfort

understood as a combination of temperature, relative humidity, air movement and radiant temperature, giving a state of mind where the person does not require any change in current conditions (ASHRAE, 2013) or a state where minimal extra energy is require to maintain the human balance (Manzano-Agugliaro *et al.*, 2015).

### Psychrometric chart

maps interrelationships of thermal conditions of the environment, zone of thermal comfort can be plotted following ASHRAE (ASHRAE, 2013) and where a condition falls outside this zone design approaches can be determined (Lechner, 2009). P chart does not show air movement or radiant temperatures

Comfort zone is where 70% are happy with the comfort levels when undertaking light activity and wearing light clothing critiqued by some as it is based on a North American Culture and developed by and industry with the goal of providing heating and cooling focused on living and working in air-conditioned buildings

Air movement on the chart is assumed to be minimal

An adaptive model acknowledges some of the limitations of the ASHRAE methods and considers how outdoor climate impacts indoor climate (de Dear, 1998). It accounts for changes in comfort preferences and expectations observed when users have access to environmental control and past exposure to local climate effects. The published adaptive standard (ANSI/ASHRAE, 2013) is based on average temperature of previous seven to thirty days with occupants in sedentary activities in naturally ventilated spaces. – other research exists that suggests an adaptive form of comfort where peoples

Givoni (1992) (Givoni, 1992)also notes how ASHRAEs standards don’t work in certain zones where locals have adapted see Colima Mexico or where higher levels of natural ventilation are typical

Global scale climate patterns Lechner (2015, p72) and micro climate are noted due to topography / proximity to water causing night day variations and in some cases anomalies.

Lechner (2015, p88) provides 17 regions each with a list of design priorities as a subset of a total 11 strategies.

Strategy example: “*Keep hot temperatures out during the summer”* (Lechner ,2015, p127)

Each strategy is then broken in to a series of related instrumental approaches eg. “*keep daytime hot air out of the building by closing all openings*” (Lechner, 2015, p126)

Methods from Climate consultant (Milne, Liggett and Benson, 2009) generates a priority ordered list 68 design guidelines are described but most not linked to specific weather conditions

## Design strategies

(Lechner, 2009) three-tiered approach:

### Tier 1 basic design

the location of the building on the site, defining building orientation and the position and size of openings on the building. Exposure to sun and wind can be both positive and negative depending on location and use.

considering the form of the building. example self-shading forms, recessed windows certain forms promote natural ventilation

Selection of materials dense materials with high thermal mass can help reduce overheating in hot climates, while light weight construction can form spaces that can be rapidly heated by solar radiation in cold climates.

### Tier 2 Passive systems

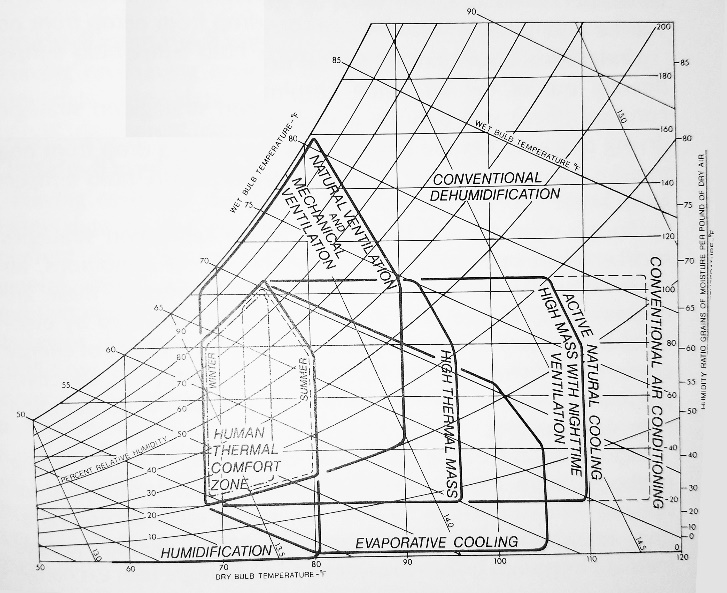
Solar, cooling and lighting

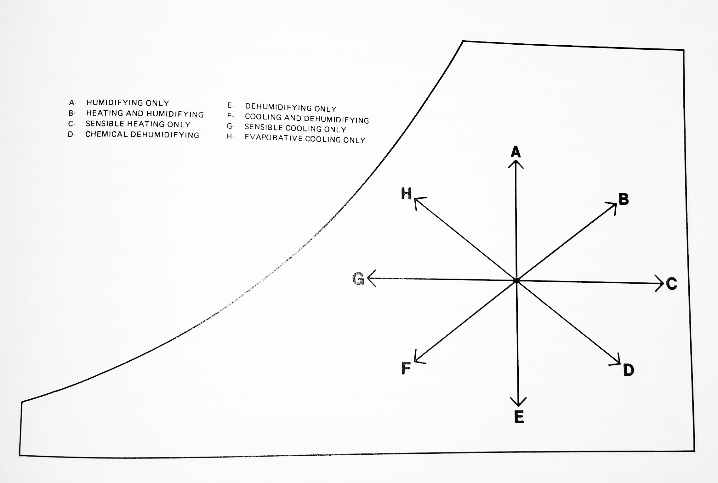
### Tier 3 Mechanical equipment

Mechanical heating and cooling systems may be eventually required in a design but they should be regarded as the final layer in developing a climate responsive low-energy design. They could involve renewable energy.

They are costly to install and costly to run and contribute to green-house gas emissions.

linked very broadly with the psychrometric chart (Milne and Givoni, 1979) high thermal mass, high mass with night-time ventilation, natural ventilation and mechanical ventilation, evaporative cooling, humidification, passive solar. Also provide a basic 8 direction navigational system for changing qualities on the psychrometric chart.





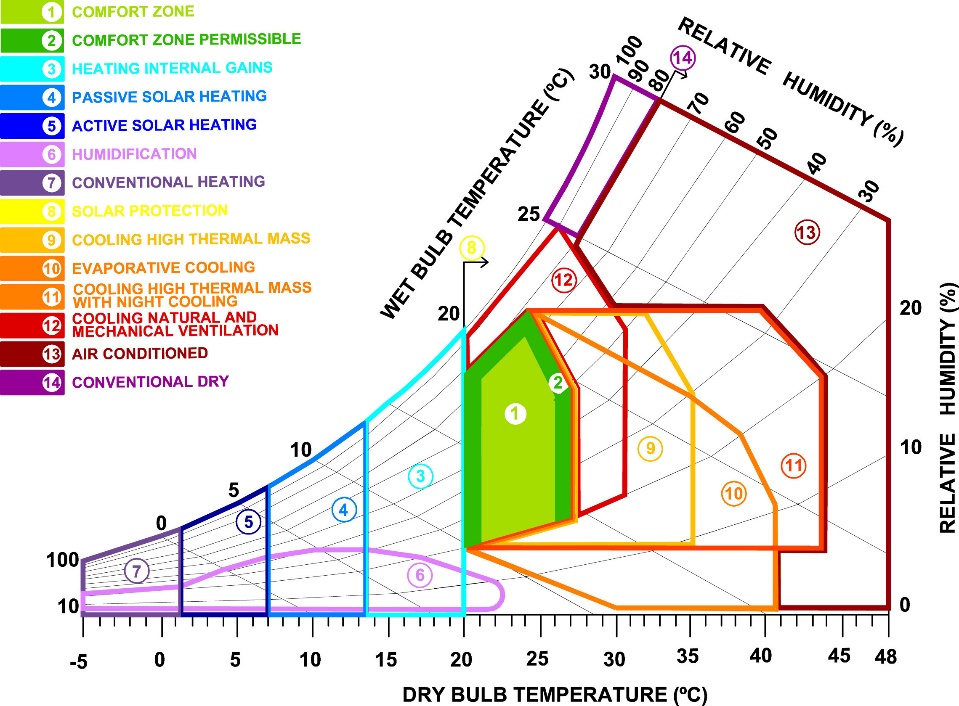
Aligning strategies with an adapted version of Givoni’s (Givoni, 1992) psychrometric chart (Manzano-Agugliaro *et al.*, 2015)

Givoni suggested an expanded comfort zone for hot developing centuries

Givoni in (Givoni, 1992) describes how the zones are defined

14 zones identified and each broken into a series of instrumental technological approaches, many of these approaches have multiple regionally specific implementations that are based on vernacular construction methods which given similar weather conditions can be transplanted. Other specific implementations are based on more experimental approaches

Boundaries are highly defined and could be viewed as arbitrary



The form of the psychrometric chart changes as air pressure changes due to the capacity of air to hold vapour this means high altitude locations need different charts than sea level locations. locating this zones in a constant way is complex

# Workflow management for scientific big data systems

Attempts to generalise big data workflows in eScience (Zhou, He and Ibrahim, 2016) provide a

## taxonomy of services:

Infrastructure: 3 types identified for eScience Grid, grid + virtualisation and cloud. HPC grid setups can be efficient for large frequent data transfers over many nodes, but grids don’t allow low cost scalability and easy access associated with clouds

Ownership: private, public, hybrid and federated

Applications: Climate and earth sciences if one of the 4 key application identified that are shifting focus to larger scale datasets – cloud platforms enable efficient analysis. Choice of platform is rarely technically justified in studies, furthermore lessons from one application may not be applicable in other projects or with other providers.

Processing tools provide workflow management systems WMS (See Pegasus and Kepler) that enable the composition and execution of analysis on distributed resources. They use Direct Acyclic Graph Manager and Condor to manage acquiring resources and schedule tasks. WMSs and resource acquisition software must be separately deployed which can be complex. Docker a container technique provides independence between applications and infrastructure. MapReduce is popular in e science data-intensive analysis. Two major storage options; files in file systems or databases. Scalable and efficient distributed file systems combine with scientific data formats to offer useful data storage.

Storage: Bigtable based databases and array based SciDB have been successfully applied to many e-science applications including environmental studies. Datasets are huge but often static and frequent updates are not required, blob based and Amazon’s S3 are ideal for e-science.

Security: balance of security of sensitive data and shared collaborative access

Service models; eScience on IaaS, PaaS and SaaS. Large data sets and long running jobs, mean cloud resources need to be carefully manged to optimised cost

Collaboration: sharing storage and or sharing computation

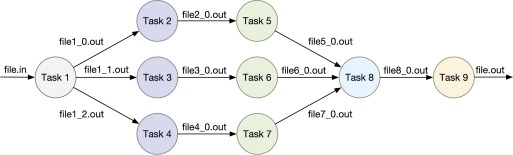
Resource provisioning – non-trivial to determine the instance type for each task of a workflow. Authors proposed a solution

## Open problems:

data lock-in S3 data cannot be used easily in azure. Performance unpredictability, Data confidentiality and auditability, lack of common infrastructures, resource management on future clouds.

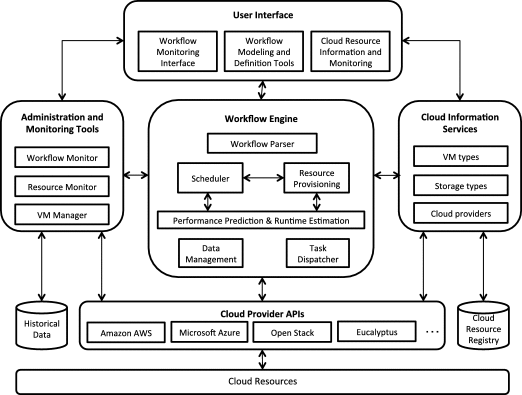
Workflow for scientific big data management applicable to weather data (Rodriguez and Buyya, 2017). WMSs provide data management and provenance, task scheduling, resource provisioning and fault tolerance

Scientific workflows support large scale complex processes and involving conducting experiments, proving hypothesis by management, analysis, simulation and visualisation of scientific data.



Workflows enable specification of loops, conditions and directed graph constructs. See Askalon, Kepler, Taverna, Pegasus motivated by the cost-effectiveness these WMSs are being extended to work with cloud platforms.

## General architecture model for WMS



UI provides interface for creating and editing workflows, Workflow Monitor allows observation of the execution progress of multiple workflows in terms of task’s status (ready, executing, staging, completed). Resource monitor information on available resources.

Workflow engines handles schedule, despatch, monitoring and management of tasks execution on remote sources

Implemented using Apache JClouds provides portable abstractions for cloud specific features

Workflow as a service WaaS (Esteves and Veiga, 2016)

Overlap between big data programming models and scientific data programming models is believed to add complexity to WMS (Deelman *et al.*, 2018). Aspects WMS must manage from extreme scale workflows identified

# Knowledge Discovery

KD involves analytic techniques that aim to extract useful knowledge from data, however, their potential has not yet been realised in climate science (Bracco *et al.*, 2017)

Knowledge Discovery in Data is broadly defined as (Begoli and Horey, 2012)1. Collection, storage and organisation of data. 2. Understanding and application of analytic methods. 3. Understanding the problem domain

Three principles for KDD (Begoli and Horey, 2012) 1. Support of analysis methods. 2. One size does not fit all. 3. Make data accessible

outlier analysis, clustering, prediction, classification and association rules mining techniques were applied to a 9 year weather data set for the Gaza strip to try and discover useful knowledge (Kohail and El-halees, 2011)

# Big data Tools

## General applications of Hadoop on Weather Data

Using Apache Spark to process at-rest weather data to determine averages for a range of variables across a set of weather stations static (Jayanthi and Sumathi, 2017).

Using Hadoop Spark to find similar weather conditions using the NOAA hourly land based data set. Used Euclidean distance of weather attributes (Rodenburg and Maria Fiore, 2017)

MapReduce compared to Spark to calculate minimum, maximum and average values of weather parameters (Chouksey and Chauhan, 2017). Using data from NOAA hurly weather data. For smaller datasets differences in performance were less pronounced, above 8GB Spark was faster.

Hadoop and MapReduce used on NOAA data (Dagade *et al.*, 2015) to compare performance of Pig and Hive to average data for each station for a single variable.

Another MapReduce used on NOAA data (Varghese and Riyaz, 2015) but the analytics used are not discussed

K-means with MapReduce runs in parallel to cluster weather data from China (Fang *et al.*, 2014). Tested different datasets to see processing speeds (250mb-2GB) and compared to another clustering algorithm.

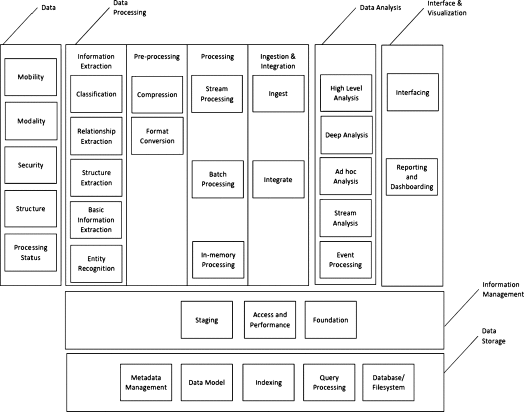
## More specific applications

Implementation of a self-organising map (SOM) (a type of artificial neural network trained using unsupervised learning) on IoT data using Apache Spark (Jayaratne *et al.*, 2017). SOM’s have also been used to extract features from data (Liu, Weisberg and Mooers, 2006) and applied to metrology and oceanography (Liu and Weisberg, 2011)(Liu and Weisberg, 2005).

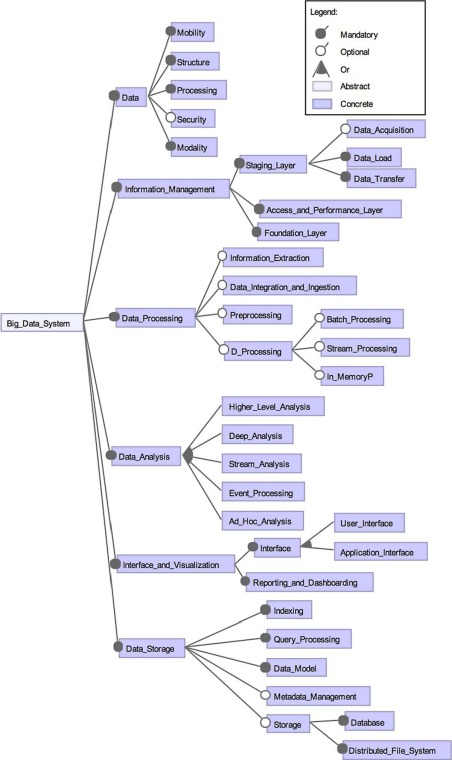
HBase for storage and MapReduce plus prediction with an artificial neural network and k-means clustering was implemented to support decision making in traffic regulations in Marrakesh based on air quality data (El Fazziki *et al.*, 2015).

# Big data application architecture and components

Big data feature model and Domain-Driven Design is discussed (Avci Salma, Tekinerdogan and Athanasiadis, 2017). A feature model, generic reference architecture and design rules are used to develop an application architecture.

**

Big data reference architecture from (Avci Salma, Tekinerdogan and Athanasiadis, 2017)



Feature model from (Avci Salma, Tekinerdogan and Athanasiadis, 2017)

## Components

Practical description of using apache spark to process at-rest weather data set and determine averages of a range of variables across a set of stations (Jayanthi and Sumathi, 2017).

Big data platforms have been developed for meteorological purposes (Chouksey and Chauhan, 2017) (Fang *et al.*, 2014),

while these provide valuable architectural descriptions for management and basic data analysis they fall short of providing platforms that offer the possibility of KD from weather data.

# Spatiotemporal data mining

## Properties

of the ST data mining that differentiate from classical data mining (Atluri, Karpatne and Kumar, 2017)

### Heterogeneity

in space and time varying ways and levels

### Auto-correlation

two nearby locations are not independent but are correlated

### Data types

event, trajectory, point reference data, raster

Raster data as fixed points but can be: regular space, irregular space, regular time, irregular time

### Data instances

classical data mining uses features with labels in ST data mining instances can be defined as points, trajectories, time series, spatial maps, and rasters.

Similarity between instances as a way of clustering, classification, pattern discovery and relationship mining

## Problems and methods

* Clustering on instances - relevant
* Predictive learning
* Frequent pattern mining
  + Motif Patterns in Time-series
  + network motif discovery – sets of distant locations experiencing similar climatic conditions with consistent temporal activity
* Anomalies ST-DBSCAN
* Change detection – transition from nino to nina
* Relationship mining

## Future work

* Novel representations of ST raster data
* Mining multi-modal data sets – in climate science different variables at the same space and time resolution
* Granularity – risk of overlooking substructures due to over or under partitioning
* To include domain theory to accelerate discovery

ST data properties are confirmed (Faghmous and Kumar, 2014) who also describe the unique characteristics of climate data that are:

Phenomena are not fixed objects but patterns that evolve over space and time

Biases exist in the sampling and measurement, some datasets are products of merged data

Highly variable due to natural fluctuations, measurement errors, model parametrization and representation.

When data is available at different spatiotemporal resolutions from different sources it is heterogenous

Climate data functions with interdependencies at multiple scales global studies cannot be used to understand long-term local impacts

Relationships in climate data may be long range and multivariate, many space-time-variable subsets where relationships can exist

Knowledge is required to define where and what to search for

Application backgrounds:

* query matching
* pattern mining – 3 types:
  + empirical orthogonal functions EOF based on PCA,
  + clustering (finding significant spatiotemporal clusters remains a major challenge because of both spatial and temporal variability)
  + user-defined patterns -storm /cyclone monitoring

Identifying events and anomalies, event causes long term change in patterns and anomaly is a short deviation from normal patterns. These differ from traditional data mining where events are unambiguous, in climate the pattern representing the event maybe unknown or base on spatiotemporal context.

### Relationship mining

linking changes in one variable to other phenomena

### Predictive modelling

beyond scope of project but interesting for future

### Network based analysis

use on gridded data sets to determine similarity and define network that can then be used for relationship mining, predictive modelling and pattern mining. Networks ca be applied to non-gridded data

Significance testing needs rethinking for climate data spatiotemporal randomization tests are required that do not break the inherent autocorrelation

Anomalies and extremes in climate data need to be understand as multivariate cumulative extremes. Performative measures are required to compare unsupervised STDM.

Opportunities for user defined pattern mining (thermal comfort patterns?) but domain knowledge is required so more unsupervised feature extraction methods to autonomously identify features may be preferable.

The need to integrate domain theory and expertise is acknowledged (Karpatne *et al.*, 2017) as a key strategy that could accelerate knowledge discovery in climate data.

## Specific ST data methods applied to climate data

Knowledge discovery applicable methods:

* Clustering methods for climate classification (Forsythe, Blenkinsop and Fowler, 2015) (Netzel *et al.*, 2016)
* Self-organising maps (Jayaratne *et al.*, 2017) (Liu, Weisberg and Mooers, 2006)
* Delta-maps (Fountalis, Bracco and Dovrolis, 2014) (Bracco *et al.*, 2017) group nodes in a network according to homogeneityapplied to precipitation and sea surface temperatures. Shows methods for robustness analysis of networks (link maps, area strength, s-core decomposition)

## Follow up on STDM

Spatio-temporal clustering (Kisilevich *et al.*, 2009) – main focus on ST trajectories but described data as georeferenced variables

Spatiotemporal data mining in the era of big spatial data (Vatsavai *et al.*, 2012) – general description of challenges some methods and specific applications – discusses climate data and the specific difficulties of multiple scales, complex dependencies, non-linearity, need for long term predictions. But states the potential for pattern discovery.

* also how the auto-correlation means that regression and weighted regression techniques need to be adapted through geographical weights or spatial and temporal constraints

and (Chandola *et al.*, 2015) – unavailable via Liverpool lib

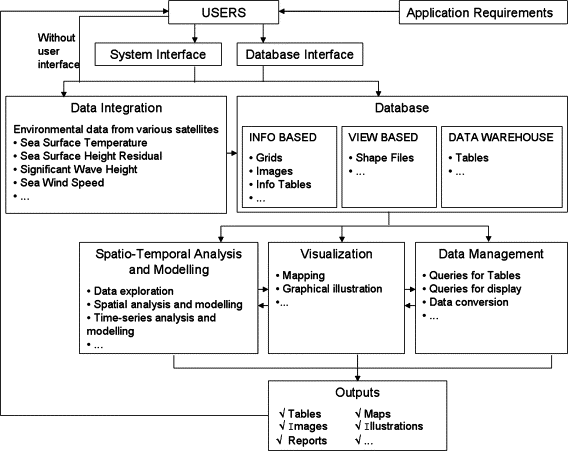
Spatiotemporal data mining (Cheng *et al.*, 2014) - STDM approaches for prediction, clustering, and visualization problems in several applications.

States that:

* artificial neural networks (ANNs) and support vector machines (SVMs) are now being successfully applied
* results are useless unless understood – need for space-time visualization so experts can access results Integration of data exploration, analysis and visualization is required
* Ref STARS: Space-Time Analysis of Regional Systems
* Autocorrelation again how geographic proximity means similarity typically
* **Ref Hsieh WW (2009) Machine learning methods in the environmental sciences: neural networks and Kernels**
* book “Machine Learning for Spatial Environmental Data”
* mention of use of nearest neighbour regression, kernel (ridge) regression, Gaussian processes, self-organizing maps (SOM), principal components analysis (PCA) and regression trees
* **Ref: Machine learning for spatial environmental data: theory, applications, and software**
* Clustering by thematic attributes, spatial, or temporal

Complex networks for climate model evaluation with application to statistical versus dynamical modelling of South American climate (Feldhoff *et al.*, 2014) – intense description of applying complex networks to model regional climate

ST-DBSCAN: An algorithm for clustering spatial–temporal data (Birant and Kut, 2007) – looks like a really use algorithm. Also an application is discussed that applies it to clustering seawater characteristics possible ref for system architecture



Extreme Learning Machines for spatial environmental data (Leuenberger and Kanevski, 2015) supervised system based on ANN, for chemicals in lake Geneva

Spatiotemporal data mining: A computational perspective (Shekhar *et al.*, 2015) good overview of problems and possibilities faced when mining ST data sets. Claims to have a statistically foundation when compared to other surveys. Could provide statistical tests for evaluation of results

* **Ref : Handbook of Spatial Statistics**
* ST couplings and tele-couplings – represent object instances that occur in close geographic and temporal proximity
* Partitioning and summarization – confirmation of use of clustering methods to partition space and time

SPARK -clustering methods:

* K-Means
* Power iteration clustering a form of spectral clustering
* Gaussian mixture see: (Lee *et al.*, 2010)
* Bisecting k-means

SEE Spark Graph X for the network possibilities

SEE The SAGE Handbook of Spatial Analysis

# References

ASHRAE (2013) *2013 ASHRAE Handbook: Fundamentals*, *ASHRAE*. doi: 10.1163/ej.9789004155947.i-937.23.

Assunção, M. D. *et al.* (2015) ‘Big Data computing and clouds: Trends and future directions’, *Journal of Parallel and Distributed Computing*. Academic Press, 79–80, pp. 3–15. doi: 10.1016/J.JPDC.2014.08.003.

Atluri, G., Karpatne, A. and Kumar, V. (2017) ‘Spatio-Temporal Data Mining: A Survey of Problems and Methods’, *ACM Comput. Surv*, 1(1). doi: 10.1145/nnnnnnn.nnnnnnn.

Avci Salma, C., Tekinerdogan, B. and Athanasiadis, I. N. (2017) ‘Chapter 4 – Domain-Driven Design of Big Data Systems Based on a Reference Architecture’, in *Software Architecture for Big Data and the Cloud*, pp. 49–68. doi: 10.1016/B978-0-12-805467-3.00004-1.

Begoli, E. and Horey, J. (2012) ‘Design principles for effective knowledge discovery from big data’, in *Proceedings of the 2012 Joint Working Conference on Software Architecture and 6th European Conference on Software Architecture, WICSA/ECSA 2012*, pp. 215–218. doi: 10.1109/WICSA-ECSA.212.32.

Birant, D. and Kut, A. (2007) ‘ST-DBSCAN: An algorithm for clustering spatial-temporal data’, *Data and Knowledge Engineering*, 60(1), pp. 208–221. doi: 10.1016/j.datak.2006.01.013.

Bracco, A. *et al.* (2017) ‘Advancing climate science with knowledge-discovery through data mining’, *npj Climate and Atmospheric Science*, 1(1), p. 4. doi: 10.1038/s41612-017-0006-4.

Chandola, V. *et al.* (2015) ‘Chapter 10 - Analyzing Big Spatial and Big Spatiotemporal Data: A Case Study of Methods and Applications’, in *Big Data Analytics*, pp. 239–258. doi: http://dx.doi.org/10.1016/B978-0-444-63492-4.00010-1.

Cheng, T. *et al.* (2014) *Spatiotemporal data mining*, *Handbook of Regional Science*. doi: 10.1007/978-3-642-23430-9\_68.

Chouksey, P. and Chauhan, A. S. (2017) ‘Weather Data Analytics using MapReduce and Spark’, *International Journal of Advanced Research in Computer and Communication Engineering*, 6(2). doi: 10.17148/IJARCCE.2017.6210.

Dagade, V. *et al.* (2015) ‘Big Data Weather Analytics Using Hadoop’, *International Journal of Emerging Technology in Computer Science & Electronics*, 14(2), pp. 976–1353. Available at: https://pdfs.semanticscholar.org/f2e4/918444be9b30f29132e93ce02d29ccf26eda.pdf (Accessed: 29 May 2018).

de Dear, R. (1998) ‘Developing an adaptive model of thermal comfort and preference’, *ASHRAE Transactions*, 104. Available at: https://cloudfront.escholarship.org/dist/prd/content/qt4qq2p9c6/qt4qq2p9c6.pdf (Accessed: 29 May 2018).

Deelman, E. *et al.* (2018) ‘The future of scientific workflows’, *The International Journal of High Performance Computing Applications*, 32(1), pp. 159–175. doi: 10.1177/1094342017704893.

Dimoudi, A. and Tompa, C. (2008) ‘Energy and environmental indicators related to construction of office buildings’, *Resources, Conservation and Recycling*, 53(1–2), pp. 86–95. doi: 10.1016/j.resconrec.2008.09.008.

Esteves, S. and Veiga, L. (2016) ‘WaaS: Workflow-as-a-Service for the Cloud with Scheduling of Continuous and Data-Intensive Workflows’, *The Computer Journal*. Oxford University Press, 59(3), pp. 371–383. doi: 10.1093/comjnl/bxu158.

Faghmous, J. H. and Kumar, V. (2014) ‘Spatio-temporal Data Mining for Climate Data: Advances, Challenges, and Opportunities’, in Chu, W. (ed.) *Springer Berlin Heidelberg*, pp. 83–116. doi: 10.1007/978-3-642-40837-3\_3.

Fang, W. *et al.* (2014) ‘Meteorological data analysis using MapReduce.’, *The Scientific World Journal*, 2014, p. 646497. doi: 10.1155/2014/646497.

El Fazziki, A. *et al.* (2015) ‘A multi-agent framework for a hadoop based air quality decision support system’, in *CEUR Workshop Proceedings*, pp. 45–59.

Feldhoff, J. H. *et al.* (2014) ‘Complex networks for climate model evaluation with application to statistical versus dynamical modeling of South American climate’, *Climate Dynamics*, pp. 1–15. doi: 10.1007/s00382-014-2182-9.

Forsythe, N., Blenkinsop, S. and Fowler, H. J. (2015) ‘Exploring objective climate classification for the Himalayan arc and adjacent regions using gridded data sources’, *Earth System Dynamics*, 6(1), pp. 311–326. doi: 10.5194/esd-6-311-2015.

Fountalis, I., Bracco, A. and Dovrolis, C. (2014) ‘Spatio-temporal network analysis for studying climate patterns’, *Climate Dynamics*, 42(3–4), pp. 879–899. doi: 10.1007/s00382-013-1729-5.

Givoni, B. (1992) ‘Comfort, climate analysis and building design guidelines’, *Energy and Buildings*, 18, pp. 11–23. Available at: https://ac-els-cdn-com.liverpool.idm.oclc.org/037877889290047K/1-s2.0-037877889290047K-main.pdf?\_tid=67644909-7d5e-4718-b47f-80732d024251&acdnat=1527603110\_130b2a2e77cd97d48361202940532374 (Accessed: 29 May 2018).

Hausladen, G., Liedl, P. and De Saldanha, M. T. A.-T. T.- (2011) ‘Building to Suit the Climate : a Handbook.’ Basel : De Gruyter,. Available at: http://ebookcentral.proquest.com/lib/columbia/detail.action?docID=1020450.

Jain, H. and Jain, R. (2017) ‘Big data in weather forecasting: Applications and challenges’, in *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*. IEEE, pp. 138–142. doi: 10.1109/ICBDACI.2017.8070824.

Jayanthi, D. and Sumathi, G. (2017) ‘Weather data analysis using spark — An in-memory computing framework’, in *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*. IEEE, pp. 1–5. doi: 10.1109/IPACT.2017.8245142.

Jayaratne, M. *et al.* (2017) ‘Apache spark based distributed self-organizing map algorithm for sensor data analysis’, in *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*, pp. 8343–8349. doi: 10.1109/IECON.2017.8217465.

Karpatne, A. *et al.* (2017) ‘Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data’, *IEEE Transactions on Knowledge and Data Engineering*, 29(10), pp. 2318–2331. doi: 10.1109/TKDE.2017.2720168.

Kisilevich, S. *et al.* (2009) ‘Spatio-temporal clustering’, in *Data Mining and Knowledge Discovery Handbook*, pp. 855–874. doi: 10.1007/978-0-387-09823-4\_44.

Kohail, S. N. and El-halees, A. M. (2011) ‘Implementation of Data Mining Techniques for Meteorological Data Analysis ( A case study for Gaza Strip )’, *International Journal of Information and Communication Technology Research*, 1(3), pp. 96–100.

Lechner, N. (2009) *Heating, Cooling, Lighting : Sustainable Design Methods for Architects*. 4th edn. Hoboken, New Jersery: Wiley.

Lee, K. J. *et al.* (2010) ‘Application of the Gaussian mixture model in pulsar astronomy-pulsar classification and candidates ranking for Fermi 2FGL catalog’, *Mon. Not. R. Astron. Soc*, 000, pp. 1–9. Available at: https://arxiv.org/pdf/1205.6221.pdf (Accessed: 8 August 2018).

Leuenberger, M. and Kanevski, M. (2015) ‘Extreme Learning Machines for spatial environmental data’, *Computers and Geosciences*, 85, pp. 64–73. doi: 10.1016/j.cageo.2015.06.020.

Liu, Y. and Weisberg, R. H. (2005) ‘Patterns of ocean current variability on the West Florida Shelf using the self-organizing map’, *Journal of Geophysical Research: Oceans*, 110(6), pp. 1–12. doi: 10.1029/2004JC002786.

Liu, Y. and Weisberg, R. H. (2011) ‘A Review of Self-Organizing Map Applications in Meteorology and Oceanography’, in Igadwa Mwasiagi, J. (ed.) *Self Organizing Maps - Applications and Novel Algorithm Design*. www.intechopen.com. doi: 10.5772/13146.

Liu, Y., Weisberg, R. H. and Mooers, C. N. K. (2006) ‘Performance evaluation of the self-organizing map for feature extraction’, *Journal of Geophysical Research: Oceans*, 111(5). doi: 10.1029/2005JC003117.

Manzano-Agugliaro, F. *et al.* (2015) ‘Review of bioclimatic architecture strategies for achieving thermal comfort’, *Renewable and Sustainable Energy Reviews*. Pergamon, 49, pp. 736–755. doi: 10.1016/J.RSER.2015.04.095.

Milne, M. and Givoni, B. (Barukh) (1979) ‘Architectural design based on climate’, in Watson, D. (ed.) *Energy conservation through building design*. New York, NY, USA: McGraw-Hill, pp. 96–113.

Milne, M., Liggett, R. and Benson, A. (2009) ‘Climate Consultant 4.0 develops design guidelines for each unique climate’, *American Solar Energy Society Meeting*. Available at: http://www.energy-design-tools.aud.ucla.edu/papers/ases09-milne.pdf (Accessed: 10 April 2018).

Netzel, P. *et al.* (2016) ‘On Using a Clustering Approach for Global Climate Classification’, *Journal of Climate*, 29(9), pp. 3387–3401. doi: 10.1175/JCLI-D-15-0640.1.

Olgyay, V. and Olgyay, A. (2015) *Design With Climate: Bioclimatic Approach to Architectural Regionalism*. Princeton University Press.

Omer, A. M. (2008) ‘Energy, environment and sustainable development’, *Renewable and Sustainable Energy Reviews*, 12, pp. 2265–2300. doi: 10.1016/j.rser.2007.05.001.

Rodenburg, B. and Maria Fiore, M. (2017) *Detecting Weather Twins using Apache Spark*, *LSDE: Large Scale Data Engineering 2017*. Available at: https://event.cwi.nl/lsde/2017/showcase\_n2.shtml (Accessed: 29 May 2018).

Rodriguez, M. A. and Buyya, R. (2017) ‘Chapter 18 – Scientific Workflow Management System for Clouds’, in *Software Architecture for Big Data and the Cloud*, pp. 367–387. doi: 10.1016/B978-0-12-805467-3.00018-1.

Shekhar, S. *et al.* (2015) ‘Spatiotemporal data mining: A computational perspective’, *ISPRS International Journal of Geo-Information*, 4(4), pp. 2306–2338. doi: 10.3390/ijgi4042306.

Varghese, S. M. and Riyaz, P. . (2015) ‘Leveraging Map Reduce With Hadoop for Weather Data Analytics’, *IOSR Journal of Computer Engineering Ver. II*, 17(3), pp. 2278–661. doi: 10.9790/0661-17320612.

Vatsavai, R. R. *et al.* (2012) ‘Spatiotemporal data mining in the era of big spatial data’, *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data - BigSpatial ’12*, pp. 1–10. doi: 10.1145/2447481.2447482.

Zhou, A. C., He, B. and Ibrahim, S. (2016) ‘Chapter 18 – eScience and Big Data Workflows in Clouds: A Taxonomy and Survey’, in *Big Data*, pp. 431–455. doi: 10.1016/B978-0-12-805394-2.00018-0.