Advances in self-organizing maps for spatiotemporal and nonlinear systems

Author:
Clark, Stephanie

Publication Date:
2018

DOI:
https://doi.org/10.26190/unsworks/20825

Advances in self-organizing maps for spatiotemporal and nonlinear systems

Stephanie Clark

A thesis in fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Mathematics and Statistics

Faculty of Science

University of New South Wales, Sydney, Australia

January 2018

THE UNIVERSITY OF NEW SOUTH WALES

Thesis/Dissertation Sheet

Surname: Clark	*Abbreviation for degree:* PhD
First name: Stephanie	School of Mathematics and Statistics
Middle name: Robyn	Faculty of Science

Title: **Advances in self-organizing maps for spatiotemporal and nonlinear systems**

Abstract

This thesis is aimed at enhancing the use of self-organizing maps (SOMs) within water-related research. A type of artificial neural network, the SOM is proficient at dimension reduction and clustering of large data sets to reveal underlying patterns. Innovations in processes such as the SOM for extracting patterns from large quantities of water-related data will provide more informative, concise and intuitively-understood messages, leading to a better knowledge of the relationships between system components and more informed management decisions. The SOM was chosen as the technique of interest as it is a nonparametric, highly-intuitive process of data analysis with an inherent visualisation of intercomponent relationships that allows messages to be immediately communicated and integrated into decision-making processes. A literature review and experimental modelling revealed potential areas for advancement in the current SOMs method with respect to:

- the representation of evolving temporal trends in cluster dynamics,
- pattern extraction from nonlinear data (more than slightly nonlinear),
- the choice of size and shape of the map grid to best represent a data set,
- identification of pathways traced by individual data items in a shifting temporal cluster structure, and
- the successful integration of new SOMs theoretical ideas into practical applications.

These gaps have been addressed through a series of five papers in this thesis. At the time of thesis submission, the first three papers have been published, the fourth is under review, and the fifth has been submitted. Advances in the SOMs method presented in each paper are demonstrated on real-world applications, focusing on the analysis of spatiotemporal global water resource dynamics at country, basin and city scales. Through this sequence of papers, this thesis contributes a closely-tied set of advancements focused on improving the extraction and interpretation of messages within large, high-dimensional, short time-step, nonlinear, water-related data sets.

Declaration relating to disposition of project thesis/dissertation

I hereby grant to the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or in part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all property rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstracts International (this is applicable to doctoral theses only).

		September 30, 2017
…………………………………………………..…	……………………………………..………………	……..……………………..…
Signature	Witness Signature	Date

The University recognises that there may be exceptional circumstances requiring restrictions on copying or conditions on use. Requests for restriction for a period of up to 2 years must be made in writing. Requests for a longer period of restriction may be considered in exceptional circumstances and require the approval of the Dean of Graduate Research.

FOR OFFICE USE ONLY Date of completion of requirements for Award:

ORIGINALITY STATEMENT

‘I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.’

Signed ……………………………………………..............

Date ……………………………………………..............

COPYRIGHT STATEMENT

Signed ……………………………………………..........................

Date ……………………………………………..........................

AUTHENTICITY STATEMENT

Signed ……………………………………………..........................

Date ……………………………………………..........................

Table of Contents

1 INTRODUCTION

Innovative methods for the analysis of water-related data sets have the potential to improve the extraction of interesting and useful information from the great quantity of collected hydrological data. Enhanced extraction of relevant features from the data will lead to better support of water resources management at local, regional and global scales. The volume of hydrological measurements gathered globally is continuously increasing as scientific advances ease the collection of remote and automatic measurements. The simplification and organisation of these immense, multivariate, hydrologic data sets is essential to allow patterns to be intuitively recognized and useful information to be extracted for easy incorporation into decision-making processes.

Organizing information, through summarizing and sorting, is the foundation of any processing and analytical task involving large data sets. The summarizing portion of this process refers to identification of the prevalent patterns and intercomponent relationships in the data, and the sorting portion entails clustering similar pieces of information together. These processes can be tied together to provide a system of data organization. Statistical learning methods, such as artificial neural networks, use such a system to organise data through the recognition of patterns within high-dimensional data sets based on the identification of the cluster structure in the data.

The self-organizing map (SOM, Kohonen, 1990) is an artificial neural network proficient at extracting and ordering the prevalent patterns in a data set, sorting the data in accordance with these patterns and conveying the information through meaningful mappings into low dimensional space. Unique combinations of multiple nonlinearly related variables, which comprise the common patterns or states of a system, are identified. Through a nonlinear projection to low-dimensional space, an order is established for the extracted patterns that best preserves the topology of the data structure. These patterns form the basis for the clustering of data items into groups sharing meaningful similarities. The clusters of similar data items become situated on the low-dimensional projection, or map, so that inter-cluster distance expresses a measure of dissimilarity. Interpretation of the map reveals information about the structure of the data set, correlations between variables and the cluster configuration in the data. The combination of a nonlinear projection from high-dimensional to low-dimensional space, the preservation of data topology and an ordered clustering provide distinct benefits of the SOM over other neural networks (Yin, 2005).

The SOM is a widely used method in water-related research due to its intuitive implementation, resilience to missing and noisy data, ability to integrate real-time data, and straightforward visual summary of the system and intercomponent relationships. Of particular benefit is the ability of the SOM to organize data with no requirement for an explicit understanding and description of any complex underlying systems that may have produced the data. This attribute is valuable in exploring multivariate environmental data for which the creation of realistic system models can be time-consuming and complex. The SOM naturally integrates cross-disciplinary data in a non-biased manner, facilitating understanding and interaction between collaborators from diverse fields. Relationships of hydrologic variables to physical, chemical, social or economic systems in multidisciplinary research can be investigated without requiring expert knowledge from each of the varied disciplines. The SOM is able to incorporate new data as it is measured, without requiring a retraining of the model, enabling real-time data analysis. The intuitively comprehended visualization allows the extracted information to be directly integrated into decision-making processes. For these reasons, the SOM is often used for a wide range of data organisation purposes by scientists, engineers and researchers interested in exploring the physical and chemical processes of hydrology and water resources, as well as social and economic associations to these processes.

The combination of data mining innovations with appropriate field expertise provides the ability to retrieve a desired amount of information through identification of an appropriate level of data summarisation. Well-produced SOMs are able to provide decision makers with an overall impression of the structure of a data set at a suitable level of abstraction and insight into the intervariable relationships. Traditionally, a low percentage of the vast amount of environmental measurements that are collected is actually used, due to a lack of efficient and effective analysis tools for processing the data (Liu & Weisberg, 2011); though new data analysis techniques are beginning to provide means to convert unprecedented amounts of data into useful information that can drive development (website [1]). Cottrell et al. (2016) note that 'as the computational complexity of the SOM algorithm is low compared with the number of data items it can process, and it is particularly well suited to stream data, the SOM appears to have a great future ahead in a big data context'. Though an intuitive statistical procedure such as the SOM that accurately allows the user to discover common patterns and relationships without the complexity of mathematical modelling is valuable in multidisciplinary environmental studies, it will be effective only if the method is appropriately applied to the specific data and the results are suitable representations of the information contained within the data set.

Developments that may lead to an improved SOM method are suggested in the literature. Yang & Wu (2006) indicate that the visualisation of changes in data structures has not yet been properly addressed in data mining research, nor has the separation of the temporal and cross-sectional structure of a data set. Yin (2008) asserts that the potential of the SOM method is not realised in current applications which are often limited to empirically chosen parameters. Van der Maaten et al. (2007) articulate that the highly nonlinear, high-dimensional characteristics of real world data require appropriate analysis techniques, and emphasise that any improvements in methods need to be accessible to researchers and engineers. Kohonen (2008) identifies the representation of dynamic phenomena with SOMs as a prominent issue to be addressed. He also states that the determination of the number of map nodes that would best represent a data set is one of the most common questions arising from users (Kohonen, 2013). Abrahart et al. (2012) call for the identification of novel applications that can only be solved with neural networks, highlighting the importance of applying the methods to appropriate data analysis tasks.

A thorough review of recent literature pertaining to applications of SOMs in hydrology and water resources has been conducted to explore the current scope for advancements to the SOM method with respect to these fields. It has become clear that the improvements explicitly called for in the literature cited above are yet to be realised. Though most hydrological and water resource systems include a spatiotemporal component (referring to data in which there is a cross-sectional structure as well as a temporal one) and many also contain nonlinear manifolds (such as the fluctuating intervariable relationships generated from diurnal or season effects), the SOM method continues to encounter various limitations when applied to spatiotemporal and nonlinear data.

It is also evident that there is a disconnect between recent theoretical statistical research in the SOM method (published in neural network, machine learning and statistical journals) and practical applications of the SOM (published in environmental and engineering journals). Most technical advances are not adopted into the commercial and research realms; instead, users tend to favour heuristics and software defaults for parameter selection in the SOM process rather than making informed choices based on their data and purposes.

In consideration of the current state of SOM knowledge and water-related applications, this thesis focuses on advances to the SOM method through providing a series of improvements in:

- the representation of dynamic spatiotemporal data
- a method for deliberate, application-specific parameter selection,
- pattern extraction from highly nonlinear data,
- the visualisation of individual paths of data items though temporal shifts in the cross-sectional structure of the data,
- the successful integration of new SOMs theoretical ideas into applied research.

The overall objective of the thesis, through a combination of these improvements, is an enhanced extraction and visualization of information from large, high dimensional data sets to reveal patterns and clusters that are an accurate representation of the data structure.

This thesis is comprised of five papers concentrating on distinct aspects of these improvements. The flow between the papers is methodical and documented, with each paper using and building on developments from other papers in the thesis as well as drawing on concurrent related literature published by other researchers. Throughout the papers, associated statistical topics are discussed, including: spatiotemporal exploratory data analysis, clustering methods, temporal cluster trends and evolution, representation of outliers, time series analysis, nonlinear manifold learning and dimension reduction.

To support the expansion of theoretical advances into applied research, each of the developments introduced here is demonstrated on the extraction of meaningful information from a real, water-related data set. Data concerning the relationships of human populations with their freshwater resources generally contain difficult-to-define dynamic relationships, and vastly differing data sources and measurement techniques, making them especially well suited for analysis with the SOM. Adequate knowledge of these relationships can bring about more successful management of water extraction and consumption, trade, land use, disaster mitigation strategies, agricultural use, pollution prevention, development, and climate change mitigation. The clustering of spatiotemporal data containing nonlinear intervariable relationships and the identification of data items with similar or diverging trends are common themes in this thesis.

A range of scales is included in the application studies in this thesis. The literature indicates that studies at the global scale in the water sector have traditionally been restricted by a lack of global datasets and methods (Jongman et al., 2012) and yet potentially important regional water resource patterns may be masked by continental and global scale summaries (Vorosmarty et al., 2000). Therefore, each study in this thesis provides a global view of relationships between regional systems. With the introduced innovations in the SOM method, the global structure of each system is determined and, within it, regional patterns are analysed and compared. Global hydrologic connectivity is recognized and associations are visualised between areas of the world sharing similar conditions, be it countries, river basins or cities.

The structure of this thesis is as follows. In Section 2, the literature review, current state of SOMs and gaps in knowledge are summarized. Section 3 discusses the approach taken to address these gaps, the flow and connectivity of the overall project, information on publications and applications, and an introduction to each of the five papers comprising the thesis. Sections 4-8 contain the published and submitted papers in their original forms. The conclusion in Section 9 summarises the relationship between the papers, the discoveries, results, and improvements in the state of SOMs and water resources, and suggests directions for future research.

2 LITERATURE REVIEW SUMMARY

A broad literature review was conducted to explore a range of recent statistical methods and data exploration topics in hydrology and water resources. This review was refined into a thorough literature review encompassing theoretical advances in the statistical methodology of the SOM, and a wide variety of SOMs applications were read to determine the current state of SOMs practice and existing gaps in the knowledge base.

2.1 BROAD LITERATURE REVIEW OF STATISTICS FOR WATER-RELATED DATA

Techniques and ideas were investigated regarding the extraction and expression of information from large amounts of spatiotemporally variable water-related data. The literature read in this preliminary portion of the review highlighted many current topics of interest, and although most of this broad review is not included in this thesis, a few topics were encountered that shaped the overall direction of research. These were: spatiotemporal self-similarity and cluster analysis (eg., Bierman et al., 2011, Ruiz-Medina, 2012); data assimilation for the improved prediction of systems, discovery of anomalies and validation of models (eg., Mendoza et al., 2002, Fekete et al., 2002, Reichle, 2008, Decharme et.al., 2008, Xia et al., 2012, Houborg et al., 2012); determination of the interrelated effects of climate change and anthropogenic activities on rivers (eg., Shanmuganathan et al., 2006, Steynor et al., 2009, Gao et al., 2013); measurement of surface water extent, discharge and inundation from space (eg., Smith, 1997, Vorosmarty et al., 2005, Alsdorf et al., 2007, Papa et al., 2008, Pan et al., 2008, Dorigo et al., 2012, Chapman & Charantonis, 2017); Bayesian space-time models to address the mismatch in data sampling scales of spatial and temporal variability (eg., Kingston et al., 2005, Wikle et al., 1998, Chiu & Lehmann, 2011, Vanem et al., 2011); and the innovative uses of neural networks for water-related systems (Abrahart et al., 2012). Recent research using neural networks for trend visualization in other fields (outside hydrology) was also explored to identify interesting methods that may have the potential for adaptation to water-related applications. For instance, neural networks are widely used to reveal trends and identify clusters in financial time series analysis and knowledge domain visualisaton (eg. Tay et al., 2001, Fung et al., 2002, Yu et al., 2005, Wu et al., 2012, Powell et al., 2008, Skupin, 2004, Mothe et al., 2006, Segev & Cantola, 2012, Lee & Chen, 2012, Abe & Tsumoto, 2011, Borner, 2003, Skupin et al., 2013, He et al., 2005, Fenn et al., 2012).

The outcome of this broad review was the perception that recent research is tending to become increasingly concerned with methods for the incorporation of data from various sources and disciplines into data-driven analyses, for defining relationships and dependencies. In particular, the potential for neural networks to provide innovative spatiotemporal estimation and trend analysis was noted, with the possibility of incorporating information from in-situ, remote measurements and anthropogenic influences. The SOM neural network, specifically, appeared to exhibit an interesting capacity for data-driven, multi-objective analysis that is well suited to hydrological data, and provides an effective visual communication of the results.

2.2 SELF-ORGANIZING MAP CURRENT KNOWLEDGE BASE AND APPLICATIONS

A focused literature review on SOMs ensued, identifying basic theory, recent theoretical advances, current SOMs usage in water-related applications and potential areas for improvement in the method and applications. This review is presented in Paper 5**,** with specific aspects incorporated into Papers 1-4. The principal findings are summarised here.

General gaps in the knowledge base of SOMs have become evident in the following key areas:

1. **Spatiotemporal clustering.** Hydrological, water quality, and climate applications to date customarily apply the SOM to the extraction of data patterns and identification of clusters in either the spatial (cross-sectional) or the temporal domain. The preferred contemporary methods in the literature for spatiotemporal analysis are the creation of either: 1) a series of SOMs, one for each time step of the data set - these must be visually compared to extract patterns and trends, or 2) a single SOM produced with the entire data set on which trajectories or subsets of data (representing different areas or time periods) are mapped. These popular methods do not account for: 1) possible differences in the data structure at each time step, 2) the incomparability of maps created from data subsets of differing structures, or 3) the subjective interpretation of each user in the comparison phase. The self-organizing time map (SOTM, Sarlin, 2012), the first spatiotemporal SOM method to provide results on a single visualization, attempts to address these issues, but gaps remain: the map size and shape of the SOTM do not adapt to the data at each time step, the SOTM cannot process data sets with missing values, and it remains difficult to track individual data items through the evolution of the dynamic cluster structure. (Further background information on spatiotemporal clustering issues is provided in Papers 1, 3 and 5.)
2. **Parameter selection.** The map configuration (number and formation of nodes in the grid) is a choice to be specified by the user before each separate application of the SOM. The grid setup influences the patterns and clusters revealed on the final map (Kohonen, 2013), though little consolidated guidance is provided in the literature informing the selection process. Popular heuristics are often used, users 'borrow' parameters from previously published SOMs applications which are likely irrelevant for their current data set, or quality measures are applied to a series of maps trained with various parameters to determine which map best represents the data. It is widely accepted that SOMs have the potential to reveal more information from a data set with the use of carefully chosen parameters, however there is no consensus on the best method for parameter selection. (Details of this portion of the literature review are provided in Papers 2 and 5.)
3. **Representation of highly nonlinear intervariable relationships**. The SOM algorithm encounters limitations when attempting to represent data with nonlinear underlying manifolds (nonlinear in the sense of being more than a simple perturbation from linear) (Demartines & Herault, 1997; Shao et al., 2015). Due to the linear principal-component-based initialization method, the SOM method can tend to flatten a data structure during the projection stage, rather than unfolding it. In general, the map setup is chosen to represent a data set before any understanding of a possible submanifold is gained. Attempts in the literature to address this issue by tacking a

nonlinear initialization method on to the beginning of the SOM algorithm result in: a restriction of the movement of map nodes to the dimensions of map space (rather than data space), and the possibility that a single node may represent data from distinct parts of the geodesic surface. (Paper 3 contains a review of discussions in the literature on this issue.)

4. **Objective function minimisation.** It has been demonstrated in the literature that the SOM algorithm does not minimize a single objective function, but instead the training method attempts to optimize two competing objectives: quantization of the data items and preservation of the topology of the data set (Erwin et al., 1992; Yin, 2008a). The lack of an objective function is a popular source of discussion in the literature, as it restricts the possibility to select parameters based on current information theory techniques using maximum likelihood estimation. The literature contains many attempts to describe an objective function that the SOM may follow, or to alter the SOM algorithm to force it to follow a certain objective function. These methods are not being adopted by researchers however, and the traditional SOM continues to be used. (This issue is discussed in detail in Paper 2 and Paper 5.)
5. **Crossover of theory into applied research.** It has become evident through the review of a substantial number of SOMs papers that statistical innovations in the SOMs methodology are not being embraced by researchers concerned with using SOMs for practical applications. The theoretical research may be inaccessible to researchers who are interested only in the use of SOMs as a data exploration tool, and do not have the time to sift through and interpret numerous statistical papers or an interest in advancing the method themselves. This lack of crossover between theoretical innovations and practical applications is causing possibilities to be overlooked for SOMs finetuning that may lead to an improved representation of each specific data set.

The current state of SOMs in relation to hydrology and water resources research is:

- SOMs are an increasingly popular method within these fields, for exploratory data analysis of large, multivariate data sets, with approximately 180 articles currently published per year (website 2).
- SOMs water-related applications involve a wide range of pattern extraction, clustering, missing data infilling, prediction and time series analysis tasks.
- SOMs are particularly well suited to noisy data or data with missing values, as are common features of environmental measurements collected remotely and in the field.
- The SOM has the potential to be tailored to best suit a specific data set through the tweaking of map configuration and training parameters. However most hydrologic applications continue to use software defaults or arbitrarily chosen parameters, forfeiting some of the potential benefits of the method.
- Variants to the traditional SOM method, such as temporal SOMs and growing SOMs, are appearing widely in theoretical papers though rarely in water-related application literature, indicating that technical advances are not being transferred into practical implementations.

2.3 Organisation of literature review

Theoretical background information on the SOMs method including algorithmic details, recent advances in theory, a general literature review on SOMs usage in water-related and environmental research, and a comparison of SOMs with related methods is contained in Paper 5. The background and method sections of Papers 1-4 contain literature reviews focused on their specific areas of concentration, reviewing and discussing the state of contemporary research in each area.

The necessity for expressing the direct context of each paper of this thesis within past and contemporary research, combined with an attempt to avoid unnecessary repetition in the thesis as a whole, has led to a fragmented presentation of the literature review. Table 1 lists specific sub-areas that were covered in the literature review process and directs the reader to the relevant areas of the thesis in which they are reported.

Table 1: Literature review table of contents

Literature review topic	Discussed in thesis paper(s)
Spatiotemporal analysis and clustering	1, 4, 5
Temporal use and extensions of SOMs	1, 4, 5
SOM initialization	2, 3, 5
SOMs for missing data	1, 5
SOMs for geographic comparisons	1, 2, 4
Water-related SOMs hydrologic applications	5
SOM size and shape selection	2, 4, 5
SOMs quality measures	2, 4, 5
Trend prediction	5
Associated variables	1, 5
Comparison with related methods	5
Nonlinear manifold learning and dimension reduction	3
General SOM development	5
Cluster theory and cluster evolution	1, 2, 3, 5
Interpretation of the output map	1, 2, 5
SOM second-level clustering methods	1, 4, 5
Objective functions of the SOM	5
Probabilistic alternatives to the SOM	5

3 APPROACH

3.1 OVERVIEW

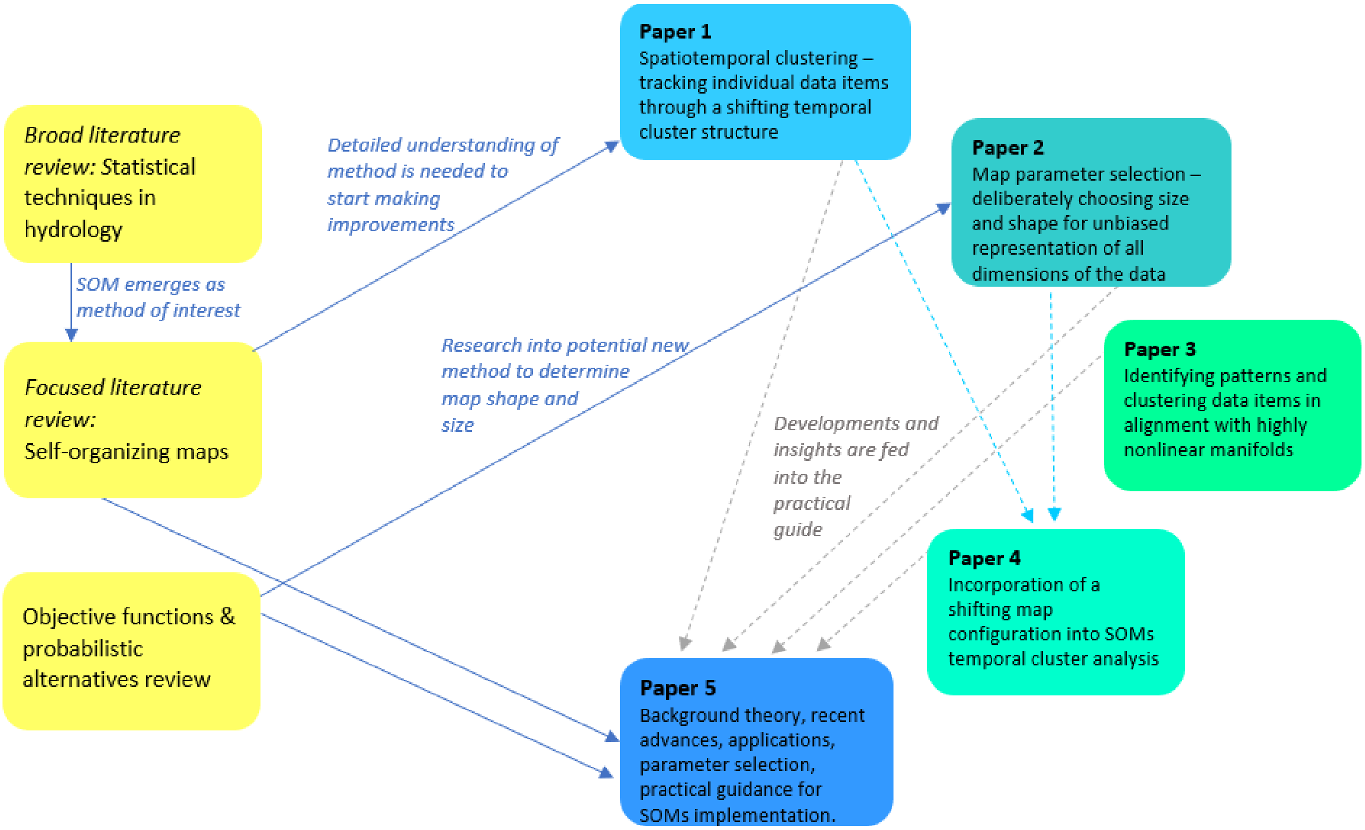
The investigations and innovations developed in this thesis were performed through the production of a series of distinct, consecutive projects, each presented in a separate paper (Papers 1 to 5). In general, the preparation of Papers 1 to 4 entailed:

- a focused literature review conducted into the state of SOMs in the area of concentration,
- a gap identified in the existing SOMs method,
- a technical innovation envisaged, troubleshooted and implemented in MATLAB, and
- a demonstration of the innovation on water resources or hydrologic data.

Paper 5 involved a different process. This paper was devised during the initial literature review, written concurrently with the other four papers to incorporate the base of knowledge gained during their production, and finalized at the completion of the thesis.

3.2 RELATIONSHIPS OF THESIS ELEMENTS

The flow of the project through the distinct elements of the overall thesis, and the links between these elements, are described in this section and depicted in Figure 1.



Broad literature review: Statistical techniques in hydrology
SOM emerges as method of interest
Focused literature review: Self-organizing maps
Objective functions & probabilistic alternatives review
Detailed understanding of method is needed to start making improvements
Research into potential new method to determine map shape and size
Paper 1
Spatiotemporal clustering – tracking individual data items through a shifting temporal cluster structure
Paper 2
Map parameter selection – deliberately choosing size and shape for unbiased representation of all dimensions of the data
Paper 3
Identifying patterns and clustering data items in alignment with highly nonlinear manifolds
Paper 4
Incorporation of a shifting map configuration into SOMs temporal cluster analysis
Developments and insights are fed into the practical guide
Paper 5
Background theory, recent advances, applications, parameter selection, practical guidance for SOMs implementation.

Figure 1: Flow of the project

The self-organizing map was selected as the subject of the thesis, after a broad review of literature concerning statistical methods in hydrology, due to its significant intuitive appeal and evident popularity amongst environmental researchers combined with the vagueness and ambiguity with which it is currently applied and the potential for improvements in both technical and applied aspects of the method.

It became apparent during the literature review that the most substantial gap in current SOMs practice concerns the representation of spatiotemporal data with SOMs. The majority of data sets in hydrology and water resources contain a temporal component as well as geographical (or otherwise cross-sectional), making this a notable issue. A goal was formed to develop a method to effectively present as much spatiotemporal information as possible on a single SOM, enabling the use of the popular SOMs attributes of data analysis and visualization without as much of the current requirement for subjective user analysis. A technique was developed in Paper 1 to produce a single output visualization that could track and compare the trajectories of individual data items through changes in a one-dimensional projection of the system cluster structure over time.

It also became evident during the literature review that many SOMs issues arise due to the lack of a best practice for determining the number and configuration of nodes in the output map. It was noted during the preparation of Paper 1 that unless a lot of care was taken, the one-dimensional map at each time step tended to span much more of the variance of one dimension than another. Also, each time step was represented by the same size of map regardless of the distribution of that subset of the data. A goal was outlined to pursue a relatively even coverage of each data dimension by the map through an update to the map size and shape selection process. It was initially anticipated this might be accomplished through the use of an objective function and maximum likelihood estimation; however, as the traditional SOM does not follow a single objective function and cannot be made to follow one whilst maintaining its fundamental goals of data quantisation and topological preservation, a novel method was required. In Paper 2, a method was developed and presented for determining an optimal number and configuration of map nodes to represent a data set with a minimum of user input. This new method quantifies the range of each dimension of the data represented by individual nodes on a series of potential maps. The information lost through the use of non-optimal SOM setups is quantified, aiding in selection of the map to best represent a specific data set.

During development of the method in Paper 2, it was found that the proposed technique worked well for generally ‘cloud shaped’ data sets in which the information of interest could be extracted through a direct ‘pressing’ of the data, but not for data sets with highly nonlinear manifolds that require a more careful ‘unrolling’. This is a considerable drawback in environmental sciences, which frequently contain nonlinear data, motivating Paper 3 which introduces a unique integration of nonlinear dimension reduction theory within the traditional SOM framework.

Paper 4 delves further into spatiotemporal and map size/shape issues, expanding on developments from Papers 1 and 2. This paper presents a high-interest water resources analysis through an updated temporal method incorporating the new map size and shape selection method at each time step. Temporal changes are revealed in the overall data structure and the usual requirement to represent the data at each time step by a map of the same size and shape is overcome.

The research and ideas in this thesis revolve around **Paper 5** in which the information needed to begin, as well as the knowledge gained throughout the process of researching the other four papers, is summarized. Consisting of the main parts of the literature review as well as a summary of practical experience gained through modelling, it was created as a publishable paper to transfer this knowledge forward and save future researchers considerable time and effort in piecing together background information, practicalities of the method, and mathematical details of the SOM from a wide variety of highly-focused papers.

3.3 APPLICATIONS

The applications included in the papers of this thesis apply the advances in SOMs techniques in each paper to statistical investigations of changing patterns of water use and water availability over time with nonlinearly related variables. A clear understanding of the intervariable relationships affecting freshwater resources is imperative to ensure sustainable management (Vorosmarty et al., 2000). The global distributions of runoff, water use, and scarcity are highly variable both spatially and temporally, and correspond poorly to the global population distribution (Postel et al., 1996), making the future adequacy of freshwater resources difficult to assess. Vorosmarty et al. (2005) express the need for new methods that are able incorporate interdisciplinary data sets to provide a complete understanding of human-water interactions and achieve sustainable environmental management.

Focussing on these issues, each project of this thesis applies the new methods to an analysis of nonlinear aspects of human/water interactions. Topics addressed include:

- water consumption and virtual water flows implicit to international trade,
- the attainment of Millennium Development Goals with respect to access to improved water and sanitation,
- water scarcity in river basins influenced by a combination of availability and management, and
- projected changes in urban flood impacts due to socioeconomic development and climate change.

Data used in the applications has been obtained from a variety of sources: UN databases (eg. website 3), satellite data/remote sensing (eg. website 4), institutional published data sets (eg. websites 5 and 6), and national hydrologic measurements (eg. website 7).

3.4 PUBLISHING OF PAPERS

The papers in this thesis were/will be published in international journals as a deliberate step to fulfil the objective of effectively influencing contemporary practice by making innovations immediately available. Ranging from purely statistical through a spectrum of numericallyinclined environmental journals, both traditional and open-source publications were chosen to provide exposure of the new developments to researchers, engineers and scientists using SOMs for research and applications. A list of the journals, with reasons for their selection, is given below.

Paper 1 *Journal:* Ecological Informatics
5-year impact factor: 2.29
Comments: This journal was chosen as it is concerned with the growing capacity of computational technology to harness complex data for the use in informing sustainable environmental management decisions. Paper 1 was published in a special issue on ecoinformatics decision support systems.

Paper 2 *Journal:* Pattern Recognition
5-year impact factor: 4.99
Comments: The official journal of the Pattern Recognition Society, this journal is dedicated to presenting papers with original contributions to theory, methodology and application of pattern recognition in fields including neural networks.

Paper 3 *Journal:* Environmental Modelling & Software
5-year impact factor: 4.98
Comments: This journal was chosen due to its aim to improve the representation and communication of the behaviour of environmental systems, with generalizable interdisciplinary techniques that provide insights into real-world applications and integrate modelling with environmental system management.

Paper 4 *Journal:* Hydrology and Earth System Sciences (requested revision stage)
5-year impact factor: 5.06
Comments: This journal is concerned with multi-disciplinary approaches concerning interactions between water, the earth and humans. It was selected due to its interactive public peer review discussion process and open access status, contributing to the goal of bridging the gap between SOMs theoretical advances and applications through improving the accessibility of new methods to researchers and commercial users.

Paper 5 Submitted

3.5 Introduction to papers

The papers are introduced in this section with paragraphs describing the motivation, methods, applications and contributions of each study. The full-length papers, as published or submitted, follow in Sections 4-8.

Research, modelling and manuscript preparation was performed by the first author. Other authors provided valuable reviews and suggestions for improvements to the manuscript. By nature, each paper needs to begin with a description of basic SOM methodology and recent relevant literature, and therefore some overlap in material will be found within the following chapters of this thesis.

Paper 1: Increasing dependence on foreign water resources? An assessment of trends in global virtual water flows using a self-organizing time map.

Journal: Ecological Informatics
Published: June 2014
http://dx.doi.org/10.1016/j.ecoinf.2014.05.012

The current use of SOMs for analysing temporal cluster evolution in spatiotemporal data is extended in this paper, with particular focus on tracking the individual data item within the changing global structure of the data. A method is introduced for following the flow of each data item through the shifting cluster structure. Motivated by the potential to improve upon the current practice of using a series of maps to represent spatiotemporal data, which requires significant effort from the user to track individual data items, the aim of this study is to provide an automated means to investigate the movement of data through the evolving data clusters with a single visualization as output.

Improvements are made to the self-organizing time map algorithm (SOTM, Sarlin, 2012) including: enabling the use of the SOTM with missing data, altering the training mechanism of each time step to provide an accurate snapshot of each sequential system state, and performing second-level clustering with a 1D SOM which ensures an ordering to the clusters and allows an indexed colour scheme to be used. A post-processing technique is created to assess changes in cluster memberships over time, determining which data items have converging and diverging circumstances, thereby providing relationship and trending information of individual data items within the context of the whole data set.

These improvements are demonstrated with an investigation into the relationship of country-level virtual water use combined with national renewable water resource assets for 172 countries. Fifty years of virtual water flow through international trade is compared to distinct national water resource situations. Countries are clustered into groups with similar states of dependence on foreign water resources, and the change in hydrologic dependency of each country is tracked and compared with the others. The association of national health, environmental, and socioeconomic variables with states of hydrologic self-sufficiency are investigated. The literature contains some regional and national virtual water balances, but this is the first visualisation of the evolution of global virtual water patterns over time in combination with information on available domestic water resources.

The primary contributions of this paper are: advancements in the SOTM and a method for identifying data items experiencing similar temporal trends through a shifting spatiotemporal cluster structure. This can be used as a decision support tool by establishing and communicating relationships and trends of individual entities in a global context.

Paper 2: A dimension range representation measure for self-organizing maps

Journal: Pattern Recognition
Published: November 2015
https://doi.org/10.1016/j.patcog.2015.11.002

In the preparation of Paper 1, it was noted that the preliminary map results based on the commonly used heuristics for choosing map parameters tended to reveal information essentially about only one variable. The map was observed to span a much higher percentage of the variance of this variable than the other, due to the attempt to reduce the highly nonlinear intervariable relationship to a single dimension. This was discovered through a tedious amount of manual checking, inspiring the search for an automated quality measure that would alert the user to this situation. A deliberate choice of appropriate map dimensions to represent a specific data set, which is commonly based on an assessment of quantization and topological map quality, could also benefit from an analysis of the spanning of the map over each data dimension.

A method is introduced in this paper for selection of the optimal number and configuration of map nodes to represent a multivariate data set, by exploring the range of each dimension that is represented by individual map nodes. A two-dimensional map grid 'draped' over a multidimensional data cloud will logically leave some edges of the data cloud uncovered, and this extent is measured here.

The real-world application in this paper is an investigation of access to improved rural and urban water and sanitation facilities in 142 countries, as provided by the Millennium Development Goals database. On this four-dimensional data set, it is demonstrated that using the measure introduced here in conjunction with other commonly used quality measures can improve the representation of data by the SOM, through ensuring that the distribution of each data dimension appropriately influences the size of the map in that direction.

The primary contribution of this paper is a quality measure that can be used to determine the optimal number and configuration of map nodes. The new measure eliminates the need for investigation of map coverage by visual comparison of the map and data, a process which quickly becomes infeasible for high-dimensional datasets. It is demonstrated that incorporating this quality measure into the selection of map setup parameters leads to an output map that more effectively reveals the characteristics of each variable.

Paper 3: Nonlinear manifold learning in natural systems

Journal: Environmental Modelling and Software
Published: March 2017
http://dx.doi.org/10.1016/j.envsoft.2016.11.028

The pattern extraction and clustering capabilities of the SOM are extended here for application to data with highly nonlinear underlying manifolds. Environmental data sets often contain cyclical, wavy or helical structures due to diurnal, seasonal and hysteresis effects, and the SOM encounters known limitations when expected to discover such nonlinear manifolds in data sets.

In this paper, the SOM is expanded into the 'SOMersault', incorporating nonlinear dimension reduction techniques and a recurrent transfer of information between high- and low-dimensional spaces. Existing clustering and visualization algorithms range from very flexible (k-means clustering) to very rigid (linear principal manifolds), with the traditional SOM (a constrained k-means algorithm based on linear principal component analysis) lying somewhere in the middle of the spectrum. The SOMersault algorithm shifts the placement of the SOM further towards the flexible end of this spectrum whilst still maintaining the distinctive ordering of clusters which is the main advantage of the SOM.

The real-world application in this paper is a demonstration of the use of the SOMersault to investigate the uneven global spatial and temporal distribution and management of water resources, which leads to water scarcity in certain river basins at certain times of the year. Spatiotemporal clusters of basin-specific monthly conditions of water scarcity and availability are produced, indicating river basins with comparable circumstances. This highlights global similarities between basins sharing similar states of water scarcity due to drought, as well as those experiencing scarcity when storage is available (though perhaps in another form such as groundwater, swamp or ice), indicating a potential for improvements through technology or revised management strategies.

The primary contribution of this paper is an expansion of the SOM technique resulting in a geodesic ordering of output patterns and clusters when representing data with highly nonlinear manifolds, creating a map on which the extracted patterns and clusters are better aligned with the curves of the manifold. Geodesic error measures are provided to assess the quality of the new mapping technique. This method is generalizable to the exploration and visualization of patterns in data from all environmental systems in which an underlying nonlinear manifold exists amongst the high number of variables collected during field measurements.

Paper 4: Patterns and comparisons of human-induced changes in river flood impacts in cities

Journal: Hydrology and Earth System Sciences
Published: March 2018
doi:10.5194/hess-2017-162

Dynamic data sets may possess distinct distributions and correlations at each time step, making direct comparisons ineffectual. This paper addresses this issue by providing a method for tracing similarities in data items based on their relationships to the directions of maximum importance in the structure of available data at each time step, as determined by the organization of the map.

Updating the temporal assessment from Paper 1 through incorporation of the map size and shape selection process from Paper 2, this study allows a different map size and shape to be selected based on the variables at each time step. As initialisations are not based on preceding time steps, each map can be created using different variables. The post-processing treatment from Paper 1 is updated to investigate transitions of individual data items with respect to the main nonlinear directions of importance in the data as determined by the SOM axes rather than changes in cluster membership.

The application in this paper responds to a recent call in the literature for 'an integrated analysis system that can represent the effects of climate and the interface with socioeconomic effects as both drivers and receptors of flood risk' (Sofia et al., 2017). It is an investigation of global patterns of urban river flooding at the city level, as influenced by urbanisation and climate change. World cities each encompass a unique set of environmental and social conditions, with global and local human activities directly and indirectly affecting their watercourses. Cities can therefore expect diverse responses to future changes, including alterations in urban hydrology due to changing rainfall patterns (from climate change) or runoff patterns (from development), and shifting exposures of population and property located in the flood zone through migration and unmanaged development. This study highlights the dependence of dynamic patterns in local conditions on global processes.

The primary contributions of this paper are: 1) the incorporation of a shifting map configuration into SOMs temporal analysis based on the actual data structure at each time step, which still allows individual data items to be traced through the overall temporal form, and 2) a demonstration of the improved method on the analysis of a current, high-interest water resources issue. This is the first study using an artificial neural network to investigate global patterns of city-scale interactions of socioeconomic development and climate change on urban flooding. It is shown that prevalent patterns from this complex data set can be successfully extracted and communicated with this method to gain insight into intervariable relationships for knowledge sharing and resource management.

Paper 5: Practical guide for SOMs implementation in environmental science and engineering

Status: Submitted September 2017

This paper is a conglomerate of the SOMs knowledge deciphered and distilled from many theoretical and application papers. Providing a summary of the basics needed to knowledgably apply a SOM, separate sections of the paper describe the background, creation and interpretation of a SOM. It was written concurrently throughout my candidature, providing both an initial literature review and a summary of knowledge that will hopefully aid other researchers and free them from expending similar time and effort.

This paper was motivated by an unsuccessful search through the literature for a cohesive guide which would allow a researcher to understand the best current standard of the SOM method, and to create a SOM to represent a data set in a timely manner. Much time and effort is currently required to sift through heavily specialized statistical literature, decompose the algorithm and MATLAB code, and understand the finetuning options relevant to each individual application - tasks which every researcher who would like to use SOMs knowledgably should not be required to repeat.

The primary contribution of this paper is a transfer of knowledge to other researchers, attempting to bridge the disparity between published SOMs theory and SOMs applications, which are concurrent, but currently vastly divided, streams of literature.

4 PAPER 1 - TEMPORAL CLUSTER DYNAMICS

This chapter has been published as:

Increasing dependence on foreign water resources? An assessment of trends in global virtual water flows using a self-organizing time map.

Clark S, Sarlin P, Sharma A, Sisson SA. *Ecological Informatics*. 2014; 26: 192-202.

4.1 ABSTRACT

Water resources are continually redistributed across international borders as a result of virtual water flows associated with global trade, where 'virtual water' is the term describing water used in the production of commodities. This transfer of virtual water allows some countries to rely heavily on the water resources of other countries without having to transport the water itself. This paper contains an investigation into the relationship between international virtual water flows and domestically available renewable water resources for a number of countries, to determine trends in national dependencies on foreign water resources over time. Countries with similar states of dependence are clustered, and changes in these clusters are tracked from 1965 to 2010 to determine country-specific and global trends. We make use of a temporal version of the self-organizing map (SOM), the self-organizing time map (SOTM), which provides the means for visualizing structural changes in spatiotemporal data. The SOTM is investigated through a second-level clustering to visualize emerging, changing and disappearing clusters in the data. A post-processing technique is introduced to facilitate interpretation of individual country trends on the SOTM. This study reveals a global trend towards an increased dependence on foreign water resources between 1965 and 2010. The method presented in this study is a workflow tool that results in a visualization of countries with similar and diverging trends of water resource dependencies. This tool can be used to inform national trade, water resources, and environmental management decisions which must take international hydrologic connectivity into account. The sustainability of current virtual water trade and water use trends can be examined with respect to the level of water scarcity experienced by individual and groups of countries.

4.2 INTRODUCTION

The water resources of a country can be significantly impacted by cross-border virtual water flows as a result of international trade (Hoekstra & Mekonnen, 2012). Virtual water is defined as the water used in the production of commodities, such as the quantity of water required to produce a tonne of apples or cereal (Allan, 1998). The transport of trade items across borders can convey large quantities of water virtually 'embedded' in the traded items (without requiring transport of the water itself). This connects the water resources of separate countries, leading to an effective redistribution of water resources between countries and significantly affecting the dispersal of global water resources (Hoekstra 2011, Tamea et al., 2013).

In general, agricultural production is the largest contributor to global water use and pollution (92%) (followed by industrial production (4.4%) and domestic water supply (3.6%)) (Hoekstra & Mekonnen, 2012). Consequently, it is not surprising that the international food trade results in large fluxes of virtual water across international borders (Hoekstra & Mekonnen, 2012). It is less complicated to import crops than to import the water required to grow them, therefore domestic water resources can be supplemented by importing water-intensive food from more hydrologically advantaged regions (Allan 1998). In this way, countries have the ability to make use of more water than they have available domestically due to the influx of virtual water through imported agricultural products (Suweis et al., 2013).

When combined with an investigation of the naturally occurring available water resources within a country, the consideration of virtual water flows allows for an appraisal of a nation's actual water scarcity (Ercin & Mekonnen, 2013). In this paper, we will investigate the relationship between virtual water imports through agricultural trade and the available domestic renewable water resources for a set of countries, in order to explore trends in dependencies on foreign water resources. Countries with high virtual water imports and low internal water resources will be considered relatively dependent on foreign water resources. These dependencies will naturally evolve over time as a result of changing national circumstances, policies, consumption patterns, and environmental factors. Considering the dynamic patterns of water dependencies in a global context will allow for the exploration and comparison of the trends of individual countries.

The literature provides several applications of statistical methods to the virtual water trade network. Konar et al. (2011) used complex network theory to investigate virtual water trade connections as a framework for network optimization, creating a model of nodes (countries) and links (virtual water flows). This study highlighted how individual countries fit into the global structure of the virtual water trade at a certain point in time. Carr et al. (2012) investigated the connections of the virtual water network using trade matrices, to describe changes in the flows to and from specific countries over time. Tamea et al. (2013) showed trends in the virtual water balance on a national basis for a selection of countries. Suweis e al. (2013) calculated the 'carrying capacity of nations' based on the domestic water currently used in food production compared with the virtual water imports of each country. Whilst current research is focused on quantifying national water footprints (that is the total volume of water used to produce goods and services consumed by a country's population) (Mekonnen & Hoekstra, 2011; Hoekstra & Mekonnen, 2012), and investigating flows between countries (Konar et al., 2011; Carr et al., 2012; Tamea et al., 2013), there is a gap in research relating virtual water flows to available domestic water resources. Investigating this relationship will enable an assessment of the actual state of a country's reliance on external water resources at specific points in time.

We consider the self-organizing map (SOM) and its temporal extension, the self-organizing time map (SOTM), to be useful exploratory approaches for this investigation, due to the difficulty in quantifying links between hydrological and other (in this case, trade) data or comparing data collected using different methods in different countries (UN Water, 2009). The SOM enables dimension and data reduction of a complex dataset through projection and clustering, and has previously been used to illustrate refined relationships between countries (Kaski & Kohonen, 1996). Countries can be grouped, and inferences drawn on their relative attributes with respect to other countries internationally, leading to the use of SOMs as a decision support tool (Kaski & Kohonen, 1996; Shanmuganathan et al, 2006). The SOM has been increasingly used in water resources applications over the past decade (Kalteh et al., 2008), but although water resources data often contain a temporal component, investigations frequently focus on either spatial structure or temporal structure, not allowing for an assessment of the changes in spatial structure over time. As the virtual water network is extremely dynamic (Carr et al., 2012), it is important to understand not only spatial, but also temporal, trends in the data. The literature has provided a number of approaches for incorporating time in SOMs (Kohonen, 1988; Kohonen, 1991; Chappell and Taylor, 1993; Guo et al., 2006). But these SOM-based approaches are not aimed at visualizing temporal changes in cluster structures, which is the key aim of the SOTM (Sarlin, 2013). The SOTM includes time as a dimension of the map, thereby providing insight into the trends of the data over time. In this study, the use of the SOTM will enable assessment of which groups of countries have experienced similar transformations in their dependence on foreign water resources over the timeline of the study.

The key focus of this paper is to develop a decision support tool to study the changing dependencies of countries on foreign water resources over time, as decisions regarding national water resources must account for international hydrologic connectivity (UN Water, 2009). Recently, interest in applying the concept of virtual water fluxes to government policy has grown, recognizing the need to understand the effects of trade on water resources. Increased understanding will produce better-informed management decisions, which have conventionally relied only on domestic water use statistics (Hoekstra & Mekonnen 2012).

This tool is exploratory in nature, as it aims to provide visual insights into data that are evolving over time. It seeks to cluster countries based on their state of dependence on foreign water resources and to investigate how this cluster structure has progressed. In order to achieve this, we develop extensions to the existing SOTM framework. In particular the response to missing values in the data (which are common in hydrological datasets) is modified, and a post-processing technique is developed to disentangle the trends of individual countries on the SOTM.

This chapter is structured as follows: Section 4.3 provides a description of the SOM and the SOTM algorithm, along with accompanying clustering and visualization tools; Section 4.4 presents the data and implementation used in this study; a discussion of the results follows in Section 4.5; and a conclusion in Section 4.6.

4.3 METHOD

4.3.1 The self-organizing map

4.3.1.1 Overview

The SOM is an unsupervised learning algorithm from the family of artificial neural networks, used for defining and visualizing non-linear relationships for high-dimensional, multivariate systems. Through training of the SOM with a data set, a topology-preserving mapping of the data is produced from a high-dimensional input space to a low-dimensional output grid (Kohonen, 1998). A key benefit of the SOM is the ability to extract unseen patterns from large quantities of data without requiring an explicit understanding of the underlying relationships.

The SOM grid is first initialized based on the overall structure of the input data set (the training data), and then trained based on the individual input data items. The initialized map units assume linearly spaced values along a set of axes aligned with the eigenvectors corresponding to the principal components of the input data space (Kohonen, 1998). As a primary purpose of the SOM is to provide visualization of a data set (Kohonen, 1998), 1D or 2D output grids are usually used.

The training of the SOM consists of applying two iterative processes: selection of the best map unit to match each item of input data, and updating of the map to better represent the input data. These processes seek the optimal map structure to represent the form of the input data (Kohonen, 1998). During the selection step of training, the map node that best matches each item of input data is selected (the best matching unit, or BMU) based on minimum Euclidean distance. In the updating stage, the BMU and its neighbouring map units (within a specified neighbourhood radius) move to become closer to the input. The neighbourhood radius decreases with each iteration of selection and updating, producing a smoothed final map. For a 2D SOM, Hastie et al. (2009) encourages the reader to consider the map units as buttons that have been sewn in a regular pattern onto the 2D principal component plane of the input data (the input data may be in two or more dimensions), and the training process of the SOM bends and stretches the plane until the buttons best approximate the distribution of the data.

4.3.1.2 Details

A more detailed description of the input, output and training process of the SOM are provided here.

The input data, X, are in vector format with a separate vector, x_j (where j=1, …, N), for each input item. All input vectors have the same number of dimensions, d, (ie. $x_j = x_{j1}, \ldots, x_{jd}$). The SOM output consists of M map nodes in a grid format, with a prototype vector, m_i (where i=1, …, M) associated with each node. The output vectors are of the same dimension, d, as the input vectors.

The selection of BMU for each input item, x_j, consists of finding the closest map node, m_c, by Euclidean distance measure, where c is the index given to the best match (Kohonen, 1998):

$$||x_j - m_c|| = \min_i\{||x_j - m_i||\}$$

The batch updating process (Kohonen, 1993) considers the entire input data set at once, over a series of iterations. At each iteration, the input data is divided into subsets that share the same BMU, m_c, and a Gaussian neighbourhood function is applied as a smoothing kernel when updating the map (Kohonen, 1998). The neighbourhood function is computed at each map unit, where $h_{ic(j)}$ is the value at map node m_i of the neighbourhood function centred around the best matching unit, m_c, of data item x_j (Kohonen, 2013). That is:

$$h_{ic(j)}(s) = \exp(\frac{-\mathrm{sqdist}(c,i)}{2\,\sigma^{2}(s)})$$

where sqdist(c,i) is the squared Euclidean distance on the map grid between nodes m_c and m_i; σ is a user specified, monotonically decreasing, neighbourhood radius; and s is the iteration index (Kohonen, 1998). The value of $h_{ic(s)}$ is the same for all data vectors sharing the same BMU, and decreases in size as iterations progress. This produces a global ordering to the map when the neighbourhood is large, followed by a fine tuning of the map when the neighbourhood has reduced. Each map unit, m_i, is updated with the weighted average of the n data items in its neighbourhood, where the weight of each item is the neighbourhood function. The updated nodes are calculated as (Kohonen, 2013; Vesanto, 2000):

$$m_i(s+1) = \frac{\sum_{j=1}^{n} h_{ic(j)}(s)\, x_j}{\sum_{j=1}^{n} h_{ic(j)}(s)}$$

Through this iterative method, a set of vectors is constructed to represent the input data, which are projected in a topology preserving manner onto a low-dimensional output grid (Vesanto & Alhoniemi, 2000). For more details on training, understanding and interpreting SOMs, refer to Kohonen (1998) or Kohonen (2001).

4.3.2 The self-organizing time map

4.3.2.1 Overview

Data often consists of multivariate samples at different points in time, and it is useful not only to analyze and visualize the dataset as a whole, but also to understand the temporal changes that have occurred along the timeline. The SOTM uses the capabilities of the SOM for the abstraction of structural changes in spatiotemporal data, providing an exploratory tool for temporally dynamic datasets. In essence, the SOTM is a series of vertical 1D SOMs arranged next to each other in order of increasing time, and connected through short-term memory.

Input to the SOTM takes the same format as input to the SOM with the distinction that it is split into a distinct number of time periods (for example a set number of months, years or decades). The output map of the SOTM is two-dimensional, with the vertical axis representing positions in dataspace and the horizontal axis denoting time. Thus, a single image is created to convey temporal changes in the data.

4.3.2.2 Details

To observe the structure of the dataset at each time unit, t (where t=1, 2, ..., T), the SOTM performs a mapping of each input item, $x_j(t)$ (where j=1, 2, ..., N(t)), from the input space Ω

(t) onto a one-dimensional array, A(t), of output units, $m_i(t)$ (where i=1, 2, …, M). To preserve the orientation of the one-dimensional arrays between consecutive timesteps, the SOTM uses short-term memory incorporated into the initialization of the map nodes. At the first timestep (t=1), the column of map nodes, $A(t_1)$, is initialized based on principal component analysis of the data belonging to that timestep, $\Omega(t_1)$, as was described for the SOM. For the remainder of the timesteps, the map units are initialized based on the output of the preceding timestep (ie. the output vectors of A(t -1) become the initial values of A(t)).

During training of the SOTM, adjustment to temporal changes is achieved by performing a SOM-type batch update for each time unit, t (ie. batch updates are performed separately for each vertical column of nodes, A(t)). The topology preservation of the SOTM is hence twofold: the horizontal direction preserves time topology and the vertical preserves data topology. The training of the SOTM follows the two standard steps from the SOM paradigm. First, the BMUs are located by a time-restricted matching of each data point to the map unit with the nearest Euclidean distance. This means, for example, that an input data item belonging to the second timestep can only search for its BMU on the second vertical column of nodes. Then, each reference vector, $m_i(t)$, is updated through a time-restricted version of the SOM batch update:

$$\mathrm{m_i(s+1,t)} = \frac{\sum_{\mathrm{j=1}}^{\mathrm{N(t)}} \mathrm{h_{ic(j)}(s,t)x_j}(t)}{\sum_{\mathrm{j=1}}^{\mathrm{N(t)}} \mathrm{h_{ic(j)}(s,t)}}$$

where s indicates the training iteration, t is the horizontal location on the timeline, and the neighborhood, $h_{ic(j)}(s,t)$, is restricted to units $m_i(t)$:

$$\mathrm{h_{ic(j)}(s,t)} = \exp\left(\frac{-\mathrm{sqdist(c,i)}}{2\,\sigma^{2}\,\mathrm{(s,t)}}\right)$$

This means that only the nodes in each vertical column are updated concurrently. As with the SOM, the radius of the Gaussian neighborhood function is initiated with a user-specified neighborhood parameter which decreases with each batch update at each timestep. Though, in contrast to the standard SOM, the neighborhood radius only includes vertical relationships.

Figure 2 provides an indication of the functioning of the SOTM.

As in Park et al. (2003), the vertical number of nodes at each timestep is determined using an empirical method of minimizing quantization and topographic errors. Quantization error of the SOTM is defined as the average distance between each input item and its BMU, m_c, at each timestep, averaged over all timesteps (Sarlin, 2013):

$$\mathrm{QE_{SOTM}} = \frac{1}{\mathrm{T}}\sum_{\mathrm{t=1}}^{\mathrm{T}} \frac{1}{\mathrm{N(t)}} \sum_{\mathrm{j=1}}^{\mathrm{N(t)}} \left\| \mathrm{x_j(t)} - \mathrm{m_{c(j)}(t)} \right\|$$

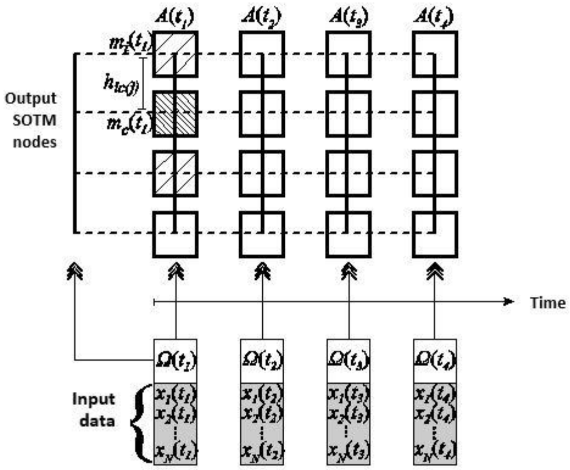


Figure 2: The functioning principles of the SOTM (adapted from Sarlin, 2013), where Ω(t) is the input data consisting of $x_j(t)$ (j=1,...,N(t)) at timestep t; $h_{ic(j)}$ is the decreasing Gaussian neighbourhood function; $m_c(t)$ is the best matching map unit for the current input data; and $m_i(t)$ are the other map units. A one-dimensional array, A(t), is created to represent the input data of each timestep. The arrays are arranged in order of ascending time to form the SOTM, with horizontally adjacent arrays connected through short-term memory.

The topographic error is the average proportion of $x_j(t)$ (at each timestep) for which first and second BMUs are non-adjacent, $u(x_j(t))$, (Sarlin, 2013). The topographic error of the entire SOTM is $u(x_j(t))$ averaged over all timesteps:

$$TE_{SOTM} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N(t)}\sum_{j=1}^{N(t)} u(x_j(t))$$

4.3.2.3 Missing data

Motivated by the large number of missing values often encountered in hydrological data, we follow the approach put forward by Kaski & Kohonen (1996), and earlier generalized for self-organization overall by Samad & Harp (1992), to allow for partial missing values in a dataset. Kaski & Kohonen (1996) assert that the SOM is robust to data with about two thirds of its values missing. In particular, when all variables are not available for an observation, only the available data are considered in SOM matching. If an input item is missing a value for one or more of its variables, this particular input item is mapped based on the values of its remaining variables. Hence, a more formal description replaces the right-hand side of the SOM BMU selection equation, as follows:

$$|\|x_j - m_c\|| = \min_i\left\{\sum\nolimits_{s\in S(x_j)} \left(x_{js} - m_{is}\right)^2\right\}$$

where $S(x_{js})$ is a set of positions in data vector x_j that are complete (whereas positions of missing values are disregarded).

4.3.3 Clustering the SOTM nodes

Clustering refers to a class of techniques that partition data into clusters (groups) with the aim that data in each cluster are more similar to each other than to data in other clusters. It is important to extract clusters from a dataset in order to fully explore its properties and produce summary information (Vesanto, 2000). The technique of using second-level clustering to group the map units of the SOM (the first level of clustering is performed by the SOM itself) was introduced by Vesanto (2000) to distinguish groups of similar output nodes. We follow Sarlin & Yao (2013) in performing cluster analysis to group the output map units of the SOTM over all timesteps. This divides the SOTM into a number of clusters which may include one or more map units at one or more points in time (ie. the cluster boundaries may traverse the map horizontally and vertically without restriction). This is a particularly useful technique for SOTM analysis as it strengthens the horizontal connection of the timesteps.

In this paper, clustering of the SOTM nodes is performed using a 1D 'second-level' SOM of the same size as a single timestep of the SOTM. Initialization of this SOM is based on output from the final timestep of the SOTM. The use of a SOM for the second-level clustering, rather than an alternative clustering method such as k-means, maintains an order to the clusters, ensuring that similar clusters are neighbours. This allows for an indexed colour scheme to be applied to the SOTM to depict separate clusters. The colouring of the clusters of the SOTM leads to a visualization of the changes in cluster structure of the underlying data, with similarly coloured clusters representing similar data. Emerging, changing and disappearing clusters will become evident through this method (Sarlin & Yao, 2013).

4.3.4 Interpreting the SOTM with component planes

As is common in the SOM literature (Vesanto, 1999), the variables of the SOTM may also be represented using component planes. Component planes are regular grids with the same format as the SOTM output array, representing the values of the individual variables contributing to the SOTM. Each node of a component plane shows the value of one variable at that node of the SOTM, as nodes at the same grid location represent the same input data on all component planes and the SOTM. The spread of values for each variable becomes evident on the component planes, as well as the temporal changes in distributions of each variable. This enables a variable-wise examination of the SOTM, indicating structural data properties at each timestep (vertically) and changes in these structures over time (horizontally).

4.3.5 Associating variables to the SOTM

Beyond standard component planes, we can associate additional variables to the SOTM. The introduction of a set of variables onto a SOM that has been trained with other variables is an innovative approach in using SOMs for water resource applications (Cereghino & Park, 2009). This process, known as associating variables, allows the relationship between the two sets of variables to be investigated. The available data can be separated into 'training variables' and 'associated variables', with the form of the map determined only by the training variables, and the associated variables incorporated during the updating stage. That is, the BMUs are found based only on the training variables, but then the map nodes are updated using both the training and associated variables. This assigns values to the map for the dimensions of the associated variables without them having an effect on the training of the map. This can be useful in depicting how external variables compare to the relationships established by the training variables.

For example, a SOM that maps countries based on the state of their environmental conditions could have latitude data 'associated' to see if there are any patterns that become evident. Creating the map using latitude as a training variable would not be useful as the relationship between countries and latitude is already well understood, and it may overshadow the environmental factors in the formation of the map structure. But it may be interesting to see if there are any patterns between groups of countries with similar environmental states, as established by the SOM, and latitude.

The associated variables can be visualized on component planes as described in Section 4.3.4, to allow comparison to the training variables.

4.3.6 Post-processing the SOTM to explore data trends

For an exploration of how individual data items are trending over time in relation to the entire dataset, trajectories of data items may be tracked horizontally across the SOTM. One way to do this would be to label each data item at each timestep of the SOTM and track its movements. But since the number of data observations is generally much larger than the number of nodes at each timestep, the labeled map has the potential to become indecipherable. In addition, a simple labeling of each node would not provide information on the movement of a data point through the SOTM; the path would also need to be linked across the timesteps, further cluttering the map.

Therefore, a new post-processing step has been developed in this study to provide a visualization of the movements of each data point through the SOTM. After the SOTM has been trained and clustered, the cluster memberships of each data point are determined at each timestep. These cluster memberships are then used as input into a post-processing SOM. As the dimensions of the input data for this SOM are timeframes, data mapping to the same output map node share a similar trend over time. Therefore, the result is a map depicting data that have trended in a similar manner, as well as those that have diverged over the course of time.

4.4 Data and Implementation

4.4.1 Data

Data for this study were assembled from a number of databases. Water resource indicators by country were obtained from the AquaStat database (http://www.fao.org/nr/water/aquastat/main/index.stm) of the UN Food and Agriculture Organization (FAO). Detailed international trade data, by country and by product, were obtained from the FAOstat database (http://faostat.fao.org/). Water footprint data, (indicating the total volume of water used to produce goods) by product and by country, were obtained from the WaterStat database (Mekonnen & Hoekstra, 2011). Population and GDP information were obtained from the World Bank databank (http://databank.worldbank.org/data/home.aspx). The data, summarized with sources and units in Table 2, occur in consecutive 5 year blocks from 1965-2010, producing a dataset of 10 timesteps containing 172 items (countries) each.

Table 2: Data, units and sources

Variable	Unit	Source
Total renewable water resources per capita (actual) [a]	m3/inhab/yr	AquaStat
Water stress: Freshwater withdrawal as % of total actual renewable water	%	AquaStat
Water content per crop	m3/tonne of crop	WaterStat
Import quantities of all agricultural products	Tonnes/year	FAOStat
Export quantities of all agricultural products	Tonnes/year	FAOStat
Population	total	World Bank
GDP per capita	constant US $2005	World Bank

[a] Total actual renewable water resources per capita is defined on the AquaStat database as 'the maximum theoretical yearly amount of water actually available for a country' considering both surface water and groundwater, including inflows from upstream countries and border waterways.

Agricultural products have a relatively high impact on global water resources, with over 90% of global water use and pollution attributed to agricultural production, and an estimated 76% of the virtual water flow between countries due to trade in crops and derived crop products (Hoekstra & Mekonnen, 2012). Therefore, only agricultural trade products are considered in this study. For more information on water footprints and the virtual water content of various commodities, refer to www.waterfootprint.org.

The net virtual water imports of each country are calculated in cubic metres (m^3) per capita imported annually (imports-exports). FAO trade data (imports and exports of all agricultural products) for each product and each country, and the total water content required to produce over 200 crops and crop products (from the WaterStat database) are used.

The calculation of virtual water fluxes follows Hoekstra et al. (2011) with the exception that global averages of water footprint per crop are used rather than specific country averages as the countries of origin of each product are unknown. The total volume of water required for production of each crop is used, including green (rainwater), blue (surface water and groundwater), and grey water (volume of water required to assimilate pollution). For each country and crop pair, the quantity (tonnes) of crop or product imported and exported per 5 year interval is determined. Next, this traded quantity is multiplied by the water footprint of the crop, providing the quantity of virtual water imported to and exported from each country by means of each specific crop in each 5 year interval. These quantities are then summed over all products for each country, and converted into net imports of virtual water (m^3). This method uses the 'top-down' approach described by Hoekstra et al. (2011) which is very reliant on the quality of the trade data (as opposed to the 'bottom up' approach which relies more on consumption data).

The data is subject to some limitations and assumptions. The estimated water footprints of the agricultural products used in this calculation are 10 year averages (1996-2005), as they are based on climate data. This necessitates the assumption that the water content required to grow each crop doesn't change over time, though in reality water requirements will differ due to climatic variations and changes in agricultural technology. Also, as global averages of crop water footprints are used rather than country-specific estimates, it is assumed that there is no significant variation in water required for a specific crop in different geographic regions. In actuality, water requirements will differ between geographic regions due to many local factors including soil conditions, precipitation interception, plant types and climate conditions. The effects of this regionalization may be included in future studies. Water content information is not available for all the items in the trade data, and so some products which appear in the trade data are not included in this study. The trade data has inconsistencies that are more likely a result of a change in record-keeping practices than actual changes in trade. No imputation of missing data has been attempted. The AquaStat data is assembled from a variety of sources which were collected intermittently, and is subject to variations in collection and estimation methods (http://www.fao.org/nr/water/ aquastat/metadata/index.stm).

4.4.2 Implementation

In this study, the SOTM is implemented using extensions to the MATLAB SOM toolbox (http://www.cis.hut.fi/somtoolbox). The SOTM is trained using 100 iterations of the SOM batch algorithm at each timestep, and a Gaussian neighbourhood kernel with an initial radius of 2.5, decreasing to 1. The following training variables are used: net virtual water imports (m^3) per capita (1965-2010), and available renewable water resources (m^3) per capita (1965-2010). Data are considered per capita rather than country totals, to prevent spatial differences in country size or temporal changes in population affecting the map. The data is transformed variable-wise, into a range from 0 to 1. Approximately 12% of the data is missing for each variable. Following the method described in Section 4.3.2 of choosing the number of nodes of the SOTM, the optimal size of the map is determined to be 10 vertical nodes at each of the 10 timesteps, for a total of 100 nodes. At each horizontal timestep, the SOTM training process maps the countries that share a similar state of reliance on external water resources (based on the training variables) to the same vertical node. Each country will map to exactly one node per timestep (column of the map), and therefore will appear 10 times on the map from left to right, though possibly in different vertical positions.

After training is complete, the SOTM nodes are clustered into similar groups to investigate the general data structure at each timestep, and to visualize trends over the timeline. It is the evolution of the clusters (and the countries which are members of them) that is of interest in this study. As described in Section 4.3.3, all the SOTM output map vectors are presented as input into a new 1D SOM containing 10 nodes, to perform a second-level clustering. This process produces 10 ordered clusters of the SOTM map nodes, allowing for an indexed colour scheme to be applied.

The trained (but unlabeled) SOTM is shown in Figure 3, with time on the horizontal axis and a vertical column of nodes for each 5 year interval. The nodes of the SOTM are coloured based on their cluster membership as determined by the SOM, with Cluster 1 in the lower right, Cluster 10 in the upper left, and a linear colour scale in between. Through this colouring, the temporal evolution of the clusters across the entire timespan becomes evident, providing a horizontal linkage to the map. Examining the cluster colours on Figure 3, it can be seen that the clusters experience a general upwards trend over time (from left to right), with the upper clusters disappearing and the lower clusters emerging and growing.

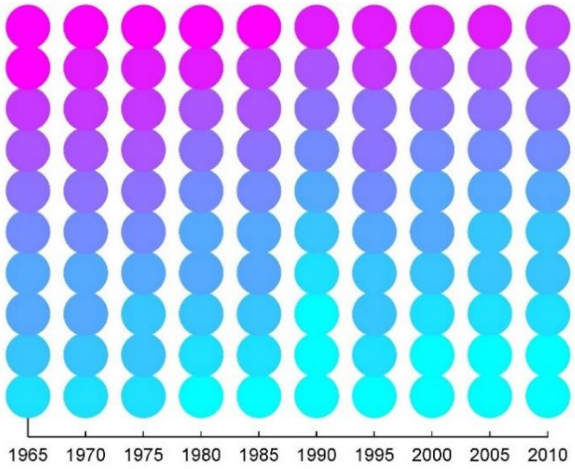


Figure 3: The trained SOTM. Each vertical column is a timestep representing a 5 year interval from 1965 to 2010. Each of the 172 countries maps to exactly one of the 10 nodes at each timestep. Clustering (indicated by colour) links similar properties across the map. The evolution of clusters through time becomes evident, as the top left cluster disappears and the lower clusters expand upwards.

At this stage, the SOTM in Figure 3 only provides an indication of the existence of a cluster structure in the data, and evidence that the cluster structure is shifting over time. Further investigation is needed to extract more useful information from the SOTM. The following discussion will provide an interpretation of the properties of the clusters in Figure 3. In particular, we will: decipher the meaning of the node colours by analyzing the contribution from each variable; explore additional general characteristics of countries mapping to each section of the SOTM; and track the course of individual countries along the timeline of the SOTM.

4.5 DISCUSSION

In this study, the non-linear relationship between national virtual water imports and available water resources is investigated, and countries with similar dependencies on foreign water resources are clustered and tracked over time through the use of a SOTM. The SOTM highlights how these clusters of countries have evolved over the period of the study. It will become evident how the water dependencies of certain countries have changed in similar manners, and which countries were initially similar but have ultimately diverged towards the end of the study period. Possible reasons for any divergence between countries may include unilateral changes in: domestic water resource availability due to environmental changes, the type or quantity of internationally traded agricultural items, agricultural policy or practices involving water use such as restrictions or technological advances, or changing water requirements with regional climate changes.

This section presents the interpretation and post-processing of the SOTM.

4.5.1 Component planes

To decipher the properties represented by the node colours in Figure 3, it is useful to explore the individual training variables using component plane visualizations as described in Section 4.3.4. Figure 4 presents a component plane for each variable, depicting changes in the distribution of each variable over time. The grids in Figure 4 correspond exactly to the grid in Figure 3, with each node representing the same group of countries on each map. As was done with the SOTM in Figure 3, a second-level clustering is used to colour the individual component planes, with lighter nodes representing low values and darker nodes representing high values (though note that the same shades do not represent equal values on each component plane).

On the first subplot of Figure 4, a clear trend is evident towards a decrease in renewable water resources per capita, as the darker clusters fade out and the lighter clusters become more prominent over time. On the second subplot, the emergence and expansion of both the dark and light clusters over time indicate that the range of net virtual water imports has expanded in both directions over time, with 2010 experiencing clusters with both larger and smaller net import values than existed in 1965.

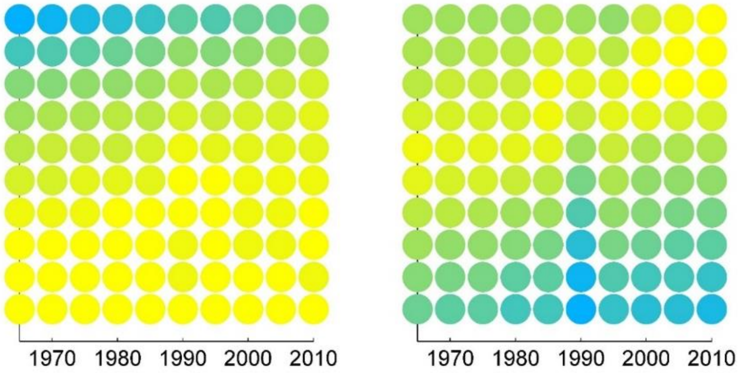


Figure 4: Component planes for the variables of the SOTM: a) Renewable water resources per capita, and b) Virtual water imports per capita. Lighter nodes represent low values and darker nodes represent high values. Each node represents the same group of countries at the same time on both grids, as well as on Figure 3. The difference in colouring between the two planes indicates the non-linear relationship between the training variables.

Since these component planes are separate visualizations of the same map (the SOTM), and the nodes on each represent the same groups of countries, general information about the clusters of the SOTM in Figure 3 can be gained from an exploration of the component planes in Figure 4. It can be determined that the lower right nodes on the SOTM in Figure 3 generally correspond to relatively low renewable water resources (lighter nodes) and relatively high virtual water imports (darker nodes) on the lower right nodes of the planes in Figure 4. Therefore, countries mapping to these nodes could generally be considered more dependent on foreign water resources than those mapping higher up on the SOTM. The upper part of the

SOTM represents countries with higher renewable water resources (darker nodes on Figure 4), but includes countries with both medium and low virtual water imports (medium to light nodes). Therefore, the upper left cluster (Cluster 10) in Figure 3 represents a more hydrologically self-sufficient state than the lower right cluster (Cluster 1), with an ordered range in between.

Consequently, the trend that is evident on the SOTM, visualized by the overall general colour shift over time, is towards greater dependency on foreign water resources.

4.5.2 Associated variables

To provide a generalization of the characteristics of countries that may share similar foreign water dependencies, information on population, GDP and water stress (where water stress is defined as freshwater withdrawal as a percent of total actual renewable water resources) has been investigated. This has been done following the method of associating variables, as described in Section 4.3.5. In this way, the map has been created based on virtual water imports and available domestic water resources, but the information on population, GDP and water stress has been added onto the map for comparison. The associated variables have been plotted on component planes in Figure 5, again with groups of countries at each node matching those of the grids in Figure 3 and Figure 4. We use the same color coding as Figure 4, ranging from light to dark (low to high), and denote empty nodes (due to a lack of water stress data for any of the countries mapping to them) with white space.

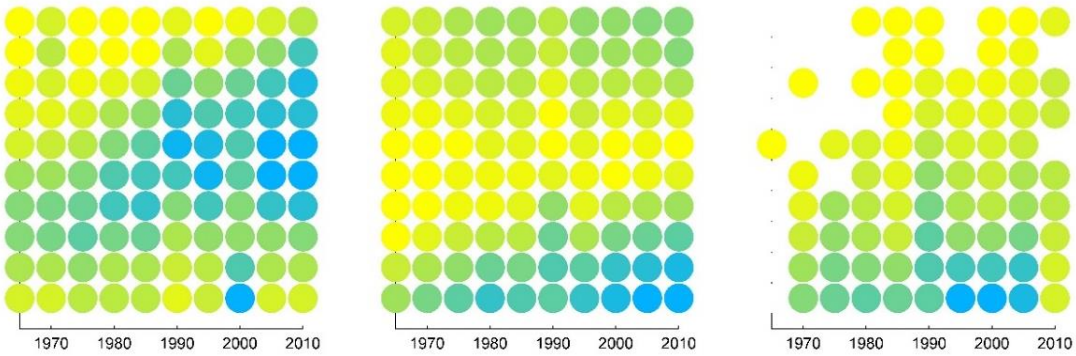


Figure 5: Associated variables: a) population, b) GDP per capita, and c) water stress. Additional characteristics of countries mapping to regions of the SOTM are investigated through these associated variables. Each node on the three planes corresponds to the same group of countries (as well as to the equivalent nodes on Figure 3 and Figure 4).

When compared with the SOTM in Figure 3, it can be seen that countries with the highest dependencies on foreign water resources (lower right nodes) tend to be countries with medium population levels and high GDP per capita. They also tend to be countries experiencing water stress. Countries that are relatively independent of foreign water resources (upper left nodes) tend to have lower populations, average GDP per capita, and of course, low water stress.

4.5.3 Post-processing SOM

Whilst the general global temporal dynamics of the data distribution can be gleaned from an investigation of the unlabeled SOTM in Figure 3, the countries mapping to each node must be identified in order to understand country-level trends. But due to the large ratio of input data items to map nodes at each timestep, labeling of the SOTM has the potential to become indecipherable.

As an example of country-level SOTM labeling, Figure 6 provides labels for each vertical node of the 2005 timestep only. The vertical nodes have been rotated to a horizontal orientation to accommodate the list of labels, with Node 1 representing the lowest node of the 2005 timestep on Figure 3, and Node 10 the uppermost node (the colouring matches Figure 3 to facilitate comparison). Overall, it can be seen that the majority of countries map to the lower nodes, which as described above, indicate a relatively higher dependence on foreign water resources than the upper nodes. This is to be expected from the 2005 data as the SOTM has indicated a trend towards greater global dependence on foreign water resources from 1965 to 2010.

Node 1	Node 2	Node 3	Node 4	Node 5	Node 6	Node 7	Node 8	Node 9	Node 10
				Turkey					
				Estonia					
				India					
				Pakistan					
				Bangladesh					
				Sri Lanka					
		Germany		Nepal					
		Finland		Iran	Czech Rep				
		Sweden	Austria	Afghanistan	Slovak Rep				
		UK	Poland	Turkmenistan	Latvia				
Belgium		Greece	Croatia	Uzbekistan	Romania	USA			
Denmark		Albania	Azerbaijan	Kyrgyz Rep	RussianFed	France			
Netherlands		Georgia	Belarus	Chad	Benin	Hungary			
Ireland		Egypt	Tajikistan	Mali	Togo	Bulgaria			
Faeroe I.		Tunisia	Mongolia	Niger	Burundi	Lithuania			
Luxembourg		Morocco	Yemen	Nigeria	CAR	Ukraine			
Slovenia		Mauritania	Eritrea	Burkina Faso	Sierra Leone	Kazakhstan			
Portugal		CapeVerde	Gambia	Somalia	Eq. Guinea	Liberia			
Spain	Italy	Maldives	Namibia	Senegal	Guinea	Cameroon			
Malta	Switzerland	Iraq	Botswana	Sudan	Ethiopia	Ghana			
Cyprus	Norway	Oman	Swaziland	Kenya	Uganda	Comoros			
Israel	Japan	Syria	Lesotho	Mozambique	Rwanda	Ecuador	Australia		
Jordan	Bosnia&H	Cuba	Venezuela	Angola	Malawi	Guatemala	Malaysia		
Libya	Algeria	Bahamas	Panama	Zimbabwe	Madagascar	Honduras	Brazil	Canada	
Djibouti	Lebanon	Jamaica	Peru	Zambia	Thailand	Nicaragua	Uruguay	Argentina	
Kuwait	Bahrain	Mexico	Chile	South Africa	Vietnam	Colombia	Bolivia	Paraguay	
Qatar	SaudiArabia	Trin & Tob	El Salvador	Haiti	Myanmar	Lao	Belize	PNG	Iceland
UAE	Mauritius	Dom Rep	Fiji	Dominica	Cambodia	Indonesia	Costa Rica	Solomon I.	Guyana
PuertoRico	Antigua&Bar	Brunei	Samoa	Philippines	Timor-Leste	Kiribati	CotedIvoire	Bhutan	Congo
Seychelles	Barbados	N Zealand	Singapore	Tonga	Grenada	Vanuatu	GuineaBissau	Gabon	Suriname

Figure 6: Labels for the 2005 timestep of the SOTM (colours correspond to cluster colours from Figure 3). This shows where individual countries map on to the SOTM in 2005, with the majority mapping to the lower nodes, indicating a high reliance on foreign water resources.

The lists of individual countries in Figure 6 indicate that many countries of Europe, the Middle East, North Africa, and the Caribbean were amongst those particularly dependent on foreign water resources in 2005, whereas countries of North America, eastern South America, the South Pacific, and western Africa were amongst the least dependent nations. A further investigation links Figure 6 to Figure 5, where it was indicated that countries mapping to Node 1 in 2005 generally have low to medium populations and high GDP per capita, whereas countries mapping to Nodes 5 and 6 generally have higher populations and medium to low

GDPs per capita (this remains a generalization as these associated factors have not been taken into account in the creation of the map, though are of interest in interpreting it).

Whilst Figure 6 provides a snapshot of the clusters of countries with similar states of dependence in 2005, these clusters will not be consistent over all timesteps. Over the years, countries will have transferred between clusters as their individual conditions changed. To investigate each country's change over the timeline of the study, as well as to see which countries changed in similar ways, it is necessary to identify the countries mapping to each cluster of the SOTM at each timestep. As indicated above, with 172 countries labeled onto only 10 nodes at each timestep, it would prove difficult to interpret the labeled map to extract information about individual country trends or to identify countries that are trending similarly.

Therefore, a technique has been developed to circumvent the need to directly label the SOTM. As described in Section 4.3.6, a 'post-processing' SOM has been used to group countries that have trended in a similar manner in terms of dependence on foreign water resources during the study period. The input to this SOM is each country's cluster membership at the beginning, middle and end of the SOTM timeline. This new method reveals patterns of movement on the SOTM and clusters of similarly-trending countries.

The resulting post-processing SOM is shown in Figure 7a. The groupings of countries represent those that have followed a similar path of dependency on foreign water resources over the years. The quantities are not necessarily comparable within groups. Rather, it is the **relationship** between the virtual water being traded across the borders and the domestic water resources available within the country that is being compared. The map is coloured based on the overall change in cluster membership (and therefore the change in dependency situation) from 1965-2010. Countries in the white region have shifted the least in terms of hydrologic dependency; countries to the right have moved towards greater hydrologic self-sufficiency, whilst countries to the left of the white region have become increasingly dependent on foreign water resources. The more intensely coloured regions at the edges represent the greatest increase (on the right of the map) and the greatest decrease (on the left of the map) in hydrological independence.

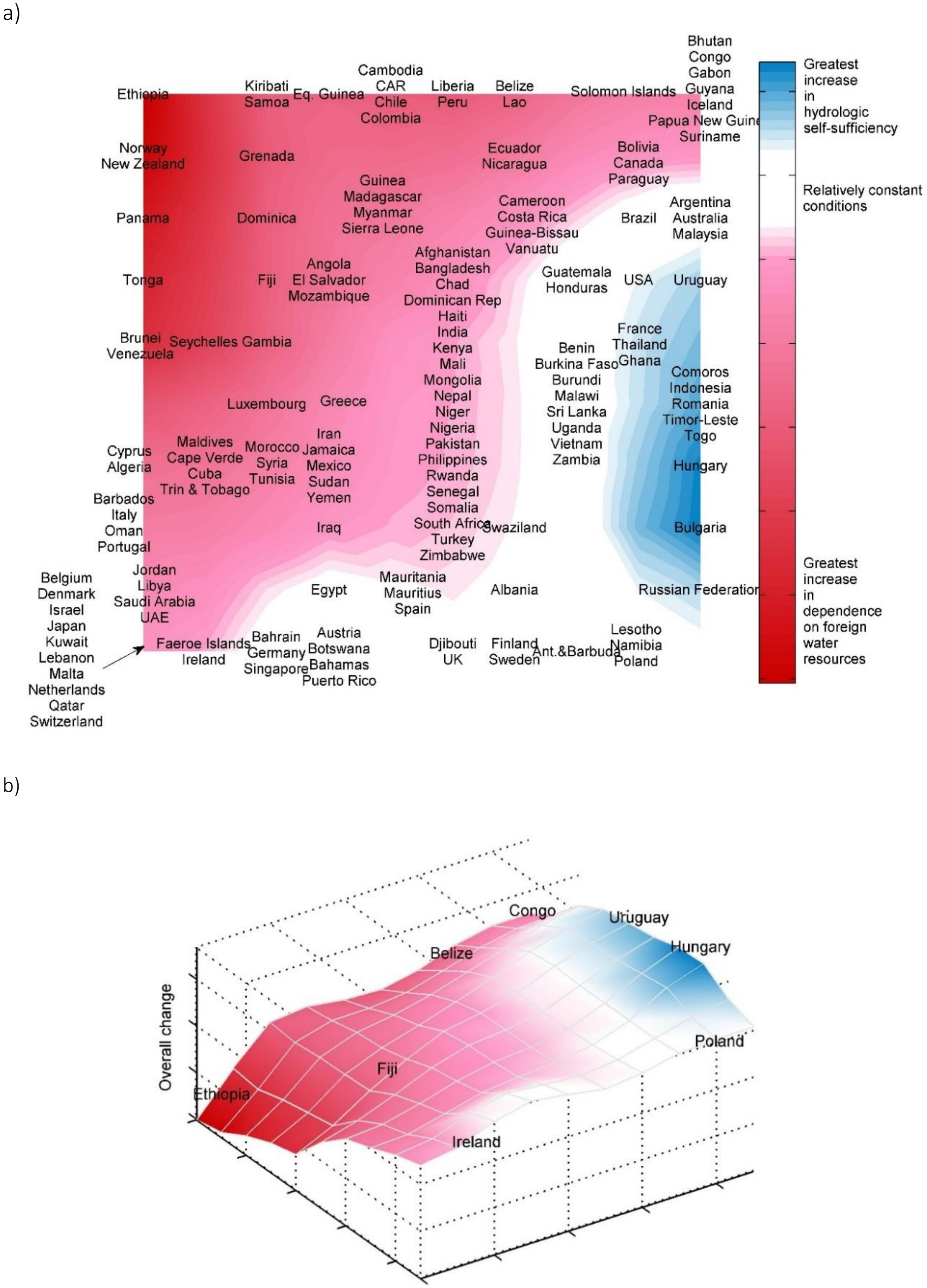


Figure 7: a) A post-processing SOM of the clustering of the SOTM in Figure 3. The groups represent countries that have moved in a similar manner through the clusters of the SOTM, indicating similar changes in states of hydrological

dependence from 1965-2010. In general: the white region represents a relatively constant state of hydrologic dependence, the region to the right represents a move towards greater independence, and the region to the left represents an increase in dependence on foreign water resources over the timeframe of the study. b) The post-processing SOM is shown with the z-axis representing overall change in hydrologic independence from 1965-2010. Ethiopia had the greatest decrease in hydrologic independence, and Hungary the greatest increase in self-sufficiency. It can be seen that the conditions in certain countries (Ethiopia/Congo, and Fiji/Uruguay) were similar in 1965 and diverged over the study period.

The vertical location on the map approximates the starting condition of each country in 1965, with the countries that were relatively hydrologically independent towards the top of the map, and more dependent on foreign water resources towards the bottom. (Though note that within each cluster the list is alphabetized and centred vertically, therefore for example Afghanistan to Zimbabwe are all at the same horizontal level as Greece.) Countries mapping to approximately the same horizontal position started with similar conditions. Those within the same cluster remained similar, whilst those in different clusters diverged over time.

Figure 7b shows a 3d representation of Figure 7a with the z-axis representing the overall change in hydrologic dependence over the timeline of the study. It can be seen that Ethiopia and the Congo started with similar conditions in 1965 and diverged, with Ethiopia becoming far more hydrologically dependent over the years than the Congo. The same can be seen when comparing Fiji and Uruguay which started with similar conditions, but Uruguay became less dependent whilst Fiji became more so.

4.5.4 Verification of the post-processing SOM

The post-processing SOM in Figure 7 has separated the countries into groups with similar trends of hydrologic dependency from 1965-2010. In order to verify that the post-processing SOM has produced successful groupings of countries, Figure 8 shows a selection of countries from the SOM coloured with their SOTM cluster colours for each timestep (from Figure 3). A legend links cluster colour to number, with higher numbered clusters representing greater hydrologic self-sufficiency, and lower numbered clusters representing greater dependence on foreign water resources.

The countries in the top right of the SOM (Solomon Islands to Suriname) were relatively independent of foreign water resources in 1965, and have only slightly increased their dependence over the years. Countries further down the right hand side of the map (Argentina to Uruguay) had slightly less initial independence than the countries above them, with a small increase in independence over the years. On the left of the map, Norway to Panama began in a similar state as Bolivia to Paraguay on the right, but have experienced a large increase in dependency. The countries located in the lower left corner of Figure 7 (Belgium to Saudi Arabia) were highly dependent on foreign water resources in 1965, and have continued to be so for the duration of the study period.

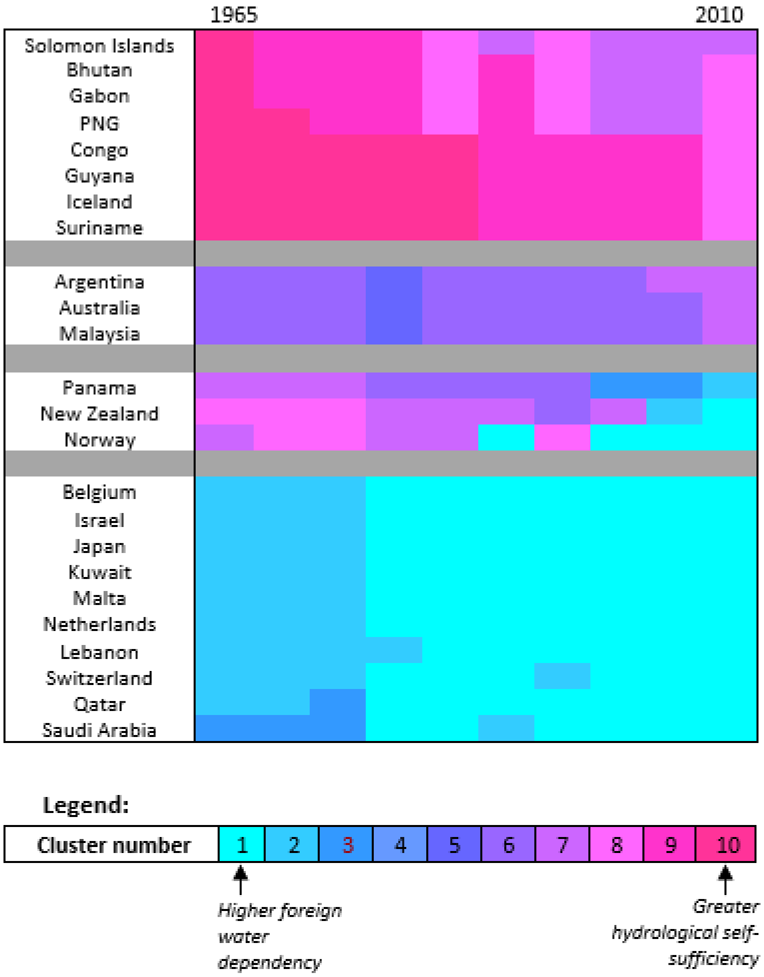


Figure 8: Cluster membership at each timestep of the SOTM in Figure 3 is shown for a selection of countries from the SOM in Figure 8. This confirms that the SOM in Figure 7 has clustered countries that are trending in the same manner across the SOTM. Colours refer to the clusters in Figure 3.

4.5.5 Workflow method summary

The method for the abstraction of dynamic cluster trends that have been developed in this study is summarized in Figure 9. This method combines the SOTM with the traditional SOM, and adds extensions to the SOTM framework through new post-processing techniques. The output is a single visualization of trends in clusters of data from large, multi-dimensional datasets with non-linear relationships between variables. Through this visualization, an indication of the relative trends experienced by each of the input observations can be attained.

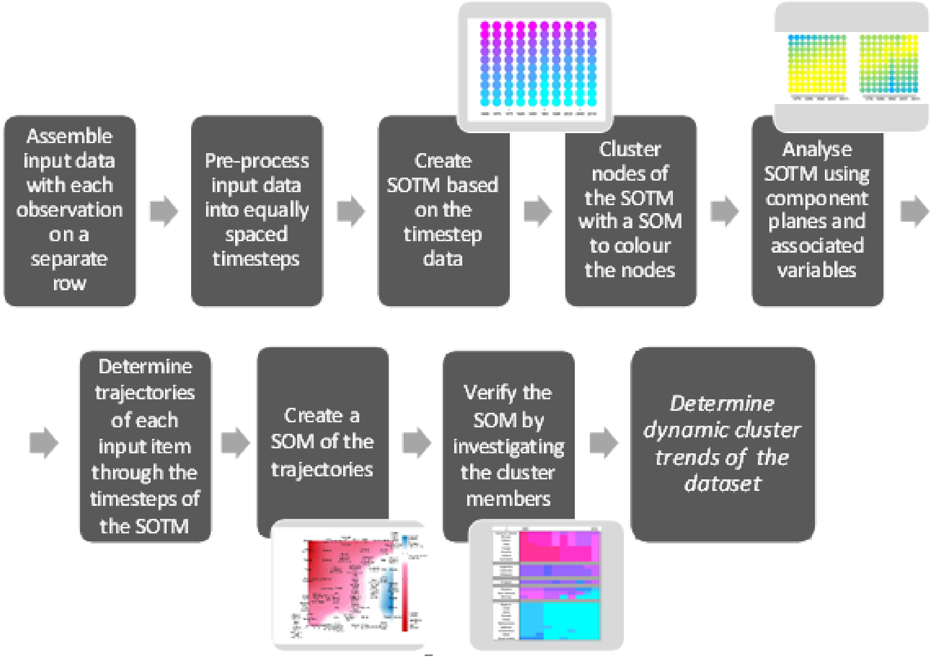

Assemble input data with each observation on a separate row
Pre-process input data into equally spaced timesteps
Create SOTM based on the timestep data
Cluster nodes of the SOTM with a SOM to colour the nodes
Analyse SOTM using component planes and associated variables
Determine trajectories of each input item through the timesteps of the SOTM
Create a SOM of the trajectories
Verify the SOM by investigating the cluster members
Determine dynamic cluster trends of the dataset

Figure 9: Summary of decision support tool method.

4.6 Conclusion

In this study, countries have been clustered in terms of their dependence on foreign water resources, as defined by the non-linear relationship between net virtual water imports through the food trade and available domestic renewable water resources. The temporal dynamics of these clusters from 1965 to 2010 have been investigated with a SOTM. Overall, a global trend towards increasing dependence on foreign water resources has become evident. Individual countries have been grouped based on similar movements through the SOTM, and these groups are presented in a single visual output that allows determination of countries with similar and diverging trends of water resource dependencies over the timeline of the study.

This study has introduced a workflow method (Section 4.5.5) resulting in a visual support tool that may inform national water resource management decisions, which must take into account the global flux of water resources due to the virtual water trade. Issues to be considered include whether this trend towards foreign water resource dependence is sustainable, and what the result of unforeseen alterations in the global food trade in the future may be.

The use of the SOTM and SOM have provided a means to reduce the large spatiotemporal hydrological and trade datasets into a specified number of representative vectors which are ordered based on similarity, providing an indication of similar data points. This allows for the possibility of further quantitative analysis of trends over the datasets, without having to manipulate the vast amounts of individual observations or define the non-linear relationships explicitly. The exploratory nature of the SOTM provides overall insights into the dynamic cluster structure of the spatiotemporal data. This method of analyzing trends, with the extensions to the SOTM outlined in this study, could be applied to a wide variety of datasets and is well suited to ecological and environmental data with missing values and data structures that are changing over time.

Future work on the decision support tool presented here may include improvements in both the technological aspects of the method, and in the data used. For example, more regionalized water footprint data could be used within this process, which would provide more accurate national results. Also, the specifications of the SOTM could be improved through a formal framework for evaluating the adequacy of a proposed configuration (such as the number of nodes at each timestep), and an extension of the SOTM for a possible projection of visible trends into the next time step could be considered.

4.7 References

Allan, J. A. (1998). Virtual water: A strategic resource global solutions to regional deficits. Ground Water, 36(4), 545-546. doi: 10.1111/j.1745-6584.1998.tb02825.x

Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/Iris]. Irvine, CA: University of California, School of Information and Computer Science.

Carr, J. A., D'Odorico, P., Laio, F., & Ridolfi, L. (2012). On the temporal variability of the virtual water network. Geophysical Research Letters, 39. doi: 10.1029/2012gl051247

Cereghino, R., & Park, Y. S. (2009). Review of the Self-Organizing Map (SOM) approach in water resources: Commentary. Environmental Modelling & Software, 24(8), 945-947. doi: 10.1016/j.envsoft.2009.01.008

Chappell, Geoffrey J, & Taylor, John G. (1993). The temporal Kohønen map. Neural networks, 6(3), 441-445.

Ercin, A. E., Mekonnen, M. M., & Hoekstra, A. Y. (2013). Sustainability of national consumption from a water resources perspective: The case study for France. Ecological Economics, 88, 133-147. doi: 10.1016/j.ecolecon.2013.01.015

Guo, D., Chen, J., MacEachren, A.M., & Liao, K. (2006). A visualization system for space-time and multivariate patterns (vis-stamp). Visualization and Computer Graphics, IEEE Transactions on, 12(6), 1461-1474.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: New York: Springer.

Hoekstra, A.Y., Chapagain, A.K., Aldaya, M.M., & Mekonnen, M.M. (2011). The water footprint assessment manual: Setting the global standard, Earthscan, London, UK.

Hoekstra, A. Y., & Mekonnen, M. M. (2012). The water footprint of humanity. Proceedings of the National Academy of Sciences of the United States of America, 109(9), 3232-3237. doi: 10.1073/pnas.1109936109

Kalteh, A. M., Hiorth, P., & Bemdtsson, R. (2008). Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. Environmental Modelling & Software, 23(7), 835-845. doi: 10.1016/j.envsoft.2007.10.001

Kaski, S., & Kohonen, T. (1996). Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world. Paper presented at the Neural networks in financial engineering. Proceedings of the third international conference on neural networks in the capital markets.

Kohonen, T. (1988). The'neural'phonetic typewriter. Computer, 21(3), 11-22.

Kohonen, T. (1991). The Hypermap Architecture. Amsterdam: Elsevier Science Publ B V.

Kohonen, T. (1993). Things you haven't heard about the Self-Organizing Map. Paper presented at the Neural Networks, 1993., IEEE International Conference on.

Kohonen, T. (1998). The self-organizing map. Neurocomputing, 21(1), 1-6.

Kohonen, T. (2001). Self-organizing maps (Vol. 30): Springer.

Kohonen, T. (2013). Essentials of the self-organizing map. Neural Networks, 37, 52-65. doi: 10.1016/j.neunet.2012.09.018

Konar, M., Dalin, C., Suweis, S., Hanasaki, N., Rinaldo, A., & Rodriguez-Iturbe, I. (2011). Water for food: The global virtual water trade network. Water Resources Research, 47. doi: 10.1029/2010wr010307

Mekonnen, M. M., & Hoekstra, A. Y. (2011). The green, blue and grey water footprint of crops and derived crop products. Hydrology and Earth System Sciences, 15(5), 1577-1600. doi: 10.5194/hess-15-1577-2011

Park, Young-Seuk, Céréghino, Régis, Compin, Arthur, & Lek, Sovan. (2003). Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. Ecological Modelling, 160(3), 265-280.

Samad, T., & Harp, S. A. (1992). Self-organization with partial data. Network-Computation in Neural Systems, 3(2), 205-212. doi: 10.1088/0954-898x/3/2/008

Sarlin P., (2013). Self-Organizing Time Map: An Abstraction of Temporal Multivariate Patterns. Neurocomputing 99(1), pp. 496--508.

Sarlin, P., & Yao, Z. Y. (2013). Clustering of the Self-Organizing Time Map. Neurocomputing, 121, 317-327. doi: 10.1016/j.neucom.2013.04.007

Shanmuganathan, S., Sallis, P., & Buckeridge, J. (2006). Self-organizing map methods in integrated modelling of environmental and economic systems. Environmental Modelling & Software, 21(9), 1247-1256. doi: 10.1016/j.envsoft.2005.04.011

Suweis, S., Rinaldo, A., Maritan, A., & D'Odorico, P. (2013). Water-controlled wealth of nations. Proceedings of the National Academy of Sciences of the United States of America, 110(11), 4230-4233. doi: 10.1073/pnas.1222452110

Tamea, S., Allamano, P., Carr, J. A., Claps, P., Laio, F., & Ridolfi, L. (2013). Local and global perspectives on the virtual water trade. Hydrology and Earth System Sciences, 17(3), 1205-1215. doi: 10.5194/hess-17-1205-2013

UN Water. (2009). The United Nations World Water Development Report 3: Earthscan.

Vesanto, J. (1999). SOM-based data visualization methods. Intelligent data analysis, 3(2), 111-126.

Vesanto, J. (2000). Neural network tool for data mining: SOM toolbox. Paper presented at the Proceedings of symposium on tool environments and development methods for intelligent systems (TOOLMET2000).

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. IEEE Transactions on Neural Networks, 11(3), 586-600. doi: 10.1109/72.846731

5 PAPER 2 - PARAMETER SELECTION FOR MAP SETUP

This chapter has been published as:

A dimension range representation measure for self-organizing maps

Clark S, Sisson SA, Sharma A. *Pattern Recognition.* 2016; 53: 276-86.

5.1 ABSTRACT

A common tool in exploratory data analysis, the self-organizing map, or SOM, is used for clustering and visualisation to discover patterns in large, high-dimensional data sets. The output map may be interpreted to gain an understanding of the structure of the original data set, correlations between variables, and the characteristics the clusters formed by placing the data on the map. However, if the map does not represent all dimensions of the data in an informative way, map interpretation may be misleading. Currently there is no measure of how well a SOM represents a data set in each dimension, and therefore how descriptive the map vectors are of the full structure of the data they represent. A dimension range representation (DRR) measure is proposed to quantify how well represented each dimension of the data set is by the map vectors of the SOM. This can be used to choose between different map size and shape options to represent a specific data set. Through examples, it is demonstrated how the DRR measure is used to inform the choice of map size and shape, leading to more informative insight into the original data set through examination of the output map.

5.2 INTRODUCTION

Similarities between individual data items are often difficult to discern when observing sizeable, high-dimensional data sets. Therefore, certain patterns inherent to the data set may evade observation. Clustering and visualizing the data items can aid in gaining insight into the characteristics and patterns in the data [1]. When doing so, however, it is valuable to understand if the clustering and visualisation represent the characteristics of all dimensions of the data set, or if some dimensions are not as well characterized as others.

A popular clustering and visualization technique, the self-organizing map, or SOM [2], is used in the process of exploratory data analysis to extract patterns and similarities from large, high-dimensional, nonlinear data sets [3, 4]. The SOM method performs data reduction (quantization), dimension reduction, and clustering of the data to produce a more manageable data set consisting of a smaller set of vectors in a lower dimension. This is accomplished by bending and stretching a (typically) one or two-dimensional map grid comprised of interconnected nodes to cover the data set. A map vector attributed to each node takes on the (high-dimensional) value of the data space occupied by the node. Each data item is assigned to one node, and the map (with assigned data) is presented as output in its one or two-dimensional form. The results are a set of groups, or clusters, of similar data arranged at each node with similar clusters close to each other on the output map [2].

The output map can be interpreted to gain a visual insight into: the shape of the data set, correlations between input variables, and any cluster structure present in the data [5]. The visualisation enables an analysis of the distribution of the individual variables in the data [6]. Investigating the placement of the data on the output map allows structure and patterns of the data set to be observed [1]. All data points allocated to the same location on the map are understood to have similar features [1], and it is also assumed that the map vectors are representative of all the data mapping to them [4].

In light of these common interpretations, it is important that the map vectors represent all variables (or dimensions) of the data in an informative way. If a map vector does not represent certain variables as accurately as others, interpretation of the map in the ways listed above may be misleading. An imbalanced representation of dimensions may occur if the map grid does not overlay the data of each variable to a similar extent. If the output map represents certain dimensions more fully than others, the resulting groups, or clusters, of data assigned to each node will effectively represent clusters based only on the more well-represented variable(s). In this case, interpretation of the map may not be as revealing about the structure of the entire data set, correlations between the variables, the range of the individual variables, or the characteristics of the data clusters assigned to these nodes.

Insight into the characteristics of the data set could be enhanced by ensuring the map vectors represent the ranges of all the variables as completely as possible. The size and shape of the output map (the number of map nodes and the length/width ratio of the nodes) influence the ability of the map to effectively cover the data. Determination of the optimal map size and shape is a key challenge in producing a SOM [3, 7], as these 'setup parameters' must be specified by the user before the data analysis begins, and different parameter choices may result in different SOM output patterns [4, 8]. Therefore, it is of particular importance in establishing good map coverage to choose the map size and shape that will best represent each dimension of the data. When analysing the output map, there is currently no indication if any dimensions are not as well represented by the map as others. Inspection of each map vector will show the dimension values of the map, but not the portions of the data range outside the map in each dimension; what is missed by the map is not evident.

In this paper, it is shown how the choice of map size and shape influences whether the map fits the data better in some dimensions than in others. A 'dimension range representation' measure (DRR) is presented to explicitly quantify the representation of the data in each dimension by the map grid. This measure can be used to inform the choice of map size and shape, leading to more accurate insight being gained from interpretation of the map.

The paper is structured as follows: Section 5.3 reviews relevant SOMs theory and literature; Section 5.4 describes in detail the issue to be investigated and introduces the DRR measure; and in Section 5.5 the measure is explored on three examples. It is demonstrated that with the use of the DRR measure, the selection of map size and shape can be refined to lead to a better fit by the map to the extents of all variables in the data set, therefore providing better representation of the data by the map vectors.

5.3 Background

5.3.1 SOM overview and map training

The SOM is from the family of artificial neural networks, and performs a type of non-linear regression on complex data sets. It allows clusters, patterns and relationships to be extracted from large amounts of data without requiring an explicit understanding of the underlying correlations between the variables. The topology of the data set is preserved by the arrangement of the nodes on the grid, with the spatial location of each map node corresponding to a certain subspace of the input data [2]. Similar input items become located close to each other on the output map. Clustering is the most common implementation of the SOM [9], as clusters and dissimilarities in the data become easily apparent on the map with nearby clusters being more similar than distant clusters.

A SOM is created by initializing a map grid based on the basic input data structure and then training (stretching and bending) the grid to better represent the data [5]. Initialisation is typically performed by creating a linearly spaced grid along the largest principal component(s) of the data set. Map training follows a nonparametric, recursive regression process [10] where each item of input data is assigned to a single map node (its best matching unit, or BMU) and then the locations of the map nodes are updated to better match the data. In the updating phase, each map node is moved to the weighted average of all the data points mapping to it and to its neighbours [2]. The weighting is performed by a neighbourhood function that decreases in size with each iteration. SOMs notation used in this paper is as follows: Input data size: N; D-dimensional input data vectors: x_i (where i=1…N); Number of map nodes: M; D-dimensional map vectors: m_j, (where j=1…M); best matching map unit (BMU) for data vector x_i: m_c.

For a more detailed explanation of SOM training, see [2, 7, 11].

5.3.2 Literature review

The literature provides limited guidance on the choice of output map size and shape [12]. Since much time and experience is required to optimize these parameters [13], the number of map nodes is often simply chosen by the heuristic $5 * \sqrt{N}$ (for example as in [14]) based only on the number of input data items and not considering the output map vectors. This method is the default in the MATLAB SOM toolbox [15]. The most recommended and commonly used method for choosing map size and shape, though, is by trial and error of a number of maps made with different parameters [4, 7]. This involves the assessment of quality measures (discussed below) based on the data and the map vectors, as in [16, 17, 18]. An evaluation of the cluster structure of the input data is also a frequently used technique to try to match the number of output nodes to the number of clusters existing in the data [3]. Map size may also be based on the desired visual outcome or degree of generalization required, as in [19, 20, 21, 22]. For map shape, Kohonen [7] recommends that the ratio of the length and width of the map is as close as possible to the ratio of the first and second eigenvalues of the data set; this is the default shape used by the MATLAB SOM toolbox. Choosing the number and configuration of map nodes presents a specific challenge as the SOM does not follow the minimization of a single objective function [23], due to the opposing SOM aims of quantization (data reduction) and topology-preserving projection (visualisation). Consequently, it is not possible to determine map setup based on a maximum likelihood specification. In the absence of an objective function, quality measures are therefore commonly used to choose between potential map setups [24, 25].

A variety of SOM quality measures exist, and some are more widely used than others in the selection of setup parameters. Most commonly used are: the quantization error, QE (how well the map matches the data) [2], and the topographic error, TE (how well the topology of the data set is preserved on the map) [26]. Usually, a combination of quantization and topographic errors is investigated, to allow for the competing goals of data reduction and topology preservation to be considered [24]. Due to the trade-off between these two competing goals, as QE decreases TE will generally increase, though not always monotonically. As it is therefore not possible to minimise both QE and TE at the same time, the user must determine the balance between them. This is often done by examining plots of QE and TE to choose an optimal map from a set of maps [24], and is a subjective exercise based on the weighting attributed to each by the user. Other measures which are less commonly used include: the distortion measure [2], the goodness measure [27], and cluster measures such as the Davies-Bouldin index [28]. A brief overview of these measures is given here.

The **quantization error**, QE, is a common measure of the ability of the relatively smaller number of SOM map nodes to represent the relatively large amount of input data. It is a measure of map resolution, and quantifies how close the map nodes are to the input data, in Euclidean distance. The optimal map for the same input data will be the one yielding the smallest average quantization error, if the main priority is vector quantization. Quantization error is defined as the average of the Euclidean distances between each data item and its BMU, for all input items: $QE = \frac{1}{N}\sum_{i=1}^{N}\|x_i - m_c\|^2$. QE is useful for comparing the SOM to other clustering or vector quantization methods, though it is not particularly useful on its own for comparing maps with different numbers of nodes as it decreases with increasing map size. **Topographic error**, TE, is a measure of the preservation of the topology of the input data on the output map. It is defined as the proportion of times the first and second BMUs are not nearest neighbours on the map grid, summed over all the data points. For each data point, the BMU and second BMU are checked to see if they are adjacent: $TE = \frac{1}{N}\sum_{i=1}^{N}u_{x_i}$, where $u_{x_i} = 1$ if the first and second BMUs of x_i are nearest neighbours, 0 otherwise. A single topographic error value represents the entire map, with a higher value indicating more imperfect topological representation of the data by the map, and a value of zero indicating perfect topology preservation. The **distortion measure**, DM, is similar to QE, but differs in that each squared distance between a data point and map unit is weighted by the value of the neighbourhood function. While QE considers distances between each input, x_i, and it's BMU, m_c, the distortion measure includes distances between each input, x_i, and all map units, m_j, with these distances weighted according to the neighbourhood function, h_{ic}, for the BMU of that particular input value. DM is defined as: $DM = \sum_{i=1}^{N}\sum_{j=1}^{M} h_{ic}\left\|x_i - m_j\right\|^2$. The **goodness measure**, C, combines QE and the distance on the map grid from the data point to the first and then the second BMU, via nearest neighbours. Distances are calculated in input space along the surface formed by the SOM, and averaged over all data points. The goodness measure can be used to choose maps that do not fold unnecessarily. **Cluster measures**, based on the ratio of within-cluster and between-cluster variation, are often used for estimating the number of clusters existing in a data set [29]. The goal is to ensure that the data within each cluster is as similar as possible, and as different as possible from data in other clusters [30]. The SOM output map size is often chosen based on the number of clusters that are believed to exist in the input data, with the number of map nodes matching the expected cluster number [3]. The Davies-Bouldin Index provides a measure of a certain clustering of a data set, with lower values indicating better clustering. This index reports the ratio of within cluster scatter to the separation between clusters, looking at each cluster and its most similar one.

The literature also offers some research defining the spread of data covered by a SOM, focussing on the fraction of variance unexplained (FVU). Lee [31] used the square root of the FVU to measure network error in neural networks. Akinduko & Mirkes [32] define FVU as the dimensionless least square evaluation of error, and use this to compare SOMs produced by different methods of initialization. Mirkes [33] uses the FVU to set the number of nodes in a growing SOM (GSOM). The numerator of the FVU is the sum of squared distances from the data to the approximating line (ie. the grid line connecting the two nearest map nodes), and the denominator is the sum of squared distances from the data to the data mean [32]. The FVU is defined as: $FVU = \frac{\sum_{i=1}^{N} d^2(x_i)}{\sum_{i=1}^{N}\|x_i-\bar{x}\|^2}$, where $d(x_i)$ is the distance from the data point to the grid (including nodes and connecting lines between nodes). The FVU is essentially the ratio of the QE to the variance of the data. It has not been developed beyond a 1-dimensional SOM, and it seems complications may be encountered in a 2D SOM when the direction to the BMU and to the mean are not always the same, and may even be in opposing directions.

5.4 PROPOSED MEASURE

To our knowledge, no measure currently exists of how well a SOM represents a data set in each dimension, and therefore how descriptive the map vectors are of the full structure of the data assigned to them. Such a measure will indicate if the map vectors provide more accurate information on some dimensions of the resulting data clusters than others.

5.4.1 Importance of ensuring a good fit of the map to each dimension

The creation of a SOM is based upon the assumption that the features of the data set can be captured with a one-dimensional line or a two-dimensional rectangle [1]. However, two issues are encountered when matching a map to a data set: 1) the map will not ever reach the edges of each dimension, and 2) the map may reach closer to the edges of some dimensions than others causing an imbalance in their representations. These issues are expanded in the following two paragraphs.

The updating step of the SOM training algorithm entails the location of each map node being updated based on the location of the data points closest to it as well as the data points closest to the neighbouring nodes. The involvement of the data assigned to the neighbouring nodes results in the nodes being drawn towards each other and towards the centre of the data, ensuring that the boundary of the SOM grid will not quite ever reach the boundary of the data [9]. Because of this, nodes near the edge of the SOM come to represent larger portions of the input space than the interior map nodes [9]. This can lead to a loss of information being described by the map vectors at these edge nodes, due to the higher dispersion of data values allocated to each node. This is an important consideration in most SOMs applications as the extremities of complex, real-world data sets often contain valuable information which may not be captured on the SOM.

Furthermore, a data set will commonly have different distributions of data in each of its dimensions, so the map may come to have a larger proportion of data mapping to a single edge node in some dimensions than in others. For example, it can be imagined that a rectangle attempting to represent a high-dimensional data set will not necessarily be able to reach the same distance towards the boundaries of all of the dimensions. If the map does not reach the extremities of some dimensions as well as others, the nodes to which the outer data is allocated will have a larger spread of values in the under-represented dimension(s). Data assigned to a node may match the node's corresponding map vector very well in some dimensions and not so well in others. When interpreting the clusters created by placing the data on the map, this leads to less information being revealed about these dimensions compared to the others.

In order to investigate how well a map represents the data in each dimension, it is currently necessary to perform a detailed visual investigation of plots of 2- or 3-dimensional subsets of the data with the map overlayed (as shown in Figure 1 for 2-dimensional data). This technique is clearly infeasible for high-dimensional data, as it involves too many plots to visually assess.

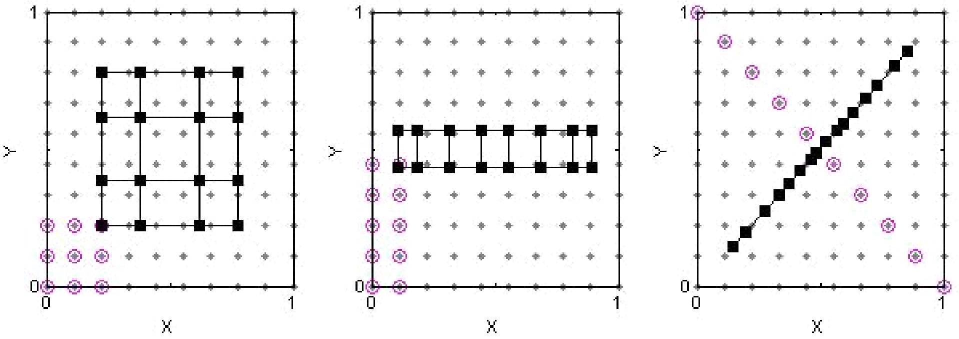


Figure 10: A synthetic two-dimensional data set representing a uniform grid (grey dots), with different SOMs overlayed (black squares). Each of the SOMs consists of 16 nodes in a variety of configurations. On each plot, the circled data are all allocated to a single node. On the plots from left to right, the data assigned to a single node becomes more dispersed. In the centre plot, the data is more dispersed in the Y dimension than the X dimension, indicating that assumptions about data characteristics based on the map vector for that node would be more accurate for the X variable. On the right hand plot, no information is gained about the values of either the X variable or the Y variable, as the node represents data from 100% of the range of both variables.

Figure 10 shows a synthetic two-dimensional data set in a uniform grid (grey dots), with a selection of trained SOMs overlayed (black squares). Each SOM consists of the same number of nodes (16) in different configurations. On each plot, the data allocated to a single node is circled. It can be seen that this circled data comprises certain (possibly different) proportions of each dimension.

On the first plot, data from 22% of the range of the X dimension and 22% of the range of the Y dimension are allocated to the corner node. On the second plot, 11% of the range of the X dimension, and 44% of the range of the Y dimension are allocated to the example node. The cluster of data created at this node is therefore more closely related to each other in the X dimension. The output map will provide more detailed information with respect to the X variable for the cluster members than for the Y values. Interpretation of this map would also imply the data set was rectangular rather than square, as the range of Y values of the map vectors is much smaller than the range of X values. On the third plot, data from 100% of the range of both dimensions is represented by the example node. Therefore, no information about either the X or Y values of the data clustered at this node would be gained by inspection of the map.

When analysing the maps to gain insight into the data set, it can therefore be seen that a map representing the extents of some dimensions better than others may lead to:

- a generalized assignment of map vector values to same-cluster data that are in fact very diverse (specifically, each node of the second map of Figure 10 will attribute the same Y value to half of the range of that variable's data), and
- an under-estimation of the size of the data with respect to the under-represented variable(s) (specifically, in Figure 10 the second map will indicate the data has a much larger range in the X dimension than the Y dimension).

5.4.2 The DRR measure

A 'dimension range representation' measure (DRR) is proposed to quantify how well represented each dimension of the data set is by the map vectors of a SOM. This measure can be used to choose between different map size and shape options to represent a specific data set.

As discussed in Section 5.4.1, due to the training process of the SOM the outer map nodes will never quite reach the outer boundary of the data, leaving a gap between the boundary of the map and the boundary of the data set. The amount of data outside the map will vary across the dimensions (Figure 10 indicates how different maps may reach closer to the boundaries of the data in one dimension than in others). No matter how diverse, the data outside the map boundaries will be allocated in a group to the closest map node.

To illustrate what is being measured by the DRR, the boxplot [34] is used in Figure 11 to depict the coverage of the map over the data, for each dimension of a synthetic 3-dimensional data set. The three dimensions of data are shown separately (white boxplots) with the corresponding map ranges for each dimension (black boxes) to the right of them. This visualization removes the need to plot the data points and overlayed map nodes in dimension pairs in order to inspect map coverage, as was done in Figure 10. The boxplot gives an indication of the dispersion and skewness present in the data, as ranked data is split into four even groups, each containing a quarter of the data. Due to the non-normal distributions of data commonly used in SOMs applications, all outliers are included within the whiskers of the data boxplots. The map ranges are represented as solid boxes next to each dimensions' data boxplot. This comparison of the data and map boxes indicates which dimensions, if any, are less well covered by the extents of the map. Data that is outside (above or below) the range of the map will be assigned to the closest boundary node, and it is this data that is being measured by the DRR.

It is evident on Figure 11 that in dimension 1 the map does not cover even the inner 50% of the data; in dimension 2 the map reaches partway into the upper and lower ranges of the data; and in dimension 3 the map reaches the lower boundary of the data and comes closer to the upper boundary than the other map dimensions. Therefore, the range of dimension 3 of the data is best represented by this particular map, and an analysis of the map vectors and the clusters of data produced on the output map will reveal more information about variable 3 than variable 2, and to a lesser extent, variable 1.

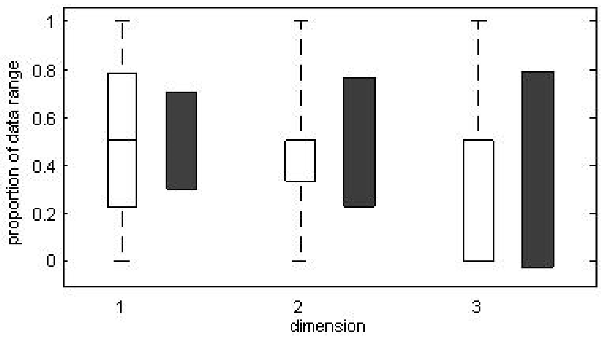


Figure 11: Boxplots of a 3D synthetic data set (white) and its map grid (black) provide a visual indication of the coverage of the data by the map, separately for each dimension. The data range is the same in each dimension, but the map does not cover the data to the same extent for each dimension (for this data set, dimension 3 is best represented by the map). In this way, many dimensions can be shown on a single output, allowing all variables to be investigated simultaneously. The data outside the range of the map (where the boxes don't overlap) becomes clustered together at the edge of the map. It is this data that is quantified by the DRR measure.

The DRR measure assesses the maximum intra-cluster spread of data in each dimension that is assigned to a single node. For each dimension, the DRR measure is quantified by investigating the maximum difference between all data points allocated to each map node. The node with the greatest range of assigned data is determined in each dimension. This range of data is then calculated as a proportion of the overall range of values in that dimension. The result is that for each dimension, d, the maximum proportion of the input data range assigned to a single node of a map is given by the DRR measure:

$$DRR(d) = \max_{j} \frac{\max_{ij}(x_{ij}(d)) - \min_{ij}(x_{ij}(d))}{\max_{i}(x_i(d)) - \min_{i}(x_i(d))}$$

where $x_i(d)$ are all the data values in dimension d, and $x_{ij}(d)$ are the data values in dimension d that are assigned to map unit j.

The DRR measure therefore provides an indication of the largest proportion of each dimension that is assigned to a single node. This measure differs from the commonly used QE in that whilst QE quantifies the average distance of all data points from their assigned node to assess overall quantization, the DRR measure is concerned with the diversity between data items assigned to the same node. This gives an indication of how well the corresponding map vector will represent the assigned data and how much insight the trained map will provide about each variable.

5.5 EXAMPLES

The DRR measure is illustrated on two synthetic examples and one real-world example. Using this measure, we investigate if an optimal number and configuration of map nodes can be selected that will ensure the maximum coverage of the data in each dimension.

The first example highlights how the DRR measure can be used to choose an optimal map size from a subset of map size options suggested by the frequently used QE and TE measures. The second example shows how the DRR measure can be used to choose an optimal map shape, for a user-determined number of nodes. These examples involve 3- and 2-dimensional data sets, respectively, mapped to 2-dimensional grids. This allows for straightforward concept visualisation, but this method is of most use when the data set is of much higher dimension than the map grid, and visually investigating the coverage of the map over the many dimensions of the data would not be feasible.

The third example applies the DRR measure to an investigation of a real-world data set in four dimensions (the progress of Millennium Development Goal 7C with respect to water and sanitation) in which it is used to improve insights into the clusters of countries with similar progress towards the four aspects of the goal.

The examples have been produced with the aid of the MATLAB SOM Toolbox, employing the batch training algorithm and a Gaussian neighbourhood function. For the first example, map configurations have been chosen to correspond to the ratio of the two largest eigenvalues of the data.

5.5.1 Example 1: Map size selection

A 3-dimensional random data set is distributed uniformly on the unit cube, as shown in Figure 12a. We aim to create a SOM with an ideal number of nodes to ensure that each of the three dimensions is represented as fully as possible by the trained map.

To begin, the QE and TE measures are used to obtain a subset of potential map sizes based on the goals of data quantization and topology preservation. Plots of QE and TE versus the number of map nodes, in Figure 12b, indicate that there are a number of potential map sizes that could be chosen for this data set based on attempting to jointly minimize QE and TE, such as in the vicinity of 90 or 120 nodes. Incorporation of the DRR measure can lead to these options being refined.

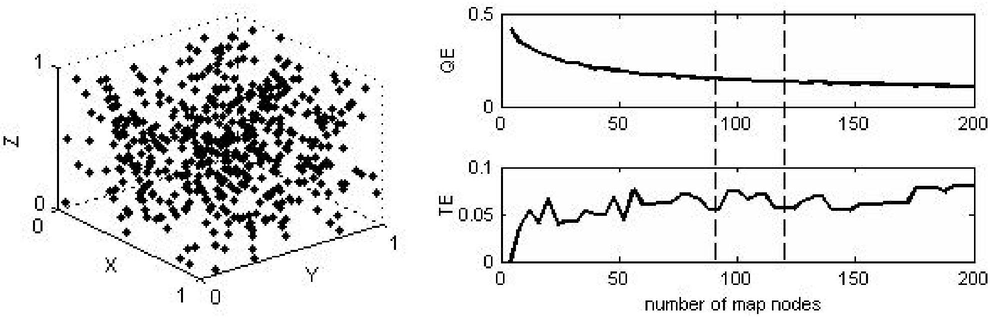


Figure 12: Random 3D data. a) 500 data points are distributed uniformly over 3 dimensions. b) Plots of quantization error (QE) and topographic error (TE) versus the number of nodes are often used together to choose an appropriate map size to best represent the data (by seeking to minimise both).

The DRR measure indicates that the maximum proportions of each dimension (X, Y, and Z) that become assigned to a single node with map size 90 are: 34%, 44%, and 36% respectively. With 120 nodes, the proportions assigned to a single node are: 39%, 52% and 32%. This implies that a map with 90 nodes will provide improved coverage (lower proportions of each variable allocated to a single node) over the X and Y dimensions than a map with 120 nodes. The map vectors corresponding to these nodes will therefore provide more information about the spread of the X and Y variables of the data.

The DRR measure is therefore able to supplement the currently available information used in map size selection, leading to a map that better reveals the characteristics of each variable. When comparing choices with similar QE and TE values, the information provided by the DRR has been informative in choosing between possible map sizes, as in this case it appears the smaller map will be preferable even though QE decreases monotonically with increasing map size.

5.5.2 Example 2: Map configuration selection

As mentioned in Section 5.3, the number of map nodes (and therefore the number of clusters in the output) is often pre-determined by the user based on criteria other than quality measures, such as the desired level of accuracy of the output information. Once the number of nodes has been selected, there is also the option for the user to specify the node configuration (the length and width of the map grid) and the specified length and width will have an effect on how much information is retained from each of the dimensions, as described in Section 5.4.1. The DRR measure can be used to inform this choice by providing information on how well the grid overlays each dimension of the data, for each grid configuration option.

This process is depicted in Figure 13. The left hand plot shows a data set of 800 points in two dimensions (data adapted from [11]). We aim to create a SOM that represents the large data set with a much smaller set of map nodes. For this example, we assume the user has specified a map size of 15 nodes in the desire to balance output accuracy and generalisation, and to provide an easily observed number of output clusters. Clearly, a grid of 15 nodes could be configured either as 1x15 or 3x5 nodes.

The centre and right hand plots of Figure 13 show the 1x15 and 3x5 map grids covering the same data set. Data points that are assigned to a single (example) node have been highlighted with circles on each of these plots.

It can be seen on the 1x15 map that the data set will be assigned to nodes roughly in vertical slices; the data assigned to each node effectively forms a vertical band from top to bottom of the data set. Data points with low, medium and high values of dimension Y are assigned to the same node. This indicates that the information that can be gained by investigating the groups of data assigned to each node pertains mostly to the differences in the values of dimension X.

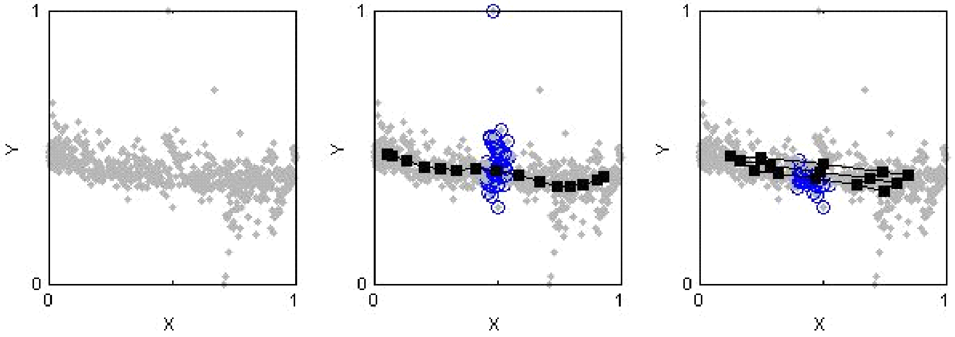


Figure 13: Data set of 800 points in two dimensions [11]. Maps with 15 nodes (black squares) are trained to represent the same data set with different configurations (1x15 in the centre, 3x5 on the right). Data that plots to a single map node are highlighted with circles on the centre and right plots. It is evident that nodes in the 3x5 configuration represent a smaller range of the Y dimension of the data set than nodes in the 1x15 configuration, and therefore the map vectors associated with these nodes will better characterize the data assigned to them.

On the right, the 3x5 map shows the data is split into more even groups horizontally as well as vertically. In this case, data assigned to a single node may only have either high, medium, or low values of dimension Y. This indicates an improvement in insight that will be gained from investigating the data assignments and the map vectors, as the map vectors will better characterize all dimensions of the data assigned to them. Now each node will provide more accurate information on both dimension X and dimension Y values of the clusters of data assigned to it, and there will be more intra-cluster similarities between data assigned to each node.

It is apparent from Figure 13 that the specified map shape will affect the representation of the data by the map vectors, even though the number of nodes in the map remains the same. Using the DRR measure, a set of map configurations may be tested to determine which provides the best coverage for all dimensions of the data. This will avoid the map being trained to effectively represent only a subset of the variables, as in the centre plot where the map vectors predominantly represent differences in variable X. Using the DRR measure, it has been determined that nodes of the 3x5 grid represent a smaller range of the Y dimension. This can be confirmed visually on the 2-dimensional example in Figure 13, but would be difficult to visually assess in higher dimensions.

5.5.3 Real-world example: Achievement of Millennium Development Goal Target 7.C

The DRR is applied to an analysis of the achievement of the United Nations' Millennium Development Goal Target 7.C: 'Halve, by 2015, the proportion of people without sustainable access to safe drinking water and basic sanitation'. The aim in creating a SOM from this data is to produce a visual clustering of countries, based on similar levels in each of the four areas of: improved rural and urban water sources, and improved rural and urban sanitation facilities. The goal is to produce clusters of countries with similar conditions to visually compare the relative state of development (with respect to water and sanitation) amongst the countries. The SOM is a useful method for this data to provide not only a high level visual overview of global conditions, but also an indication of countries in similar stages of development with regards to water and sanitation.

Data has been obtained from the Millennium Development Goals Database of the United Nations Statistics Division [35] (a detailed description of the data can be found on the website of the WHO/UNICEF Joint Monitoring Programme for Water Supply and Sanitation [36]). The data set, as shown in Figure 14, consists of 142 countries and the following four variables (2012 data):

- Proportion of rural population using improved drinking water sources
- Proportion of urban population using improved drinking water sources
- Proportion of rural population using improved sanitation facilities
- Proportion of urban population using improved sanitation facilities

Determining groups of countries with similarities in each of the four variables is infeasible by visual inspection alone (as is evident by the complexity of Figure 14). Therefore, a grouping and visualisation method such as the SOM is beneficial for finding countries with similar patterns across all four variables.

A SOM trained with this data using default setup parameters would consist of 64 map nodes in a 16x4 configuration. The SOM constructs groups at each of the 64 nodes, consisting of countries with similar levels in each of the four variables. In the output map created with default parameters, though, 49 out of the 64 groups are assigned only 0, 1 or 2 countries.

For this particular application, larger clusters would be desirable in order to draw conclusions about countries with similar states of water and sanitation development. In order to obtain a more substantial number of countries in each group to allow for comparisons, whilst also maintaining enough groups to allow differences to be evident, we aim for approximately 24 nodes. This user-specified number of nodes should provide larger groups of countries, whilst maintaining a sufficient number of groups to show the various patterns in the data (though of course the user may have chosen any number of nodes that they preferred).

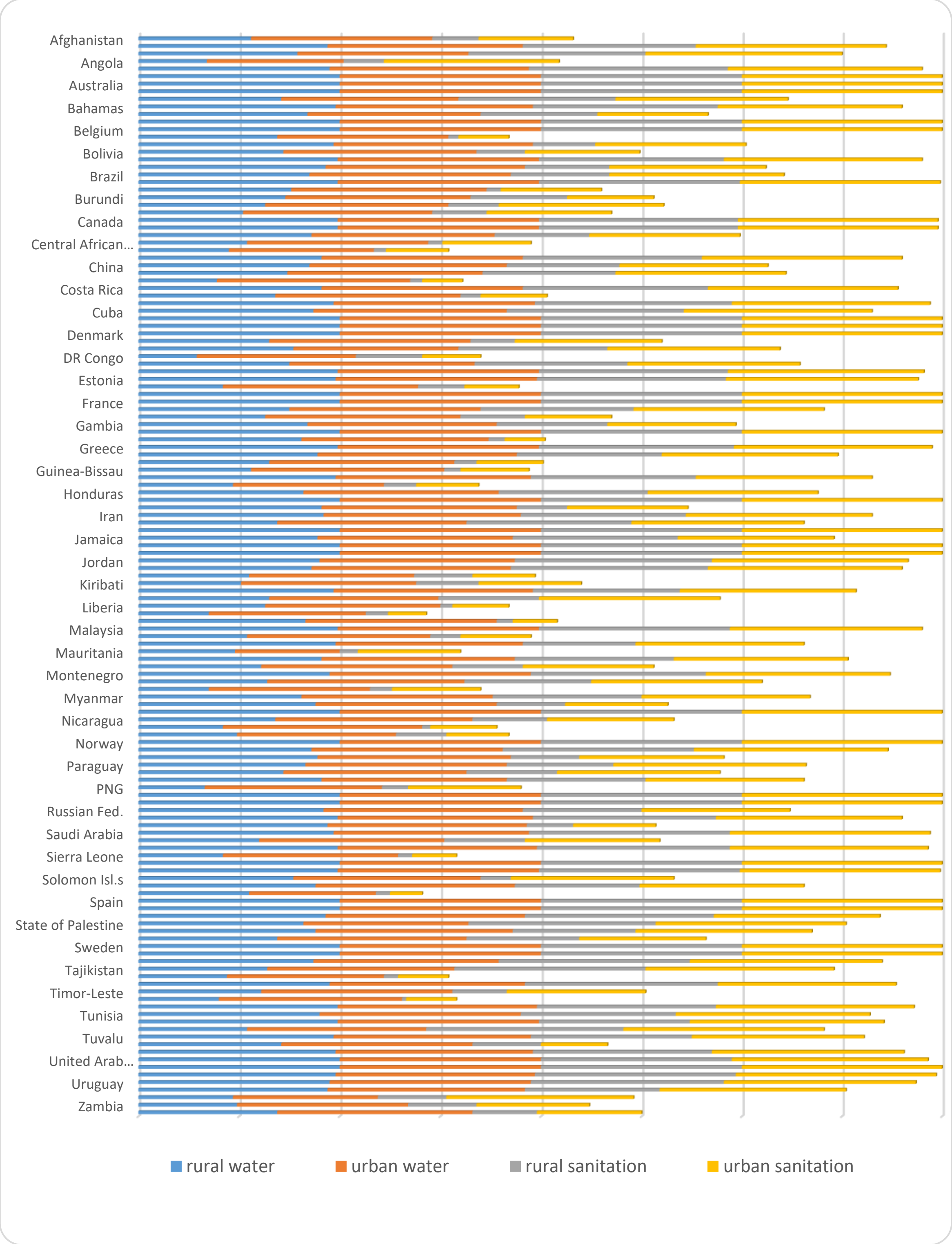


Figure 14: United Nations' Millennium Development Goals data consisting of 142 countries and 4 variables. Groups of countries with similar states of development with regards to usage of improved rural and urban water and sanitation facilities are not obvious from a visual inspection of the data set. Due to minimum text sizes, not all data is labelled.

It is now important to find the best configuration of the 24 nodes (4x6 or 8x3) to provide the most explanatory groupings with regards to the four variables. This will be done using the DRR measure.

We will start with the 4x6 map. Figure 15 shows the SOM (black squares) plotted over the data in dimension pairs (grey dots) for a map with 24 nodes in a 4x6 configuration. It can be seen that the map does not come close to the boundaries of the data for any dimension, except in the top right corner which represents the highest values of each variable. In particular, the map does not spread far in the 'urban water' dimension, compared with the other dimensions. It appears that the map only covers approximately the upper 40% of the range of urban water values.

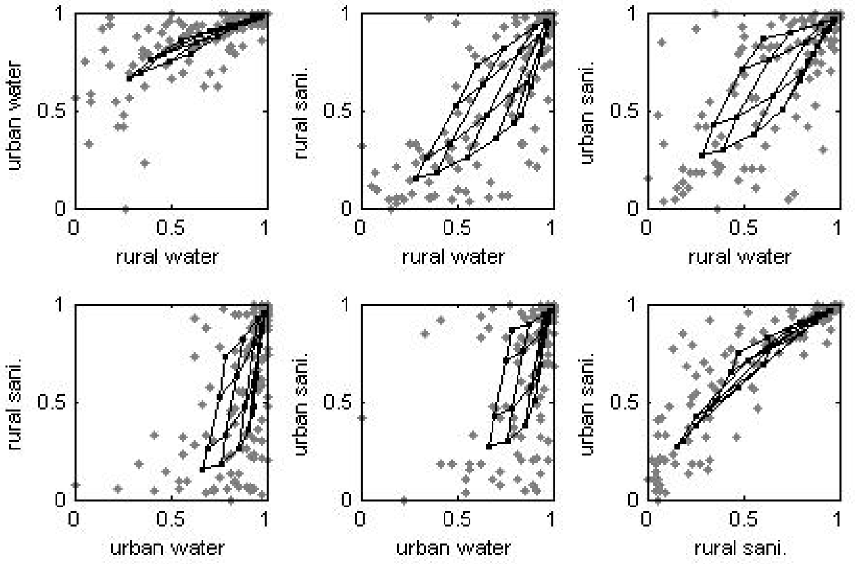


Figure 15: A SOM (black squares) with 24 nodes in a 4x6 configuration overlaying the input data (grey dots) in plots of all combinations of dimension pairs. It is clear that the map does not come close to the outer boundaries of the data set in any direction except the top right corner. The urban water dimension is the least well covered by the map.

Figure 16 shows boxplots representing the coverage of the map over the data for each of the four dimensions (rural water, urban water, rural sanitation, and urban sanitation), giving a visual representation of the DRR measure. The white boxplots represent the (standardized) data for each dimension, and the black boxes are the corresponding map ranges. From Figure 16, it can again be seen that the urban water dimension is not well covered by the map. A large portion of the lower data range is not overlapped by the map range. The data in this lower region will all be assigned together to the nearest map node. In comparison, rural and urban sanitation, and rural water have more complete coverage by the map.

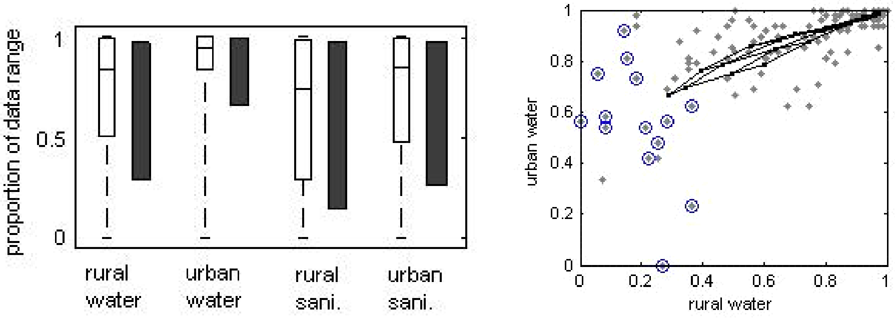


Figure 16: a) Boxplots of the data (white) and the map (black). The four dimensions are shown from left to right: rural water, urban water, rural sanitation and urban sanitation. b) The 4x6 SOM (black squares) overlayed on two dimensions of the data (grey dots). Circled data is all represented by a single node. Interpretation of the cluster at this map node will not provide much information on the state of urban water for the member countries.

The second plot of Figure 16 highlights with circles the set of data points that are assigned to a single node of this 4x6 map. The DRR measure indicates that the proportions of each dimension (rural water, urban water, rural sanitation and urban sanitation) that map to this node are: 41%, 92%, 42% and 48% respectively. Information gained from interpreting the output map will not contain much insight into the state of urban water for the countries assigned to this node, as the proportion of the urban water range covered is so large. These countries and their respective values for each variable (% of population with access to improved water and sanitation facilities) are listed in Table 3.

Table 3: Percentages of the population with access to improved urban and rural water and sanitation facilities, for countries that are assigned to a single node. The DRR measure indicates that this node covers 92% of the range of urban water values in the input data. When interpreting the map, little information will be gained about the state of urban water in this group of countries.

Country	Rural Water	Urban Water	Rural Sanitation	Urban Sanitation
Chad	45	72	6	31
Congo	39	96	6	20
DRC	29	79	33	29
Haiti	47	75	16	31
Kenya	55	82	29	31
Madagascar	35	78	11	19
Mauritania	48	52	9	51
Mozambique	35	80	11	44
Nigeria	49	79	25	31
Papua New Guinea	33	88	13	56
Sierra Leone	42	87	7	22
South Sudan	55	63	7	16
Togo	40	91	2	25
United Republic of Tanzania	44	78	7	25

This single map node has data mapping to it from 92% of the range of the urban water component (Mauritania, with 52% access to improved urban water sources is in the same grouping as the Congo with 96% access). Therefore, the map vector corresponding to this node is not as descriptive of the urban water dimension of its allocated data points as it is of the other dimensions, and useful information about the state of urban water amongst its member countries will not be revealed. The cluster formed at this node is effectively a cluster based only on the other three components.

This is not the case with other nodes of the same map, though, such as the node detailed in Table 4 for which the low DRR values indicate a good clustering in all dimensions.

Table 4: Low DRR values in all variables indicate this is an example of a node which provides good information about each of the 4 variables of the member countries.

Country	Rural Water	Urban Water	Rural Sanitation	Urban Sanitation
Estonia	98	100	94	96
Serbia	99	99	96	99
Egypt	99	100	94	98
Bosnia & Herzegovina	99	100	92	99
United Arab Emirates	100	100	95	98
Malaysia	99	100	95	96

The DRR measure shows that the coverage of the map over the data can be improved by changing the node configuration to 8x3. This configuration will reduce the proportion of the range of each dimension plotting to individual nodes. Table 5 shows the results of applying the DRR measure to each map configuration. The maximum proportions of the data range of each variable assigned to a single node are given for the 4x6 configuration and the 8x3 configuration. In particular, the spread of the urban water component represented by a single node is reduced from 92% to 58%. Inspection of the map vector corresponding to this node will now provide more accurate information about the state of urban water in these countries.

Table 5: The DRR measure (maximum proportion of the range of each dimension of the data that becomes represented by a single node) on a 4x6 map and an 8x3 map. Lower values indicate the map is spread more effectively over the data. The coverage of all dimensions by the map is improved by changing from a 4x6 configuration to 8x3, with the most notable improvement being for the urban water variable.

Configuration	Rural water	Urban water	Rural Sanitation	Urban Sanitation
4x6	41%	92%	42%	48%
8x3	38%	58%	33%	43%

Figure 17 indicates in circles the data assigned to this single node with the 8x3 configuration. In comparison to Figure 16b, is can be seen this is a considerable improvement in the representation of the urban water dimension by the map.

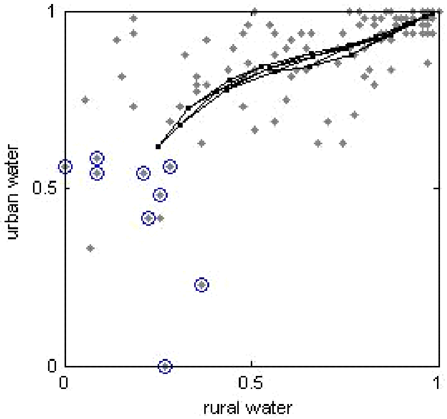


Figure 17: The 8x3 SOM (black squares) overlayed on two dimensions of the data (grey dots). Circled data is all assigned in a group to a single node. The 8x3 configuration provides better segmentation of the data than 4x6, as a smaller proportion of the range of urban water is now represented by this node. The map vector related to this node will now provide more information about the urban water dimension of the data points it represents.

As confirmation of the improvement in the clustering of the data as a result of applying the DRR measure, Figure 18 shows the Davies-Bouldin cluster index on the nodes of the 8x3 configuration (right) compared to the 4x6 configuration (left). The 8x3 map leads to fewer large values of the index, indicating better clustering.

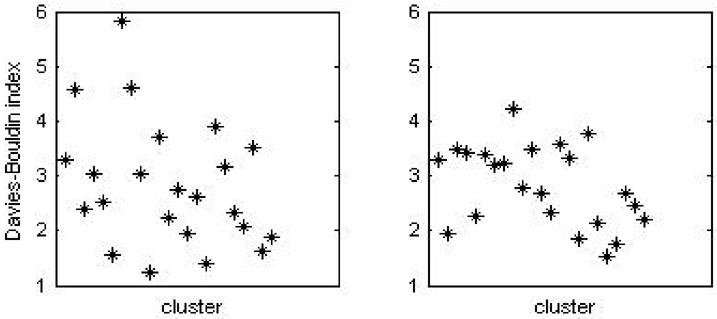


Figure 18: Davies-Bouldin indices for the 24 nodes in 4x6 (left) and 8x3 (right) configurations. Lower values indicate better clustering.

5.6 CONCLUSION

The DRR measure has been introduced to quantify how well a SOM represents each dimension of a data set, and therefore how representative the resulting clusters at each node are of the structure of the data. This will indicate if any dimensions of input data are under-represented in the creation of the map. We have shown that the data may match the allocated map vector better in some dimensions than others, and yet the map vector will be interpreted as being representative of the data in all dimensions.

The measure may be used in conjunction with existing SOM quality measures to inform the choice of the number and configuration of output map nodes. Incorporation of the DRR measure in the map setup process allows for comparison between competing map sizes and shapes attempting to represent the same data set, and will reassure the user that the extents of the data are being reached by the map grid as fully as possible in each dimension.

This study has been conducted under the assumption that each dimension of the data is of equal interest to the user. The choice to incorporate this measure into the selection of map setup parameters is dependent on the importance of representing as much of the extremities of the data of each variable as possible. This must be determined by the user through knowledge of the data set.

Future work may include the use of the DRR measure to aid map size and shape selection in an automated way (in combination with other quality measures). It may also be used to assist in map size selection for variants to the SOM. Automatic, or user-specified, weighting of the importance of dimensions could be investigated, as could the response of the measure to noisy and redundant dimensions.

5.7 References

[1] C. Fyfe, 'Topographic Maps for Clustering and Data Visualization', in Computational Intelligence: A Compendium (Springer, 2008), pp. 111-53.

[2] T. Kohonen, 'Self-Organizing Maps'. Vol. 30 (Springer, 2001).

[3] A. Flexer, 'On the Use of Self-Organizing Maps for Clustering and Visualization', in Principles of Data Mining and Knowledge Discovery (Springer, 1999), pp. 80-88.

[4] C.A. Astudillo and B.J. Oommen, 'Topology-Oriented Self-Organizing Maps: A Survey', Pattern Analysis and Applications, 17 (2014), 223-48.

[5] J. Vesanto, 'Som-Based Data Visualization Methods', Intelligent Data Analysis, 3 (1999), 111-26.

[6] S. Kaski and T. Kohonen, 'Exploratory Data Analysis by the Self-Organizing Map: Structures of Welfare and Poverty in the World', in Neural Networks in Financial Engineering. Proceedings of the Third International Conference on Neural Networks in the Capital Markets (Citeseer, 1996).

[7] T. Kohonen, 'Essentials of the Self-Organizing Map', Neural Networks, 37 (2013), 52-65.

[8] Y. Liu, R.H. Weisberg, and C.N.K. Mooers, 'Performance Evaluation of the Self-Organizing Map for Feature Extraction', Journal of Geophysical Research: Oceans (1978–2012), 111 (2006).

[9] P. Agarwal and A. Skupin, 'Self-Organizing Maps: Applications in Geographic Information Science' (John Wiley & Sons, 2008).

[10] T. Kohonen, 'The Self-Organizing Map', Neurocomputing, 21 (1998), 1-6.

[11] S. Clark, P. Sarlin, A. Sharma, and S.A. Sisson, 'Increasing Dependence on Foreign Water Resources? An Assessment of Trends in Global Virtual Water Flows Using a Self-Organizing Time Map', Ecological Informatics (2014).

[12] A.M. Kalteh, P. Hjorth, and R. Berndtsson, 'Review of the Self-Organizing Map (SOM) Approach in Water Resources: Analysis, Modelling and Application', Environmental Modelling & Software, 23 (2008), 835-45.

[13] R. Gopakumar, K. Takara, and E. James, 'Hydrologic Data Exploration and River Flow Forecasting of a Humid Tropical River Basin Using Artificial Neural Networks', Water Resources Management, 21 (2007), 1915-40.

[14] A. Adeloye and R. Rustum, 'Self-Organizing Map Rainfall-Runoff Multivariate Modelling for Runoff Reconstruction in Inadequately Gauged Basins', Hydrology Research, 43 (2012), 603-17.

[15] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas, 'Som Toolbox for MATLAB 5' (Citeseer, 2000).

[16] Y. Park, R. Céréghino, A. Compin, and S. Lek, 'Applications of Artificial Neural Networks for Patterning and Predicting Aquatic Insect Species Richness in Running Waters', Ecological Modelling, 160 (2003), 265-80.

[17] Y. Morioka, T. Tozuka, and T. Yamagata, 'Climate Variability in the Southern Indian Ocean as Revealed by Self-Organizing Maps', Climate Dynamics, 35 (2010), 1059-72.

[18] A. Dejean, R. Cereghino, J.M. Carpenter, B. Corbara, B. Herault, V. Rossi, M. Leponce, J. Orivel, and D. Bonal, 'Climate Change Impact on Neotropical Social Wasps', PloS one, 6 (2011), e27004.

[19] K. Hsu, H.V. Gupta, X. Gao, S. Sorooshian, and B. Imam, 'Self-Organizing Linear Output Map (Solo): An Artificial Neural Network Suitable for Hydrologic Modeling and Analysis', Water Resources Research, 38 (2002), 38-1-38-17.

[20] E. Toth, 'Classification of Hydro-Meteorological Conditions and Multiple Artificial Neural Networks for Streamflow Forecasting', Hydrology and Earth System Sciences, 13 (2009), 1555-66.

[21] A.C. Steynor, B.C. Hewitson, and M.A. Tadross, 'Projected Future Runoff of the Breede River under Climate Change', Water SA, 35 (2009), 433-40.

[22] Y. Wang, and C. Feng, 'Patterns and Trends in Land-Use Land-Cover Change Research Explored Using Self-Organizing Map', International Journal of Remote Sensing, 32 (2011), 3765-90.

[23] E. Erwin, K. Obermayer, and K. Schulten, 'Self-Organizing Maps: Ordering, Convergence Properties and Energy Functions', Biological Cybernetics, 67 (1992), 47-55.

[24] R. Cereghino, and Y. S. Park, 'Review of the Self-Organizing Map (SOM) Approach in Water Resources: Commentary', Environmental Modelling & Software, 24 (2009), 945-47.

[25] M. Pena, W. Barbakh, and C. Fyfe, 'Topology-Preserving Mappings for Data Visualisation', in Principal Manifolds for Data Visualization and Dimension Reduction (Springer, 2008), pp. 131-50.

[26] K. Kiviluoto, 'Topology Preservation in Self-Organizing Maps', in IEEE International Conference on Neural Networks (1996), pp. 294-99.

[27] S. Kaski, and K. Lagus, 'Comparing Self-Organizing Maps', in Artificial Neural Networks—Icann 96 (Springer, 1996), pp. 809-14.

[28] D.L. Davies, and D.W. Bouldin, 'A Cluster Separation Measure', Pattern Analysis and Machine Intelligence, IEEE Transactions on (1979), 224-27.

[29] R. Xu, and D. Wunsch, 'Survey of Clustering Algorithms', Neural Networks, IEEE Transactions on, 16 (2005), 645-78.

[30] W. Härdle, and L. Simar, ‘Applied Multivariate Statistical Analysis’ (Springer Science & Business Media, 2007).

[31] J. Lee, H. Lee, J. Kim, D. Nam, and C.H. Park, 'Self-Organizing Neural Networks by Construction and Pruning', IEICE TRANSACTIONS on Information and Systems, 87 (2004), 2489-98.

[32] A.A. Akinduko, and E.M. Mirkes, 'Initialization of Self-Organizing Maps: Principal Components versus Random Initialization. A Case Study', arXiv preprint arXiv: 1210.5873 (2012).

[33] Mirkes, E.M. Principal Component Analysis and Self-Organizing Maps: applet. University of Leicester, 2011.

[34] J.W. Tukey, 'Exploratory Data Analysis', (1977).

[35] http://mdgs.un.org/unsd/mdg/Metadata.aspx

[36] www.wssinfo.org

6 Paper 3 - Nonlinear manifold representation

This chapter has been published as:

Nonlinear manifold representation in natural systems

Clark S, Sisson SA, Sharma A. *Environmental Modelling & Software.* 2017; 89: 61-76.

6.1 Abstract

Natural systems often contain rhythmically fluctuating individual components which, when combined, can result in nonlinear patterns such as cycles, helixes, and parabolas. The self-organizing map (SOM) is a widely used artificial neural network for exploratory data analysis of high dimensional, multivariate data sets, however it encounters limitations when dealing with such highly nonlinear patterns. The SOMersault method is an expansion of the SOM, effective for gaining an understanding of patterns and clusters in natural data sets containing a low dimensional nonlinear manifold set amongst complex high dimensional data measurements. With the SOMersault, data clusters become ordered with respect to the nonlinear degrees of freedom in the data, and patterns extracted are closely related to the data they represent. Results are shown on synthetic data and a real world data set involving water scarcity and storage relationships in a global set of river basins, with clustering and pattern extraction improvements displayed visually and quantified through a new set of geodesic error measures.

6.2 Introduction

Natural systems (e.g. environmental, ecological, meteorological, and biological) are often characterised by nonlinear relationships between the individual system components. In nature, countless physical and chemical factors simultaneously exert their influences on the state of structures and organizations. Out of this complexity, however, simpler natural spatial and temporal patterns can emerge: waves, parabolas, v-shapes, oscillations, spirals, temporal cycles, and helixes are each the product of combinations of innumerable complex interactions. In the analysis of natural formations, rhythms and organizations, it is therefore beneficial to extract information from these more basic patterns (Adam (2006), Stewart (2008), Frank (2009), Streit (2015)).

The self-organizing map, or SOM (Kohonen, 2001), is an exploratory data analysis technique that is particularly suitable for finding patterns and producing clusters in data with complex, and potentially undefined, relationships between variables. A type of artificial neural network, the SOM is able to analyse large amounts of high dimensional data without requiring an explicit understanding of the underlying relationships. The SOM is also resilient to data sets with noisy or missing data (Kohonen, 2001), issues that are common to data collected in the field. These attributes make the SOM a popular technique for exploring natural spatiotemporal data sets (Park (2003), Agarwal & Skupin (2008), Steynor (2009), Cereghino (2009), Morioka (2010), and Adeloye (2012)). A data set is summarized and visualized with the SOM method into a low dimensional representation of the predominant patterns present in the data. To discover these patterns, a mesh of connected nodes (usually rectangular) is effectively stretched over the set of input observations and then trained until the location of the map nodes best represents the distribution of the data. A prototype vector associated with each map node takes on the value of the location of the data space that the node occupies. Each data point is matched to its most similar (nearest) node. Generally, there are far fewer map nodes than data points, and the data becomes clustered into subsets that share the same nearest node. Therefore, without requiring any prior knowledge of potential patterns existing in the data, each prototype vector comes to represent a prevalent pattern, the data points are clustered into groups sharing the same patterns, and the characteristics that are shared by similar data points can be recognized.

The performance of the SOM is known to have some limitations, particularly when the data set is characterized by an underlying nonlinear, low dimensional pattern (Demartines (1997), Zhang (2004), Ota (2011), Shao (2015)). These limitations can affect the pattern extraction and clustering results of the SOM. Many seemingly complex natural systems do have an intrinsic low dimensionality, even though the data measurements may be of high dimension. In these cases, few variables have much variation in the data set compared to the number of measured variables, and the high dimensional measurements are essentially indirect measurements of the underlying low dimensional source that cannot be, or has not been, measured (Tenenbaum (2000), Zhang (2004), Ghodsi (2006)).

This low dimensional source (or geodesic surface or manifold) may be determined using dimension reduction, or manifold learning, techniques. The representation of the data set obtained through these techniques characterises the data with the minimum number of parameters required to explain its properties (Van der Maaten, 2009). The coordinate axes of the low dimensional reduction represent the dimensions that vary meaningfully in the high dimensional data (Tenenbaum, 2000). Dimension reduction aids in classification and visualization of high dimensional data (Van der Maaten, 2009), which are the two main goals of the SOM. A variety of nonlinear dimension reduction (manifold learning) techniques exist for discovering low dimensional manifolds within high dimensional data (see e.g. Van der Maaten, 2009, for an overview). Popular methods include the ISOMAP (Tenenbaum, 2000), local linear embedding (LLE) (Roweis, 2000), and local tangent space alignment (LTSA) (Zhang, 2004). These techniques all provide a low (usually 2) dimensional projection with the same number of points as in the original data. Dimension reduction is performed, but not a clustering or summarizing of the data set.

The traditional SOM framework performs a type of nonlinear dimension reduction, however this is based on an initial linear approximation of the low dimensional manifold. This first approximation of the location of the SOM grid is typically performed with principal component analysis (PCA). PCA discovers linear principal manifolds in the data by minimizing the sum of squared distances to the data points. A set of orthogonal principal components (vectors) is found in the directions of maximum variance of the data, and the traditional SOM map uses these vectors as the initial axes of the grid. However, PCA cannot adequately find nonlinear structures in data (Tenenbaum, 2000, Van der Maaten, 2009) and therefore a SOM based on this initialisation will only discover nonlinear manifolds that are a minor perturbation of the initial linear approximation (Shao, 2015). This can impede the applicability of the SOM when fitting it to a data cloud with complex structures (Lee & Verleyson (2005), Huang (2013)). It has even been claimed that when SOMs are successful at nonlinear mapping, it is only by chance (Demartines, 1997), and that any information gained by the SOM about the global structure of the data set is obtained from analyzing overlapping structures of the entire data set (Zhang, 2004). Therefore, upon encountering data sampled from a nonlinear low dimensional manifold embedded in a high dimensional space, the traditional SOM tends to generate topological defects or become set in locally optimal solutions as a result of the linear initialisation (Shi (2006), Ota (2011), Shao (2015)). As many natural data sets contain nonlinear structures which are invisible to PCA (Tenenbaum, 2000) and hence also invisible to SOMs, an alternative approach is needed to ensure the initial SOM grid topology approximately corresponds to the intrinsic shape of any underlying nonlinear data manifold. This would support the application of the fundamental SOM functions of clustering and visualization to data with essential nonlinear structures.

While the SOM is useful for clustering and visualizing multivariate data sets but cannot find highly nonlinear structure in the data, typical dimension reduction or manifold learning methods can discover the intrinsic nonlinear structure of the data but cannot cluster or summarize it (Shi (2006), Ota (2011), Shao (2015)). A pairing of the SOM with a nonlinear manifold learning technique would therefore broaden the applicability of the SOM with respect to typically complex data sets regularly encountered in natural systems.

The literature contains some endeavors to combine nonlinear manifold learning within the traditional SOM framework. The ISOSOM (Guan, 2006) combines the SOM with the ISOMAP in a two-step process, with the SOM applied to a low dimensional projection found with the ISOMAP. Although the trained SOM is transferred back into high dimensional space, the training has already been finalized in low dimensions and therefore the prototypes are not able to move off the geodesic surface into the high dimensional space. There is also no restriction on matching the high dimensional data points to the trained map through a constraint along the geodesic surface and so the final map may be found to stretch between unrelated parts of the surface. The GDBSOM (Shi, 2006) first constructs a neighbourhood graph of geometric distances between all data points and then sets up the basic (rectangular) SOM with corners based on the largest geometric distances between points. The map nodes are updated towards the data in Euclidean distance, which may cause a transfer of map nodes between separate parts of the geodesic surface. The DSOM (Shao, 2015) searches for the underlying manifold by starting with a small amount of training data and a small grid, and expanding the training data set and the map according to the identified geodesic structure. Map size is increased by adding either a row or a column in a complicated expansion process with a number of user-specified parameters. Real world applications for these methods are so far restricted to the field of image analysis.

In this paper we introduce a method termed the ‘SOMersault’, a type of SOM in which the map essentially unrolls itself along the geodesic surface of the data, allowing the discovery and representation of nonlinear manifolds by the map. The SOMersault integrates LTSA within the traditional SOM process in a method that incorporates both high and low dimensional representations of the data. In constructing the SOMersault, adaptions have been made to the basic SOM framework including: modification of the initialization stage to ensure the map is globally ordered in alignment with the geodesic surface; alteration of the size and shape of the neighbourhood kernel during map training to restrict map movement to localised high dimensional areas near the geodesic surface; and the development of geodesic error measures. These error measures are based on the most popular measures conventionally used to evaluate SOMs, modified to assess the alignment of the trained map with respect to the low dimensional manifold. We illustrate examples on synthetic data and a real world natural data set, with highly favourable results compared to those obtained using the traditional SOM on the same data.

6.3 METHOD

The traditional SOM technique is described below, followed by a description of the SOMersault.

6.3.1 Traditional SOM

A SOM (Kohonen, 2001) consists of a number of nodes connected in a predetermined regular grid formation, which become organized (or trained) to represent the patterns and clusters in a data set. The map is first placed on the data set in an initial approximate location, and then trained to better fit the data. Usually the number of map nodes, M, is much smaller than the number of observed data points, N. Training consists of a number of iterations, *T* (with each iteration denoted by *t* where *t=1…T*), of first matching each data point to the nearest map node, and then updating the locations of all the map nodes to become closer to their matching data, as follows.

1. **Initialisation:** The initial approximation for the location of the map nodes is conventionally based on the linear principal components of the data set. The directions of the map axes correspond to the first and second eigenvectors of the data, thereby following the linear, orthogonal directions of maximum variance in the data. The lengths of the axes are set proportional to the ratio of the two greatest eigenvalues. The nodes are placed at uniform intervals along the axes, producing an approximately uniform lattice spacing for the initial grid.

2. **Matching:** A prototype vector (m_i, where $i=1 \cdots M$) associated with each map node takes on the value of the data space in which it is located. Individual data points (x_j, where $j=1, \cdots, N$) are matched to their nearest map node (or best matching unit, BMU). The BMU, m_c, is chosen for each data item by minimizing the Euclidean distances between all data and all map nodes:

$$c = arg\,\min_i\{\|x_j - m_i\|\}$$

3. **Updating:** The location of each prototype vector, m_i, is updated to become a weighted average of the data assigned to that node and its neighbouring nodes, whilst maintaining the original connections of the grid. The weighting is based on a neighbourhood smoothing kernel, $h_{ic}(t)$, which defines the stiffness of the map. At each iteration all nodes are updated concurrently, however, only the data points assigned to nodes within the user defined neighbourhood radius, $\sigma(t)$, measured in map space around each node, influence the updating of that node. The neighbourhood kernel describes the weighting of data matched to node *c* on the updating of node *i* at iteration *t*. The default neighbourhood kernel for the traditional SOM is:

$$h_{ic}(t) = \exp\left\{\frac{-(\|r_c - r_i\|^2)}{2\sigma^2(t)}\right\}$$

   where r_c and r_i are the locations of nodes m_c and m_i in map space (Vesanto et al., 2000). Denoting prototype vector m_j at iteration *t*, by $m_j(t)$ the updating of the prototype vectors is given by:

$$m_i(t+1) = \frac{\sum_{j=1}^{N} h_{ic}(t)\, x_j}{\sum_{j=1}^{N} h_{ic}(t)}$$

   where *c* is the BMU index of data point *j* and *N* is the number of data points. The neighbourhood radius, $\sigma(t)$, begins relatively large at iteration t=1, and diminishes in size at each subsequent iteration according to some user defined schedule. This allows for a *global ordering* of the map nodes along the data set at the start of training, and *fine-tuning* of their locations at the end of training.

After training, the prototype vectors will each be an average of the subsection of data assigned to them, and will therefore represent a characteristic pattern of the data set. The data matched to each node form clusters that share these characteristics. Patterns have been extracted and clusters have been found without any prior knowledge of the patterns in the data. The nodes have maintained their connections throughout training, ensuring the extracted patterns and clusters are ordered based on the overall topology of the map.

6.3.2 SOMersault

When an underlying nonlinear lower dimensional manifold exists in a data set, globally ordering the map grid based on the main directions of variance of the data set as a whole (as performed in the traditional SOM technique) will cause the map to disregard any bends or twists in the manifold that are more complicated than simple perturbations of the initial linear approximation provided by the principal components (Demartines (1997), Zhang (2004), Ota (2011), Shao (2015)). The global ordering step is therefore modified in the SOMersault method to create a map alignment that corresponds to the main degrees of freedom of the geodesic surface rather than that of the data set as a whole. The fine-tuning of the node locations is completed within localized neighbourhoods of the input data space to produce high dimensional prototypes, allowing the data to join the most representative high dimensional cluster. The SOMersault creates an alignment of the map grid along the low dimensional nonlinear manifold in the following manner:

1. **Initialisation:** The high dimensional input data set (x_j, where $j=1, \cdots, N$) is projected into low dimensional space using the local tangent space alignment (LTSA) nonlinear dimension reduction technique (Zhang, 2004), creating a corresponding low dimensional data item ($x_{j,LD}$, where $j=1, \cdots, N$) for each original data item. With LTSA, a manifold is unfolded by finding an alignment of all the tangent hyperplanes of the manifold. The tangent space is computed in the local neighbourhood of each point, and then the alignment of the spaces is optimized. A new set of coordinates is defined to represent the principal manifold of the high dimensional data in the low dimensional space by finding a linear mapping from both the high dimensional data points and their corresponding low dimensional data points to the same local tangent space (see Zhang, 2004, for further details). A grid of map vectors ($m_{i,LD}$, where $i=1 \cdots M$) with dimensions matching the number of dimensions in the projection, is placed over the projected data with axes corresponding to the length and directions of maximum variance of the *unraveled principal manifold* determined with LTSA. These directions correspond to the main nonlinear degrees of freedom in the high dimensional data set. For the initial grid layout, map nodes are distributed uniformly along these axes.

2. **Global ordering:** The low dimensional grid is partially trained (~2 iterations) on the low dimensional projected data using a large neighbourhood kernel that includes the entire map. On the unraveled geodesic surface, the nodes find their *approximate* locations amongst the data points. This produces an overall ordering of the nodes as they shift into the general areas where they will eventually permanently settle. For the global ordering stage, the SOMersault uses the Gaussian neighbourhood kernel which updates all map nodes based on a weighted average of all of the data points:

$$h_{ic}(t) = \exp\left\{\frac{-(\|r_c - r_i\|^2)}{2\sigma^2(t)}\right\}$$

   where $\|r_c - r_i\|$ is the Euclidean distance on the map grid between nodes $m_{c,LD}$ and $m_{i,LD}$, and σ is the neighbourhood radius, or bandwidth of the kernel. This is the neighbourhood kernel of the traditional SOM as described above which updates the low dimensional map nodes based on all data points in the projection, weighted by the distance of their BMU from the updating node. Use of this neighbourhood kernel produces a global ordering of the map along the geodesic surface, as it becomes stretched over the LTSA projection of the data set.

3. **Fine-tuning:** The roughly trained grid is projected into input data space to be used as the initial grid for map training in the high dimensional space. The map now consists of nodes ($m_{i,HD}$, where $i=1 \cdots M$) with the number of dimensions matching those of the input data. The map is further trained (~100 iterations) in data space with a small,

localized neighbourhood kernel to fine-tune node locations to lie amongst the high dimensional data. At this stage, the prototypes are allowed to leave the embedded geodesic surface. This will improve the placement and connections of the nodes within data space. For this fine-tuning of node locations in high dimensional space, the SOMersault uses a truncated Gaussian neighbourhood kernel rather than the Gaussian kernel. This restricts the influence of the data points on the updating of the map nodes to a localized area, so that data outside the specified radius have no influence on the updating of the nodes. The truncated Gaussian kernel is:

$$h_{ic,HD}(t) = \exp\left\{\frac{-(\|r_c - r_i\|^2)}{2\sigma^2(t)}\right\} I(\ \|r_c - r_i\|^2 \leq \sigma)$$

where r_c and r_i are the locations of nodes $m_{c,HD}$ and $m_{i,HD}$ on the map grid. This modification is necessary in high dimensional space since the use of the Gaussian neighbourhood kernel would lead to the updating of the entire map based on all data points, pulling the nodes towards the data based on Euclidean distance and not geodesic distance. For highly bent or twisted manifolds, this could result in the movement of nodes between sections of the manifold that may be close in Euclidean distance but distant in geodesic distance. Since the global ordering of the map nodes has already been achieved by the initial training on the low dimensional projection, the fine-tuning of the map on the high dimensional data does not require a large neighbourhood. The localised neighbourhood radius at this point includes only the data points that can be expected to lie on a linear subspace around the updating node, and the data outside this should not be updated linearly. The updating of the high dimensional map nodes proceeds as:

$$m_{i,HD}(t+1) = \frac{\sum_{j=1}^{N} h_{ic,HD}(t)\, x_j}{\sum_{j=1}^{N} h_{ic,HD}(t)}.$$

This method will lead to a better correspondence of the size and shape of the map grid with the main directions of variance of the nonlinear manifold, thereby creating a map that better represents the intrinsic structure of the data set. The SOMersault algorithm is summarized in Section 6.3.2.1 and demonstrated in Figure 19 on the Swiss roll data set, a 2D manifold embedded in 3D data space (data from Wittman, 2005).

6.3.2.1 SOMersault Algorithm

1. The data set (Figure 1a) is projected into low dimensional space by performing LTSA on the high dimensional data set (Figure 1b),
2. The directions of maximum variance of the low dimensional projection are discovered and the map grid is draped over the low dimensional data, aligned in these principal directions (Figure 1c, black nodes),
3. The map is coarsely trained to stretch over the low dimensional data (Figure 1d, pink grid),

4. The nearest low dimensional data point is found for each low dimensional map node (by Euclidean distance) (Figure 1e, red circles),
5. The corresponding high dimensional data point is determined for these nearest points (Figure 1f, black nodes),
6. These high dimensional data points are joined together to form a mesh retaining the connections between the map nodes in low dimension (Figure 1g, black mesh),
7. This mesh is used as the initial high dimensional map for the fine-tuning stage,
8. The map is trained in high dimensional space using a narrow, truncated neighbourhood kernel (Figure 1h, pink mesh is trained map).

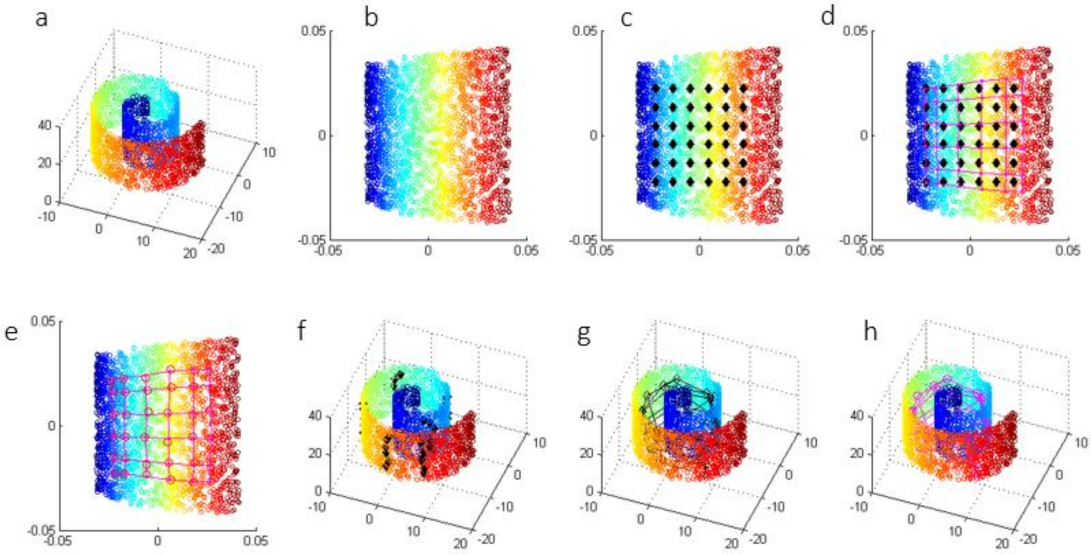


Figure 19: Demonstration of the SOMersault algorithm on the Swiss roll data set (data from Wittman, 2005). The trained map in the lower right plot (pink) closely follows the underlying manifold of the curved data. a) data; b) LTSA low dimensional projection; c) initialized grid on low dimensional projection; d) globally trained grid; e) nearest neighbours of globally trained grid; f) high dimensional corresponding data points of nearest neighbours of low dimensional trained grid; g) mesh of high dimensional initial points; h) map trained in high dimensions.

6.3.3 Quality assessment measures

To provide meaningful information, the metrics used in the error measures must match the distance metrics of the quantization and clustering methods used on the data (Hardle & Simar, 2007). Accordingly, we develop new measures for quantifying the quality of the data representation based on geodesic distance, to replace the conventional SOM error measures which are based on Euclidean distance. These geodesic error measures rely on the assumption that the low dimensional projection is a correct representation of any geodesic surface present in the data.

6.3.3.1 Quantization

Map training performs vector quantization on the data set by producing a smaller, representative set of vectors (Vesanto, 1999). The quantization error, QE, is conventionally used in SOM applications to measure how closely these vectors represent the input data set.

This measure can be used to compare the representation of a data set by maps with equal numbers of nodes. QE is the average sum of squares of the distances between each data point and its BMU in Euclidean distance:

$$QE = \frac{1}{N}\sum_{i=1}^{N}\|x_i - m_c\|^2.$$

We introduce the geodesic quantization error, GQE, to measure the data quantization whilst taking the shape of the manifold into account. GQE measures the average sum of squares of the distances between each data point and its closest node *along the geodesic surface*:

$$GQE = \frac{1}{N}\sum_{i=1}^{N}\left\|x_{i,LD} - m_{c,LDT}\right\|^2$$

where $x_{i,LD}$ is the low dimensional projection of data point x_i , and $m_{c,LDT}$ is the low dimensional projection of the high dimensional BMU of (high dimensional) data point x_i, as described below.

The BMUs of the data points are found in input data space, defining the Voronoi sets, or groups of data points that share the same nearest node. The Voronoi set of data points sharing node $m_{i,HD}$ is R_i:

$$R_i = \left\{j\colon \left\|x_{j,HD} - m_{i,HD}\right\| \leq \left\|x_{j,HD} - m_{k,HD}\right\|, \qquad \forall\ k \neq i\right\}$$

The high dimensional, trained map is then projected into low dimensional space by setting the location of each node to the mean of the low dimensional data points whose corresponding high dimensional points comprise the Voronoi set for each node:

$$m_{i,LDT} = \frac{1}{|R_i|}\sum_{j\in R_i} x_{j,LD}$$

where $|R_i|$ is the number of data points in R_i.

Note that $m_{i,LDT}$ is a low dimensional projection of a node on the final map that has completed training in high dimensions, whereas $m_{i,LD}$ is a node of the original low dimensional initial map that has not yet been trained.

The distance from the low dimensional projection of each data point to the low dimensional projection of its closest trained map node is found, and averaged over all data points. Determining these distances on the unravelled manifold provides a measure of how well the data is quantized with respect to the geodesic surface. A higher GQE may indicate that nodes are picking up data from distant parts of the geodesic surface that become close in Euclidean distance through the bending or twisting of the manifold in high dimensional space.

The GQE may be expected to be lower for the SOMersault than the traditional SOM as the map is designed to align with the geodesic surface rather than with the overall directions of maximum variance of the data, in which case the map may cut through the geodesic surface. Data points will be matched to nodes that are on the same portion of the geodesic surface, even if another portion of the surface is also nearby due to the curvature of the surface. The map nodes are drawn towards the data points in both methods, though in the SOM they are moved according to Euclidean distance, and with the SOMersault they are drawn along the geodesic surface. The alignment of the SOMersault grid with the geodesic surface also means the nodes are less likely to settle in the empty spaces between areas of high data density, as can happen with the SOM when it follows the linear principal components.

6.3.3.2 Topographic preservation

The frequently used topographic error, TE, indicates how well the topology of the input data is preserved by the output map. This error is defined as the proportion of data points for which the first and second BMUs are not nearest neighbours on the trained map grid. The TE indicates if the map is folded, bent or twisted. With a low dimensional nonlinear manifold present in the data, however, the map may become folded, bent and twisted in order to best represent the manifold. In this case, the TE will indicate a higher error than a linear map though it may be providing a more true characterization of the topography of the data. The standard TE is therefore not ideal for use with such nonlinear manifolds.

We define the geodesic topographic error, GTE, to evaluate the topological ordering of the high dimensional map nodes of the SOMersault. The GTE indicates if the map smoothly follows the manifold, whilst allowing for bending and folding of the map if required by the topology of the data it aims to represent. It assesses if the local structure of the geodesic surface of the data is retained on the map by evaluating whether adjacent map nodes represent closely situated data. This is done by measuring if the first and second closest nodes to each data point are adjacent along the geodesic surface. The GTE is calculated as:

$$GTE = \frac{1}{N}\sum_{i=1}^{N} u_{x_i} \, ,$$

where $u_{x_i} = 0$ if the nearest and second nearest nodes to data point x_i *measured along the geodesic surface* are adjacent on the map grid, 1 otherwise.

A GTE of zero indicates perfect geodesic topological preservation by the trained map. The GTE is expected to show an improvement in the ordering of nodes by the SOMersault, as the grid is designed to follow the geodesic surface. Though the nodes of the traditional SOM may end up in similar locations of data space, the SOMersault nodes will be linked in a manner that more closely represents the manifold.

6.3.3.3 Clustering

Clustering of the data, with the map nodes defined as the cluster centres, is evaluated using the Davies-Bouldin (DB) clustering index (Davies & Bouldin, 1979) (as in Vesanto & Alhoniemi, 2000). This is a general clustering index (not specific to the SOM) which assesses the ratio of within-cluster scatter to between-cluster separation, and is not dependent on the number of clusters being analysed or the method of clustering. The goal is to ensure that the data within each cluster is as similar as possible, and as different as possible to data in other clusters (Hardle & Simar, 2007). Lower DB values indicate a better clustering of the data set. The DB index on the trained map projected to the geodesic surface, GC, is defined here as:

$$GC = \frac{1}{M}\sum_{i=1}^{M} \max_{k \neq i} \left\{ \frac{S_i + S_k}{d_{ik}} \right\}$$

where $S_i = \frac{1}{|R_i|}\sum_{j \in R_i} \left\| x_{j,LD} - m_{i,LDT} \right\|$ is the within-cluster scatter, and

$d_{ik} = \left\| m_{i,LDT} - m_{k,LDT} \right\|$ is the between-cluster separation.

The GC index is expected to indicate significantly improved clustering of the data set with the SOMersault compared to the traditional SOM, as the prototypes of the SOMersault will represent data drawn from a localised region of the underlying manifold, and therefore more similar data.

6.4 Results

We demonstrate the performance of the SOMersault on two- and three-dimensional synthetic examples and a real world hydrologic application. Comparisons of the SOMersault output are made with the output of the traditional SOM through error measures and visualisations. The quantization, topological representation and clustering of the data by the maps is evaluated with the error measures, for which lower values indicate better data representation.

6.4.1 Synthetic examples

In this section, the SOMersault and SOM are applied to a variety of data sets in two and three dimensions. The SOMersault is implemented in MATLAB, with the LTSA component performed by the Nonlinear Toolbox (Van de Maaten, 2007) and mani.m (Wittman, 2005).

6.4.1.1 2D examples

Figure 20 illustrates three synthetic two-dimensional data sets with shapes commonly found in natural dynamic and organizational systems. These systems could include, amongst many other possibilities: population dynamics, hydrologic systems and events, geological forms, diurnal or seasonal meteorological systems, orbits, fluid dynamics, or biological structures (Stewart, 2008). The one-dimensional manifolds embedded in the two-dimensional space are shown in the first column, the traditional SOM linear initialisation (black) and trained map (blue) in the middle column, and the SOMersault nonlinear initialisation (black) and trained map (red) in the right column.

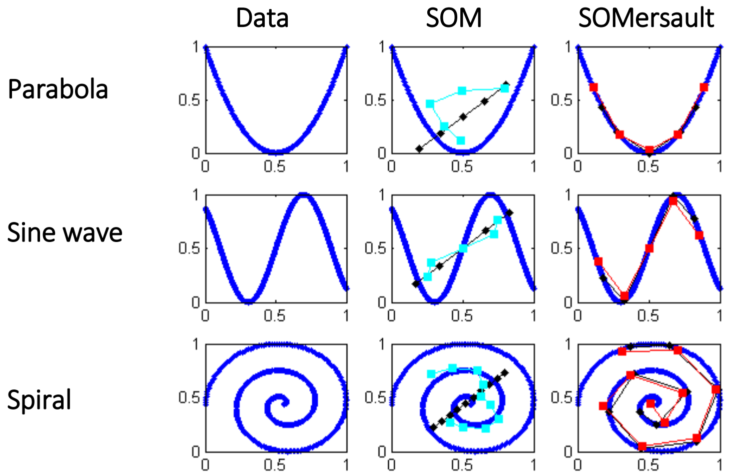


Figure 20: The SOMersault and SOM methods implemented on 1D manifolds embedded in 2D space. On the left is the data set, in the middle is the linear initialization (black) and resulting traditional SOM (blue), on the right is the non-linear initialization (black) and resulting SOMersault (red). It can be seen that the map on the right better discovers the structure of the 1D manifold present in the data, the prototype vectors corresponding to each node have values that are close to the data points, and the nodes are ordered (linked) according to the manifold.

The quantization, topological preservation, and clustering of the 2D data sets in Figure 20 are assessed with the geodesic error measures, GQE, GTE and GC. The results are listed in Table 6. The GQE is improved for each of the examples with use of the SOMersault, as the map nodes are positioned closer to the data they represent, and are not located away from the dense areas of data (in the white space) as has occurred with some of the SOM nodes due to the Euclidean updating process. The GTE indicates perfect topological preservation on each of the SOMersault maps, as the grids follow the data surface without crossing it, bending, or jumping from one area of data to another. The GC shows the data is clustered well with the SOMersault, as the data assigned to each node originates from the same local region of the data surface. These noiseless synthetic examples demonstrate that even the tidiest nonlinear manifold may not be well discovered with the use of a linear map initialisation.

Table 6: 2D results. Geodesic error measures on the representation of the three data sets in Figure 20 after applying the SOMersault and the traditional SOM.

	Method	GQE (quantization)	GTE (topological preservation)	GC (clustering)
Parabola	SOMersault	0.014	0.00	0.6
	SOM	0.019	0.44	5.5
Sine wave	SOMersault	0.005	0.00	0.5
	SOM	0.006	0.42	0.7
Spiral	SOMersault	0.008	0.16	5.1
	SOM	0.028	0.56	19.3

6.4.1.2 3D examples

Figure 21 displays three data sets with lower dimensional manifolds embedded in three-dimensional space (left column). The centre column shows the SOM linear initialisation (black) and trained map (magenta), and the right column illustrates the SOMersault nonlinear initialisation (black) and trained map (magenta). While all maps have the same number of nodes, it is visually apparent that the SOMersault maps follow the underlying manifolds more smoothly, with fewer nodes caught in empty space between regions of the manifold.

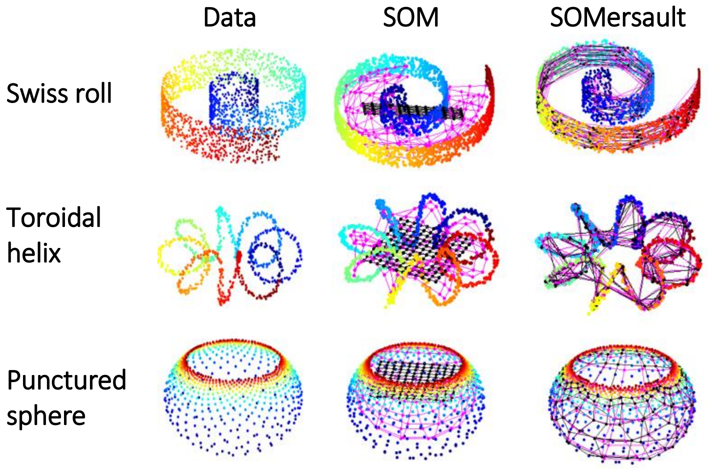


Figure 21: The SOMersault and SOM methods implemented on data in 3D space: a) data sets (from mani.m, Wittman, 2005); b) linear initialization (black) and resulting traditional SOM (magenta); c) nonlinear initialisation (black) and resulting SOMersault (magenta). Visual comparison of the final SOMs and SOMersaults indicates that the SOMersaults better discover and follow the underlying manifold of each data set.

The quantization, topological preservation, and clustering of the 3D data sets in Figure 21 are assessed with the geodesic error measures, GQE, GTE and GC. The results are listed in Table 7. With the use of the SOMersault compared to the SOM, for all examples the data quantization is improved or maintained, topology of the data surface is better preserved, and clustering is improved or maintained. Due to the already decent clustering and quantization of the punctured sphere by the linearly initialized SOM, this example shows the least improvement with use of the SOMersault.

Table 7: Geodesic error measures on the representation of the three data sets in Figure 21 after applying the SOMersault and the traditional SOM.

	Method	GQE (quantization)	GTE (topological preservation)	GC (clustering)
Swiss Roll	SOMersault	0.030	0.22	0.9
	SOM	0.050	0.34	21.2
Toroidal helix	SOMersault	0.005	0.33	0.3
	SOM	0.010	0.55	121.7
Punctured sphere	SOMersault	0.026	0.42	0.9
	SOM	0.026	0.45	0.9

6.4.1.3 Discussion of synthetic results

We investigate the results on the Swiss roll data set in further detail with Figure 22, in which the SOM output is in the left column and the SOMersault is in the right column. From top to bottom, the rows represent: a) the three-dimensional input data with the trained map grid overlaid; b) the unraveled low dimensional manifold with the map overlaid; c) the 12x12 output map grid with each node populated by the actual data points assigned to it; and d) the output map grid with each node coloured in the mean colour of its subset of data points (black boxes represent nodes which do not have any data points assigned to them - that is they are positioned in areas of no data). The four rows of Figure 22 help us visualize the differences in quantization, topological preservation and clustering between the SOMersault and the SOM.

In row a, we can see that the SOMersault map grid follows the surface of the data more smoothly than the SOM grid. The connections between the nodes do not cross the areas where no data lies, between parts of the geodesic surface. On the left, the SOM grid has connections which stretch between distinct parts of the surface (such as the red and blue areas), and nodes which are situated in the gaps between sections of the surface.

Row b shows the map grids laid on the unraveled low dimensional surface. The SOMersault grid is aligned smoothly on the surface, without node connections crossing each other. The map is not twisted or folded in relation to the geodesic surface. However, three large regions of the SOM have been pulled out of place and towards the left of the map, indicating that the map jumps from one part of the manifold to another in high dimensional space rather than running smoothly along it. The node connections cross each other, and nodes that are neighbours on the grid are not neighbours when it is placed on the data. This is reflected in the GTE error measure, which measures if the map follows the topology of the data.

Row c shows that the data assigned to each of the SOMersault nodes corresponds well with the location of the node on the grid. Like the data set, the colour of data assigned to each SOMersault node transitions smoothly across the grid from dark blue to red. However, for the SOM, three isolated areas are evident. The map begins in the light blue region, and the dark blue data is assigned to patches of the map situated amidst the other colours indicating that the trained map jumps between sections of input data space that are not adjacent. We can also see that the data mapping to each node is more consistent in the SOMersault than the SOM, which has a few nodes in the lower left area of the grid containing data of varied colours. Some nodes of the SOM do not contain any data points at all, as they are located in empty space between areas of high data density. This quality is reflected in the GQE measure.

In row d, the nodes are coloured by the average colour of their assigned data. In this way, the colours represent the typical ‘patterns’ extracted from the data set. Cluster ordering is shown by the differences in colours along the map. More black areas appear on the traditional SOM than the SOMersault, indicating that the SOM has nodes positioned in regions with no data. Some colours appear on the SOM that do not appear in the input data, indicating that the SOM is ‘extracting’ patterns that do not actually exist in the data set.

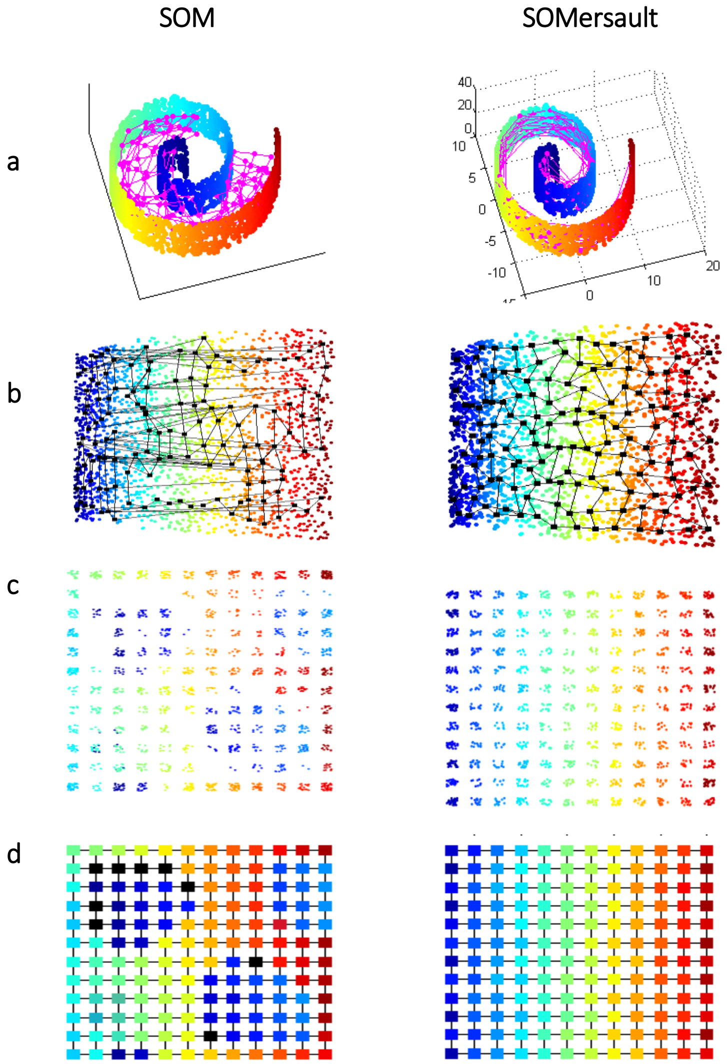


Figure 22: Detail of the Swiss roll data set analysis using the traditional SOM and the SOMersault. a) 3D data set and fitted 12x12 maps; b) Same as above but the maps are 'unravelled' along the geodesic surface. c) The 12x12 map grid, showing each node populated by the actual data points that map to it. d) The 12x12 map grid with nodes coloured by the mean colour of the data assigned to them. Further information is provided in the text.

The visualisations and error measures indicate that the quantization, topological preservation and clustering of the data by the SOMersault are improved compared to the traditional SOM. The improvement in quantization is a consequence of the grid more closely aligning with the geodesic surface and not stretching over gaps between distinct regions of the manifold. Data from one region does not tend to get assigned to nodes from another region. Topological preservation is also improved, showing a smooth transition of node values along the grid. The SOMersault nodes follow the unravelled colour scheme of the geodesic surface without any cross-over of connections between nodes. The significant improvement in the clustering of the data results from the data within each cluster of the SOMersault being more consistent as it is drawn from the same region of the geodesic surface.

6.4.2 Real world application – Water scarcity in global river basins

We now use the SOMersault method to investigate the relationship between the quantity of water present in a river basin in all forms (groundwater, surface water, soil moisture, snow, and ice) and the water scarcity experienced within the basin due to anthropogenic water consumption. Seasonal fluctuations of this relationship are explored for a global selection of river basins representing a variety of latitudes, geographies, population densities and climates.

The motivation behind this study stems from the uneven global spatial and temporal distribution of water resources as highlighted by Postel et al. (1996) and Oki (2006), who state that even though more than enough fresh water exists globally and annually to provide for all of humanity's needs, spatial and temporal variations of the resource lead to water scarcity in certain regions at certain times.

Here we use the SOMersault method to make a spatiotemporal assessment of the patterns of scarcity and water availability in a global context, which could assist in promoting the sharing of modifications in practices, trade, and agricultural management amongst basin authorities. Spatiotemporal clusters are produced, indicating river basins with similar circumstances regarding the amount of water actually existing within a catchment compared to the scarcity experienced for a given month, without requiring an explicit account of differences in basin sizes, runoff magnitudes, river lengths, geographic locations, and the degree of urbanization.

6.4.2.1 Application of SOMersault to hydrological data

In this application, we explore the use of the SOMersault with hydrological data. As with data from most environmental fields, hydrological data contains many forms of fluctuations resulting from diurnal, seasonal, or longer term variations in climate related variables. Temporal fluctuations in hydrological data could be linked to precipitation, snow melt, river discharge, groundwater storage, water chemistry, suspended loads, temperature (air and water), surface water level, soil wetting and drying, or water table oscillations. Spatial fluctuations may be due to changing geography and climates over the region of study, such as variations in altitude or vegetation. Anthropogenic influences on the water cycle, such as seasonal groundwater extraction for agricultural use, could also lead to temporal or spatial fluctuations.

When multivariate data sets include these types of rhythmic fluctuations of more than one variable, hysteresis loops may form. Hysteresis behaviour is characterised by the output of a system forming a loop due to a dynamic lag between the input and output variables in systems with alternately increasing and decreasing inputs. For example, in a system with two variables this leads to a situation where two possible values of one variable exist for a single value of the other, with the actual relationship at any time depending on the current stage of the cycle. This is a common occurrence in hydrological data, in which output (e.g. runoff) during the latter part of a process (e.g. a rainfall event) may not exactly mimic the beginning of the process due to changes in the physical system (e.g. available storage areas filling up with precipitation) during the time difference.

In a data set consisting of multiple fluctuating variables with differing frequencies and amplitudes, it can be challenging to identify any similarities between individual data items. The SOMersault can be useful for clustering such data; the clusters will discern between opposing limbs of a cycle, loop or parabola, providing a separation in the data and the extraction of patterns from both regions. A traditional SOM may not differentiate between the distinct portions of the relationship, placing data items from both sides of a loop into the same cluster and thereby not providing a complete extraction of the pertinent features of the data set. Furthermore, the SOMersault will not only discover the prevalent patterns, it will order the clusters to naturally follow the curve that describes the relationship between the variables.

6.4.2.2 Data

The data set for this analysis consists of a global set of river basins, in which two variables of interest each undergo seasonal fluctuations separated by changeable time lags. These variables are:

1) terrestrial water storage, and
2) anthropogenically induced water scarcity.

The annual cycles of these variables exhibit phase differences and hysteresis behaviour. Differing frequencies and amplitudes are present for both variables across the data set due to geographical, climatic, soil, vegetation, and anthropogenic differences among the basins. These variables will be referred to as ‘storage’ and ‘scarcity’ in this paper.

River basin water storage data has been obtained from the GRACE (Gravity Recovery and Climate Experiment) satellite system, which produces spherical harmonic coefficients describing time-variable gravity field variations (Landerer & Swenson, 2012) (websites 1 and 2). The redistribution of water, snow and ice is the main source of changes in Earth’s gravity field on a monthly timescale, and therefore these gravity changes can give an understanding of global hydrological processes (Swenson, 2002). The average GRACE monthly terrestrial water storage (TWS) values for each basin (in units of ‘mm of equivalent water thickness’) are available for 2002-2012, though some months are missing. The data was downloaded by individual basin using Total Runoff Integrating Pathways (TRIP) basin boundaries with a Gaussian smoothing kernel of 300km radius, from data centre CSR RL05 DS. ‘Scaled’ (rather than raw) TWS GRACE data, which has been corrected for signal modification due to filtering and truncation, has been downloaded and used in this study based on the recommendation of Landerer & Swenson (2012). Monthly values have been averaged over the time series to produce a typical year of average monthly data for use in this study.

Water scarcity data by river basin was obtained from Hoekstra et al. (2012). This data indicates the ratio of surface water and groundwater consumption to the estimated water availability in a river basin, where water availability is defined as the runoff that would occur in the basin under natural (uninhabited) conditions minus the flow estimated to be required for maintaining critical ecological functioning within the basin. The seasonal variability of water scarcity has been captured in this data by producing monthly estimates, which are a ten year average of data from 1996-2005. Water use has been measured in terms of consumptive use rather than water withdrawals to account for water that is typically returned to the basin and is available for reuse. Hoekstra et al. (2012) adopted a value of 20% of the natural runoff being available for consumptive use (the remaining 80% can be used as long as it is returned for reuse and not permanently depleted). Low scarcity is defined as less than this available amount being depleted, with moderate, high, and severe scarcity levels indicating that more than this amount has been consumed and therefore environmental flow requirements are not met. Severe scarcity is defined as twice as much water removed from the river than has been estimated as being available for consumption without impinging on environmental flows, a value which coincides with the definition of 'severe water stress' used by Oki and Kanae (2006).

The data set consists of 40 river basins of various sizes from 6 continents, representing differing discharge patterns, basin sizes, climates, topographies, vegetation zones, population densities, and water consumption patterns. A list of the basins and their characteristics is given in Table 8. This global selection of river basins incorporates a broad variety of flow regimes and scarcity issues. Each river in this study experiences a range of scarcity levels throughout an average year, from low to severe, meaning that during at least some months the environmental flow requirements are met and in some months they are not.

The mean annual time series of storage and scarcity for each river are shown in Figure 23. The values for each river follow roughly sinusoidal patterns within each variable. Basin total water storage data has been normalised to a mean of 0 and variance of 1 to allow the comparison of river basins with vastly differing sizes. To identify periods when the relationships between the variables are similar for more than one river is a difficult task using these separate curves. It can be seen that some rivers follow similar patterns of monthly storage or scarcity, though due to the various lags it is not easy to discern which rivers are similar in both variables at a given time. Grouping the data with a clustering technique will aid in this task.

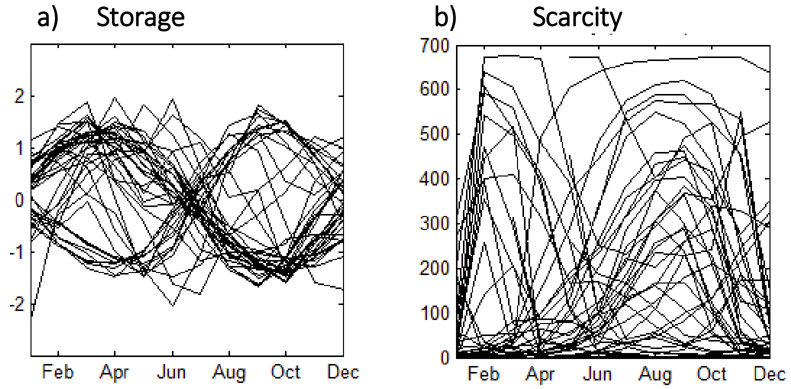


Figure 23: Average 12 month time series for all rivers: a) storage (normalised) and b) scarcity (raw).

Figure 24 shows normalised storage (blue) and scarcity (red dashed) time series for a few sample river basins. This indicates that although all storage and scarcity curves rise and fall during the year, the patterns and ratios between them vary greatly between rivers for reasons unique to each basin.

Table 8: River basins, sorted by continent and population density. Units of population density are 'people/km²'; units of scarcity are '% of water available for consumptive use that has been depleted' (see Section 6.4.2.2 for further information); and units of storage are 'mm of equivalent water thickness'. Sources: river basin area, population, and scarcity data are from Hoekstra et al. (2012); storage data is from GRACE (website 1).

River basin	Region	Area (km²)	Population	Pop density	Annual average scarcity	Annual storage variation
Chira	Peru	16,700	651,347	39	212	55
Biobio	Chile	24,109	655,158	27	25	213
Lempa	Guatemala, Honduras, El Salvador	11,780	141,848	12	210	314
Solo	Indonesia	15,146	11,102,900	733	113	200
Ganges	India, Bangladesh	1,024,463	454,094,000	443	241	315
Krishna	India (southeast)	269,869	76,933,400	285	334	344
Indus	Tibet, Pakistan	1,139,075	212,208,000	186	271	69
Dead Sea Basin	Jordan, Israel, West Bank, Lebanon, Egypt	35,444	6,149,610	174	328	34
Huang He (Yellow)	China	988,063	160,715,000	163	205	36
Chao Phraya	Thailand	188,419	26,782,400	142	132	405
Ishikari	Japan	13,783	1,941,950	141	116	189
Sakarya	Turkey	62,483	5,654,860	91	176	181
Mekong	China, Burma, Laos, Thailand, Cambodia, Vietnam	787,257	57,932,400	74	135	350
Tigris & Euphrates	Turkey, Syria, Iraq, Iran, Kuwait	832,579	49,255,700	59	180	148
Volga	Russia (central)	1,408,279	61,273,800	44	56	150
Ural	Russia, Kazakhstan	339,084	4,062,630	12	53	121
Ob	Russia (western Siberia)	2,701,041	29,372,200	11	33	105
Tarim	China (northwest)	1,051,731	9,311,040	9	346	50
Thames	England	12,359	9,674,080	783	63	45
Escaut (Schelde)	France, Belgium, Netherlands	21,499	9,448,070	439	102	86
Seine	France (Paris basin)	74,228	15,598,100	210	83	98
Guadalquivir	Spain (southwest)	56,955	3,947,090	69	238	70
Douro	Spain, Portugal	96,125	3,744,450	39	101	99
Ebro	Spain (eastern)	85,159	2,922,480	34	83	77
Sebou	Morocco	36,201	5,479,260	151	188	48
Nile	Uganda, South Sudan, Sudan, Ethiopia, Egypt	3,078,088	162,346,000	53	85	92
Pangani	Tanzania	50,365	2,174,380	43	186	124
Limpopo	South Africa, Botswana, Zimbabwe, Mozambique	415,623	15,637,400	38	214	79
Niger	Guinea, Mali, Niger, Benin, Nigeria	2,117,889	76,930,900	36	36	210
Orange	South Africa	972,388	12,665,700	13	146	35
Senegal	Senegal, Mauritania	436,981	5,134,070	12	9	208
San Joaquin	USA	34,366	1,681,380	49	323	174
Sacramento	USA	77,209	3,015,150	39	172	204
Brazos	USA	117,853	2,820,050	24	269	56
Mississippi	USA	3,196,605	74,637,300	23	67	86
Nelson	Canada (Manitoba)	1,099,380	5,565,740	5	95	45
Klamath	USA	40,040	137,158	3	54	171
Murray	Australia (southeastern)	1,059,508	2,348,090	2	234	26

Ord	Australia (northwestern)	55,686	2,473	0	495	278

For example, the San Joaquin River provides drinking water to more than 4.5 million people including the city of San Francisco as well as irrigating one of the most productive agricultural regions in the world, generating hydropower and supporting the habitat of many endangered and declining species. Outdated water management approaches have made it America's most endangered river (website 3), with severe scarcity occurring from May to November, coinciding with intense agricultural needs and low precipitation. The basin of the 1300km long Tarim River in northwest China is inhabited by 9.3 million people in desert conditions. Severe scarcity is experienced for 9 months of the year, from February to October. Even though January has low scarcity, very low river flow in February increases scarcity levels to severe which then decline into the summer, with high water in July caused by snow melt. A recent World Bank project to restore the waterway has led to environmental revitalization, increased agricultural output, water conservation and poverty reduction (website 5). In the Thames, even though storage starts to increase in June, scarcity increases throughout the summer months due to increased consumption. The principal river of western Africa, the Niger, floods for 9 months of the year and incorporates an immense inner delta of marshes, lakes braided streams. Though the Niger basin experiences very low scarcity for most of the year due to this flooding, in February and March when there is no flow, severe scarcity occurs. The Indus River, flowing from Tibet through Pakistan, sustains over 212 million people, and experiences 8 months per year of severe scarcity leading to unsustainable groundwater depletion (Hoekstra et al., 2012). In the Murray basin, high levels of consumption lead to aquifer depletion during certain parts of the year. The Pangani River, flowing from Mt Kilimanjaro to the Indian Ocean, has been extensively dammed for irrigation in the highlands. The reduced outflow is affecting coastal communities by the depletion of fish stocks. Even though basin water storage levels are above average, low river flows from January to March create conditions of severe scarcity. The Nelson River in northern Canada experiences low scarcity except when frozen at the beginning of the year. The gentle increase in storage from October onwards represents snowfall which does not alleviate scarcity until the melt begins in March.

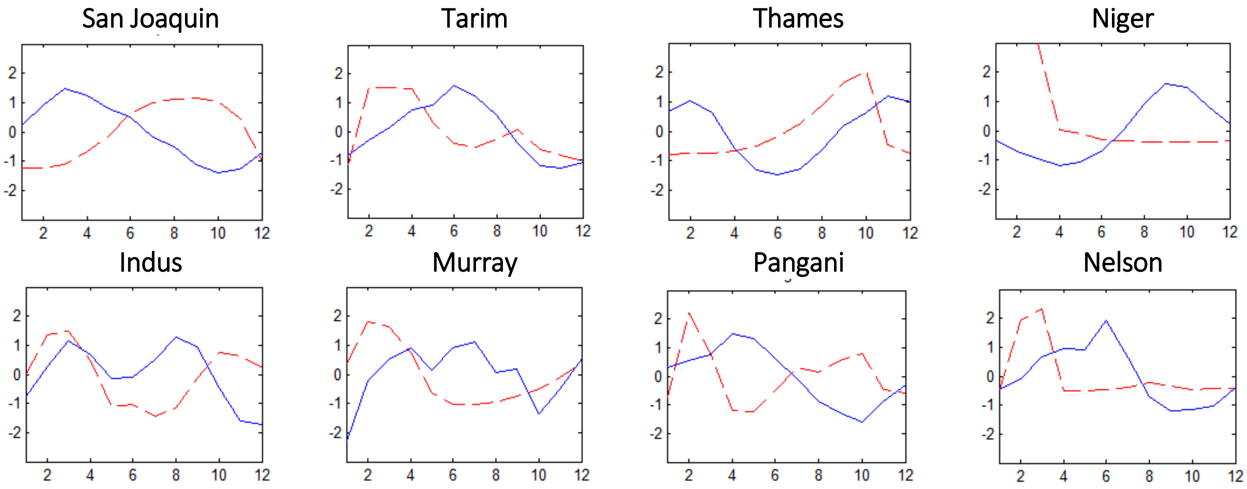


Figure 24: Normalised storage (blue) and scarcity (red dashed) average 12 month time series for a selection of rivers.

When combined, the storage and scarcity annual time series form hysteresis loops. Figure 25 shows these loops for a selection of rivers, with the time series of each separate river forming a cyclical path throughout the year. Normalised data for both scarcity and storage is used.

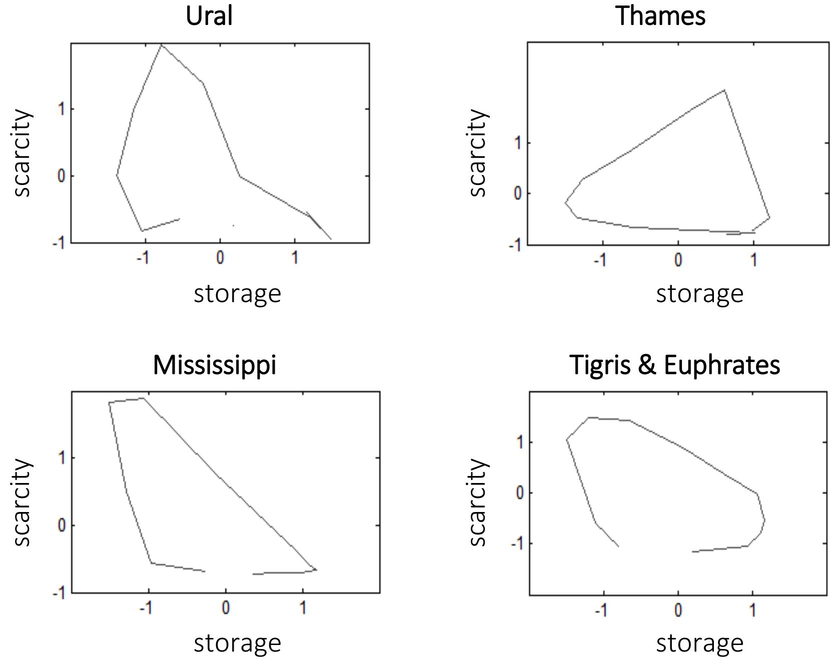


Figure 25: Storage-scarcity annual hysteresis loops for a selection of rivers.

6.4.2.3 Implementation

We first convert the cyclical annual data in Figure 25 to polar coordinates to align the individual river loops. The use of polar coordinates allows data from more than one cycle to be overlaid, and is recommended for analyzing patterns in cyclic data and identifying characteristics of spatiotemporal systems [Streit (2015), Andrienko (2006), Cheng (2001) and Wickham (2008)]. The axes of the ellipse formed by the data points in polar coordinates correspond roughly to the standard deviation of each variable, with the divergence of each point from the origin based on the level of scarcity of that river-month (further out data has a higher level of scarcity). The alignment of the hysteresis loops of the rivers through polar coordinates will allow for a comparison across the multiple cyclic patterns (Streit, 2015).

An artefact of the scarcity calculations leads to river-months with no flow (through aridity or ice) having extremely high scarcity values, regardless of the amount of water that is sought to be withdrawn. Data from these river-months has therefore been omitted from this study as it is the relationship involving actual water consumption that we are investigating.

The SOMersault will be used to produce spatiotemporal clusters of river-month combinations that experience similar conditions in the relationship between basin water storage and water scarcity. This will provide information on basins that experience similar conditions, even though the basins themselves may be dissimilar in many other ways. Clusters will be ordered in alignment with the curves of the data and will discern between the rising and falling limbs of the hysteresis loops produced by the combination of the variables.

The SOMersault method has been applied in MATLAB with modifications to the SOM Toolbox (Vesanto, 2000). The LTSA reduction was performed with MATLAB code mani.m (Whitman, 2005), with 6 nearest neighbours and sigma of 10.

6.4.2.4 Results

The application of the SOMersault and the SOM to the set of global river-months is shown in Figure 26. The map nodes (squares) are coloured in smoothly transitioning colours, and linked with a black line, to indicate the ordering of the clusters on each map. The data points mapping to each node are indicated by colour.

The SOMersault has performed a spatiotemporal clustering of the data whilst maintaining the cyclic nature of the data set. Adjacent nodes of the SOMersault represent adjacent items of data, and the nodes are positioned amid the data items rather than in empty space away from areas of high data density. The values of each variable flow smoothly along the length of the SOMersault. Starting at the left of the SOMersault in Figure 26 and travelling clockwise, the nodes represent: low storage and average scarcity, high scarcity and average storage, high storage and average scarcity, and average storage and low scarcity.

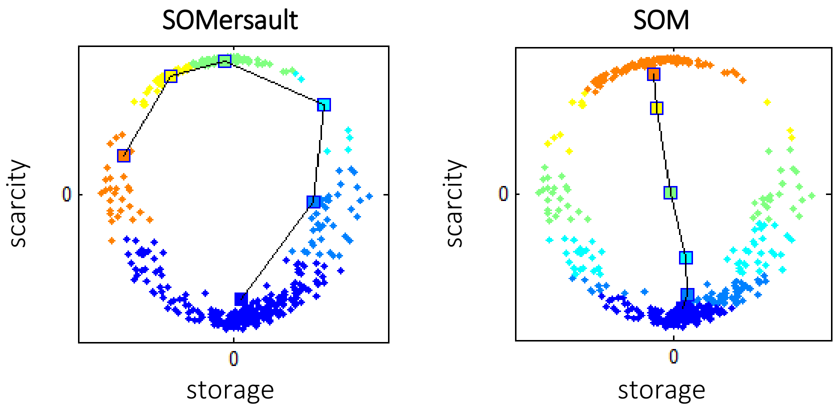


Figure 26: The hydrological data set is represented with the SOMELSTA on the left, and the traditional SOM on the right. Data points are coloured to match the node they are assigned to. It can be seen that the SOMersault separates the data into clusters closely following the curve of the data, whilst the SOM only clusters the data vertically, with both high and low storage data represented by the same node. The SOM has nodes in areas of white space, in which case the prototypes will have values that are not actually present in the data set. The SOMersault nodes, on the other hand, are located in close proximity to the individual data items.

The traditional SOM trained to represent the same data set with the same number of map nodes has also clustered the data, but the clusters do not discern between data located on the left and right reaches of the loop. It can be seen from the coloured data items that for a given level of scarcity, river-months with both high and low storage values are assigned to same map node. This means that basins experiencing above average scarcity, for example, with very little or no water storage available and those that actually have a lot of water present (perhaps in other forms such as snow, swamp, or groundwater) are represented by the same node (yellow). Most nodes are located in the empty space between areas of high data density and therefore the prototype vectors associated with the nodes, which take on an average value of all the data assigned to them, will not be similar to any of the data items they represent. The map never reaches high or low values of the storage variable.

In order to simulate the annual cycle with 12 readily understood (~monthly) divisions, the SOMersault is applied again, this time with 12 nodes. The results are shown in Figure 27.

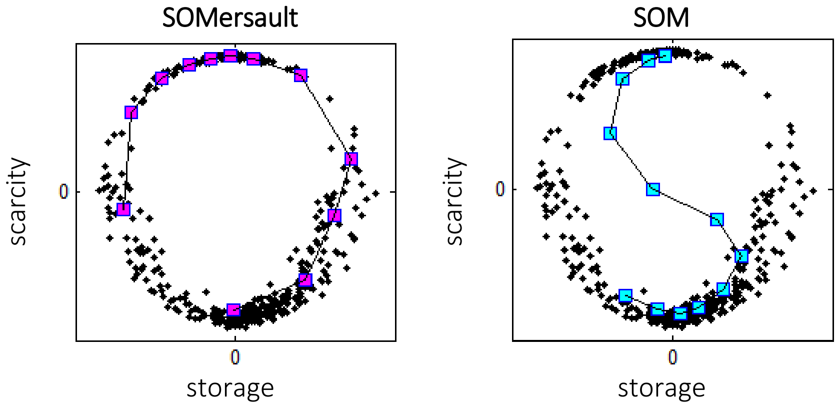


Figure 27: Annual data is divided into 12 clusters. The SOMersault is on the left, and the traditional SOM is on the right.

Error measures calculated for the representation of the river basin data by the SOMersault and the traditional SOM are given in Table 9. The error measures indicate that for this data set, the quantization, topographic preservation and clustering are improved by use of the SOMersault over the traditional SOM.

Table 9: Error measures for maps created with the SOMersault and traditional SOM methods on the river basin data. The quantization, ordering and clustering of the data set by the prototypes are evaluated. It can be seen that for each criteria, the SOMersault produces the lower result, indicating a better representation of the data set by the map.

Method	GQE (quantization)	GTE (topological preservation)	GC (clustering)
SOMersault	0.037	0.00	0.5
SOM	0.044	0.04	1.2

6.4.2.5 Discussion of real world application

A spatiotemporal clustering of the river basins has been performed with the SOMersault. Each month of the year at each river is attributed to a cluster. The clusters represent specific relationships between the total water available in the basin as measured by satellite, and the scarcity experienced in the river basin due to water consumption. These clusters have been determined without explicitly defining the complex relationships that climate, natural geographic features, and human development have on the water in the basins.

The ordering of the clusters by the SOMersault smoothly follows the varying conditions of the river-months around the hysteresis loop, with more distant clusters containing more different data. The SOMersault nodes exist in close proximity to the data items rather than in empty space on the interior of the loop as the SOM nodes do, and therefore the SOMersault vectors are more similar to the data they represent. In the SOM, the middle clusters are ordered by scarcity alone regardless of the state of storage within the basins, and the storage values of the prototypes are therefore an average of the high and low values on each side of the loop.

To further analyse the clusters of Figure 27, the nodes are numbered starting at the left with node 1 and proceeding clockwise to node 12. Each node is associated with a prototype vector that is representative of the section of data assigned to it. The smooth transition between the prototype vector values for each variable along the length of the SOMersault is shown in Figure 28. From one node to the next, the characteristics of the neighbouring clusters evolve smoothly around the hysteresis loop.

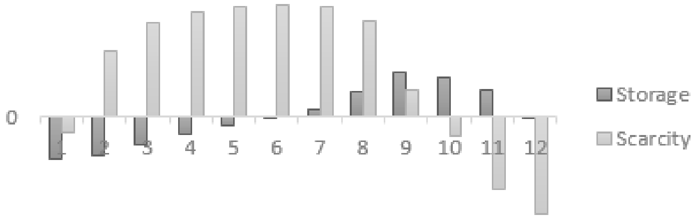


Figure 28: Prototype values by node for the SOMersault in Figure 27 (Node 1 is the first on the left, with numbering proceeding clockwise to Node 12 at the bottom). The smooth transition of values for each variable can be seen along the set of ordered map nodes. That is, neighbouring nodes have similar values for each variable. This allows the characteristics of data clustered to each map node to be compared based on the distance between nodes along the map.

Analysing the cluster characteristics described by the prototype values of each node provides information on the river-months assigned to each cluster. For example, in the SOMersault, average storage is a characteristic of both clusters 6 and 12, but cluster 6 is associated with high scarcity whereas low scarcity is a characteristic of cluster 12. The Nelson River is attributed to cluster 6 of the SOMersault in February and March (during frozen conditions) and the rest of year to cluster 12. The Dead Sea basin has very low variability in annual storage, and yet is in cluster 12 from December to April and cluster 6 the rest of the year when consumptive requirements are higher. The traditional SOM, however, effectively only clusters the data by scarcity in the mid-portion of nodes and storage in the lower nodes. Node 8 of the SOM (counting from the bottom), for example, is characterised by average scarcity but both high and low values of storage are assigned to it. The Ishikari River, in Japan, experiences low storage for the months of May, June, July and August and higher storage for the autumn and winter months. With the SOMersault, the months are divided up with these four months (May to August) attributed to nodes 1 and 2, and the rest of the year attributed to nodes 9 and 10. However, on the SOM all the months are attributed to nodes 7, 8 and 9 in the middle of the map, providing no distinction between months with high or low values of storage.

The characteristics of Node 8 of the SOMersault are above average storage and severe scarcity. This is an interesting combination as it shows that perhaps there is water available from sources other than direct river withdrawal which could potentially be used to alleviate the scarcity experienced in the basin. River-months in this cluster are: the Ord in May, the Tigris and Euphrates, Sakarya and Sacramento in June, the Yongding He in September, and the Krishna in December.

A further benefit of this process is the ability to identify rivers following similar patterns throughout their annual cycles. For this we use a post-processing SOM with each cluster number as input for that river-month, as in Clark et al. (2015). This results in an identification of groups of rivers that are relatively similar to each other in their storage and scarcity characteristics throughout the year. Some examples are shown in Figure 29: the Volga and Pangani (Russia, Tanzania); the Lempa, Niger, Krishna and Biobio (Central America, West Africa, India, Chile); the Guadalquivir, Limpopo, Ord, and San Joaquin (Spain, South Africa, Australia, USA); and the Chira, Solo and Orange (Peru, Indonesia, South Africa).

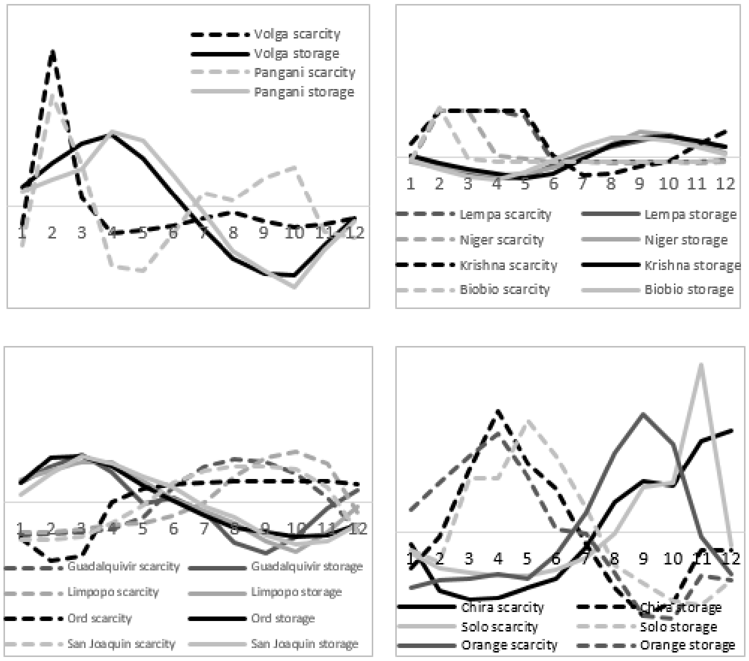


Figure 29: Clusters of rivers whose scarcity and storage relationships are similar throughout the annual cycle. The SOMersault has identified rivers with similar annual patterns in both variables.

The ordered clustering information produced by the SOMersault determines patterns of water resource conditions across basins of varying geographic, climatic and anthropogenic influences, revealing similarities experienced during certain phases of the annual cycle. This understanding could support: the sharing of management strategies between basins, decision making regarding the allocation of funding for water resources projects, agricultural planning, and the identification of further possibilities for water acquisition and water efficiency. The concept of sharing management principles between basins has been exemplified with ideas implemented in the Tarim River Basin project in China having been based on previous experience gained in the Murray basin in Australia (website 4). Consideration of alleviating water stress by close attention to the virtual water trade could also be informed, such as importing products that require high consumption during times of low storage. Certain basins may also benefit from attention to agricultural planning, such as the Tigris and Euphrates, Indus, Ganges, Tarim, Murray, and Limpopo in which 50-85% of consumption is a result of irrigation of three different crops or less (different crops by basin) (Hoekstra & Mekonnen, 2011) usually coinciding with times of low natural water storage in the basins.

Improvements in the application of the SOMersault to this real world data could include a method of linking the similarity between the end nodes representing the cyclic data (such as nodes 1 and 12 on Figure 27). In general before applying the SOMersault to real world data one should ensure that there is a low dimensional manifold present in the data set that is able to be discovered by LTSA.

6.5 DISCUSSION

The initialisation of the SOM grid is important to ensure an optimal clustering of the data (Abbas, 2008). Conventionally, initialisation is based on linear principal components which correspond to the overlapped structure of the entire data set, regardless of the shape of any manifold that might exist (Demartines, 1997). Due to their fundamental linearity, however, principal components are a suitable approximation for low dimensional manifolds only when the manifolds are embedded linearly, or nearly linearly, in the input space (Tenenbaum (2000), Guan (2005), Gorban (2008)).

When embedded manifolds are not simply perturbations of linear approximations of the data set, nonlinear techniques are useful as they are able to discover a low dimensional representation of the data in a coordinate system that captures the intrinsic degrees of freedom of the nonlinear data. Many nonlinear dimension reduction techniques exist, most of which are suitable to the investigation of a particular amount and type of data (for an overview of popular techniques refer to Wittman (2005) and Van der Maaten (2009)). Most nonlinear techniques guarantee global optimality due to the convexity of the cost function, though they cannot deal with data sets with high intrinsic dimensionality, or discontinuous manifolds (Van der Maaten, 2009).

The LTSA dimension reduction technique is used in the SOMersault, though others could also have been used. LTSA is a ‘local’ nonlinear technique which was chosen for the SOMersault as it draws inspiration from and improves upon the popular techniques of ISOMAP and LLE (Zhang, 2004). The ISOMAP, a ‘global’ technique, attempts to estimate and preserve the global properties of the original data by using geodesic distances to find the shortest paths between distant points (Tenenbaum, 2000). The ISOMAP is guaranteed for manifolds that are a convex region of Euclidean space but may be folded or twisted in high dimensional space. The algorithm involves finding the k nearest neighbours of each data point, graphing the shortest distances between all points with edges connecting neighbouring points, and then performing multidimensional scaling (MDS). LLE, a ‘local’ technique, attempts to estimate and preserve local properties of the input data by assuming the data and neighbours lie on the same patch of the locally linear manifold (Roweis, 2000). This eliminates the need to estimate distances between far data points. The LLE algorithm consists of finding the k nearest neighbours of each point and determining reconstruction weights assuming these neighbourhoods are linear.

For successful visualization and clustering of data, only the local structure of the input data needs to be retained in the reduction (Van der Maaten, 2009), and since the main functions of the SOM are visualization and clustering we have chosen to match it with a local nonlinear technique. As LTSA cannot be used directly for classification (Zhang, 2004), it is practical to combine it with a clustering or classification method such as SOMs.

As we have seen, the integration of LTSA and the SOM in the SOMersault method is not merely a two-step combination of performing LTSA and then creating a SOM. A recurrent interchange of information between the high and low dimensional spaces occurs during the SOMersault process. The map is created in a combination of low dimensional and high dimensional spaces, and interpreted in low dimensions based on the high dimensional clustering. Projecting the map into high dimensional space for the fine-tuning stage of training allows the map vectors to take on high dimensional values. The vectors can therefore be used as a smaller representative data set to work with during analysis, rather than the entire high dimensional data set. This step also allows the nodes to move off the low dimensional manifold and become closer to the data whilst retaining connections to ensure the structure of the low dimensional manifold is preserved, providing more accurate and informative clustering. High dimensional map vectors enable the visualization of the output map in input space alongside the original data, which would otherwise not be possible.

Like the SOM, the SOMersault is effective at determining clusters in nonlinear data. However, the improvement in clustering provided by the SOMersault is the ability to provide a geodesic ordering to the clusters. This ordering aligns the clusters with the underlying manifold, so that adjacent clusters on the map contain data that are neighbouring along the manifold. As the SOMersault clusters are linked along the geodesic surface, each contains data from only the local area rather than sectioning through overlapping layers of the nonlinear surface. As clustering is the most common use of the SOM, and providing an order to the clusters is a quality unique to the SOM (Agarwal & Skupin, 2008), this improvement is fundamental to the use of SOMs on nonlinear data.

The SOMersault also has specific benefits for the discovery of representative patterns in a data set. Due to the initial alignment of the SOMersault grid with the geodesic surface, the final locations of the nodes are expected to be on or near the geodesic surface of the data, and not located between layers of the manifold. The closer (more similar) the prototype vectors are to the data they represent, the more characteristic the patterns described by each vector are of the actual prevalent patterns of the data. Any nodes that settle in empty space (between layers of the manifold, with no data assigned to them) provide patterns which do not actually exist in the data set. In our examples, we have seen that nodes of the traditional SOM frequently rest in areas of data space with no nearby data points, since the linear initialisation is based on principal components which can span directions where no data exists.

The qualities of the SOMersault that offer benefits over the traditional SOM as discussed here are also expected to provide benefits when compared with similar methods combining manifold learning techniques with the SOM, such as the ISOSOM, GDBSOM and DSOM. The fine-tuning of the prototype vectors in high dimensional input space has advantages over the ISOSOM in which the training is finalised in low dimensional space: the vectors can leave the low dimensional manifold to move closer to (and become more representative of) the high dimensional data, and the final vectors are themselves high dimensional. The restriction of the updating neighbourhood in the fine-tuning stage to a localised area of the geodesic surface can be expected to provide more accurate geodesic quantisation than might result from the ISOSOM which has no such restriction or the GBDSOM which uses Euclidean updating. Also, the SOMersault method does not introduce a number of new parameters which the user must specify, as in the DSOM; the traditional SOM framework is used with the familiar parameters for neighbourhood selection in both the global ordering and fine-tuning stages of training.

In the process of exploratory data analysis, it often becomes necessary to add new data to a map that has previously been trained with other data, in order to sort the newly acquired data into the clusters already determined to exist in the data set. This is known as an out-of-sample extension. As with the SOM, this is also possible with the SOMersault. The new data point will need to first find a place on the low dimensional projection, and then the BMU in high dimensions searched from a subset of nodes in the neighbourhood of the low dimensional BMU. To find the low dimensional projection of the new data item, Teng (2005) provides a nonparametric out-of-sample extension method that can be applied to all nonlinear dimension reduction techniques. This method finds the nearest neighbor of the new data item in high dimensional input space, computes the linear mapping to the corresponding low dimensional representation, and applies the same linear mapping to the new data to find its low dimensional representation. This process will lead to some estimation errors in the new data embedding (Van der Maaten, 2009), but this will not matter as the final data assignment will be completed with the actual new input data item in high dimensional space.

As with the SOM, careful consideration must be given to the selection of the number of nodes for the SOMersault as this will determine the balance between generalization and accuracy of pattern extraction in the results (Liu, 2006). With the SOMersault this choice must be based on the optimal coverage of the low dimensional manifold.

The intrinsic dimensionality of the data set, and therefore the target dimensionality for the low dimensional projection to be used in the SOMersault, can be estimated from the decrease in residual variance as the dimensionality of the projection is increased (Tenenbaum, 2000), or by means of a maximum likelihood intrinsic dimension estimator (as in Levina, 2004). However, the existence of self-consistent principal manifolds is not guaranteed for arbitrary distributions (Demartines, 1997), and if no low dimensional manifold is actually present in the data set, LTSA will not produce a lower dimensional projection and the SOMersault method would not be required.

6.6 Conclusion

The SOMersault method has been specifically designed for the exploratory data analysis of natural data sets in which commonly encountered nonlinear patterns require an analysis technique able to extract underlying nonlinear relationships from the complexity of the high dimensional data measurements. The tendency of natural systems to involve a rhythmic rise and fall of individual components make them particularly suitable for use with the SOMersault method. Patterns in the data could appear in the form of a lone peak or trough, a continuously recurring wave, or a combination of variables that rise and fall with associated amplitudes, frequencies and dynamic lags to create spirals, helixes or hysteresis loops when combined. The trained SOMersault map will essentially roll itself along the natural curves or bends of the geodesic surface of the data to enable pattern extraction, clustering and visualisation.

In the SOMersault technique, the SOM framework has been expanded to enable the characterization of these highly nonlinear manifolds. For the effective representation of nonlinear manifolds with SOMs, the first approximation of the map node locations in data space is fundamental to the outcome. This is because the initial shape and topological structure of the map grid remain throughout the training. The map can be bent and stretched, but not reshaped. It is therefore imperative to ensure the pre-imposed grid represents the main degrees of freedom in the data set, even if they are nonlinear. To achieve this, the SOMersault method involves an initial projection of the high dimensional data into a low dimensional space with LTSA, allowing the initial map to become spread smoothly on the geodesic surface of the underlying manifold. This transfers the global ordering aspect of map training into low dimensional space, so that bends and folds in the manifold do not become represented on the output map as overlapping structures. The map is further trained in high dimensional data space with a localized neighbourhood kernel. This limits the influence of each data item to the nodes directly around it, thereby avoiding any effect that data points on more distant areas of the manifold may have on nodes that are close in Euclidean distance.

We have shown that the output SOMersault grids generally have benefits in cluster ordering and pattern extraction when compared with conventional SOM grids. The clusters produced by the SOMersault are well ordered along the unravelled surface of the data set. In contrast, the clusters produced by the standard SOM do not follow the same smooth ordering. The ‘extraction’ of patterns that do not actually exist in the data is evident on the SOM maps which contain black nodes or colours that are not present in the input data, whereas in the SOMersault patterns emerge that are more representative of the data items as the map does not become stretched between distant areas of the manifold.

These benefits of the SOMersault in clustering and pattern extraction have been demonstrated through examples. Visualisations and quality measures have indicated the improvements in the clustering, quantization and topological representation of the data. A real world application on a global set of river basins with relationships of water scarcity and availability too complicated to explicitly define (for our purposes), has produced well-ordered clusters of rivers in space and time.

Future work may include an improved method for mapping the prototype vectors back and forth between the low and high dimensional spaces, for instance in calculation of the error measures in which the low dimensional locations of the vectors are currently approximated using the Voronoi sets of the high dimensional data. The geodesic topographic error defined in this paper has been based on the conventionally used TE, in which an error is registered only for data in which the first or second BMUs are out of place, but not if both are out of place. This could be improved to take into account large groups of map nodes shifted together to distant areas of the geodesic surface, as we have seen on the three dark blue sections of the

Swiss roll data set with the SOM. Any refinements to the nonlinear dimension reduction performed with LTSA will be included automatically.

6.7 Software and data availability

The SOM Toolbox for MATLAB (Vesanto, 2000) is available for free download from The Adaptive Informatics Research Centre of the Helsinki University of Technology at: http://www.cis.hut.fi/somtoolbox/. The mani.m manifold learning code for MATLAB (Whitman, 2005), including LTSA, is available for free download from: http://ocw.mit.edu/courses/earth-atmospheric-and-planetary-sciences/12-s990-quantifying-uncertainty-fall-2012/tools/mani.m. The Swiss roll, toroidal helix and punctured sphere data sets are part of this code. LTSA code for MATLAB is also contained in the NL Toolbox (Van der Maaten, 2007) available at: https://lvdmaaten.github.io/drtoolbox/. Water scarcity data is available in spreadsheet format at: http://waterfootprint.org/en/resources/water-footprint-statistics/#CP4, and water storage data is available for download by river basin at: http://geoid.colorado.edu/grace/dataportal.html.

6.8 References

Abbas, O.A. (2008). Comparisons between data clustering algorithms. International Arab Journal of Information Technology, 5(3), 320-325.

Adam, J. A. (2006). Mathematics in nature: Modeling patterns in the natural world. Princeton University Press.

Adeloye, A. J., & Rustum, R. (2012). Self-organizing map rainfall-runoff multivariate modelling for runoff reconstruction in inadequately gauged basins. Hydrology Research, 43(5), 603-617.

Agarwal, P., & Skupin, A. (2008). Self-organizing maps: Applications in geographic information science. John Wiley & Sons.

Andrienko, N., & Andrienko, G. (2006). Exploratory analysis of spatial and temporal data: A systematic approach. Springer Science & Business Media.

Céréghino, R, & Park, Y.S. (2009). Review of the self-organizing map (SOM) approach in water resources: commentary. Environmental Modelling & Software, 24(8), 945-947.

Cheng, T., & Wang, D. (2001). Visualizing cyclic spatio-temporal patterns in polar coordinate systems. icaci.org. 2460-2467.

Clark, S., Sarlin, P., Sharma, A., & Sisson, S. A. (2015). Increasing dependence on foreign water resources? An assessment of trends in global virtual water flows using a self-organizing time map. Ecological Informatics, 26, 192-202.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. Pattern Analysis and Machine Intelligence, IEEE Transactions, (2), 224-227.

Demartines, P., & Hérault, J. (1997). Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. Neural Networks, IEEE Transactions, 8(1), 148-154.

Frank, S. A. (2009). The common patterns of nature. Journal of Evolutionary Biology, 22(8), 1563-1585.

Ghodsi, A. (2006). Dimensionality reduction: A short tutorial. Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada.

Gorban, A. N., & Zinovyev, A. Y. (2008). Elastic maps and nets for approximating principal manifolds and their application to microarray data visualization. Principal Manifolds for Data Visualization and Dimension Reduction, pp. 96-130: Springer Berlin Heidelberg.

Guan, H., & Turk, M. (2005). 3D hand pose reconstruction with ISOSOM. Advances in Visual Computing (pp. 630-635): Springer.

Härdle, W., & Simar, L. (2007). Applied multivariate statistical analysis: Springer Science & Business Media.

Hoekstra, A. Y., & Mekonnen, M. M. (2011). Global water scarcity: The monthly blue water footprint compared to blue water availability for the world's major river basins.

Hoekstra, A. Y., Mekonnen, M. M., Chapagain, A. K., Mathews, R. E., & Richter, B. D. (2012). Global monthly water scarcity: blue water footprints versus blue water availability. PLoS One, 7(2), e32688.

Huang, Y., Zha, X. F., Lee, J., & Liu, C. (2013). Discriminant diffusion maps analysis: A robust manifold learner for dimensionality reduction and its applications in machine condition monitoring and fault diagnosis. Mechanical Systems and Signal Processing, 34(1), 277-297.

Kneubühl, F. K. (2013). Oscillations and waves. Springer Science & Business Media.

Kohonen, T. (2001). Self-organizing maps. Volume 30 of Springer Series in Information Sciences: Springer Berlin.

Landerer, F. W., & Swenson, S. C. (2012). Accuracy of scaled GRACE terrestrial water storage estimates. Water Resources Research, 48(4).

Lee, J. A., & Verleysen, M. (2005). Nonlinear dimensionality reduction of data manifolds with essential loops. Neurocomputing, 67, 29-53.

Levina, E., & Bickel, P. J. (2004). Maximum likelihood estimation of intrinsic dimension. Advances in Neural Information Processing Systems, 777-784.

Morioka, Y., Tozuka, T., & Yamagata, T. (2010). Climate variability in the southern Indian Ocean as revealed by self-organizing maps. Climate Dynamics, 35(6), 1059-1072.

Oki, T., & Kanae, S. (2006). Global hydrological cycles and world water resources. Science, 313(5790), 1068-1072.

Ota, K., Aoki, T., Kurata, K., & Aoyagi, T. (2011). Asymmetric neighborhood functions accelerate ordering process of self-organizing maps. Physical Review E, 83(2), 021903.

Park, Y. S., Céréghino, R., Compin, A., & Lek, S. (2003). Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. Ecological Modelling, 160(3), 265-280.

Postel, S. L., Daily, G. C., & Ehrlich, P. R. (1996). Human appropriation of renewable fresh water. Science, 271(5250), 785-788. doi: 10.1126/science.271.5250.785

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science, 290(5500), 2323-2326.

Shao, C., Wan, C., & Hu, H. (2015). Manifold learning and visualization based on dynamic self-organizing map. Neural Network World, 25(2), 175.

Shi, C., Zhang, S., & Shi, Z. Z. (2006). Geodesic distance based SOM for image clustering. International Conference on Sensing, Computing and Automation, 2483-2488.

Smakhtin, V., Revenga, C., & Döll, P. (2004). A pilot global assessment of environmental water requirements and scarcity. Water International, 29(3), 307-317.

Stewart, I. (2008). Nature's numbers: the unreal reality of mathematics. Basic Books.

Steynor, A. C., Hewitson, B. C., & Tadross, M. A. (2009). Projected future runoff of the Breede River under climate change. Water SA, 35(4), 433-440.

Streit, M., & Gehlenborg, N. (2015). Points of view: Temporal data. Nature Methods, 12(2), 97-97.

Swenson, S., & Wahr, J. (2002). Methods for inferring regional surface-mass anomalies from Gravity Recovery and Climate Experiment (GRACE) measurements of time-variable gravity. Journal of Geophysical Research: Solid Earth, 107(B9).

Swenson, S., Wahr, J., & Milly, P. C. D. (2003). Estimated accuracies of regional water storage variations inferred from the Gravity Recovery and Climate Experiment (GRACE). Water Resources Research, 39(8).

Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500), 2319-2323.

Teng, L., Li, H., Fu, X., Chen, W., & Shen, I. F. (2005). Dimension reduction of microarray data based on local tangent space alignment. Fourth IEEE Conference on Cognitive Informatics (ICCI 2005), 154-159. IEEE.

Van der Maaten, L., Postma, E., & Van den Herik, J. (2007). MATLAB toolbox for dimensionality reduction. MICC, Maastricht University.

Van der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: a comparative review. Journal of Machine Learning Research, 10, 66-71.

Vesanto, J. (1999). SOM-based data visualization methods. Intelligent Data Analysis, 3(2), 111-126.

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. Ieee Transactions on Neural Networks, 11(3), 586-600. doi: 10.1109/72.846731

Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (2000). SOM toolbox for MATLAB 5. Citeseer.

Wickham, H. A. (2008). Practical tools for exploring data and models. ProQuest.

Wittman, T. (2005) "Manifold learning MATLAB demo." Department of Mathematics, University of Minnesota.

Zhang, Z., & Zha, H. (2004). Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. Journal of Shanghai University (English Edition), 8(4), 406-424.

Websites:

1) http://geoid.colorado.edu/grace/dataportal.html
2) http://www.csr.utexas.edu/grace /overview.html
3) http://www.americanrivers.org/endangered-rivers/2014-report/san-joaquin/
4) http://www.worldbank.org/en/news/feature/2007/05/29/restoring-chinas-tarim-river-basin

7 PAPER 4 – EXTENSIONS TO TEMPORAL CLUSTER ANALYSIS AND PARAMETER SELECTION

This chapter is published as:

Patterns and comparisons of human-induced changes in river flood impacts in cities

Stephanie Clark, Ashish Sharma, Scott A. Sisson. *Hydrology and Earth System Sciences.* 22, 1793–1810, 2018.

7.1 ABSTRACT

This study investigates patterns of current conditions and anticipated future changes in city-level flood impacts driven by urbanisation and climate change. Global patterns relating urban river flood impacts to socioeconomic development and changing hydrologic conditions are established, and world cities are matched to these patterns. Comparisons are provided between 98 individual cities. We use a novel adaption of the self-organizing map method to establish and present patterns in the nonlinearly-related environmental and social variables. Output maps of prevalent patterns compare baseline and changing trends of city-specific exposures of population and property to river flooding, revealing relationships between the cities based on their relative map placements. Cities experiencing high (or low) baseline flood impacts on population and/or property that are expected to improve (or worsen), as a result of anticipated climate change and development, are identified and compared. This paper condenses and conveys large amounts of information through visual communication to accelerate the understanding of relationships between local urban conditions and global processes, and to potentially motivate knowledge transfer between decision makers facing similar circumstances.

7.2 INTRODUCTION

Through urban development and climate change, humans are progressively generating (and being on the receiving end of) increased hydrologic impacts, with these anthropogenically induced changes becoming particularly evident in cities (Revi et al., 2014, Mills 2007, Kreimer et al., 2003; Willems et al., 2012). With high densities of urban populations, infrastructure, property and industry, cities are both substantial drivers and receivers of environmental impacts. River flooding, the environmental event affecting more people than any other natural hazard (Doocy et al., 2013; Desai et al., 2015; Sofia et al., 2016), currently poses a threat to almost 380 million urban residents (UN-Habitat, 2014). Globally, hydrologic regimes leading to urban flooding are varying with climate change (Desai et al., 2015; UNEP, 2016; Willems et al., 2012), and locally, socioeconomic factors associated with urban development (variations in population growth, development, land use and urban density) are uniquely altering each city's individual response to these changing flood levels and frequencies (Desai et al., 2015). In the next few decades, cities will need to anticipate and adapt to this combination of shifting quantities of water and city features (Revi et al. 2014; Doocy et al., 2013). In this study, we aim to develop an understanding of the prevalent global patterns of human-environmental relationships influencing city-level river flooding, and discover how a global set of individual cities fits into these patterns.

Climate change and urbanization are combining to force more frequent flooding and higher flood peaks in cities, though the influence of each factor varies spatially and temporally (Desai et al., 2015). Historically, cities have formed near rivers and population density is still highest, globally, where the closest water feature is a large river. As cities grow, the proximity of population and property to these water courses increases (Kummu et al., 2011). It is estimated that 70% of the world's population will live in cities by 2050 (UN-Habitat, 2010), up from 54% in 2015 (UN-DESA, 2015). With this rapid urbanization, highly populated areas are experiencing an increase in flood vulnerability (Kreimer et al., 2003), as unplanned expansion often leads to migration into urban flood plains (Jongman et al., 2012; Revi et al., 2014). Global urban land cover is increasing at a rate over double that of urban population growth (Angel et al., 2010a) and is projected to increase three-fold by 2030 (Pachauri et al., 2014). More impervious areas and encroachment into the surrounding countryside are forcing faster concentrations of rainfall in urban rivers during storm events, as well as higher flood peaks (Desai et al., 2015; Doocy et al., 2013; Kreimer et al., 2003).

Hydrology in cities is also affected by increased surface temperatures associated with climate change. Already, increases in the frequency and intensity of precipitation (Frich, et al. 2002; Desai et al., 2015; UNEP, 2016), changes in spatial and temporal storm patterns (Wasko & Sharma, 2015) and changing snow melt conditions (Schiermeier, 2011; Barnett et al., 2005; Immerzeel et al., 2010) are leading to variations in the magnitude, frequency and timing of urban river floods, with higher peak flows and shorter response times (Shiermeier, 2011; Cunderlik, 2009). These changing patterns of precipitation and runoff are complex and not uniformly spatially distributed (Meehl et al., 2005; Desai et al., 2015; Wentz et al., 2007; Frich et al., 2002). In the future, cities in particular are predicted to become even more vulnerable to extreme hydrologic events as a result of climate change (Pachauri et al., 2014; Willems et al., 2012; Revi et al., 2014, Sofia et al., 2016). Increases in rainfall intensity at urban hydrology scales of up to 60% are anticipated by 2100 (Willems et al., 2012), and the micro-climates of cities are expected to interact with climate change in a variety of ways, potentially exacerbating flood effects (Revi et al., 2014).

In this paper, a comparison is made amongst a selection of cities based on their current and projected future urban river flood impacts on population and property, resulting from an anticipated combination of climate change and development. It should be noted that fluvial flooding is the only type of flooding that is considered here, and this study does not include an analysis of cities subject to coastal or pluvial flooding. Analysing data with city-specific projections of changes in hydrology, population and development levels (based on future climate scenarios, projected development pathways, and a best assumption of flood protection standards) we produce an analysis and visualisation of the patterns of baseline conditions and anticipated changes in city-level river flooding impacts to the year 2030. We establish the prevalent global spatial and temporal patterns of urban flood impacts, explore these impacts as resulting from both developmental and hydrological drivers, and match the cities to their most similar pattern. The patterns are established through dimension reduction, clustering and visualisation of multivariate data with an adaptation of the self-organizing map (SOM) technique. The SOM is an artificial neural network useful for exploring nonlinearly related variables, and is popular for investigating potentially difficult-to-define environmental responses to human influences (e.g. Shanmuganathan et al., 2006; Vaclavik et al., 2013; Clark et al., 2016b) as well as providing comparisons between geographic areas (Kaski & Kohonen, 1996; Clark et al., 2015; Clark et al., 2016). We begin by presenting analyses of patterns of urban flood conditions (as measured by the amount of population affected and urban damages costs) for a baseline global snapshot (2010), then investigate projected temporal changes (up to 2030), and finally combine this information into a global temporal analysis of the cities. As individual cities are matched to their closest patterns at each stage, we discover clusters of cities with similar urban flooding characteristics and projected trends.

A growing body of research is investigating the impact of anthropogenic changes on urban flooding at regional and global scales, however we have found no literature comparing specific cities in terms of changing city-level flood impacts on populations and property. The Intergovernmental Panel on Climate Change's 5th Assessment Report Chapter 8 'Urban Areas' (Revi et al., 2014) discusses the vulnerabilities and resilience of cities to climate change in general, noting that the analysis is based on economic losses and would differ if a human component is included. Jongman et al. (2012) investigated global trends of coastal and river flooding based on changing regional population densities and land use. Increased vulnerability to flooding is attributed to population growth or increases in wealth, though the modelling does not include changing hydrology due to climate change. Jongman et al. (2015) estimated regional trends in human and economic river flooding vulnerabilities by income level, through hazard and exposure calculations. Kunkel et al., (1999) investigated the increasing trend of economic losses and fatalities in the USA due to increasing vulnerability to floods, however the climate change contribution to this increase was not possible to quantify due to a lack of data. Winsemius et al. (2016) produced the first projections of global future flood risk that consider separate impacts of climate change and socioeconomic development, with results discussed by geographic region (river basin) and economic level. The investigation of the connection between coastal flooding and climate change (increasing storms combined with sea level rise) is more common in the literature than the connection between river flooding and climate change (Nicholls et al., 2008; Nature, 2016) due to better data availability. Most existing river flood assessments are at a local or regional scale (as in Muis et al., 2015), limiting the possibility to compare between multiple cities, as studies at a global scale have traditionally been limited by a lack of datasets and methods. Sofia et al (2016) emphasize that analyses of climate change and socio-economic development as both drivers and receptors of flood risk is needed. Muis et al. (2015) call for an investigation between the combination of land use change and hydrologic change on future flood risk. (Jongman et al., 2012) highlight that due to population growth and climate change, global methods incorporating both spatial and temporal dynamics to investigate inland flooding at the city scale are necessary for global development studies and estimating costs associated with climate change. To date, a global examination of changing flood conditions at the city level resulting from urban development and climate change, including a direct comparison between specific cities, has not been made. The analysis we present here corresponds directly to this gap in the literature.

General patterns as well as specific relationships can be extracted from the output maps in this paper. In the interest of channelling the 'potential of visual communication to accelerate social learning and motivate implementation of changes' (Sheppard, 2005) the aim of the method used here is to discover and demonstrate potentially interesting global patterns and relationships that would not otherwise be evident in the data, for example: clusters of cities which are currently experiencing high flood impacts that are projected to greatly increase in the future, and to what extent this may be due to climate change (or socioeconomic development) within each city; which cities not currently experiencing notable effects of flooding may expect to in the future; which cities are projected to mitigate potentially adverse flood effects from climate change with reductions in flooding due to socioeconomic factors; which cities are projected to experience an increased flood vulnerability driven by socioeconomic factors alone; and the relationship between the changes in vulnerability of the population and urban damages costs for each city.

The comparison of individual cities in this study (rather than river catchments) allows a blending of environmental and social information which reinforces the co-dependence of humans and their natural environment, a relationship which is often easily overlooked by urban dwellers. Explicitly visualising the role that urbanisation may have on the environmental conditions experienced by urban citizens is an essential reminder of this connection. Cities potentially facing similar circumstances and challenges are identified in this study, suggesting possibilities for a sharing of strategies. As climate change, development, and urban administrations transcend river basin boundaries, an investigation of impacts and determination of potential mitigation strategies at the city level as well as the basin level expands the potential for decision makers to be presented with all the available, relevant data for consideration.

7.3 DATA AND METHOD

7.3.1 Data

The data set used in this study combines city-level estimates of annual expected urban river flood impacts on population and urban damages costs (2010), projections of future changes in flood impacts attributed to climate change and/or development (up to 2030), and socioeconomic data for a globally distributed set of cities.

The selection of cities used here is based on a list provided by the Lincoln Institute of Land Policy's Atlas of Urban Expansion (Angel et al., 2010, website 1), spanning all continents except Antarctica, encompassing four economic levels and four population levels. City population data (2010) and future population estimates (2030) are from the UN Department of Economic and Social Affairs (UN-DESA, 2015), and GDP per country are from the World Bank's World Development Indicators database (website 2).

Annual river flood impact estimates are obtained from the global dataset of fluvial flood risk published in the World Resources Institute's Aqueduct Global Flood Analyzer Tool (herein referred to as Aqueduct) (Winsemius et al., 2013; Ward et al., 2013; website 3). Released in 2015, this data set comprises the first unified global set of fluvial flood risk data at the city level. As this data is solely related to the influence of fluvial flooding on metropolitan areas, it does not include coastal or pluvial flood risks. In this data set, Aqueduct provides separate estimates of annual impacts on the number of affected population (people exposed to flood waters) and urban property damages costs (in US dollars), which will be referred to in this paper as 'population' and 'damages' impacts.

Global hydrologic and hydraulic models, inundation modelling, and spatial data sets of population, land use and infrastructure are used within Aqueduct to quantify flood risk in each city. Aqueduct identifies future anticipated changes in urban flood vulnerabilities as driven by climate change (altered hydrology), socioeconomic development (population, land use and economic changes), or in most cases a combination of both. Either of these drivers may increase or decrease the frequency and intensity of flooding, and the resulting flood impacts, for a given city. Three separate scenarios of climate change and socioeconomic development (optimistic, business-as-usual, and pessimistic) are given in Aqueduct, and in this study we use data from the business-as-usual case for our future flood impact scenario. Future hydrologic and hydraulic estimates in Aqueduct are based on global circulation model data from the ISIMIP project (website 4) and changes in population and economic development are based on Shared Socioeconomic Pathways data with a downscaling procedure that differentiates between urban and rural growth (website 5; Samir & Lutz, 2014). Recent papers published with this data include Winsemius et al. (2016), Jongman et al. (2015) and Muis et al. (2015).

Expected flood impacts are provided by Aqueduct for nine possible levels of city-wide flood protection, from protection against the 2-year average return interval (ARI) flood to the 1000-year ARI flood. This protection level indicates how well protected the area is against flood damage, based on the standard or capacity of flood protection measures such as dikes, levees or dams. In this study, we assign an assumed flood protection level to each city based on the country's World Bank income level (as in the World Resource Institute's Aqueduct Global Flood Risk Country Rankings, website 6) due to a lack of information on each city's actual protection level. This method follows recommendations based on the rational that higher standards of protection against flooding may be expected in higher income countries (Jongman et al., 2012; Nicholls et al., 2008), and findings by Doocy et al. (2013) that flood impacts are significantly associated with classification of income level by the World Bank. We assume each city's flood protection level remains the same during the timeline of this study.

To allow for a comparison between cities of greatly differing sizes and hydrologic conditions, the wide-ranging data values were log-transformed. The data set was then standardized by transforming these values linearly into the range 0-1 (with the lowest value becoming 0 and the highest value becoming 1) for each variable (population affected, urban damages, etc). The data is log transformed, following recommendation by Agarwal & Skupin (2008) that highly skewed variable distributions may benefit from log transformation before use in the SOM. Cities with no flood impacts in both 2010 and 2030 were removed (22 cities), though cities with no flood impacts in 2010 but with flood impacts in 2030 have been kept in the study. The final list of cities is presented in *Table 10*.

Table 10: City list - alphabetically by region.

Eastern Asia & the Pacific	
Anqing	China
Ansan	Rep. of Korea
Beijing	China
Changzhi	China
Chinju	Rep. of Korea
Fukuoka	Japan
Guangzhou	China
Leshan	China
Pusan	Rep. of Korea
Seoul	Rep. of Korea
Shanghai	China
Sydney	Australia
Tokyo	Japan
Ulan Bator Mongolia	
Yiyang	China
Yulin	China
ZhengzhouChina	

Southeast Asia	
Bandung	Indonesia
Bangkok	Thailand
Ho Chi Minh City	Vietnam
Kuala Lumpur	Malaysia
Manila	Philippines
Palembang	Indonesia
Songkhla	Thailand

South Asia	
Dhaka	Bangladesh
HyderabadIndia	
Jalna	India
Kanpur	India
Kolkata	India
Mumbai	India
Puna	India
Rajshahi	Bangladesh
Vijayawada	India

Western & Central Asia	
Ahvaz	Iran
Astrakhan	Russian Fed.
Baku	Azerbaijan
Gorgan	Iran
Istanbul	Turkey
Kuwait City	Kuwait
Malatya	Turkey
Moscow	Russian Fed.
Oktyabrsky	Russian Fed.
Sanaa	Yemen
Shimkent	Kazakhstan
Teheran	Iran
Tel Aviv	Israel
Yerevan	Armenia
Zugdidi	Georgia

North Africa	
Alexandria Egypt	
Algiers	Algeria
Aswan	Egypt
Cairo	Egypt
Casablanca	Morocco
MarrakechMorocco	
Port SudanSudan	
Tebessa	Algeria

Sub-Saharan Africa	
Accra	Ghana
Bamako	Mali
Harare	Zimbabwe
Ibadan	Nigeria
Johannesburg	South Africa
Kampala	Uganda
Kigali	Rwanda
Ouagadougou	Burkina Faso

Latin America & the Caribbean	
Buenos Aires	Argentina
Caracas	Venezuela
Guadalajara	Mexico
Ilheus	Brazil
Jequie	Brazil
Mexico City	Mexico
Montevideo	Uruguay
Ribeirao Preto	Brazil
Santiago	Chile
Sao Paulo	Brazil
Tijuana	Mexico
ValleduparColombia	

North America	
Chicago	United States
Cincinnati	United States
Houston	United States
Los Angeles	United States
Minneapolis	United States
Modesto	United States
Philadelphia	United States
Pittsburgh	United States
Springfield United States	
St. Catharine's	Canada
Tacoma	United States

Europe	
Budapest	Hungary
Castellon	Spain
Le Mans	France
Leipzig	Germany
London	UK
Madrid	Spain
Paris	France
Sheffield	UK
Thessaloniki	Greece
Warsaw	Poland
Wien	Austria

7.3.2 Method

We use an extension to the self-organizing map method to determine patterns and similarities in the impacts, changes and drivers of urban flooding amongst the cities. The self-organizing map (SOM, Kohonen, 2001) is an unsupervised learning algorithm from the family of artificial neural networks that discovers patterns in multivariate data sets with nonlinear inter-variable relationships.

The SOM reduces the dimensionality of the data set by creating a (in this case) two-dimensional map grid which, through an iterative process, is essentially bent and stretched over the data set until it best characterizes the shape of the data cloud. The numerous data items become represented by a (usually) much smaller number of map nodes, known as prototypes. The map nodes, or prototypes, move iteratively into position amongst the data whilst maintaining their grid formation, establishing a higher density of prototypes in areas of higher data density. Once in position, the prototypes represent the most prevalent patterns in the data. Each data item is then matched to its closest prototype, creating clusters of similar data items.

The SOM algorithm consists of a two-step iterative process of comparing the map and the data, and then updating the map to better represent the data. The method begins with a calculation of distances in data space (in this case we use Euclidean distance) between each data item, x_i (where $i = 1{:}N$), and each map node, m_j (where $j = 1{:}M$). Data and map nodes vectors are all of the same dimension, d. The goal of the comparison stage is to find the nearest map node to each data item (commonly referred to as the best matching unit, BMU), which is then given the index c, using the following calculation:

$$\|\mathrm{x_i} - \mathrm{m_c}\| = \mathrm{min_j}\{\|\mathrm{x_i} - \mathrm{m_j}\|\}\ .$$

This partitions the data into subsets of items sharing the same nearest node, $\mathrm{m_c}$. Next, the locations of the map nodes are adjusted to become closer to their nearby data items. Application of a smoothing 'neighbourhood' kernel during this stage produces a smoother map by updating neighbouring nodes to a similar extent based on the nearby data. That is, the location of each map unit, m_j, becomes updated based on a weighted average of the data items matching itself as well as its neighbouring nodes, where the weighting is given by the neighbourhood kernel. The size of the kernel decreases with each iteration to include fewer nodes. We use a Gaussian shaped neighbourhood kernel, where h_{ij} (the neighbourhood kernel element indicating the influence of each data item, x_i, on the updating of node m_j) is defined at iteration t as:

$$\mathrm{h_{ij}(t)} = \exp(\frac{-(m_c - m_j)^2}{2\sigma^2(\mathrm{t})})$$

where σ is the kernel radius. At each iteration (t), the updated node locations are found as in (Kohonen, 2013):

$$\mathrm{m_j(t+1)} = \frac{\sum_{\mathrm{i=1}}^{N} \mathrm{h_{ij}(t)\,x_i}}{\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{h_{ij}(t)}}.$$

After map training is complete, the map node vectors each represent a unique combination of variables in the data, according the final location of the map nodes in data space. Each of these unique combinations of variables represent a characteristic pattern in the data. The data items are once again matched to their closest map node, forming clusters of data that best match each pattern.

In this study, the 'patterns' are the key characteristics represented by each map node vector (such as specific baseline and/or projected flood conditions, and the drivers of change). The 'cluster' members are the cities that match the pattern represented by their nearest map node better than they match the patterns of any other nodes.

As the SOM is an unsupervised learning algorithm, there is no subjectivity in the resulting cluster memberships. The iterative training process discovers the principal curves of the data set (the nonlinear directions of maximum variance) and aligns the map coordinate system with these, so that the two axes of the map generally follow the first two principal curves of the data. When the map is presented in its two-dimensional form, with data items located at their nearest map node, similar data ends up in close proximity on the map and dissimilar data is far apart. Through the SOM creation process the prevalent data patterns are identified by the nodes, data items become grouped into clusters around these patterns, and the clusters are ordered by similarity on the map. For a more detailed summary of the SOM method, refer to e.g. Clark et al. (2015).

In this study, the data set is split into two subsets ('baseline' data and 'projected future changes') for each city, allowing a progressive investigation of spatial and temporal patterns of urban flooding. A series of three separate SOMs (also referred to as maps) are created with prevalent global patterns and city similarities established separately on each map through colouring and labels, as follows:

- SOM1 explores the spatial properties of the baseline data set, enabling a comparison of the state of urban river flood impacts in each city at a snapshot in time (2010).
- SOM2 explores patterns of projected temporal changes in impacts of urban flooding on population and property (to 2030), incorporating the drivers of climate change and urban development, and
- SOM3 portrays the temporal relationships between the cities in a type of longitudinal exploratory data analysis, clustering cities that are similar in the baseline situation and are also projected to trend similarly in response to each driver in the future.

SOM1, the baseline map, depicts prevalent global spatial patterns and identifies urban flooding conditions in each city based on two variables: 1) the total population affected annually by river flooding, and 2) annual urban property damages costs incurred by river flooding. The map is created based on these two variables, though by projecting new variables onto the trained map it is also used to show: 3) the percentage of each city's population affected, and 4) the percentage of the country's GDP affected. Usually used with higher-dimensional input data, the SOM method is useful here for creating a map with two variables as the nonlinear projection establishes the relationships between cities in alignment with the directions of maximum variance (ie. the directions of most importance) in the data. It also allows for the results to be used as input into SOM3 later.

SOM2, the future projected changes map, describes the anticipated alterations in urban river flooding in each city by 2030. This map is based on four variables of projected changes and their associated drivers: 1) the projected change in population affected annually, 2) the projected change in annual urban damages costs, 3) the proportion of change in population affected that is anticipated to be attributable to climate change, and 4) the proportion of change in urban damages costs that is anticipated to be attributable to climate change. The remainder of the increase or decrease in impacts is attributed to socioeconomic causes (such as population change, urban density change, increased city footprint, and changes in urban land cover).

SOM3, the temporal map, uses the location of each city along the axes of the two-dimensional baseline and future projected changes maps (which essentially delineate the first two principle curves in each higher dimensional data subset) as input data. In creating SOM1 and SOM2, the baseline and future data subsets have already been reduced to their two most prominent dimensions respectively (which have become the axes of these maps), and each of these four dimensions is considered equally when placing the cities on the temporal map. This method is based on the method used in Clark et al. (2015) to investigate individual data items transitioning through a self-organizing-time-map, and has been modified for the comparison of patterns on two-dimensional maps of differing sizes and shapes that have been created separately based on different variables.

Distinct patterns that have emerged through the process of training the three maps are represented by the nodes of SOM3. These patterns are the most relevant combinations of dynamic city flood impacts, socioeconomic, and climate change characteristics in the overall data set. SOM3 is clustered, coloured and labelled to indicate the relationships between the cities in terms of similar or differing baseline situations *and* projected changes. Cities with relatively close locations on both the baseline and future projected changes maps are considered to have parallel temporal paths, and will be found close together on the temporal map. Those with converging trends (dissimilar baseline conditions, but similar future projected changes) and diverging trends (close baseline conditions, but dissimilar future projected changes) are also identifiable on this map.

In the creation of each map, grid size and shape have been determined using quantization, topographic and dimension range representation error measures (QE, TE, and DRR) with comparisons between the data set and the map.

The QE (Kohonen, 2001) measures how well the map nodes represent the data items using the sum of squared Euclidean distances between each data item, x_i, and the node closest to it, m_c, averaged over all data points:

$$QE = \frac{1}{N}\sum_i \|\mathrm{m_c} - \mathrm{x_i}\| = \frac{1}{N}\sum_i \sqrt{(m_c{}^2 + x_i{}^2 - 2m_c x_i)}\,.$$

The TE (Kiviluoto, 1996) indicates how well the topography of the data set is preserved on the map, giving higher error values for maps that are unnecessarily bent or twisted. The BMU and second BMU for each data point are checked to determine if they are adjacent ($u_{x_i} = 1$ if the first and second BMUs of x_i are neighbours, 0 otherwise), and TE is calculated as:

$$TE = \frac{1}{N}\sum_{i=1}^{N} u_{x_i}$$

The DRR (Clark et al., 2015) measures how well the map represents each variable of the data set to ensure even coverage of the dimensions. The maximum intra-cluster spread of data items in each dimension, d, that become represented by a single map node, x_i (as a proportion of the overall data range in that dimension) is determined. The DRR is calculated as follows, where $x_i(d)$ are data values in dimension d, and $x_{ij}(d)$ are the data values in dimension d that are assigned to map unit j:

$$DRR(d) = \max_j \frac{\max_{ij}(x_{ij}(d)) - \min_{ij}(x_{ij}(d))}{\max_i(x_i(d)) - \min_i(x_i(d))}$$

For the baseline map, a 10*7 grid is found to be the optimum shape to represent the data based on the error measures. An 8*8 map is fitted to the future projected changes data set. After finding these optimum side ratios, the maps are increased in size preserving their side ratios (to 20*14 and 18*18) to allow the data items to spread out until most cities are placed individually, allowing the relationships between all cities to become evident (as in Skupin & Hagelman, 2005). The temporal map is sized at 25*17 nodes. Whilst the input data for the baseline and future projected changes maps were standardized into the range 0-1 before training, the input data for the temporal map is not standardised in order to preserve the ratios between the lengths of the first two principal curves in each of the first two data subsets.

Prevalent cluster characteristics are determined using a 'second level' clustering of the nodes of the SOM (as in Vesanto & Alhoniemi, 2000; Skupin & Hagelman, 2005), performed using Ward's clustering method (Ward, 1963) with the number of clusters determined using the Davies-Bouldin index (Davies & Bouldin, 1979). The Davies-Bouldin index reports the ratio of within cluster scatter (S_j for cluster j) to inter-cluster distances, looking at each cluster and its most similar one, (M_{jk}), with a lower ratio ($^S/_M$) indicating a better estimate of the number of clusters of interest present in the data. Ward's minimum variance method is a hierarchical clustering algorithm based on minimizing the total within-cluster variance. With this second-level clustering, each data item of the original data set becomes a member of the same final cluster as its closest node (Vesanto & Alhoniemi, 2000).

The final clustering is visually verified with a SOM 'U-matrix' (Ultsch, 2003). The U-matrix visualises distances in data space between immediately neighbouring nodes, indicating these distances by colour on a grid of the same size as the SOM. By computing how close adjacent map nodes are in data space, the U-matrix is able to provide an indication of cluster boundaries based on large dissimilarities between neighbouring nodes. A greater change in relative distance between the locations of the nodes in data space than in map space is displayed in a lighter colour on the grid, and lesser distances in darker shades. The darker regions of the grid then indicate the cluster centres, separated by lighter coloured boundary areas.

By reducing the information from this multivariate data set into the two most prominent dimensions and finding relationships between the data items at each of these three stages, spatial and temporal information about global patterns of urban flooding is abstracted, and similarities and differences between the cities are clearly portrayed. This method extracts two levels of information:

(1) the most characteristic socio-environmental patterns in the data are found, and
(2) cities are compared to each other with respect to their relative flooding conditions.

The simulations are run in MATLAB with use of the SOM Toolbox (website 7) with variables and map sizes as described above.

7.4 RESULTS

Three SOMs are presented sequentially to reveal three unique sets of patterns in the data, where the term ‘patterns’ refers to combinations of variables that characterise a specific set of conditions. The cities are clustered into groups with conditions matching these patterns, based only on the given data. The maps each have different sizes, shapes and colours as they represent different subsets of input data.

7.4.1 SOM1: Baseline urban flood impacts

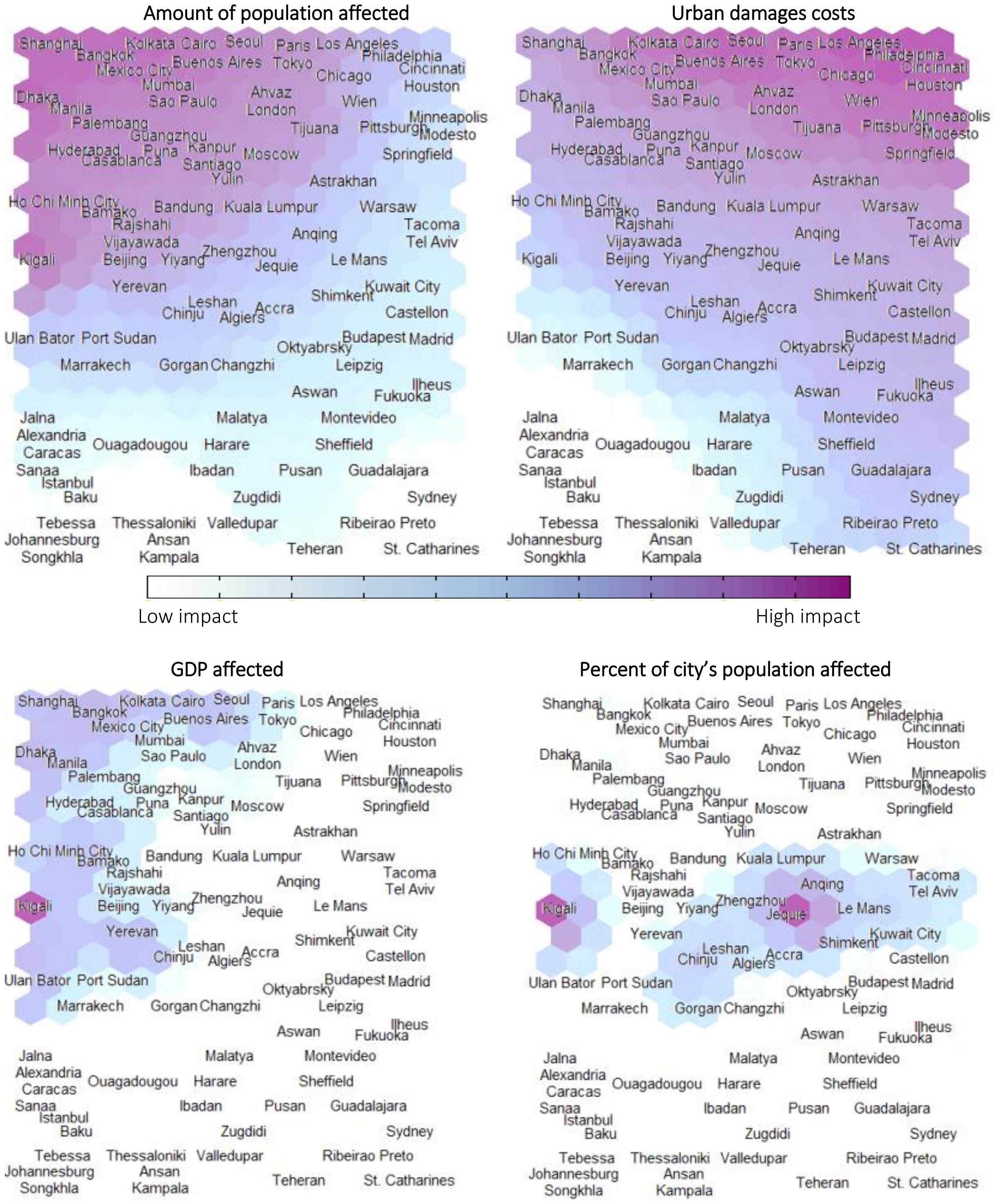
Patterns of urban flood conditions in 2010 are shown on the baseline map, SOM1, in Figure 1. The placement of city labels indicates the relationship of each city to each other in terms of river flood impacts on population and urban damages costs. The map is created by organizing the cities with respect to each other based on both of these factors. Cities close together are more similar in the amount of population affected and urban damages costs, and cities located far apart are less similar.

The relative placement of the cities on the map is the main map characteristic providing insight into the features of the data, indicating differences in a *combination* of the variables which can be discerned from the colouring of Figure 30(a). Each map node has a four-component vector (representing the value of each of the four variables at the location of the node in data space). The four images in Figure 30(a) show SOM1’s city labels over grids coloured separately by the values of each of the four variables (white is low, purple is high). For each city, the relative value of each of the variables can be seen. For example, Cincinnati (top right) incurs high material damages costs, and medium population affected, whereas Ulan Bator (mid left) has similar population affected to Cincinnati, but much lower material damages costs.

The nonlinearity of the relationships between the variables is evident, as is the smooth transition of the values of each variable along the map. General information about the prevalent baseline global patterns and the relative flood conditions in the specific cities can be gained from inspection of these map labels and coloured grids.

Each area of the grid represents a general pattern, or combination of variables in the data, some of which are indicated by annotations on Figure 30(b). In general, higher amounts of population affected and urban damages costs resulting from river flooding are represented by areas towards the top of the map, and these variables decrease in value down the map. Values of affected population are lowest just in from the lower left corner and undulate along the bottom of the map, sweeping upwards to a maximum at the upper left corner. Urban damage values are lowest in the lower left corner and increase in concentric arcs up to the upper right corner. Generally, the left of the map contains patterns involving higher impacts on populations than on property, and the right of the map higher impacts on property than on populations.

a)


Amount of population affected
Urban damages costs
GDP affected
Percent of city's population affected
Low impact
High impact
Shanghai
Bangkok
Kolkata
Cairo
Seoul
Paris
Los Angeles
Philadelphia
Mexico City
Buenos Aires
Tokyo
Chicago
Cincinnati
Mumbai
Houston
Dhaka
Manila
Sao Paulo
Ahvaz
London
Wien
Palembang
Minneapolis
Guangzhou
Tijuana
Pittsburgh
Modesto
Hyderabad
Casablanca
Puna
Kanpur
Santiago
Moscow
Springfield
Yulin
Astrakhan
Ho Chi Minh City
Bamako
Bandung
Kuala Lumpur
Warsaw
Rajshahi
Tacoma
Anqing
Vijayawada
Tel Aviv
Kigali
Beijing
Yiyang
Zhengzhou
Jequie
Le Mans
Yerevan
Kuwait City
Shimkent
Leshan
Accra
Chinju
Algiers
Castellon
Ulan Bator
Port Sudan
Budapest
Madrid
Oktyabrsky
Marrakech
Gorgan
Changzhi
Leipzig
Ilheus
Aswan
Fukuoka
Jalna
Malatya
Montevideo
Alexandria
Caracas
Ouagadougou
Harare
Sheffield
Sanaa
Ibadan
Pusan
Guadalajara
Istanbul
Baku
Zugdidi
Sydney
Tebessa
Thessaloniki
Valledupar
Ribeirao Preto
Johannesburg
Ansan
Kampala
Songkhla
Teheran
St. Catharines

b)

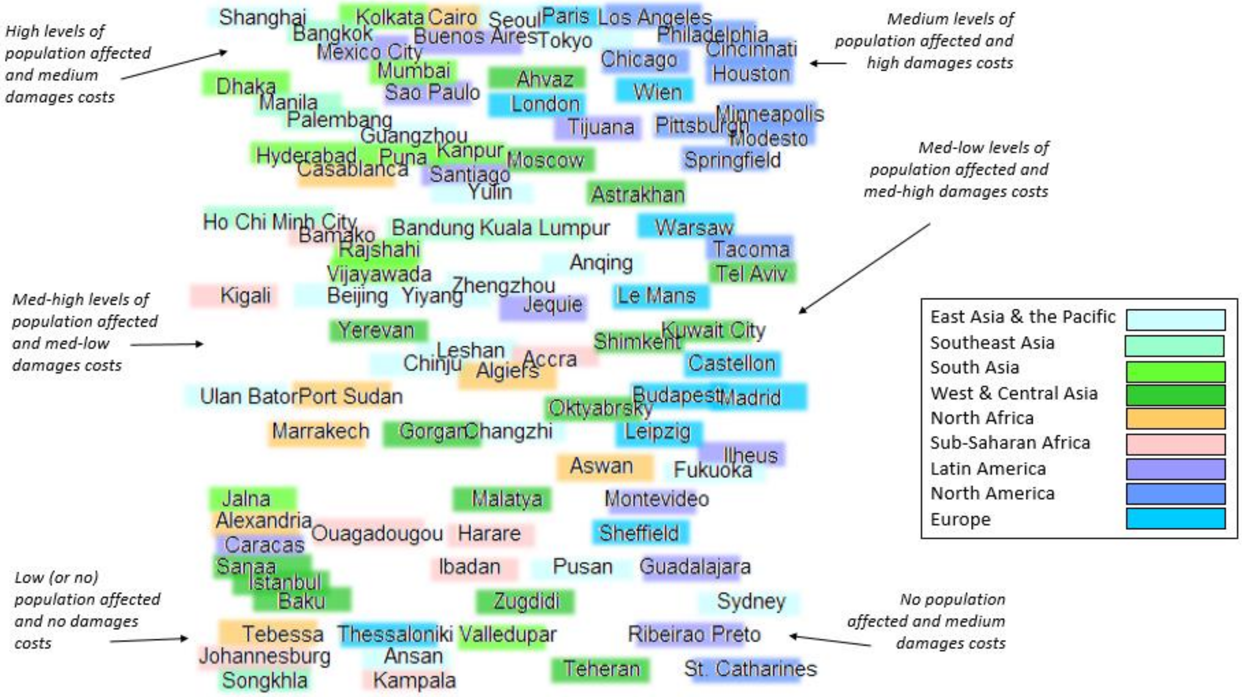


Figure 30: SOM1 - Baseline (2010) urban flood conditions. Cities are placed relative to each other based on annual river flooding impacts on population and urban damages costs. a) The same map is repeated for each of four variables, with colouring indicating low (white) and high (purple) values. b) The city labels are coloured by region (see Table 1), and characteristic patterns of general areas of the map are annotated. The reader may refer to the online version to zoom in on text if required.

From Figure 30(b), relationships can be discerned between regions, as well as between cities in the same region. For instance, cities in North Africa, Sub-Saharan Africa and West & Central Asia are predominantly located in the lower portion of the map, corresponding to a prevalent pattern of low flood impacts on both population and property. Cities in Southeast and South Asia generally correspond to the patterns of high impacts on population and property found in the upper left of the map. Cities in Europe stretch from the top to the bottom of the map, ranging from high overall flood effects (Paris) to no flood effects at all (Thessaloniki). North American cities are matched to patterns that represent more significant impacts on property than on population (down the right side of the map), and are split between those with high property damages (Philadelphia, LA, etc. – in the top right) and those with low damages (St. Catherine's – in the bottom right).

Impacts on GDP and the proportion of the cities' populations affected are shown in the two lower maps of Figure 30(a), though these variables were not used to position the cities on the map. Cities in which river-related urban flooding is estimated to highly affect the country's GDP are coloured on the lower left map. Kigali, in particular, which incurs medium-high flood impacts, sees a large impact on Rwanda's GDP, perhaps because Kigali is the main city in this relatively small country (Kreimer et al., 2003). GDP is most affected by flooding in: Kigali, Bangkok, Yerevan, Dhaka, Bamako and Cairo. Cities in which the flood-affected population forms a significant proportion of the city's population are coloured on the lower right map, predominantly in a horizontal strip across the centre. The highest proportions are in: Jequie (15%), Kigali (7%), Chinju (6%), Le Mans (5%) and Tacoma (3%).

7.4.2 SOM2: Projected changes in urban flood impacts (to 2030)

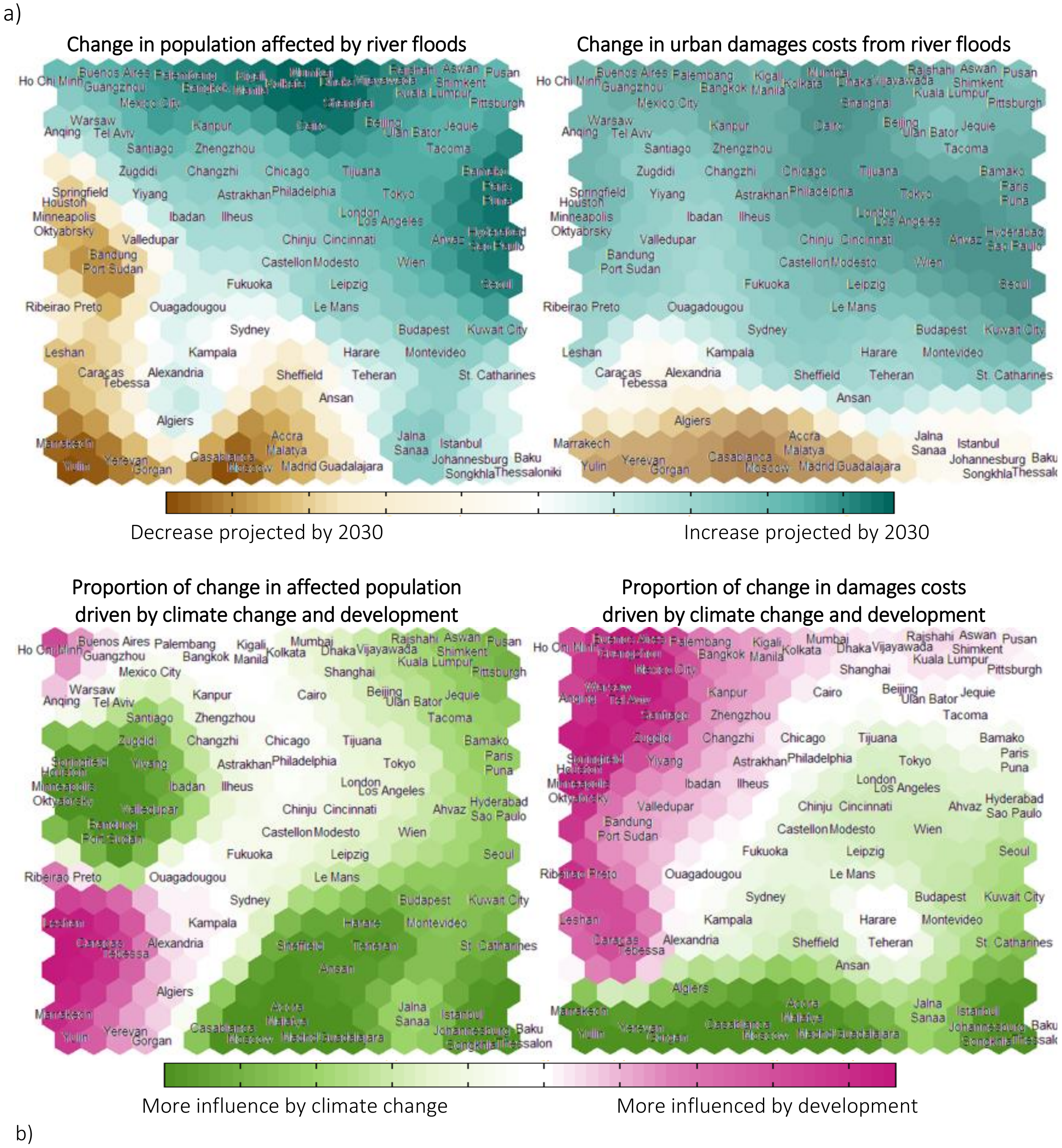
SOM2 identifies the projected patterns of evolving river flood conditions in the cities (between 2010 and 2030), based on city-specific projections of increasing or decreasing flood impacts on population and damages costs, and whether these changes are anticipated to be driven more by climate change or development (Figure 31).

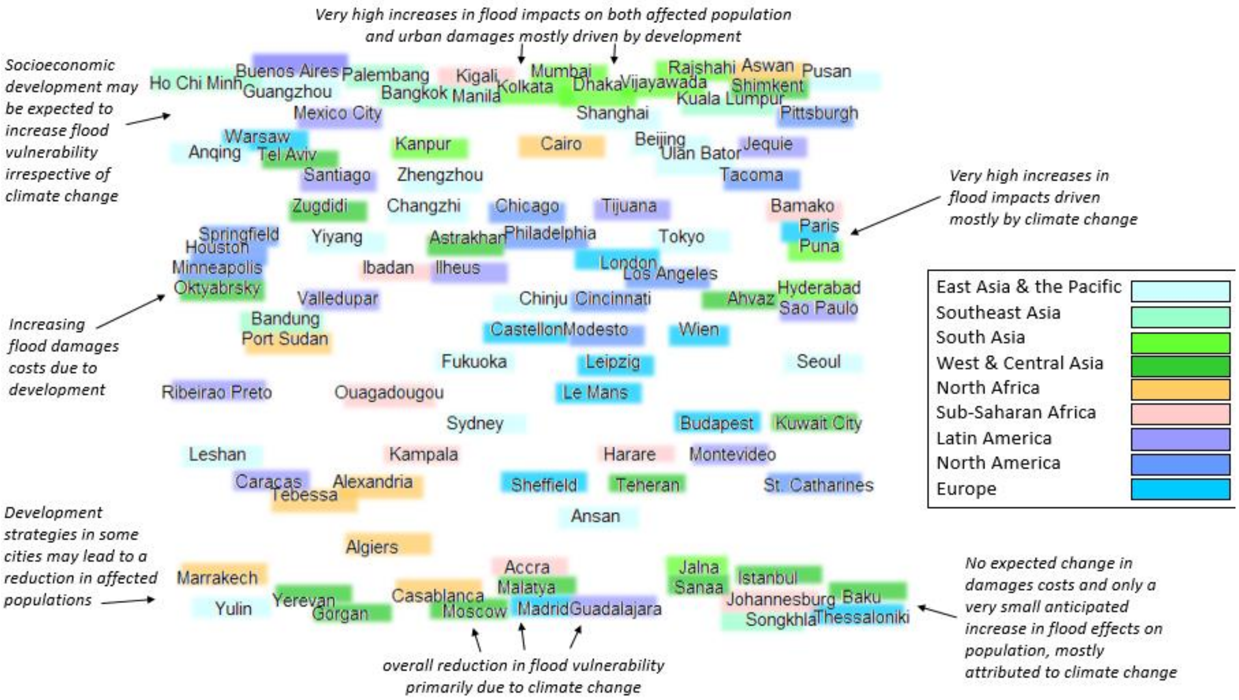
In Figure 31(a), regions of the map representing projected increases in flood impacts on either populations or damages costs are coloured blue and reductions in flood impacts are coloured brown (in the top row), with white indicating no projected change. Projected changes primarily driven by socioeconomic development are coloured purple (in the lower row), and green indicates that the primary driver is climate change. White represents a mid-point in which both climate change and development are predicted impact future flood conditions relatively equally. Areas of the map representing patterns of increased flood impacts predominantly due to climate change or development can be located on Figure 31(b).

Investigating SOM2, we see that climate change is projected to be predominantly responsible for increases in population vulnerability in all cities besides those in the top left corner (around Ho Chi Minh City). Climate change is anticipated to decrease flood damages costs in cities located at the bottom of the map (around Madrid), and decrease impacts on populations in cities in the mid-left (around Minneapolis) and mid-lower (again around Madrid) portions of the map. Socioeconomic development is projected to be the main driver increasing flood damages costs in cities on the upper-left triangle of the map (roughly from Mumbai down to Tebessa). Only in Ho Chi Minh City is development anticipated to be almost completely responsible for all increases in river flood impacts, all other cities in this study are at least partially affected by climate change. Development is not projected to play any part in a decrease in flood damages costs in any cities in this study (Caracas and Tebessa have no change in damages costs on the upper map, though it is attributed to development on the lower map).

Geographic regions are shown on Figure 31(b) with coloured text backgrounds. Cities in Southeast Asia are almost all found at the top of the map indicating high projected increases in overall flood impacts. South Asian cities are mostly located in the two areas of the map with patterns of very high increases in flood impacts, split between those most affected by development (around Mumbai, top middle) and those most affected by climate change (around Puna, mid right). Many North African cities are located in the lower left, indicating anticipated reductions in flooding due to socioeconomic development. North American cities are spread across the middle of the map indicating a wide range of projected changes.

Climate change and development may lead to opposing changes in a city’s flood impacts on population and property. A number of cities are predicted to have affected populations decreasing due to climate change, whilst damages costs increase due to socioeconomic factors (around Springfield and Port Sudan, in the mid-left). A decrease in flood effects on urban damages due to climate change, but an increase in affected population largely due to development is, out of the cities in this study, only projected for Algiers (in the lower left portion of the map).





Very high increases in flood impacts on both affected population and urban damages mostly driven by development

Socioeconomic development may be expected to increase flood vulnerability irrespective of climate change

Very high increases in flood impacts driven mostly by climate change

Increasing flood damages costs due to development

Development strategies in some cities may lead to a reduction in affected populations

No expected change in damages costs and only a very small anticipated increase in flood effects on population, mostly attributed to climate change

overall reduction in flood vulnerability primarily due to climate change

Figure 31: SOM2 - PROJECTED CHANGES IN RIVER FLOOD IMPACTS WITH ASSOCIATED DRIVERS. River flooding in individual cities will be affected separately by climate change and development between 2010 and 2030. Cities that are anticipated to experience similar pressures and responses in terms of river flooding impacts are located nearby on the map. a) City labels are placed over coloured copies of the map showing the relative values of each variable. b) City labels are coloured by region, and characteristic patterns of general areas of the map are annotated. The reader may refer to the online version to zoom in on text if required.

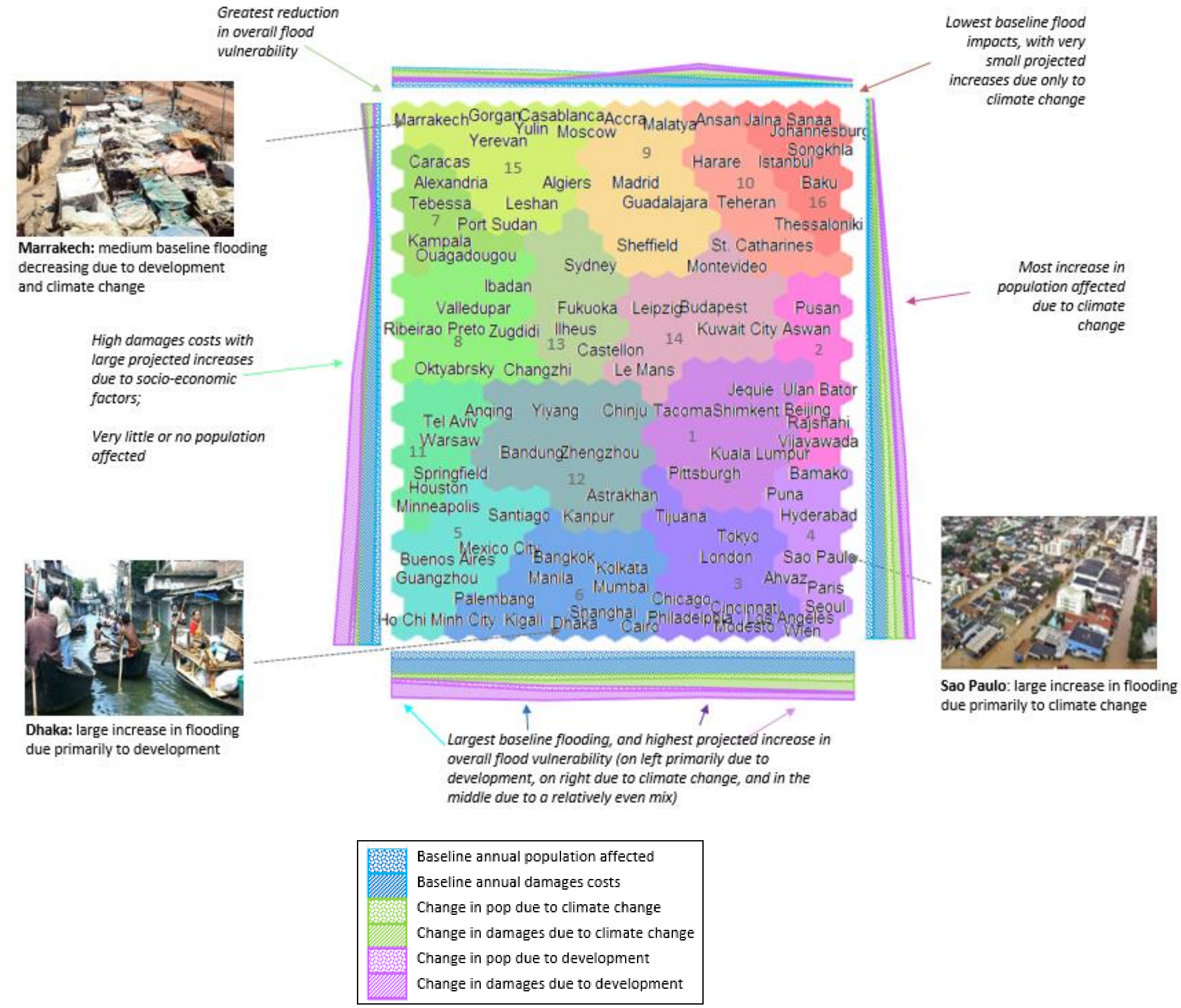
In some cities, both drivers may generate changes in the same direction. For instance, in Marrakech, Yulin, Yerevan and Gorgan, climate change is projected to be responsible for a decrease in damages costs whilst socioeconomic development is anticipated to play a major role in the decrease in population affected, suggesting that the reduction of population vulnerability due to development is complementing the direction of change instigated by climate change. In certain cities near the upper left of the map (Santiago, Zugdidi and Yiyang), an overall increase in flood impacts is expected, with increases in affected population almost completely attributed to climate change and increases in damages costs almost completely attributed to development.

7.4.3 SOM3: Temporal patterns

Relationships between the baseline characteristics and projected future changes of urban flooding in the individual cities are shown in Figure 30 and Figure 31 respectively, however potentially similar temporal patterns between the cities are not evident from these maps. To link the information abstracted from the first two maps, we create a temporal map, SOM3, shown in Figure 32. SOM3 identifies which cities experience similar baseline flooding, are expected to incur comparable future hydrologic pressures from climate change and/or development, and are projected to respond in similar ways (or which cities may diverge in the future from similar baseline conditions).

Following the creation of SOM3 and the positioning of cities with respect to each other, we perform a second level clustering to colour the nodes, giving a visual separation to groups of more similar data. Clusters are numbered from 1 to 16 for reference. As the cities are placed on the temporal SOM based on their locations on the baseline and future projected changes SOMs (in which the values of the variables vary smoothly though not monotonically along the axes), again the characteristics of the cities will flow smoothly along the map though multiple peaks and troughs of each variable are possible. The gradients of the cluster characteristics are indicated along the axes in Figure 32(a), which are nonlinear in data space.

a)



Greatest reduction in overall flood vulnerability

Lowest baseline flood impacts, with very small projected increases due only to climate change

Marrakech: medium baseline flooding decreasing due to development and climate change

Most increase in population affected due to climate change

High damages costs with large projected increases due to socio-economic factors;

Very little or no population affected

Dhaka: large increase in flooding due primarily to development

Sao Paulo: large increase in flooding due primarily to climate change

Largest baseline flooding, and highest projected increase in overall flood vulnerability (on left primarily due to development, on right due to climate change, and in the middle due to a relatively even mix) b)

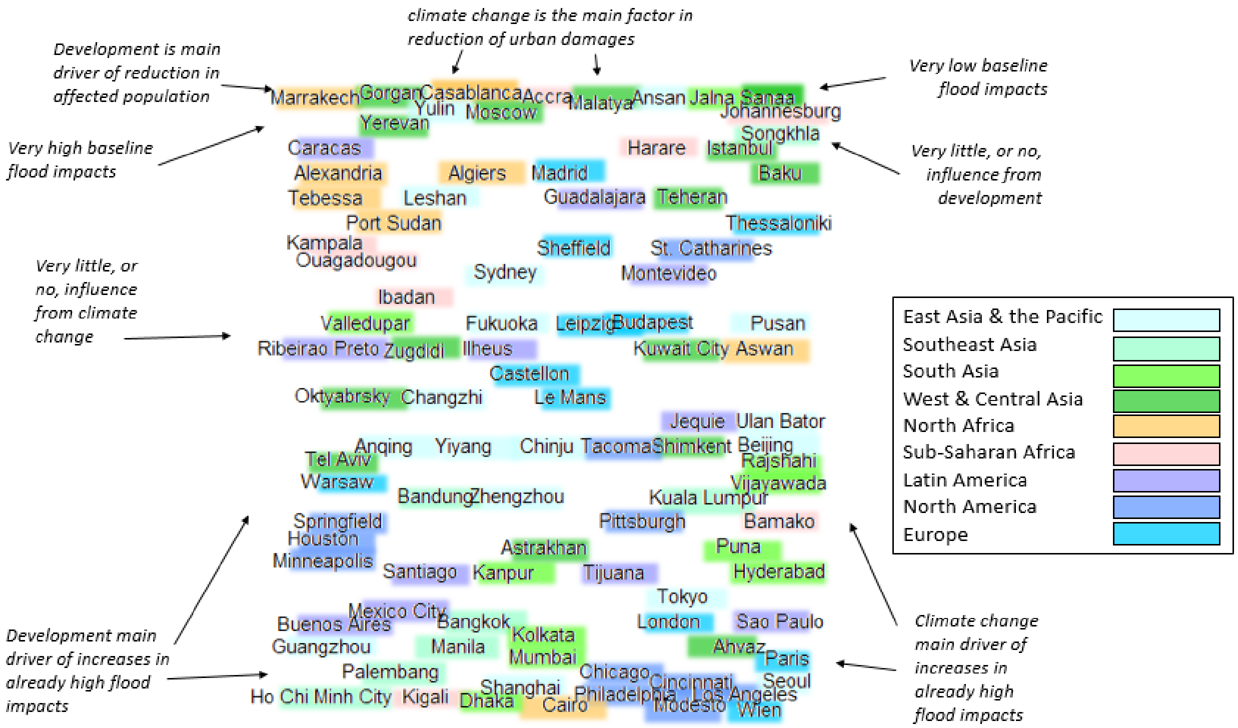

climate change is the main factor in reduction of urban damages
Development is main driver of reduction in affected population
Very low baseline flood impacts
Very high baseline flood impacts
Very little, or no, influence from development
Very little, or no, influence from climate change
Development main driver of increases in already high flood impacts
Climate change main driver of increases in already high flood impacts
Marrakech Gorgan Casablanca Accra Malatya Ansan Jalna Sanaa
Yulin Moscow Johannesburg
Yerevan
Songkhla
Caracas Harare Istanbul
Alexandria Algiers Madrid Baku
Tebessa Leshan Guadalajara Teheran
Port Sudan Thessaloniki
Kampala Sheffield St. Catharines
Ouagadougou
Sydney Montevideo
Ibadan
Valledupar Fukuoka Leipzig Budapest Pusan
Ribeirao Preto Zugdidi Ilheus Kuwait City Aswan
Castellon
Oktyabrsky Changzhi Le Mans
Jequie Ulan Bator
Anqing Yiyang Chinju Tacoma Shimkent Beijing
Tel Aviv Rajshahi
Warsaw Vijayawada
Bandung Zhengzhou Kuala Lumpur
Springfield Pittsburgh Bamako
Houston
Astrakhan Puna
Minneapolis
Santiago Kanpur Tijuana Hyderabad
Tokyo
Mexico City
Buenos Aires Bangkok London Sao Paulo
Kolkata
Guangzhou Manila Ahvaz
Mumbai Paris
Palembang Chicago Seoul
Cincinnati
Shanghai Los Angeles
Ho Chi Minh City Kigali Dhaka Philadelphia
Cairo Modesto Wien
East Asia & the Pacific
Southeast Asia
South Asia
West & Central Asia
North Africa
Sub-Saharan Africa
Latin America
North America
Europe

Figure 32: SOM3 - Temporal patterns. Cities are clustered close together that share similar baseline (2010) flood vulnerabilities as well as similar anticipated changes driven by climate change and development on population and urban damages costs by 2030. a) Locations of the cities are based on their individual relationships to the principal curves in the baseline and future projected changes data subsets - therefore, the axes represent the most important nonlinear gradients of flood vulnerabilities in the data set. Coloured bars along the axes indicate the average levels of each variable around the edges of the map. Cities are grouped into coloured clusters based on similarities. b) City labels are coloured by region, and characteristic patterns of general areas of the map are annotated. The reader may refer to the online version to zoom in on text if required.

Broad overviews of the patterns represented by certain regions of the map are identified on Figure 32(b) with arrows. The largest increases in flood effects are generally represented by nodes in the lower half of the map, whilst the largest decreases in flood effects are represented by nodes in the top left. Climate change is predicted to be the main driver of changes in population vulnerability along the top and down the left and right sides of the map, and in urban damages on the top and right of the map; therefore, climate change is the leading driver of changes in flood impacts on both population and damages costs at the top of the map. Development is the main driver of changes in flood impacts on populations in the lower and upper left side of the map, and on urban damages in the lower left area of the map; therefore, development is the leading driver of changes in flood impacts on both population and damages costs in cities on the lower left side of the map.

On Figure 32(b) the city labels are coloured by geographic region. We see the cities of each geographical region are more spread out on SOM3 than on SOM1 where each region was generally contained in one or two broad areas of the map. For example, on SOM3 Cairo and Aswan are noticeably separated from other North African cities which are located close together. Although the cities of this region have differing baseline flood levels (as shown on SOM1), most are projected to incur some reduction in future flood impacts (as shown on SOM2), with the exception of Cairo and Aswan. These cities both have forecasts of increased flood impacts - for Aswan increased impacts on the population due to climate change and impacts on property due to development, and for Cairo future impacts are projected to increase due to a relatively even mixture of both drivers. For another example, cities in the USA (all of which have similar starting conditions) are in two well-separated clusters on SOM3 - those around Houston and those around Los Angeles. The cities clustered around Houston are characterised by low impacts on population but high damages costs projected to elevate due to development, implying the possibility for local redemption due to better planning or mitigation strategies. The cities clustered around Los Angeles, however, are characterised by high overall impacts projected to get higher predominantly due to climate change. Further, in Sub-Saharan Africa we see Kigali and Bamako (which have similar medium-high baseline flooding conditions) are both expected to see increased impacts, but the cities are separated by SOM3 as these flood increases are attributed to development in Kigali and climate change in Bamako.

To further analyse the characteristics of each cluster and the patterns found on SOM3, the properties of each city in the 16 clusters are shown in a radial plot in Figure 33. Baseline values of population affected (blue, units = number of people) and damages (orange, units = $US) are shown on a symmetrical logarithmic scale ranging from -8 (ie. signifying a value of -100,000,000) to 11 (100,000,000,000) with the region between -1 and 1 on the plot set as linear to avoid logarithmic discontinuities in the vicinity of zero. Zero is indicated by a dashed circumference, and each progressive ring is an exponentially higher (or lower) value. Changes in population affected and damages costs are shown on the same scale, in grey and yellow respectively. Values inside the dashed (zero) circle represent decreases in flood impacts, and values outside represent increases, with the size of the increase or decrease indicated by the distance from the dashed circle. The influence of climate change is shown (light green for population and dark green for damages) on a linear scale from the same zero circumference, in units of 'percentage of projected change attributable to climate change' (each progressive ring is 10%). Green lines closer to the outer ring than the centre therefore indicate that the flood impacts on the city are anticipated to be more influenced by climate change than by development. If the green lines are both in the middle of the segment, this indicates a relatively equal influence of both drivers on both population and property. Diverging green lines indicate that either population or damages costs are more influenced by climate change, and the other by development.

From Figure 33, we can see the differences between neighbouring clusters, such as 10 and 16 located in the top right of the map. Both clusters are characterized by low baseline impacts of flooding on the population, with small increases in population impacts projected primarily due to climate change. However, cities in cluster 16 incur no flood damages costs at all in the baseline or future cases, yet in cluster 10 damages costs are projected to increase due to climate change and development. Therefore, development has little or no impact on cities in cluster 16 but does play a role in the increase in damages in cluster 10. In the top left of SOM3, we can now also discern the difference between clusters 9 and 15. In both clusters, development is projected to have no impact on the reduction of flood damages costs in most cities. Development does however play a strong role in the reduction of flood impacts on populations in cluster 15 (except for Moscow and Casablanca) but none on populations in cluster 9.

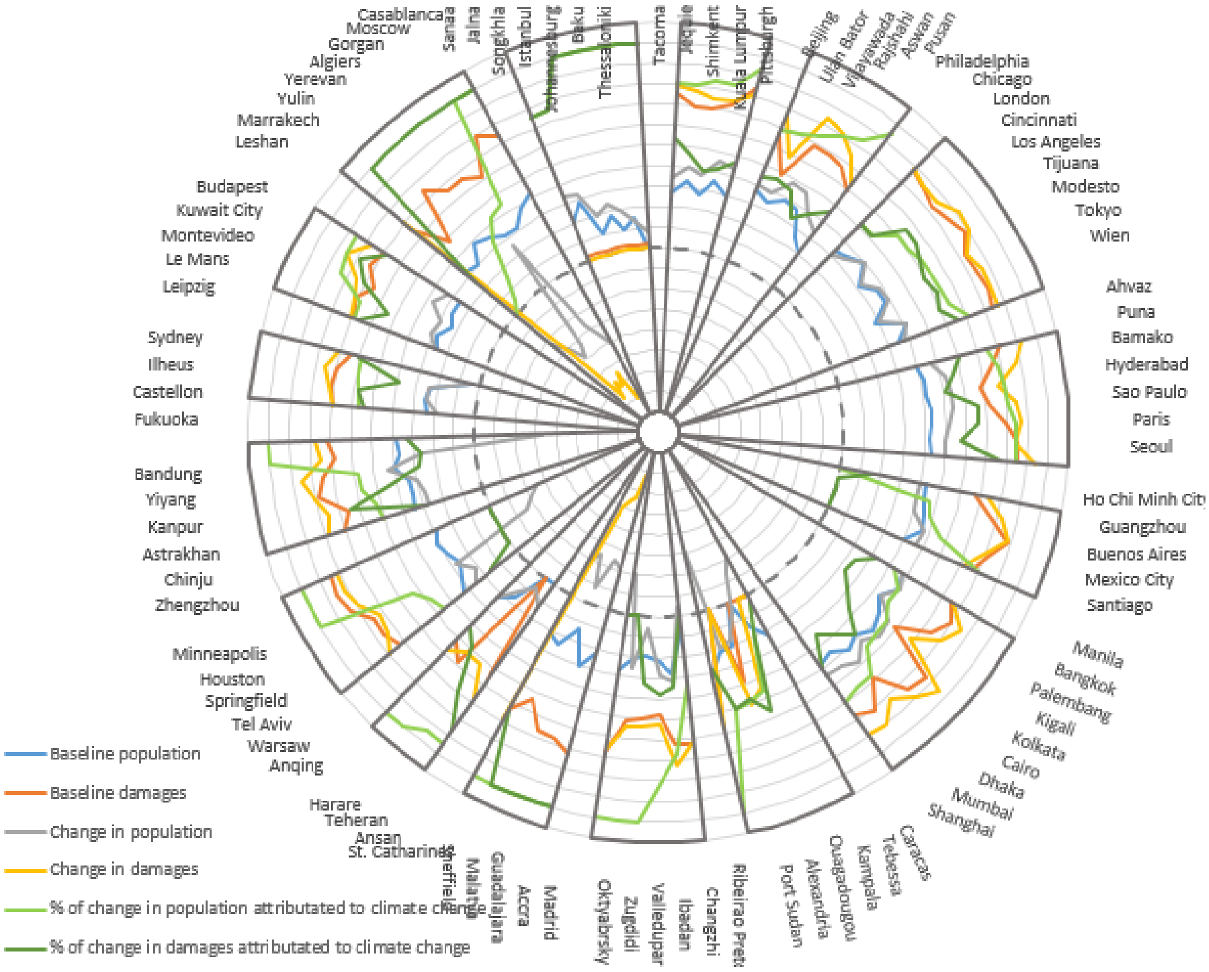


Figure 33: Radial plot of clusters of Figure 32 – The city members of the 16 clusters of Figure 32 are shown with their individual variable values. The scale is logarithmic for baseline and changes in population and damages, and linear for the percent of change attributed to climate change, with the dashed circle representing zero.

The relationship between the two drivers, climate change and development, can discerned from Figure 32 and Figure 33. Climate change is projected to impact populations more than urban damages costs in clusters stretched across the centre of the map (clusters 8, 11, 5, 6, 12, 14, 1, 10, 1, 2, and 4 - in cities in these clusters, the proportion of change in the population affected attributed to climate change is higher than the proportion of change in damages costs attributed to climate change). In cluster 15, the population is projected to be more influenced by development than damages costs will be (a higher proportion of the change in damages costs is attributed to climate change than for population). In the remaining clusters, climate change (and development) are projected to affect the population and damages costs relatively similarly (clusters 7, 13, 9, 16 and 3). Some examples of diverging impacts on population and damages costs stand out on the radial plot in Figure 33. For instance, in Port Sudan, Sheffield and Bandung, significant reductions in affected population are projected to be 100% due to climate change, however large projected increases (~300 to 400%) in damages are due mostly to development. In Leshan, development is projected to slightly lower the amount of affected population and also to increase damages costs more than three-fold.

7.5 Discussion

In this study, the 'patterns' and 'clusters' in the data have been identified. The patterns, depicting key combinations of variables that are characteristic of the data set, have been extracted at three separate levels on SOM1, SOM2 and SOM3. For example, each pattern of SOM3 is a separate combination of levels of baseline and projected future flood conditions as well as projected influences of climate change and development. The clusters consist of groups of cities whose conditions are anticipated to be similar to these patterns, based on the given data. A discussion of a selection of these patterns and clusters is provided here.

Some cities already experiencing large flood effects are anticipated to incur great flood increases influenced predominantly by socioeconomic factors (migration, changing land use and unplanned development in flood zones). In the lower left region of SOM3, we see examples of cities in which climate change is playing a large role, and yet it is overshadowed by the magnitude of regional economic growth (UNEP, 2016; website 8). Many of these cities are in Asia, where the climate is experiencing warming trends, increasing temperature and precipitation extremes, and rapid glacial melting resulting from climate change (Pachauri et al., (2014) chapter 24: 'Asia'). However, socioeconomic growth in this area is projected to have even more of an impact on urban floods than climate change is. Flood risk and human and material losses are already heavily concentrated in India, Bangladesh and China (Pachauri et al., (2014), chapter 24: 'Asia'), and Jongman (2012) estimates the largest current and future economic exposure to river floods to be in Asia. As an example, we take a closer look at Dhaka which, with a GDP per capita of $1212 in 2015, already has one of the highest levels of population affected annually by flooding (over 130,000) and this number is projected to increase almost five-fold to over 630,000 by 2030. The greatest change predicted for Dhaka, though, is an almost 22-fold increase in annual damage costs (from $8 million to $175 million). Dhaka is subjected to regular flooding from surrounding rivers, with peak flows in the Brahmaputra and Ganges Rivers coinciding to exacerbate flood impacts. In the past, most low-lying areas of western Dhaka were infilled for residential and commercial use, causing a reduction in areas for flood water storage. Furthermore, uncontrolled and unplanned urban expansion is spreading rapidly across the floodplains in the east of the city placing more people in flood hazard zones (Kreimer et al., 2003). These hasty developmental changes are having more of an impact on the urban hydrology of Dhaka than the climate change is. Other examples of cities in similar situations include Kolkata (with the highest baseline affected population in this study), Mumbai (with a seven-fold increase in both population affected and damages due 40% and 60%, respectively, to development), Bangkok (with large increases 50-75% of which are attributed to development) and Ho Chi Minh City (with a 50% increase in affected population and an over five-fold increase in damages costs, almost entirely attributed to development).

Globally, migration trends are seeing more people moving into informal settlements in urban flood zones – the population exposed to river flooding increased by 2.6% more than total global population growth between 1970 and 2010 (Jongman et al., 2012). Most global population growth in the near future is projected to occur in cities of lower income countries, organically and through migration (Kreimer et al., 2003), with urban populations in these countries growing at a rate five times faster than in higher income countries (UN-DESA, 2015) and predicted to double in the next 30 years (Angel et al., 2010). The same regions experiencing such high urban population growth are also projected to triple their urban footprint in the same timeframe (Angel et al., 2010). These developmental changes are leading to, and will continue to produce, substantial effects on urban hydrology if not countered.

Developmental changes in some cities, however, appear to be effectively reducing impacts from river flooding. Marrakech, in cluster 15, is an example of this. The affected population level is projected to decrease mostly due to socioeconomic factors (website 3; Ward et al., 2013; Winsemius et al., 2013). Morocco is taking responsibility to make efforts countering global climate change, and through an 'Integrated Disaster Risk Management and Resilience Program for Morocco' (World Bank, April 2016-Dec 2021), is making its population more resilient to climate change, less vulnerable to natural hazards and ensuring a rapid transition to a low-carbon economy. Through Morocco's National Strategy for Sustainable Development, a commitment has been made to reduce national greenhouse gas emissions by 32% by 2030, through an increase in renewable energy sources to 50% by 2025, a reduction in energy consumption by 15% by 2030, as well as various agricultural, water, waste, forest, industry and housing initiatives (website 9). These housing initiatives in Marrakech include a slum clearance and relocation project, which has become part of urban policy (Ibrahim, 2016), reducing the amount of people inhabiting flood hazard zones. Alert systems in the valleys of the Atlas region above Marrakech have been improved, and the proportion of the population living in slums has decreased from over 8% in 2004 to less than 4% in 2010 (UN-Habitat website). The urbanization rate in Morocco is also projected to slow down towards 2030 (UN-Habitat website). This risk-prevention approach combining early warning systems, relocation of inhabitants out of the flood zone, and less urban expansion is expected to combine to reduce the impact of floods on the population of Marrakesh.

The analysis in this paper is based solely on the data provided in Aqueduct, regardless of the extent to which on-the-ground flood management measures are incorporated into the socioeconomic models which produced this data. A discussion characterizing individual cities is included here as a point of interest to relate the data to current national conditions, providing possible reasons why these cities may fit into the map where they do.

Current high flood impact conditions projected to get much greater primarily due to climate change are anticipated for cities in the lower right of SOM3, with high magnitude changes expected for impacts on both population and property. One of these cities, Sao Paulo, for instance, is expected to experience an almost seven-fold increase in both the number of population affected (to over 140,000 annually) and urban damages costs (to over $500,000,000 annually) by 2030. 15% of the change in population and 35% of the change in damages is attributed to development, but the majority of the change is projected to be caused by climate change. Sao Paulo is the largest city in Brazil, and the city footprint is projected to increase over 38% by 2030, by which time 22% of the urban area may be located in flood zones (Young, 2013). The IPCC (Pachauri et al., (2014) chapter 14 'Latin America') predicts the increase in temperature in central and south Brazil to be the largest projected increase in Latin America, which will be combined with a +10 to +15% increase in autumn precipitation, greatly affecting the hydrologic cycle in the region. The substantial change in development is therefore expected to be eclipsed by the even greater projected change in climate in Sao Paulo.

The anticipated reduction in flood damage costs caused by climate change (evident in Cluster 15) may be a result of changing snow melt conditions upstream of these cities. It has been shown that some global regions will experience a decreasing trend in the magnitude and frequency of snow melt floods as the climate warms, as well as a shift in the timing of these floods (Schiermeier, 2011; Barnett et al., 2005; Immerzeel et al., 2010). Although changing climate in some areas is projected to lessen regional flooding, development within urban flood zones may be severe enough to offset any reductions in flood impacts. This can be seen most prominently in a strip on the left of SOM3 stretching from Port Sudan down to Santiago.

Many high-income cities with already high current flood vulnerabilities have projections for large elevations in damage costs, but not increased levels of affected population. This can be seen in cities on SOM3 centred around London, Tokyo, LA and Vienna (cluster 3), and Sydney and Castellon (cluster 13). Through high levels of planning, preparedness and infrastructure, prosperous regions generally have systems in place to minimize flood impacts on the population, even though they may incur large economic losses (Desai et al., 2015; Kreimer et al., 2003). Almost half of the projected increases in these clusters are attributed to development, suggesting that these cities may have the capacity for lessening potentially elevated flood damage costs by concentrating on planning and mitigation policies.

Though this study does not consider coastal flooding, it may be noted that due to their locations near river mouths, many of the cities in the lower left of the map that are projected to experience high increases in impacts from river flooding are also at risk of increased coastal flooding from intensified storms and sea level rise due to climate change. Mumbai, Guangzhou, Shanghai, Ho Chi Minh City, Kolkata, Bangkok, and Dhaka are 7 of the top 14 cities (out of 136) ranked by current population exposure to coastal flooding. These same cities also comprise the top 7 cities (in this order: Kolkata, Dhaka, Mumbai, Guangzhou, Ho Chi Minh City, Shanghai, Bangkok) ranked by future (2070) estimated population exposed to coastal flooding (UNEP, 2016; Nicholls et al., 2008).

Almost all projected changes in flooding in this data set are of a relatively similar order of magnitude to the original effects, as can be observed on Figure 33. That is, most cities that are only marginally affected by flooding in 2010 are projected to experience only small increases by 2030, whereas cities with larger flood effects can expect greater changes. A significant correlation exists between the magnitudes of the cities' baseline flooding effects and the changes projected by 2030 (log-transformed absolute values for both variables) – an 88% correlation exists in the number of population affected and a 94% correlation for property damage costs. This supports the findings of Milly et al. (2002) who observed that the frequency of large flood events in large basins had increased substantially in the 20th century, but smaller floods had not.

7.6 Conclusion

Global patterns of urban flood responses to global and local changes in hydrology driven by climate change and development have been identified and visually communicated. Cities have been matched to these global patterns, and relationships between the individual cities have been discerned with respect to baseline flooding conditions and expected future changes. Information has been extracted from a large, recently released, global data set of city-level flood impacts relating hydrology and urban development, and combined with city-specific demographic information. The analysis and visual interpretation in this study has revealed interesting city-level patterns that are otherwise unobservable in the complex data set, and provides a comparison and distinction between individual cities that is not apparent in regional- or economic-level projections.

We have performed dimension reduction and clustering with a series of self-organizing maps to identify changing global patterns of city-level flood risks. The maps provide an indication of the predominant characteristics which determine the differences in urban river flood impacts between cities, and the cities occupy positions on the maps signifying their relative conditions. The method used here incorporates adaptions to the self-organizing map technique for map shape selection and temporal pattern extraction, allowing two levels of information to emerge: the characteristic patterns of dynamic global urban flood vulnerabilities, and a comparison between the cities with respect to flood characteristics and trends.

A shortcoming of the method used here is the assignment of flood protection level based on an assumption of proportionality with national income level. As standardised, current information on the real flood protection levels of all the cities in the data set is not readily available, this assumption has been necessary and has been made in line with current practice. This limitation has been recently acknowledged in the literature, with Winsemius et al. (2016) noting that 'currently installed flood protection is an important missing link in the assessment of global flood risk'. Future studies may aim to include specific flood protection levels for each city.

Whilst the timeline of this study is short, it is restricted by the data that is available. Studies at a global scale have been traditionally limited due to lack of cohesive data sets, and therefore the data set provided by Aqueduct is valuable for the fact that it spans a global set of cities and provides a rare opportunity for comparison. As the data is only provided for 2010 and 2030, there was no prospect for a longer analysis. Whilst this analysis may not provide a long-term outlook, at the very least an important insight into the current and near-future conditions can be gained.

Cities have major implications for climate change mitigation and adaptation (Revi et al., 2014). Unplanned development and urban migration are increasing vulnerabilities to natural hazards (UNEP, 2016) and land cover change and greenhouse gas emissions are intensifying urban hydrology. Understanding the relationship between flood impacts and social vulnerability is a necessary step for prioritizing flood mitigation and prevention strategies (Doocy et al., 2013). Whether the main driver of increased urban flood impacts is development or climate change, cities will benefit from development restrictions and planning standards for urban expansion, sustainable land development, management of population distribution and migration, and early warning systems and preparedness (Revi et al., 2014; UN-DESA, 2014; Doocy et al., 2013).

This study adds to the understanding of natural hazards in a global context, which is an important aspect of regional disaster risk management due to the dependency of local situations on global processes (Desai et al., 2015). The complex nonlinear socio-environmental relationships make it difficult to foresee local responses to global changes (Desai et al., 2015), and therefore this study focuses on risk communication (the process between risk perception and adaptation planning (Cardona et al., 2012)) to provide a visual analysis of the global patterns of evolving flood impacts, socioeconomic development and climate change, and the local city-level consequences of these changes.

Future work may include the addition of greenhouse gas emissions data, geographic location, city sizes and densities to this study, to discern the relationships of these factors with urban flood changes. Greenhouse gas emissions are the largest contributor to global warming, leading to alterations in the intensity of the hydrologic cycle (Pachauri et al., 2014, Barnett et al., 2005; Wentz et al., 2007; Schiermeier, 2011), and cities are the major contributors of greenhouse gases, with a large proportion of global emissions produced by a small global land area (Mills, 2007; Angel et al., 2010; Revi et al., 2014). The addition of these elements could highlight the essential role cities could play in climate change mitigation and the reduction of urban flood impacts.

7.7 References

Agarwal, P, and Skupin, A. (2008). Self-organizing maps: Applications in geographic information science, John Wiley & Sons.

Angel, S, Parent, J, Civco, D, & Blei, A. (2010a). Atlas of Urban Expansion, Cambridge MA: Lincoln Institute of Land Policy.

Angel, S, Parent, J, Civco, D, Blei, A, & Potere, D. (2010b). A Planet of Cities: Urban Land Cover Estimates and Projections for All Countries, 2000-2050. Lincoln Institute of Land Policy Working Paper.

Barnett, T, Adam, J & Lettenmaier, D. (2005). Potential impacts of a warming climate on water availability in snow-dominated regions. Nature, 438(7066), 303-309.

Clark, S, Sarlin, P, Sharma, A, & Sisson, SA. (2015). Increasing dependence on foreign water resources? An assessment of trends in global virtual water flows using a self-organizing time map. Ecological Informatics, 26, 192-202.

Clark, S, Sisson, SA, & Sharma, A. (2016). A dimension range representation (DRR) measure for self-organizing maps. Pattern Recognition, 53, 276-286.

Clark, S, Sisson, SA, & Sharma, A. (2016b, in press). Nonlinear manifold representation in natural systems. Environmental Modelling and Software.

Cunderlik, JM, & Ouarda, T. (2009). Trends in the timing and magnitude of floods in Canada. Journal of Hydrology, 375(3), 471-480.

Davies, DL, & Bouldin, DW. (1979). A cluster separation measure. Pattern Analysis and Machine Intelligence, IEEE Transactions (2), 224-227.

Desai, B, Maskrey, A, Peduzzi, P, De Bono, A, & Herold, C. (2015). Making Development Sustainable: The Future of Disaster Risk Management, Global Assessment Report on Disaster Risk Reduction. Geneva, Switzerland: United Nations Office for Disaster Risk Reduction.

Doocy, S, Daniels, A, Murray, S, & Kirsch, TD. The human impact of floods: A historical review of events 1980–2009 and systematic literature review. PLoS Curr, 16, 12.

Frich, P, Alexander, LV, Della-Marta, P, Gleason, B, Haylock, M, Tank, A, & Peterson, T. (2002). Observed coherent changes in climatic extremes during the second half of the twentieth century. Climate Research, 19(3), 193-212.

Ibrahim. (2016). Slum eradication policies in Marrakech, Morocco. World Bank Conference on Land and Poverty, Washington DC.

Immerzeel, WW, Van Beek, LPH, & Bierkens, MFP. (2010). Climate change will affect the Asian water towers. Science, 328(5984), 1382-1385.

Jiang, L, & O'Neill, BC. (2015). Global urbanization projections for the Shared Socioeconomic Pathways. Global Environmental Change.

Jongman, B, Ward, PJ, & Aerts, JC. (2012). Global exposure to river and coastal flooding: Long term trends and changes. Global Environmental Change, 22(4), 823-835.

Jongman B, Winsemius HC, Aerts JC, de Perez, EC, van Aalst, MK, Kron, W, Ward, PJ. (2015) Declining vulnerability to river floods and the global benefits of adaptation. Proceedings of the National Academy of Sciences. May 5 ;112 (18): E2271-80.

Kaski, S, & Kohonen, T. (1996). Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world. Proceedings of the third international conference on Neural Networks in the Capital Markets.

Katz, RW, Parlange, MB, & Naveau, P. (2002). Statistics of extremes in hydrology. Advances in water resources, 25(8), 1287-1304.

Kiviluoto, K. (1996). Topology preservation in self-organizing maps. IEEE International Conference on Neural Networks.

Kohonen, T. (2001). Self-organizing maps, Volume 30 of Series in Information Sciences.

Kreimer, A, Arnold, M, & Carlin, A. (2003). Building safer cities: the future of disaster risk: World Bank Publications.

Kummu, M, De Moel, H, Ward, PJ, & Varis, O. (2011). How close do we live to water? A global analysis of population distance to freshwater bodies. PLoS One, 6(6), e20578.

Kunkel, KE, Pielke Jr, Roger A, & Changnon, SA. (1999). Temporal fluctuations in weather and climate extremes that cause economic and human health impacts: A review. Bulletin of the American Meteorological Society, 80(6), 1077.

Meehl, GA, Arblaster, JM, & Tebaldi, C. (2005). Understanding future patterns of increased precipitation intensity in climate model simulations. Geophysical Research Letters, 32(18).

Mills, G. (2007). Cities as agents of global change. International Journal of Climatology, 27(14), 1849-1857.

Milly, PCD, Dunne, KA, & Vecchia, AV. (2005). Global pattern of trends in streamflow and water availability in a changing climate. Nature, 438(7066), 347-350.

Milly, PD, Wetherald, RT, Dunne, KA, & Delworth, TL. (2002). Increasing risk of great floods in a changing climate. Nature, 415(6871), 514-517.

Muis, S, Güneralp, B, Jongman, B, Aerts, JC, Ward, PJ. (2015) Flood risk and adaptation strategies under climate change and urban expansion: A probabilistic analysis using global data. Science of the Total Environment. 538: 445-57.

Nature. (2016). Waters encroaching. Nature Climate Change, 6(7), 635, editorial.

Nicholls, RJ, Hanson, S, Herweijer, C, Patmore, N, Hallegatte, S, Corfee-Morlot, J, Muir-Wood, R. (2008). Ranking port cities with high exposure and vulnerability to climate extremes.

Pachauri, RK, Allen, MR, Barros, VR, Broome, J, Cramer, W, Christ, R, ... & Dubash, NK. (2014). Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (p. 151). IPCC.

Revi, A, Satterthwaite, DE, Aragón-Durand, F, Corfee-Morlot, J, Kiunsi, RBR, Pelling, M, Roberts, DC & Solecki, W. (2014). Urban areas. In: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. United Kingdom and New York, NY, USA: Cambridge University Press, Cambridge.

Samir, KC, & Lutz, W. (2014). The human core of the shared socioeconomic pathways: Population scenarios by age, sex and level of education for all countries to 2100. Global Environmental Change.

Schiermeier, Q. (2011). Increased flood risk linked to global warming. Nature, 470(7334), 316.

Shanmuganathan, S, Sallis, P, & Buckeridge, J. (2006). Self-organizing map methods in integrated modelling of environmental and economic systems. Environmental Modelling & Software, 21(9), 1247-1256. doi: 10.1016/j.envsoft.2005.04.011.

Sheppard, SRJ. (2005). Landscape visualisation and climate change: the potential for influencing perceptions and behaviour. Environmental Science & Policy, 8(6), 637-654.

Skupin, A, & Hagelman, R. (2005). Visualizing demographic trajectories with self-organizing maps. GeoInformatica, 9(2), 159-179.

Sofia, G, Roder, G, Dalla, FG, Tarolli, P. (2017) Flood dynamics in urbanised landscapes: 100 years of climate and humans' interaction. Scientific reports.

Ultsch, A. (2003). U-matrix: a tool to visualize clusters in high dimensional data. Marburg: Fachbereich Mathematik und Informatik.

UN-DESA. (2015). World Urbanization Prospects: The 2014 Revision. New York: United Nations Department of Economic and Social Affairs.

UNEP. (2016). Summary of the sixth global environment outlook regional assessments: Key findings and policy messages. United Nations Environment Programme.

UN-Habitat. (2010). State of the world's cities. Earthscan.

UN-Habitat. (2014). Urban Equity in Development - Cities for Life. World Urban Forum 7. April 2014, Medellin, Colombia.

Václavík, T, Lautenbach, S, Kuemmerle, T, & Seppelt, R. (2013). Mapping global land system archetypes. Global Environmental Change, 23(6), 1637-1647.

Vesanto, J, & Alhoniemi, E. (2000). Clustering of the self-organizing map. IEEE Transactions on Neural Networks, 11(3), 586-600. doi: 10.1109/72.846731.

Ward, JHJr. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58, 236-244.

Ward, PJ, Jongman, B, Weiland, FS, Bouwman, A, van Beek, R, Bierkens, MF, ... & Winsemius, HC (2013). Assessing flood risk at the global scale: model setup, results, and sensitivity. Environmental research letters, 8(4), 044019.

Wasko, C, & Sharma, A. (2015). Steeper temporal distribution of rain intensity at higher temperatures within Australian storms. Nature Geoscience 8.7: 527-529.

Wentz, FJ, Ricciardulli, L, Hilburn, K, & Mears, C. (2007). How much more rain will global warming bring? Science, 317(5835), 233-235.

Willems, P, Olsson, J, Arnbjerg-Nielsen, K, Beecham, S, Pathirana, A, Gregersen, IB, & Madsen, H. (Eds.). (2012). Impacts of climate change on rainfall extremes and urban drainage systems. IWA publishing.

Winsemius, HC, Van Beek, LPH, Jongman, B, Ward, PJ, & Bouwman, A. (2013). A framework for global river flood risk assessments. Hydrology and Earth System Sciences, 17(5), 1871-1892.

Winsemius, HC, Aerts, JC, van Beek, LP, Bierkens, MF, Bouwman, A, Jongman, B, Kwadijk, JC, Ligtvoet, W, Lucas, PL, van Vuuren, DP, Ward, PJ. (2016) Global drivers of future river flood risk. Nature Climate Change, 6(4): 381-5.

Young, AF. (2013). Urban expansion and environmental risk in the São Paulo Metropolitan Area. Climate Research, 57(1), 73-80.

Websites:

1. Atlas of Urban Expansion: http://www.lincolninst.edu/subcenters/atlas-urban-expansion/Default.aspx
2. World Bank's World Development Indicators database: http://data.worldbank.org/data-catalog/world-development-indicators
3. World Resources Institute's Aqueduct Global Flood Analyzer Tool: http://floods.wri.org/#/
4. ISIMIP: https://www.pik-potsdam.de/research/climate-impacts-and-vulnerabilities/research/rd2-cross-cutting-activities/isi-mip/about
5. Shared Socioeconomic Pathways: https://tntcat.iiasa.ac.at/SspDb/dsd?Action=htmlpage&page=about
6. World Resources Institute's Aqueduct Global Flood Risk Country Rankings: http://www.wri.org/resources/data-sets/aqueduct-global-flood-risk-country-rankings
7. SOM Toolbox: http://www.cis.hut.fi/somtoolbox
8. Scientific American: https://www. scientificamerican.com/article/extreme-rain-may-flood-54-million-people-by-2030/
9. United Nations Framework Convention on Climate Change: www4.unfccc.in

8 PAPER 5 - LITERATURE REVIEW AND PRACTICAL APPLICATION GUIDE

Literature involving the details of the SOM method and uses of SOMs with spatiotemporal and nonlinear data sets, particularly those emerging from water-related and environmental systems, is discussed here. The literature review presented in this chapter has been prepared in the format of a publishable paper and has been submitted to an academic journal for review. This has been done to fill a gap in the existing literature, which has yet to include an explanatory document sufficient for new SOMs users to easily understand and create a SOM.

This literature review has been submitted as:

Practical guidelines for the application of self-organizing maps in environmental science and engineering

Clark S, Sisson SA, Sharma A.

8.1 ABSTRACT

Environmental measurements, often obtained over vast spatial areas at high temporal resolutions, produce great volumes of information from which meaningful messages may be extracted through appropriate summarisation. The self-organizing map (SOM) is an artificial neural network popular for extracting patterns and finding clusters in large multi-variate data sets. It is well-suited to noisy, high-dimensional measurements with nonlinear intervariable relationships as are often encountered in environmental measurements taken remotely or in the field. Though the SOM is a broadly applicable method, we have found that information regarding SOMs theory and implementation is currently widely scattered throughout theoretical and application literature, making the creation of a SOM for a first-time user a challenging and arduous task. Researchers are currently required to sift through widespread technically detailed advances that are documented in algorithmically-focused statistical papers to piece together a method that ensures a decent representation of their data with a SOM. Instead of doing this, we have noted that researchers are tending to revert to heuristic or software default parameter sets, or 'borrowing' parameters from SOM models that have been used in other applications but are not particularly relevant to the specific data set at hand. This paper draws the available information together into a cohesive guide, providing researchers with a tool to create a SOM and explore their data using techniques relevant to their particular data set. The effect that parameter selection and training choices have on the level of extracted information is discussed, practical guidance is provided for altering MATLAB code to appropriately modify parameter sets, and comparisons are made with closely-related methods. Recent examples from the literature are cited in each portion of this manuscript.

8.2 Introduction

Research in environmental sciences often involves the analysis of large, high-dimensional data sets resulting from the collection of short time-step, multi-variable measurements. Data sets amassed through automatic sensors or remote (satellite-based) measurements are frequently 'patchy' as collection is regularly interrupted by technological and meteorological issues. For meaningful messages to be extracted from these large volumes of information, the data must be reduced and summarised into a manageable number of characteristic patterns and intervariable relationships.

Dimension reduction and clustering are therefore key components of the exploratory data analysis stage of modern environmental research. The self-organizing map (SOM, Kohonen, 1990) algorithm is frequently chosen for these tasks due to its inherent resilience to noisy and missing data (Vesanto, 1999), and applicability to high-frequency, multi-dimensional data. The SOM extracts the most prevalent patterns in a data set and clusters the data items around these patterns, organizing the results into an intuitively understandable low-dimensional visualisation. SOMs applications involve the representation of a large number of high-dimensional observations with a usually much lower (and therefore more manageable) number of low-dimensional vectors which form centroids for clusters of the original data. This facilitates analysis of the properties of the data set through analysis of the ordered low-dimensional cluster structure and cluster members.

Over the past decade, approximately 1000 papers published each year in various fields of environmental science have used SOMs for data analysis (website 1). The literature provides little accessible and cohesive guidance, however, to lead non-statistical researchers through the process of SOM implementation and interpretation. We have found no paper unifying the information needed to knowledgably create a SOM relevant to a specific data set and interpret the results. Review articles generally focus on examples of the application of SOMs in certain fields (as in Kalteh et al., 2008; Liu & Weisberg, 2011; Agarwal & Skupin, 2008), reviews on SOMs general theory (as in Kohonen, 2013; Yin, 2008; Astudillo & Oommen, 2014; and Cottrell et al., 2016), or reviews of particular aspects of SOMs theory (as in Barreto, 2007 and Fyfe, 2008). Our goal lies in sifting through and drawing this information together, combining practical aspects of general theory with guidance on how to effectively apply a SOM to data. Examples are cited, but our aim is not to provide a comprehensive list of all SOMs applications, rather to help the reader to intelligently create their own with reference to what others have done.

Each application of the SOM method requires the choice and evaluation of appropriate parameters. SOM parameters cannot be optimised by maximum likelihood estimation, though, as the SOM training algorithm does not attempt to optimise any particular objective function (Erwin et al., 1992). Instead, parameter selection requires ample time and experience, making this a great disadvantage of SOM implementation for the non-expert (Gopakumar et al., 2007). The final results of the SOM are influenced by the choice of parameters and arbitrary parameter selection has the potential to lead to maps that fail to reveal portions of the data structure and intervariable relationships (Principe et al., 1998; Flexer, 1999; Vesanto, 2000;

Kohonen, 2001 & 2013; Liu et al., 2006; Cereghino & Park, 2009; Liu & Weisberg, 2011; Wang et al., 2013; Astudillo & Oommen, 2014). Deliberate parameter choices must therefore be based on an understanding of available options and the benefits of each choice.

Most recently published SOMs applications have been relying on default parameter choices that are built into software or user-specified values based only on the preference for the degree of generalisation of the final map visualisation (selecting the output that the user finds most manageable regardless of any structure present in the data set). Another frequently used method is to simply adopt the parameters used in a previous, though potentially unrelated, publication. Abrahart et al. (2012) emphasize that though default options of parameter selection for neural networks may often produce reasonable results, 'a more comprehensive assessment is needed to ensure proper guidance and support for modellers'.

SOMs literature tends to come in two streams: 1) statistical analyses of the SOM algorithm published in statistics or neural networks journals, not easily accessible to researchers from other fields; and 2) applications using SOMs as a black-box tool with little or no informed input from the user. Liu & Weisberg (2011) claim that the 'SOMs algorithm is like a black-box for most application researchers, which may prevent some potential new users from pursuing further SOM applications'. Yin (2008a) states 'the SOM may have more potential than implied by current practice, which often limits the SOM to empirically chosen parameters'. Kalteh et al. (2008) note a 'lack of comprehensive literature review for SOMs, data handling procedures and applicability' and that 'SOM applications are generally dependent on ad-hoc approaches characterised by guesswork'.

This paper is provided as a guide to aid researchers in understanding current best practices for the informed application of SOMs within a context of SOMs background theory. The reader will be guided through the production and interpretation of a suitable SOM for the exploratory analysis of their specific multivariate, nonlinear data set. Relevant applications, parameter choices, code implementation and interpretation of output maps are discussed. Key references are provided in each section to lead the reader to further resources.

The structure of this paper is as follows: Section 8.3 provides background information on the SOM mechanisms for pattern identification and clustering, and common contemporary uses of the SOM in environmental research; Section 8.4 is a step-by-step guide for the creation of a SOM from a raw data set; Section 4 offers basic examples for the modification of freely available MATLAB code; Section 8.5 imparts guidance for extracting information through the visual interpretation of a SOM; and Section 8.6 provides concluding remarks.

8.3 Background

8.3.1 SOM description

Developed in Finland in 1981 by Teuvo Kohonen, the self-organizing map has steadily gained popularity since its introduction, with explanatory papers (Kohonen, 1982; Kohonen, 1990; Kohonen 1998; and Kohonen 2013) receiving over 38,800 citations combined. The book 'Self-

Organizing Maps' (Kohonen, 1995; Kohonen, 2001) has been cited over 3400 times (website 2).

The SOM is an artificial neural network that organises data by recognising patterns in the input. Complex data sets are transformed into more readily interpretable forms by visualising clusters in the data structure whilst maintaining the overall topological structure of the data set. Through dimension reduction and nonlinear projection, the most prevalent intervariable combinations in the data set are identified as the characteristic patterns of the data. Each input item is uniquely related to one of the patterns, clustering the data into groups sharing important similarities based on the key pattern features. A low-dimensional visualisation (or map) is produced depicting the nonlinear relationships in the high-dimensional data. The prevalent patterns and data clusters are defined on the map.

Pattern identification, clustering and data visualisation, together, reveal information that may be otherwise unobservable in large data sets. Other techniques exist for each of these processes separately, but the combination leads to the SOMs' unique analysis capabilities. For example, clustering methods that are not combined with dimension reduction are not easy to visualise and therefore can be less readily interpreted and conveyed.

The general SOM method entails arranging a map grid (consisting of a regular rectangular grid of connected nodes) in an initial approximate location over a data set. Through an iterative training process, the map nodes move amongst the data items (self-organise) whilst maintaining their grid structure, until their locations provide the best possible coverage of the data set without the map becoming unnecessarily twisted. The grid is stretched and bent until the position and orientation best represent the data structure. Figure 34 shows a synthetic data set (black) consisting of three gaussian clusters, with an initial SOM grid (green) placed in a preliminary location, and the trained SOM grid (blue) better following the structure of the data. For details of map training, see the Algorithm section on the next page and Appendix 1.

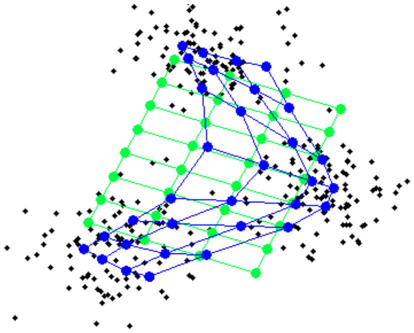


Figure 34: Data points (black, synthetic data set) with an initialised SOM (green) oriented in directions of maximum variance, and the SOM after training (blue) in which nodes have been relocated to better represent the data set.

Algorithm for map training

Network structure: A set of N input data items are listed in vectors, x_i, where $i = 1{:}N$. The output map grid consists of M map units, or nodes, with a vector, m_j (where $j = 1{:}M$), associated with each map unit. All input and output vectors are of the same dimension, d. Each input vector is connected to each map node in parallel through a set of adjustable scalar weights. These weights are the component values of the map vectors, m_j.

Training process: Map training consists of an iterative process of 1) finding the map node that best matches the current input data item, and then 2) updating this best matching map node (and its neighbours) to become closer to the input, as follows (Kohonen, 1993):

1. **Matching:** Each input item, x_i, is compared to all map units, m_j, by some distance measure (usually Euclidean) to find the closest map node (or best matching unit, BMU), which is given the index c:

$$\|\mathrm{x_i} - \mathrm{m_c}\| = \mathrm{min_j}\{\|\mathrm{x_i} - \mathrm{m_j}\|\} .$$

This process partitions the input data into subsets known as *Voronoi sets*, v_c, consisting of data items sharing the same nearest map node, m_c, at each iteration.

2. **Updating:** At each iteration, the locations of the map nodes are adjusted closer to the data items in their Voronoi sets. Each data item effectively draws on its BMU, as well as nodes within a specified neighbourhood of the BMU. Inclusion of the neighbours in the updating process is accomplished through the application of a 'neighbourhood' kernel, maintaining the smoothness of the map. Each map unit, m_j, is updated with a weighted average of the data items matching the map nodes in its local neighbourhood, where the weighting is given by the neighbourhood kernel, or function, $H = [h]_{ij}$. At each iteration (t), the updated node locations are calculated as in (Kohonen, 2013):

$$\mathrm{m_j(t+1)} = \frac{\sum_{\mathrm{i=1}}^{N} \mathrm{h_{ij}(t)\, x_i}}{\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{h_{ij}(t)}}$$

The elements of the neighbourhood function centred around the BMU (m_c) of data item x_i (eg. h_{ij}) indicate the influence of each data item (x_i) on the updating of node m_j. At iteration t, a Gaussian neighbourhood kernel is:

$$\mathrm{h_{ij}(t)} = \exp(\frac{-(m_c - m_j)^2}{2\sigma^2(\mathrm{t})})$$

where the radius, σ, of the neighbourhood kernel decreases with each training iteration to include fewer neighbouring nodes. As the neighbourhood kernel is based only on the map size and is the same regardless of the data to be represented, it can be calculated before training begins. During training, all data belonging to a Voronoi set is weighted by the same value of the neighbourhood function.

The SOM has two main organisational goals: pattern identification and topology preservation. These are realised through the separation and iteration of two interacting subtasks during the training process - matching and updating. During the matching stage, the closest map node (or best matching unit, BMU) is identified for each item of data. Then the map nodes are drawn closer to their nearby data items during the updating stage, with neighbouring nodes on the map grid moving towards the same data items. As the map nodes settle closer to the data items in data space, the result is more nodes situated in higher-density areas of the data.

The specific location in data space that each map node occupies at the conclusion of training (given by the local values of each variable) are recorded in the high-dimensional vectors of the map nodes. This unique combination of variables represents a particular characteristic pattern of the data set.

Following training, the individual data items are matched to their closest map nodes on the trained map. This produces clusters of data sharing a common nearest node, and therefore sharing common key characteristics as identified by the variable values of that node. Similar data items become matched (or mapped) to the same or nearby map nodes, and more different data mapped to more distant nodes.

The patterns and clusters extracted from the data during training are visualised on a low-dimensional output map, organised according to their similarity. The degree of similarity between each cluster is proportional to the topological distance between them, with closer areas of the map representing similar patterns in the data domain. Areas of data space with higher densities of data are represented on larger areas of the SOM.

When presented in two-dimensional space, the map is a nonlinear projection of a reduction of the data set, as the high number of high-dimensional data items become represented in a lower dimension by a usually lower number of map nodes. The output map can therefore be more manageably explored than the original data set, giving insight into the overall structure and prevalent patterns existing in the data.

Figure 35 depicts the representation of a data set with a SOM, both in data space and map space. The trained map is not twisted and data items that are near each other in data space are represented by the same or nearby nodes.

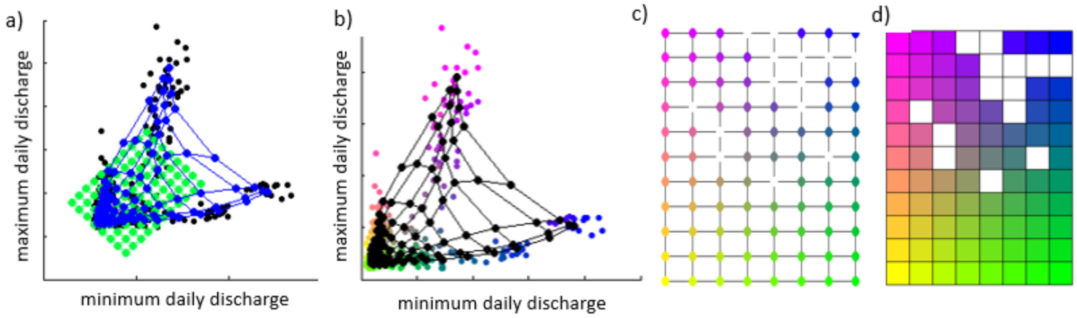


Figure 35: An example representation of a data set with a SOM: a) the data (comprised of daily minimum and maximum streamflow values per year (2000-2016) for river stations in Environment Canada's HYDAT database (website 3)) is shown with the initial map grid (green) draped in the main directions of variance and the trained SOM

in blue; b) the final trained SOM (black) is shown in data space, and the data items have been coloured so that similar items are similar colours (by spreading a two-dimensional colourmap over the principal component projection of the data); c) the SOM grid is depicted in map space with each node coloured by the mean colour of the data items matched to it (white nodes have do not have any matching data items); and d) the SOM output map is ready for labelling with the data items.

The SOM provides various benefits over other methods for the exploration of environmental data. The number of clusters present in the data does not need to be specified before hand. The SOM has an intrinsic ability to handle patchy (noisy and missing) data. The resulting clusters are ordered on the output visualisation indicating similarities and dissimilarities, allowing for further analysis if desired. Creation of a SOM does not require an explicit understanding of the complex and potentially undefinable processes and relationships within the system that produced the data; the analysis is conducted based only on the data presented to it. The SOM can therefore be used to explore complex multi-disciplinary or human-environmental data sets in which the intervariable relationships are difficult to explicitly quantify. In these cases, the creation of an elaborate mathematical system model would be complex and time consuming. Another distinct benefit of the SOM is the ability to incorporate new data after the map is created, providing a tool for online, real-time data analysis which is useful for the real-time processing of sensor data.

The nonlinear regression performed by the SOM is considered ‘nonparametric’ as it is based on fitting a number of ordered, discrete map vectors to describe the distribution of input samples. The SOM is deemed an ‘unsupervised’ process as it searches for unknown structure in unlabelled data. Unsupervised learning is closely related to density estimation in that it constructs an estimate of an unobservable probability density function (pdf) based on the data that is presented. If the input samples have a well-defined pdf, the SOM map nodes will eventually come to approximate it in an ordered way, producing a nonlinear projection of the high dimensional input pdf onto a low dimensional display (Kohonen, 1990 & 1995). Yin (2008a) points out, though, that a connection between the map nodes and the pdf of the input data has not been derived in general, but only for a one-dimensional case (by Erwin et al., 1992).

The SOMs algorithm exists in a theoretical region of knowledge between principal component analysis (PCA) and the k-means clustering method, with PCA providing the most rigid analysis of the three, and k-means the most flexible. PCA (a type of factor analysis) transforms a data set into a small number of linear, uncorrelated variables or principal components (PCs) which become axes of the projection space. The set of data points are rotated around their mean to align with the new set of axes, which point in the directions of maximum data variance. The first principal component is defined by the eigenvector of the data set with the highest eigenvalue, and the axis in this direction accounts for as much of the variance in the data as possible. PCA separates clusters of data items well, but is not ideal for representing nonlinear data as the PCs are always linear. The SOM algorithm uses PCA during the initialisation stage, aligning the initial axes of the map grid along the principal component plane. SOMs are considered a constrained form of the k-means clustering algorithm, with the nodes constrained to a two-dimensional manifold (Hastie et al., 2009). The SOM method becomes equivalent to the k-means algorithm at the end of training (when the SOM neighbourhood kernel is so small it only contains a single node). The K-means algorithm is proficient at clustering, though it does not retain information on cluster ordering, and encounters issues in visualisation when the dimensions of input data are high. The SOM is beneficial over the k-means algorithm if the retention of topological information is important, and for real-time analysis of online data (Laerhoven, 2001). Sturn et al. (2002) compare results of SOMs, k-means and PCA, determining that for all the outcome of clustering depends on the method of normalisation, the similarity measure and parameter values used. Cornford et al. (2009) found that PCA does not perform well with missing data. Hastie et al. (2009) provide an example comparing SOMs with k-means. Yin (2008b) elaborates on the relationship between SOMs and PCA, and Liu & Weisberg, 2011 found SOMs to be advantageous over PCA for pattern extraction tasks.

Figure 36 depicts a comparison of k-means, SOM and PCA results on a single data set (source: Bache & Lichman, 2013). It can be observed that all the techniques separate the two main clusters in the data, k-means and SOMs both place a user-specified number of cluster centres amongst the data, and the SOM results are similar to k-means except that the grid structure retains an ordering to the clusters. The techniques each perform well when compared with certain aspects of the SOM, however if clustering, nonlinear projection, *and* an ordering of cluster information is important, the SOM provides these attributes within a unified technique.

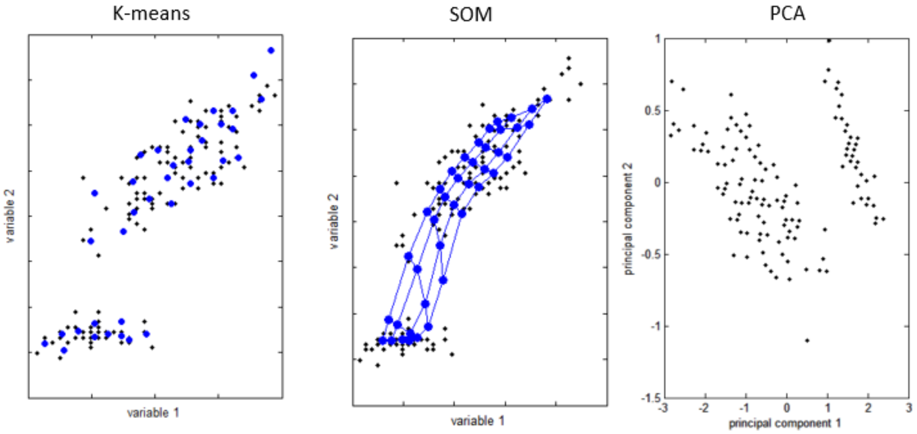


Figure 36: Comparison of k-means, SOMs and PCA representations of the same data set (data are black, nodes/cluster centroids are blue). All three methods separate the clusters, k-means and the SOM reduce the number of data points to cluster centres, but only the SOM retains topographic information through the linking of the cluster centres.

8.3.2 SOM use in environmental sciences

SOMs applications entail the summarizing of large amounts of data into representative intervariable relationships or patterns, determining ordered clusters of similar high-dimensional data items, and producing a low-dimensional visualisation of the results. With environmental data sets, the pattern extraction, clustering and visualisation capabilities of the SOM are typically applied to spatiotemporal analysis, time series analysis, infilling missing data, and prediction.

Patterns are extracted as representative system states, with each node of the trained map representing a specific pattern. These patterns are used as the basis to cluster common system states and determine the nonlinear relationships between variables. Clustering consists of identifying the number of clusters in the data and uniquely assigning each input item to one of the clusters. The clusters are based around the most common patterns in the data as identified by the nodes vectors, which form the cluster centroids.

Applications using SOMs to determine sets of typical system state patterns have included investigations of: sea surface temperature (Liu et al., 2006), meteorological patterns (Reusch et al., 2007), regional frequency analysis (Lin & Chen, 2006), wind patterns (DuVivier et al., 2016), wave climate states (Barbariol et al., 2016), spatial patterns of groundwater properties (Nguyen et al., 2015 and Choi et al., 2014), and soil quality (Rivera et al., 2015). Frequency and transition matrices can be created by determining the percentage of data matching each pattern, as in Hewitson & Crane (2002), Reusch et al. (2007), Falcieri et al. (2014), Nguyen et al. (2015) and Swales et al. (2016). Newton et al., (2014) further the use of SOMs for state frequency analysis by including the persistence of states and synaptic type frequency anomalies for each variable at each node.

Nonlinear relationships have been investigated using a SOM, between: streamflow regimes and fish communities (Tsai et al., 2016), atmospheric circulation and surface climate (Newton et al. 2014; da Anunciacao, 2014), reservoir water quality in relation to watershed land cover types (Park et al., 2014), modern and medieval climate circulation patterns (Edwards et al., 2017), regional precipitation and large scale atmospheric dynamics (Liu et al, 2016), atmospheric circulation and arctic sea ice extent (Lynch et al., 2016), rainfall and runoff (Hsu et al., 2002), and water vapour transport and mass loss in the Greenland Ice Sheet (Mattingly et al., 2016). Shanmuganathan (2006) uses SOMs to investigate patterns of complex human-environmental interactions, by plotting a trajectory of regional river water quality response as human influence increases, assuming downstream stations incur more anthropogenic influence, inferring water quality response to humans even though no human data is used. Matic (2017) discovers patterns of salinity and temperature in oscillating Adriatic Sea regimes. Rodriguez-Alarcan & Lozano (2017) use the SOM as a decision support system for reservoir regulation. Vaclavik et al. (2013) identify generic global patterns of land pressures and environmental threats by clustering data sets based on intensity of use, environmental conditions and socioeconomic indicators. Vereecken et al. (2016) investigate patterns of water, mass and energy in soil-vegetation/atmosphere interactions.

Abundant environmental applications use the SOM for clustering data into similar subsets, such as: climate regions (Morioka et al., 2010), catchments (Ley et al., 2011), segments of coastline for ecological classification (Ramos et al., 2016) and niveograph patterns (Wang et al., 2013), as well as measurement stations, satellite imagery data, seasonal patterns, sea level pressure, and precipitation.

Spatiotemporal analysis

SOMs have most often been used in recent environmental research to investigate patterns and clusters in the spatial *or* temporal aspect of the data separately. Either spatial patterns occurring at different times or temporal patterns occurring in different locations have been investigated and compared on a single map. For example, Takala et al. (2008) use the SOM to estimate the beginning of seasonal snow-melt, creating both a timeseries for subsequent days and a spatial estimation for a single day.

A popular method of spatiotemporal data analysis with SOMs involves the creation of a series of maps placed next to each other, each representing clusters of data at different timesteps (as in Skupin & Hagelman, 2005; Ellis et al., 2014; Ye et al. 2015; Lukacs et al., 2015), requiring the reader to visually extract and analyse the temporally changing spatial patterns.

Another method is to create a single map with all the available data, as in Mihanovic et al. (2015); Sharif et al. (2015); Li et al. (2015); Kim et al. (2016); Hong et al. (2016); and Jutagate et al. (2016). Changes over time can then be visualised by plotting batches of consecutive data (ie decades) onto the trained map (as in Wang & Feng, 2011; Wang, 2015), or joining consecutive BMUs into trajectory lines, as in Skupin & Hagelman (2005) either for the entire data set or as separate trajectories for specific data segments. Olkowska et al., (2014) assess patterns of seasonal anthropogenic pollution in a specific catchment, using a single map of all measurements at all spatial locations and analysing the seasonal effect through seasonal accounts of cluster memberships.

However, these popular methods may not adequately capture and express the structure of the data. If a separate SOM is produced for each time period, these maps may not be directly comparable to each other due to differences in the distributions and correlations in the data at each time step. If a single SOM is created from all data, with separate time periods mapped to it, the overall map may not accurately describe the finer structure of the data at each time step.

Attempts have been made to apply the SOM to more innovative visualisations of spatiotemporal data. Some of these involve using the traditional SOM in novel ways (as in Wang et al., 2013), and some involve extending the SOM algorithm itself (as in Sarlin, 2012). Wang creates a single SOM of twenty years of snow accumulation and melt patterns with high temporal and spatial resolution, and then clusters the nodes and plots trajectories for specific spatial locations to reveal cyclical and long-term trends. Wang simultaneously determined spatial differences in time series (snow accumulation/ melt patterns between mountain ranges) as well as changes in patterns over time at specific locations (using trajectories). Sarlin developed the self-organizing time map (SOTM) to visualise the evolution of the cluster structure across space and time. The SOTM is made up of a 1D SOM at each timestep, arranged in order of ascending time. Multivariate temporal and cross-sectional aspects are visualised at the same time. Changing, emerging and lost clusters in the data structure become apparent. Clark et al. (2014) introduce a post-processing technique to track the transition of individual data items through a changing global cluster structure. Newton et al., (2014) plot a series of time periods (months) onto a SOM, determining frequency histograms for each node.

Time series analysis

Time series analysis, or trend visualisation, using SOMs entails the identification of temporal patterns and clustering of data items where each data item is an individual time series. This produces groups of data items that trend in similar ways (as in Mothe et al., 2006; Clark et al., 2014). SOMs are practical in time-series analysis for outlining the boundaries of existing conditions, clustering based on global characteristics extracted from the time series, and identifying future directions. For long time series, most clustering techniques are impractical due to missing or too much data, however these are not restrictions for the SOM (Wang et al., 2006).

Many variants on the basic SOM exist for time series analysis, generally incorporating some form of short-term memory. Detailed information is given in Barreto (2007). These variants rarely appear in the application literature possibly due to the complexity of implementation. The simplest method of incorporating time series data is to convert the sequential items into static form using data windows, and then apply the traditional SOM (Kohonen, 2001). An innovative method is used by Wang et al., (2006) in which Euclidean distance is not used for clustering, rather clusters are based on features of the time series such as trend, seasonality, periodicity, chaos, and self-similarity.

Infilling missing data

The SOM algorithm is quite insensitive to missing values in the data, and even well suited to infilling missing multivariate data. Data items with missing values for certain variables are matched to their nearest node based only on the values that are present. The data vector then adopts the node vector's value for the missing variable (Wehrens & Buydens, 2007).

Mwale et al. (2012) found this method to provide reliable estimates and reduce uncertainties associated with insufficient data. It has been applied to infilling runoff data in inadequately gauged basins based on rainfall measurements (Adeloye & Rustum, 2012; Mwale et al., 2012; Nkiaka et al., 2016), physiochemical parameters in water samples (Folguera et al., 2015) and estimating water quality in unmonitored streams based on relationships between spatiotemporal watershed attributes and water quality in monitored streams (Gamble & Babbar-Sebens, 2012). Toth (2013) has clustered the time series of flow at gauged sites, creating groups without using any geographical, morphological or climatological information as input. By associating pluviometric and morphometric attributes, ungauged stations were then added to the clusters based on climate and landscape characteristics facilitating the transfer of information from gauged to ungauged stations. Rustum & Adeloye (2007) found SOMs perform better than most widely used neural networks in water resources for infilling missing data.

Prediction

Patterns in future data may be estimated using the knowledge of inherent states determined to exist in the current data (states that occur with certain frequencies, and the transitions between states). Harnessing information from the established nonlinear relationships allows for the prediction of a future variable based on its association with easily predicted variables.

An extension of the 'infilling missing data' technique, this is done through the assumption that for the same conditions (combinations of variable values) the same patterns will occur. Simon et al. (2005) suggest the SOM enables the prediction of an entire vector of future components with the same precision for each component.

Steynor et al. (2009) link climate states to observed runoff, and then predict future flows based on future global climate model predictions. By plotting future data onto a map trained with current data the change in cluster membership indicates changing state frequencies predicted for the future. Streamflow prediction is performed by Gopakumar et al. (2007), Chang et al. (2007); Huang et al. (2011) and Tiwari et al. (2013). Toth (2009) discerns future runoff events using the SOM. Sarlin & Marghescu (2011) create a SOM-based early warning system, using the SOM for prediction by identifying precursor conditions of an event. Takala et al., (2008) predict the onset of snowmelt through the identification of homogenous regions for transferring information from gauged to ungauged sites. Chang et al. (2016) incorporate the SOM into a monthly basin-wide prediction of groundwater levels to be used in sustainable basin management.

The relationships determined by the SOM may also be used to predict what could be expected from new data given the presence of a certain system state. For example, Dejean et al., (2011) create a map based only on climatic variables over a number of years, and then apply the number of wasp nests per year as an associated variable to determine the correspondence between the number of wasp nests and climate states.

8.3.3 Review papers

Many review papers are available in the literature, summarising SOMs theory or applications in specific fields.

SOMs technical reviews: Kohonen (2013) provides a concise summary of the SOMs technique and biological background of the SOMs 'brain map' analogy. Yin (2008a) provides a review that includes the biological background, SOMs early development as well as convergence and cost function theories. Astudillo & Oommen (2014) reiterate background theory and provide a significant section describing SOM variants. A comparison of SOMs with other topographic mapping algorithms is given by Fyfe (2008). Barreto (2007) and Hammer et al. (2005) review SOMs use for time series analysis. Cottrell et al., 2016, discuss recent theoretical advances in the mathematical theory of SOMs.

Reviews of SOMs use in specific environmental fields: Lek & Guegan (1999), Kalteh et al. (2007), Liu & Weisberg (2011) and Agarwal & Skupin (2008) provide SOMs reviews which list and discuss a number of ecological, water resource, meteorological/oceanographic and GIS applications respectively. Cerghino & Park (2009) discuss map size selection, error measures, and innovative ways for uncovering relationships between biological and environmental water resources variables. Abrahart et al. (2012) discuss neural networks (including SOMs) in rainfall-runoff and streamflow modelling. The authors advocate moving away from unstructured incremental technical improvements or the repeated application of the same techniques to different data, with the call for ‘novel applications that can only be tackled with artificial neural networks’.

8.3.4 Variations

Dynamically-sized SOMs have been proposed in the literature to avoid the need for determining map size in advance. The Growing SOM (GSOM, Alahakoon et al., 2000) uses a ‘spread factor’ to measure and control the growth of the map. New nodes are grown from boundary nodes in all free directions simultaneously. Nodes with no hits are removed. The Growing hierarchical SOM Toolbox (GHSOM, 2002, Chan & Pampalk) uses a small initial map to represent the entire data set then refines levels of granularity only where needed. Growth occurs by inserting a row or column of new nodes between existing nodes. The Growing Bayesian SOM for data clustering (GBSOM, Guo et al., 2012) adds new neurons through a process of identifying the neuron with the lowest log-likelihood. The GWR (Marsland et al., 2002) is a self-organizing network that ‘grows when required’, with new nodes added whenever the current state of the network does not sufficiently match the input. Growing cell structures (GCS, Fritzke, 1994) insert cells between the cell with the most hits and its most distant neighbor, in a process related to fractal growth.

In **Bayesian SOMs** (BSOMs) each node is a Gaussian distribution, with the mean vector, covariance matrix and prior probability being the weights. The winning node is selected by having the maximum posterior probability. The mean, covariance, and prior probability weights are updated within the neighbourhood identified by the posterior probability. The map can converge to overlapping mixture distributions. Luttrell (1994) provided a Bayesian derivation of the properties of the SOM, using folded Markov chains. Yin & Allinson (1997) replace the distance measure and neighbourhood function with estimated postierior probabilities of the nodes. Guo et al. (2011) investigate the impact of learning rates, initial values, covariance matrices, input order and number of iterations, determining that the BSOM is sensitive to learning rates, covariance matrices and input order.

The set of **Pareto-optimal solutions** for multi-objective optimization decisions are visualized in a number of ways in the literature using SOMs. Each member of the set is potentially the best solution depending on the relative priorities of the objectives. Chen et al. (2013) introduces a visual-interactive approach using SOMs to visualize a set of multi-objective optimal points, in which each objective is represented by a corner of the map. Pareto front optimization has been used with SOMs to optimise design with a number of competing objective functions (Obayashi & Sasaki, 2003). A SOM is made of the values of four objective functions, to visualize the trade-off between the design objectives and indicate which variables have similar influences on design tradeoffs. The map edges represent the Pareto solution if only two objective functions are used. Okamoto et al. (2014) visualize almost 28,000 Pareto-optimal solutions using a SOM, with the solution mapped to as many SOMs as the number of objective functions in which the same coordinate on each map represents the same solution. Kurasova et al. (2013) cluster Pareto-optimal solutions into groups and evaluate only the representatives of each cluster given by the map node vector. A spherical SOM is used for Pareto solution visualisation by Yoshimi et al. (2012) in which each Pareto solution is an input and the dimensions are the objective function values for each solution.

8.4 CREATION OF A SELF-ORGANIZING MAP

This section is a comprehensive guide for the creation of a self-organizing map for the representation of a data set. The following steps are followed to create a SOM:

1. Construct a matrix of the input data samples
2. Pre-process the data matrix (standardise/normalise/transform)
3. Determine the size and shape of the map grid to best represent the data structure
4. Initialise the map over the data (drape it over in the principal directions)
5. Train the SOM (bend and stretch the map to better match the data)
6. Place data items on the map by matching to their closest map node

8.4.1 Step 1: Consolidate input data

8.4.1.1 Input matrix construction

The input data must be in a specific form to be read by a SOM: matrix formation with rows consisting of the data items (each separate observation or measurement makes up a row) and columns consisting of variables (dimensions), as in Table 11.

Table 11: Input data matrix

	variable 1	variable 2	variable 3	variable 4	…	variable d
data item 1						
data item 2						
…						
…						
data item N						

This format ensures that each of the N items of input data is in vector format, x_i where $i = 1{:}\,N$. All vectors are of the same dimension, d. Each dimension is an observation variable such as precipitation, temperature, pH, population level, or observations of a certain variable at multiple spatial locations or times. In artificial neural network terms, the values in the input matrix become numerical weights for the variables with respect to each input sample.

There is no rule for the amount of training data needed for SOM creation. However, due to the stochastic nature of environmental data it cannot be assumed that a model made from one set of training data will represent all underlying relationships in the system, and results will improve as the quantity of training data increases (Kingston et al., 2005).

8.4.1.2 Missing data

The SOM is able to function with a large proportion of missing data, for instance if some measurements are missing information for certain variables. In this case, data is matched to the map nodes based only on the variables for which data is available; when calculating the distances between the input item and the map nodes, the missing data is excluded from the distance calculations and the data is mapped to the nodes that are closest in Euclidean distance based only on the variables that have values present in that data vector. As the distance calculations for a data item to all nodes will omit the same variable of data, the results are comparable (Vesanto, 2000).

8.4.1.3 Categorical data

The SOM is designed to represent data in which the magnitude of the values has meaning. Categorical data can be incorporated by mapping it to ordinal data. This is done by labelling each sample with a number, ensuring that the numeric labels are in a logical order and not arbitrarily assigned. For example, categories labelled 2 and 3 must be more similar than categories 2 and 10 for the resulting self-organization to have a meaningful interpretation.

8.4.2 Step 2: Preprocessing

Environmental variables naturally consist of different measurement scales and types of data. Normalising the columns of the input data matrix before map training ensures that variables with greater magnitudes or variances do not overshadow variables that may be less diverse, of smaller magnitude, or measured in different units. This roughly equalises the contribution of each variable to the results (Kohonen, 2001). Overlooking the preprocessing step risks causing the main axis of the map to be principally aligned with the variable of largest magnitude, thereby producing a map that is mainly representative of this variable (as discussed in Clark et al., 2016).

Common preprocessing methods are transformations of the input matrix that equate either the variances or minima and maxima of each dimension:

1. Normalising the columns, by scaling the variances to 1 around a mean of 0, or
2. A linear transformation equalising the minima and maxima of each variable.

Another method, though far less common, is to scale the variance of all variables separately to reflect their perceived relative importance (Kaski & Kohonen, 1996). Highly skewed variable distributions may benefit from logarithmic transformations (Agarwal & Skupin, 2008).

8.4.3 Step 3: Parameter selection

A number of parameters require specification prior to the creation of a SOM. Parameter choices include: map size (the number of nodes), map shape (the configuration of nodes), the initial and final radii of the smoothing kernel and the smoothing kernel shape.

Common techniques for selection of these parameters range from the application of heuristics, the minimization of various error measures, or trial and error to produce the most agreeable visualization for the user. These decisions are discussed in this section. Default software parameter values are described, as well as more tailored options, and the benefits are outlined for choosing values beyond the software defaults.

8.4.3.1 Objective function

Before beginning to investigate how parameters are chosen, first it is important to realise why parameter selection for the SOM is particularly challenging. In the parameter selection phase of model building, objective functions are commonly optimised to aid the choice as they are able to provide a quantifiable assessment of the optimality of a set of parameters. However, it has been proven that the SOM training process has no objective function that is optimized exactly (Erwin et al., 1992; Pampalk, 2001; Yin, 2008a). The SOM training method cannot be quantified in a single mathematical expression, instead it follows the gradient descent of a separate set of energy functions for each node (Erwin et al., 1992). For this reason, parameters must be chosen by other methods such as quality measures applied to maps created with different parameter sets, as will be described below. For more information on the lack of a SOMs objective function, see Yin (2008b) and Appendix 2.

8.4.3.2 Map structure

8.4.3.3 Grid configuration

The grid of map nodes is generally connected in either a hexagonal or rectangular lattice, as shown in Figure 37, with nodes located at each vertex. This configuration determines the number of nearest (equally close in map space) neighbours for each node.

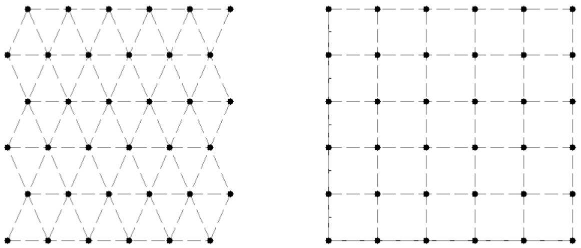


Figure 37: Hexagonal and rectangular 6x6 map grid lattices. Connections to nearest nodes are shown. Nodes on hexagonal grids have up to 6 nearest neighbours and nodes on rectangular grids have up to 4 nearest neighbours.

The nodes of a rectangular lattice have up to four nearest neighbours. In a hexagonal lattice, each map node has up to six nearest neighbours. This difference influences the results of the SOM training process in which the locations of the nearest neighbours are updated by the same amount at each iteration. A larger number of nearest neighbours leads to greater topological preservation of the input data structure and a more uniform final map, and for this reason hexagonal lattices are often considered more effective (Kalteh et al., 2008). Rectangular lattices are popular, however, due to the easy presentation of the final output map on a simple rectangular shape.

8.4.3.4 Map size (number of nodes)

The number of map vectors used to represent the input data (and therefore the size of the output SOM) is an important choice to be made by the user. The size of the output map will affect the final visualisation of the SOM, including the level of information extracted, as each node of the output map represents a characteristic pattern from the input data (Vesanto, 2000; Liu et al., 2006). Important differences between data items may be missed if the map size is too small, and yet distinctions between map vectors may be insignificant if the map size is too large (Cereghino & Park, 2009). This is analogous to common model issues of oversmoothing and overfitting.

The number of nodes also influences the applicability of the SOM for either clustering or visualisation, with a smaller number of nodes producing larger clusters, and a larger map size leading to a more spread-out visualisation of the topological structure of the data (Flexer, 1999). Liu et al. (2006) evaluated the sensitivity of SOMs to parameter selection and determined that larger maps lead to more accurate results by virtue of less pattern smoothing;

the extracted patterns are more similar to the actual patterns in the data. The resolution of the map determines which clusters become visible (Kohonen, 2013). A larger map size may produce a finer distinction between structures in the data and more accurate local estimation (Principe et al., 1998; Kohonen, 2013; Wang et al., 2013), yet larger maps may contain nodes with no data matching them, indicating that the patterns represented by these node vectors do not actually exist in the data set (Clark et al, 2017). Smaller maps compress the data into a smaller (possibly more manageable) number of patterns for analysis. Therefore, a trade-off exists between the accuracy (of representation of the data vectors) and generalization of the extracted information, when deciding on the number of map nodes to use.

In Figure 38, maps of differing sizes are shown over the same data set, with data items matching the same node coloured the same colour. This shows the range of data represented by a single node of the smaller maps compared to the larger ones. Clark et al. (2016) further discuss the implications of large ranges of data represented by individual nodes.

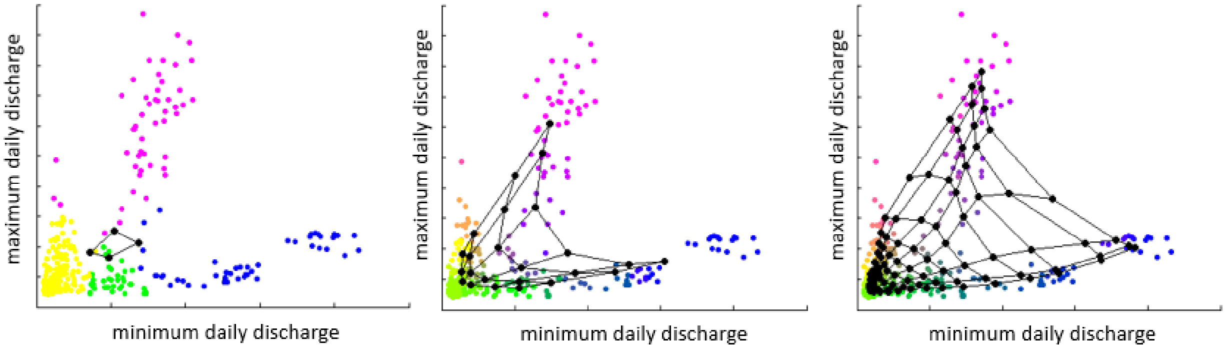


Figure 38: Maps of differing sizes (4, 24, and 88 nodes) representing the same data set. Data items represented by the same node are coloured the same.

Kohonen (2013) states that though the choice of map size is the most common question asked with regards to SOMs, it is not possible to determine it beforehand. The choice is often made with the use of quality measures or selecting the map that the user finds most interpretable - these methods require estimating parameters through training the map multiple times and comparing the results (Astudillo & Oommen, 2014; Cereghino & Park, 2009).

SOM software generally specifies a default value for the number of map nodes based only on the number of input samples available (see 'heuristics' below), however this method does not consider the possible cluster structure in the specific data set nor any user requirements for visualisation or analysis. In many cases, the number of nodes is set by estimating the cluster structure of the data set (see Section 0) and equating the number of nodes to the number of expected clusters.

Quality measures

Map size is often determined based on quality assessments on a series of output maps, quantifying the accuracy of the maps in describing the input data (Cereghino & Park, 2009). This method entails minimising some combination of error measures over the set of maps based on the primary objective, or set of objectives, of the user. The necessity for more than one quality measure arises from the two competing goals of the SOM algorithm: the approximation of the input data by the map vectors, and the preservation of the input topography by the interconnected grid of map vectors (dimension reduction and visualisation).

A number of quality measures can be used. The most common are the:

- quantization error (QE, Kohonen, 1995) - a measure of the ability of the SOM to represent the input data. QE quantifies map resolution, measuring how closely the map vectors match the data vectors; and
- topographic error (TE, Kiviluoto, 1996) - a measure of the preservation of the topology of the input data structure on the output map.

A combination of quantization error and topographic error is often used. Due to the trade-off between vector quantisation and topology preservation, as the QE decreases, TE will generally increase (though not always), so the user must determine the desired balance between them. Care must be taken with the selection as Fyfe (2008) states 'sometimes the two conflicting criteria produce a visualisation which does not accurately reflect all the features of the data'. Figure 39 shows the use of plots of QE and TE vs number of map nodes (assuming side ratios as described below under the 'map shape' heading) to aid the choice of map size, requiring a compromise to be made between the competing processes.

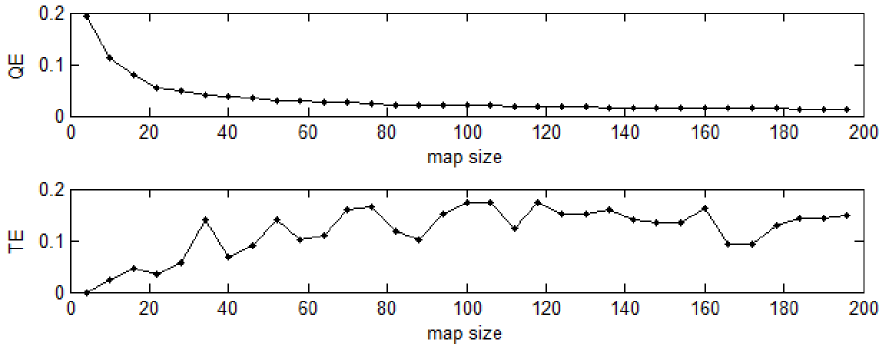


Figure 39: QE and TE plots (shown here for the data set from Figure 35) can be used to aid map size selection. Generally, QE decreases and TE increases with increasing number of map nodes, requiring a compromise to be made.

Quantization error: The SOM algorithm chooses the BMU for each input by minimizing squared Euclidean distances between the input items and map nodes. The QE is the difference between each data point, x_i, and its closest map unit, m_c, averaged over all data points:

$$QE = \frac{1}{N}\sum_i \|\mathrm{m_c} - \mathrm{x_i}\| = \frac{1}{N}\sum_i \sqrt{(m_c{}^2 + x_i{}^2 - 2m_c x_i)}$$

The optimal map for representing a data set, in terms of vector quantization, yields the smallest quantization error. QE is useful for comparing the SOM to other clustering or vector quantization methods, though it cannot be used to compare maps of different sizes (unequal numbers of map nodes) as QE will decrease as map size increases, nor for comparing maps with different neighbourhood shapes since it favours maps with specific neighbourhood radii (Kaski & Lagus, 1996).

Topographic error: The proportion of nodes for which the first and second best matching map units are not nearest neighbours on the map grid is summed over all inputs. For each data point, the BMU and second BMU are checked to see if they are adjacent:

$$TE = \frac{1}{N}\sum_{i=1}^{N} u_{x_i}$$

where $u_{x_i} = 1$ if the first and second BMUs of x_i are neighbours, 0 otherwise. One topographic error value represents the entire map. A value of zero indicates perfect topology preservation. Note that the topographic error does not consider diagonal neighbours of the rectangular lattice, and so a hexagonal lattice gives a lower TE due to having more neighbours for each unit (Pena et al., 2008).

The distortion measure (DM, Kohonen, 1995) is often encountered in the literature in discussions on parameter selection, though is seldom actually used as an error measure. It is included here for information purposes.

Distortion measure: A measure of the pull that the data items are exerting on the map, the DM can be thought of as either: the amount that each map unit is pulled towards its influencing data points summed over all the map units, or the amount each data point pulls on each of the map nodes, combined. Distortion measure incorporates the neighbourhood function into the calculation of distances between each map unit and each of the data points:

$$\mathrm{DM} = \sum_{\mathrm{i}=1}^{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{M}} \mathrm{h_{ij}}\left\|\mathrm{x_i} - \mathrm{m_j}\right\|^2$$

where h_{ij} is the value of the neighbourhood kernel centred on the BMU of x_i at the location of m_j. The distortion measure differs from the quantization error in that each squared distance is weighted by the value of the neighbourhood function. QE and DM are equivalent when the neighbourhood size includes only a single node. Distortion is larger for larger neighbourhood sizes. As the neighbourhood function and the distances from each data item to its BMU decrease with each training iteration, the distortion measure decreases as training progresses. Eventually the plot of distortion flattens out - the map is still distorted but no longer updating. DM is useful for comparing maps of equal size, but not for comparing between differing map sizes.

Heuristics

Heuristics, or rules of thumb, are commonly used to choose map size as they provide easy and quick results. The most commonly used heuristic for determining the number of map nodes, M, recommends that it should be approximately $5\sqrt{N}$ where N is the number of samples in the input data set (Vesanto, 2000). Though this method relates map size only to the amount of input data and not to the actual data values or structure, it is the default method used in the MATLAB SOM Toolbox code. Other heuristic recommendations include: ‘the number of neurons [nodes] should usually be as big as possible’ (website 4) ‘one should try for about 50 hits per node on average' (Kohonen, 2013); or if an unlimited number of inputs is available, 'one may try to use as big an array as one is able to compute' (Kohonen, 2013). It is noted that each of these methods contrast (sometimes greatly) with each other, however their use remains popular in the literature.

Cluster structure

Map size can also be specified based on the number of clusters that are determined to exist in the input data, attempting to provide one node to represent each estimated cluster. An alternative would be to create a series of different sized maps, choosing the one that produces the lowest cluster validation measure. The existence and number of clusters can be determined through cluster theory, as discussed in Section 0.

In practice, the most popular, currently used method for map size selection in environmental application papers is the production of a series of maps of different sizes, followed by a graphical or visual comparison of the output maps. There is no overall preferred method of evaluation shared by all SOMs users, though the choice appears to be most commonly based on the degree of generalisation and number of clusters desired in the output. Next in popularity is the use of a combination of quality measures, then default software heuristics, followed by a variety of individual ad-hoc methods. Many papers do not give any information about the rationale of map size selection. Inclusion of this information would allow the reader to understand if there is a reasonable basis to believe the number of nodes accurately represents the cluster structure of the data set or if other map sizes may reveal a different cluster structure.

8.4.3.5 Map shape (ratio of grid side lengths)

The best representation of the data will be obtained when the shape of the grid roughly corresponds to the shape of the data structure. For example, a two-dimensional square-shaped data set would not be best represented by a rectangular grid (with one direction much longer than the other), as illustrated in Figure 40.

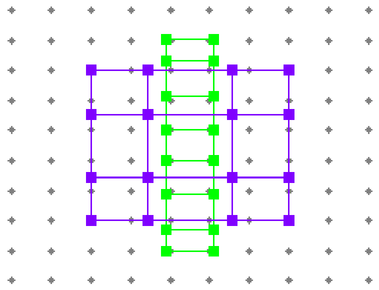


Figure 40: Choosing map shape to correspond with the data structure: a two-dimensional square set of regularly spaced data (grey dots) are represented with two 16-node SOMs (4x4 (purple) and 8x2 (green)). It is apparent that the purple SOM will provide a better representation of the data as the shape corresponds better to the shape of the data manifold.

Establishing the ratio of map side lengths based on the main intervariable relationships in the input data is the method recommended by Kohonen (2001), and is the most commonly used procedure. This is done by initialising the axes in alignment with the most important linear correlations, given by the first and second principal components (further discussed in Section 8.4.4). Through training, the axes will be bent and stretched, eventually coming to follow the most important nonlinear correlations in the data. The primary axis will represent the most significant relationship between data dimensions (the nonlinear line of best fit), and the secondary axis the next most important relationship.

This method of assigning the side length ratio is performed with the following steps (Vesanto, 2000):

1. Determine the eigenvectors and eigenvalues in the data from the autocorrelation matrix,
2. Set the ratio between the two sides of the grid equivalent to the ratio between the two largest eigenvalues, and
3. Scale the side lengths so that their product ($L_1 x L_2$) is as close as possible to the number of map units determined above.

8.4.3.6 Training parameters

8.4.3.7 Neighbourhood function

The neighbourhood function, $H = [h]_{ij}$, is a smoothing kernel applied to the map grid during training, as described in Section 8.3.1. The kernel controls the smoothness and generalisation of the mapping by defining its rigidity. A matrix item, h_{ij} , is the value at m_j of the neighbourhood kernel centred on the BMU of x_i. For example, the value of $H_{2,5}$ in Table 12 gives the influence on node 5 of data for which the BMU is node 2.

Table 12: Sample neighbourhood matrix, H (for a 3x3 map grid with a Gaussian shaped kernel of radius 1). The neighbourhood matrix does not depend on the data and can be calculated beforehand.

H								
0.33	0.16	0.04	0.16	0.08	0.02	0.04	0.02	0.01
0.20	0.26	0.20	0.10	0.12	0.10	0.02	0.04	0.02
0.04	0.16	0.33	0.02	0.08	0.16	0.01	0.02	0.04
0.20	0.10	0.02	0.26	0.12	0.04	0.20	0.10	0.02
0.12	0.16	0.12	0.16	0.20	0.16	0.12	0.16	0.12
0.02	0.10	0.20	0.04	0.12	0.26	0.02	0.10	0.20
0.04	0.02	0.01	0.16	0.08	0.02	0.33	0.16	0.04
0.02	0.04	0.02	0.10	0.12	0.10	0.20	0.26	0.20
0.01	0.02	0.04	0.02	0.08	0.16	0.04	0.16	0.33

Each column of H represents the influence on map node m_j of the data items matching all map units m_i (each in a separate row). The influence that the data in the Voronoi set of m_i exerts on other nodes is given in the rows. Each row consists of values for a surface with a peak at $m_i = m_j$. This indicates that data with BMUs closest to node m_j will have the most influence on the updating of m_j. The neighbourhood kernel is generally normalised so that each column sums to 1, equalising the sum of the influence exerted by all data items on each map node.

At each training iteration, the location of each map node, m_j, is updated based on all the data items that are matched to nodes within the specified neighbourhood radius centred at this node. The neighbourhood size, and therefore the extent of influence of the data items on the map units, decreases linearly over the training iterations, though the shape of the neighbourhood remains constant throughout training. This decrease in kernel size leads to an increased smoothing of the map. The size and shape of the neighbourhood kernel must be determined by the user.

Neighbourhood size: A large neighbourhood kernel results in a stiff map by overstressing topological ordering, and a small kernel results in freer movement of the nodes toward the data (Vesanto et al., 2003). The increased topologocial ordering of maps created with larger neighbourhoods comes at the expense of data quantisation (how close the nodes are to the data they represent), which improves as neighbourhood size decreases. For this reason, a compromise is made - a large neighbourhood kernel is used at the beginning of training to induce a global ordering of the map nodes, and the kernel diminishes in size with each training iteration. The node locations are eventually finetuned within a small neighbourhood at the end of training. The starting and finishing neighbourhood sizes can be specified by the user.

The neighbourhood radius is measured in map space, not data space. Figure 41 shows the group of nodes contained within a neighbourhood of radius 0, 1 and 2 around a node of a hexagonal and rectangular grid. On a hexagonal grid, a neighbourhood kernel of radius 2 will incorporate data from 19 nodes, whereas on a rectangular grid the same size kernel would incorporate data from 13 nodes.

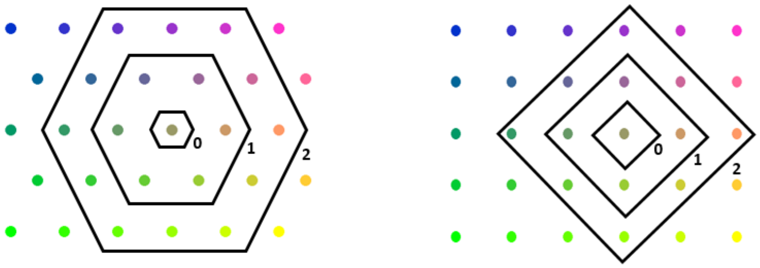


Figure 41: Neighbourhood sizes on hexagonal and rectangular grids (after Vesanto et al., 1999). Data have been coloured by 'similarity colouring' (see Section 8.5.3.1). The same size neighbourhood (>0) will encompass more map nodes on a hexagonal grid than on a rectangular grid.

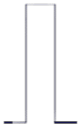
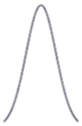
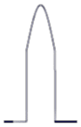
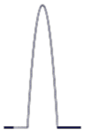
The choice of initial neighbourhood size has an impact on the results. If it is too small, the map may not achieve an appropriate global (overall) ordering (Kohonen, 1990; Hastie et al., 2007). Kohonen (2001) recommends setting a starting neighbourhood approximately half the largest side length of the map to prevent the risk of ending in a local minimum.

The final neighbourhood radius usually only includes a single node (Kohonen, 1993). The map loses its spatial interaction at this point, and the SOM becomes equivalent to k-means clustering (Hastie et al., 2007). If global ordering has been successful, this will still produce the desired results as the grid connections are maintained. Kohonen (2005) explains that equating the SOM to the k-means algorithm at the end of training (through diminishing the neighbourhood to include only a single node) guarantees the most accurate approximation of the probability density function of the input and should also eliminate any issues the neighbourhood function may encounter at the borders of the map.

Neighbourhood shape: Four shapes are commonly used for the nieghbourhood function: uniform, Gaussian, cut (truncated) Gaussian, and Epanechnikov (parabolic). These shapes are used in the MATLAB SOM Toolbox, with Gaussian as the default (Vesanto et al., 2000b).

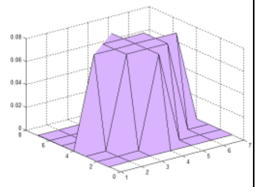
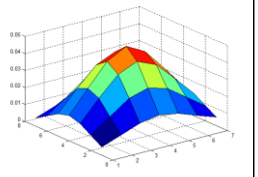
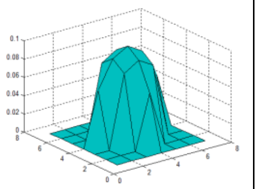
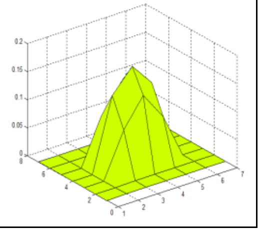
Table 13 describes the four shapes (where σ_t is the neighbourhood radius at iteration t, $d_{ci} = ||r_c - r_i||$ is the distance between map units m_c and m_i on the map grid, and $1(x)$ is a step function (taking a value of 0 if $x < 0$ or 1 if $x \geq 0$) updating only the nodes for which the function is nonzero.

Table 13 Neighbourhood function shapes

Shape	Function	Description
Uniform (bubble) 	$h_{ij}(t) = 1(\sigma_t - d_{ij})$	The value of $h_{ij}(t)$ is 1 if the distance between the units is less than or equal to the neighbourhood radius at iteration t, otherwise it is 0. All map units within the radius are updated the same amount.
Gaussian 	$h_{ij}(t) = e^{-d_{ij}^2 / 2\sigma_t^2}$	The Gaussian kernel updates all map nodes (not just inside the radius) by an amount descending from the kernel centre to the edge of the map.
Cut Gaussian 	$h_{ij}(t) = e^{-d_{ij}^2 / 2\sigma_t^2} 1(\sigma_t - d_{ij})$	The 'cut Gaussian' kernel is same shape as Gaussian, but does not update nodes outside the radius boundary.
Epanechnikov 	$h_{ij}(t) = max\{0, 1 - (\sigma_t - d_{ij})^2\}$	This kernel only produces values greater than zero when the function is between -1 and 1. Therefore only nodes that are within a radius of 1 from the kernel centre are updated.

The Gaussian kernel is the only one that incorporates the entire input data set to update each of the map nodes (and conversely uses each data item in the updating of all of the nodes); the other three kernels only update within the specified radius. The Gaussian neighbourhood therefore produces the smoothest SOM patterns (Liu et al., 2006), whilst the others have various degrees of smoothing (Epanechnikov the least) for a fixed σ_t common to all kernels. Table 14 gives sample values of neighbourhood kernels for each shape centred around the middle nodes of a 3*3 and 7*7 SOM. Higher values indicate which nodes would be updated (influenced by the data) by a greater amount.

Table 14: The effect of neighbourhood shape and size on map updating. The degree of updating of each map node is specified for a neighbourhood kernel centred on the centre node of a 3x3 and 7x7 SOM. Values have been normalised for comparison.

Shape	3x3 map (rectangular grid)			7x7 map (rectangular grid)	
	$\sigma_t = 0$	$\sigma_t = 1$	$\sigma_t = 2$	$\sigma_t = 2$	kernel shape
Uniform (bubble)	0 0 0 0 1 0 0 0 0	0.0 0.2 0.0 0.2 0.2 0.2 0.0 0.2 0.0	0.11 0.11 0.11 0.11 0.11 0.11 0.11 0.11 0.11	0 0 0 0 0 0 0 0 0 0 0.08 0 0 0 0 0 0.08 0.08 0.08 0 0 0 0.08 0.08 0.08 0.08 0.08 0 0 0 0.08 0.08 0.08 0 0 0 0 0 0.08 0 0 0 0 0 0 0 0 0 0	
Gaussian	0 0 0 0 1 0 0 0 0	0.08 0.12 0.08 0.12 0.20 0.12 0.08 0.12 0.08	0.10 0.12 0.10 0.12 0.13 0.12 0.10 0.12 0.10	0.005 0.009 0.013 0.015 0.013 0.009 0.005 0.009 0.017 0.025 0.028 0.025 0.017 0.009 0.013 0.025 0.036 0.041 0.036 0.025 0.013 0.015 0.028 0.041 0.047 0.041 0.028 0.015 0.013 0.025 0.036 0.041 0.036 0.025 0.013 0.009 0.017 0.025 0.028 0.025 0.017 0.009 0.005 0.009 0.013 0.015 0.013 0.009 0.005	
Cut Gaussian	0 0 0 0 1 0 0 0 0	0 0.18 0 0.18 0.29 0.18 0 0.18 0	0.10 0.12 0.10 0.12 0.13 0.12 0.10 0.12 0.10	0 0 0 0 0 0 0 0 0 0 0.06 0 0 0 0 0 0.08 0.09 0.08 0 0 0 0.06 0.09 0.10 0.09 0.06 0 0 0 0.08 0.09 0.08 0 0 0 0 0 0.06 0 0 0 0 0 0 0 0 0 0	
Parabolic (Epanech-nikov)	0 0 0 0 1 0 0 0 0	0 0 0 0 1 0 0 0 0	0.08 0.13 0.08 0.13 0.17 0.13 0.08 0.13 0.08	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.08 0.13 0.08 0 0 0 0 0.13 0.17 0.13 0 0 0 0 0.08 0.13 0.08 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	

Erwin et al. (1992) found that the SOM's convergence rate is heavily dependent on the shape of the neighbourhood function, and the training algorithm is more effective when a convex neighbourhood function is used rather than a concave one. Ota et al. (2011) describe the use of an asymmetric neighbourhood function to remove topological defects which frequently emerge during training and inhibit the global ordering of the map.

Training length

The final statistical accuracy of map (how well the data is represented) depends on the number of iterations, since learning is a stochastic process (Kohonen, 1990). However, with modern computational resources this should no longer be an issue and software defaults should be adequate. There is no upper limit to the number of iterations that can be used. In principal, the global ordering stage with a large neighbourhood radius can be relatively short compared with the finetuning stage with the smaller neighbourhood radius.

Mask

A 'mask' may be applied during the map training process to weight the influence of each variable in the distance calculations for determining BMUs. The mask is a vector with the same number of dimensions as the input data. Mask values indicate the relative importance of each variable (usually with 0's and 1's). It can be used to 'hide' certain variables, or make others more influential in map training if the user would like to accentuate the significance of certain variables over others. Note that the mask is only used for finding BMUs, and is not used in the initialisation stage.

Associated variables

Introducing a new set of variables onto a trained SOM that has been created with a different set of variables may enable researchers to discover interesting intervariable relationships (Cereghino & Park, 2009). These new variables are known as 'associated variables' as they are linked to the map after training. The trained map can be clustered or labelled based on values of the associated variable to visualise the relationships. See Deboeck (1998) for more information on associating variables.

8.4.4 Step 4: Initialise the map

The map is generally initialized with a regular linear array set in the directions of highest variance of the input data vectors, as determined with principal component analysis. The initial values of the map weight vectors are set at uniform intervals along the first and second principal components of the input data set, which come to form the axes. If the axes lengths are proportional to the two largest eigenvectors of the data, this should produce an approximately uniformly-spaced lattice. This linear form of initialisation is usually used as it ensures the map is already aligned with the most significant linear intervariable relationships before map training begins. Figure 42 illustrates the alignment of the axes with the principal components.

As SOM training is an iterative process of multi-dimensional nonlinear optimisation, it has the potential to lead to multiple optimal solutions (Kingston et al., 2005), meaning that for the same input data the possible output maps include rotations or inversions of each other. The orientation of the final map is dependent on the initial values assigned to the nodes, with different sets of initial node locations leading to rotations, mirror images, or symmetric inversions of the final map (Kohonen, 1990). Yin (2008) highlights the need for good initialisation as it 'can help guide to a faster or even better convergence'.

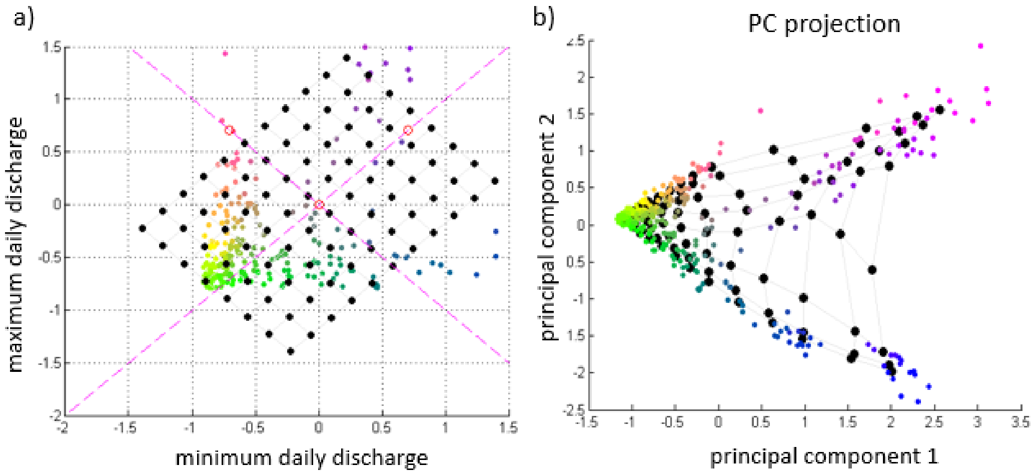


Figure 42: Linear initialisation of the data (coloured dots) from Figure 35. a) The initialised map grid (black dots) follows the principal components of the data set (main directions of variance as shown with magenta dashed lines). b) The trained map grid is shown on the principal component projection (axes of this plot are the first and second PC). Though the map has been fitted to the data, the main directions of the grid remain generally aligned with the principal components.

8.4.5 Step 5: Run the training algorithm

Running the training algorithm through the iterations of the two-step matching and updating process re-organises the linearly initialised grid of map nodes into a nonlinear arrangement amongst the data items, while maintaining the grid connections. Hastie et al., 2009, describe the initialised map nodes as 'buttons' sewn onto the principal component plane in a regular pattern; the training process of the SOM then bends and twists the plane so the buttons best approximate the data. The map is smoothed by the updating process in which the new node locations are computed based on the previous locations and the locations of the data items. Through the maintained grid connections, the map nodes organise themselves based on their similarity to each other.

The trained map will now better approximate the data set than the initialized map, with more map units positioned in areas of higher density input space. The vectors describing the location of each map node (of the same dimensions as the input data vectors) have come to represent the most prevalent patterns of unique variable combinations in the data.

At this stage, the algorithm can be run for a number of parameter sets (map size and shape, neighbourhood kernel size and shape), and the results compared through the use of quality measures (Section 8.4.3) to choose the map that best preserves the topology, quantisation, or clustering of the input data, or any combination of user objectives.

8.4.6 Step 6: Place data items on the map

After the map has been created, each data item finds a place on the map by matching it to its closest (most similar) map node. Because the nodes are organized based on their similarity, similar input data will become mapped to the same, or nearby, map nodes. The matching is based on a high-dimensional similarity measure, usually Euclidean distance. As in the matching stage of map training, each data item will have a unique best matching map node, but each map node may be matched to more than one data item, or none at all. Placement of the data items on the map leads to the identification of clusters and discovery of relationships between data items. Section 8.5 describes how to interpret the map once this stage is reached.

8.5 INTERPRETATION OF A SELF-ORGANIZING MAP

The aim in analysing a SOM is to identify the key characteristics comprising the predominant patterns in the data set (pattern extraction) and discover which data items are similar to each other with respect to these characteristics (clustering). Characteristics of the predominant patterns are revealed by the high-dimensional vectors associated with each map node. As the map becomes organised in data space during training, the location of each map node is defined by the combination of dimension values that make up its vector. As the data items are matched to their nearest map nodes after training, clusters are formed of data items sharing similar characteristics. These characteristics are identified by the pattern of the common node. Map nodes may also be grouped together to form larger clusters of data in their Voronoi sets.

Interpretation of a SOM generally includes a compilation of information through the visual investigation of the labelled map in one- or two-dimensional output space and the component plane for each variable. The visual investigation will reveal the prevalent patterns indicated by the individual node vectors, and the clusters of the data nearest to each node. Investigation of the output map should also disclose a good approximation of the input data distribution, including the overall shape and cluster structure in the data, characteristics of the clusters, and the relationships between variables. The results will allow for trend visualisation and infilling of missing data.

8.5.1 Visualisation

8.5.1.1 Output map

The output map is usually displayed as a regularly-spaced grid in one- or two-dimensional map space, labelled with the data items or the main characteristics of the data items that pertain to each node. The output map may also be displayed in data space if the dimension of data space is low. In map space, the distribution of the projected data items across the map is evident, whereas in data space, the distribution of the map nodes amongst the data items is evident. The differences between nodes are often shown by coloured markers or a surface plot.

An output map of the example data from Figure 35 is shown in Figure 43 in both data space and map space. In data space (a), the interconnected map grid is shown in black over the coloured data points, revealing the placement of the nodes amongst the data. In map space, the nodes are coloured by similarity with empty nodes remaining white (b), and data item labels are placed onto the relevant areas of the map (c).

Labels

Labelling the nodes on the output map (as in Figure 43c) gives an indication of the characteristics of data items represented by each region of the map. Labelling can involve listing each data item over the node it is assigned to, or choosing a representative label for each node based on the group of data it represents. In the latter case, nodes may be grouped into larger clusters before labelling based on predominant cluster attributes.

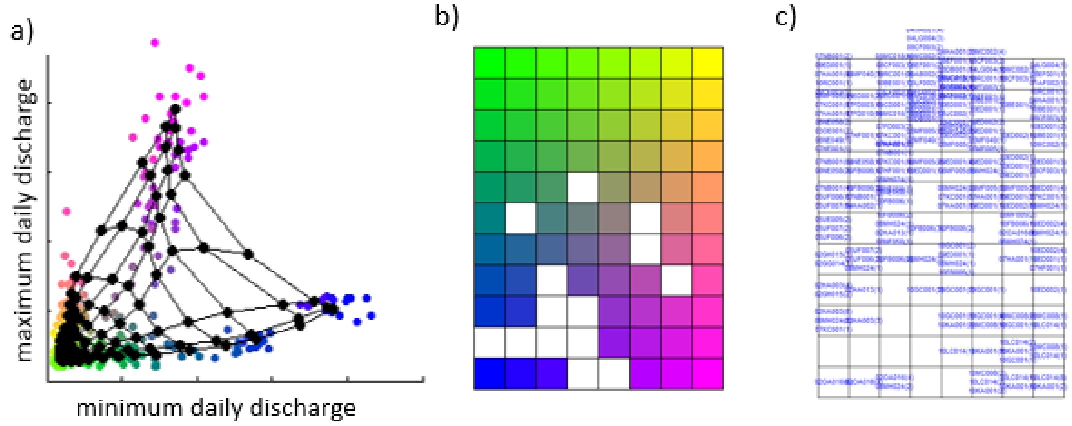


Figure 43: The output map is shown in data space (black grid, in (a)) and map space (b and c). Nodes are coloured in (b) to indicate similarity to each other, and each data point in (a) is the colour of the node it matches (in (b)). Labels (on c) indicate the placement of each data item on the map. Clusters and gaps in the data structure are discernible on this figure.

To label the nodes or clusters with representative labels in consideration of the cluster centroid (map node) variable weights, plots of cluster vs. dimension and dimension vs. cluster are useful to indicate which dimensions are most accounted for in each cluster. Figure 44 demonstrates how these plots can be used, with a 24-dimensional, 8-cluster example. In the plot on the left (a), each line represents a separate cluster with the y-axis indicating the values (weights) of each dimension at each cluster centroid. It is possible to pick out, for example, that dimension 19 has low weights in all clusters, whereas dimension 4 is prominent in 3 separate clusters. On the right (b), each line represents a dimension, with the y-axis indicating the dimensions' value at the cluster centroid for the 8 clusters of this example. It is possible to pick out, for example, which dimension has by far the largest weight in cluster 6. Wang & Feng (2011) use this method to choose the top three weighted dimensions for labelling each cluster.

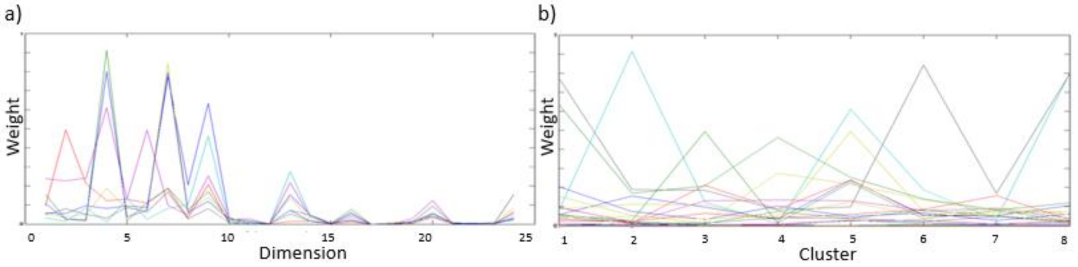


Figure 44: Plots to aid in labelling clusters on the map. a) The clusters are represented by coloured lines (8 clusters in this example), and 8 values are shown for each dimension (one for each cluster). This indicates which dimensions are prominent in certain clusters. b) Each dimension is represented by a coloured line, showing the weight of each dimension at each cluster centroid. The highest peaks in each cluster indicate the dimensions of greatest influence, which could be used for cluster labelling.

Another option is to label a map with associated data (new data that has not been used in the training) as discussed in Section 8.4.3. This shows where the new data would plot on a map trained with other data, defining the relationships between the data sets.

Axes

Through the self-organisation process, the axes of the output map establish a meaningful nonlinear coordinate system for the various features of the input data (Kohonen, 2001). The axes begin the training stage as the first and second linear principal components of the data set, and then gain nonlinearity as iterations progress. While Vesanto (1999) states that the 'axes of the map grid rarely have any clear interpretation', it is possible to form a general perception of their meaning through investigation of the component values of the map vectors along the edges of the map. This can be done by careful analysis of the component planes.

8.5.1.2 Component planes

Relationships between the individual variables can be explored with the use of component planes. Colouring is used to indicate dimension weights (values) at each node, with a separate component plane displaying values of each dimension of the SOM. The axes and grid nodes correspond exactly to those of the SOM output map.

Inspection of component planes indicates the spread of values in each dimension (Vesanto, 1999). The presence of interesting relationships between variables can be visually determined from the component planes, allowing these relationships to be further investigated with scatterplots of the subset of variables of interest. Plotting component planes of associated variables (those not used in map training) shows the relationship of new variables to those used to create the map.

In Figure 45, component planes produced from the SOM in Figure 43 indicate that data items located on the lower right of the map have high maximum discharges and medium minimum discharges, whereas those located in the lower left have more moderate maximums and relatively high minimums.

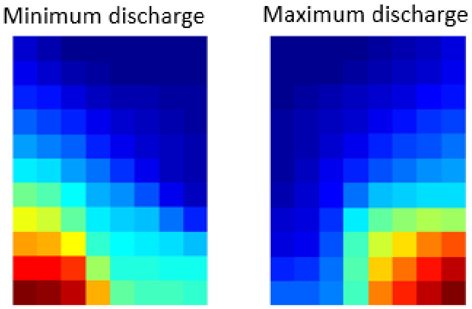


Figure 45: Component planes show the relative values of each variable at each node (high values are red, low are blue). Node locations correspond exactly to those on the SOM in Figure 43. As the values of each variable are revealed on the different regions of the component planes (and therefore on the corresponding regions of the map), the nonlinear intervariable relationships become evident.

The component planes provide a meaningful interpretation of the axes. The axes of the SOM follow the main nonlinear directions of variance in the data set, and by identifying regions of the component planes with high and low values of each variable, the general gradient of individual variables along the axes should become evident. These gradients will be continuous along the axes, though not necessarily monotonic in direction.

8.5.2 Pattern extraction

The prevalent patterns in the data set are exhibited by the vectors of each map node. The unique combination of variables making up each vector is a characteristic pattern. These combinations can be analysed via the component planes.

An analysis of frequencies of occurrence of each pattern is obtained by looking at the percentage of input data assigned to each map node. Each matching of a data item to a node is known as a 'hit'. Transitions between patterns can be observed by matching the data items to the map in a sequential order and following the trajectories of the hit locations.

8.5.3 Cluster identification

Clustering is the most frequent reason for implementing a SOM (Agarwal & Skupin, 2008), and a number of methods exist for finding clusters in the data through the use of a SOM.

Basic, or first-level, clustering involves treating each SOM node as a cluster centroid. As each input data item is uniquely related to one of the nodes, clusters are created of data that share similarities based on the features of the extracted patterns. Each node of the output map comes to represent a cluster. To use this method effectively, it is best to determine the approximate number of clusters that exist in the input data (using a cluster validity measure, see Section 0) before setting the number of nodes. The number of nodes should be set equal to or greater than this to ensure that each cluster is mapped to a separate node.

Second-level clustering is used to find groups of nodes that themselves make up a cluster. This is useful when creating a map with a large number of nodes compared to the number of clusters in the data (which may be done to gain a good separation between data items on the map). It is known as second level clustering since the data has already been clustered with the SOM (first-level clustering). Second-level clustering is useful for producing summaries or descriptions of SOM results.

Second-level clustering groups the SOM nodes, either with another SOM or with a different technique as discussed below. The high-dimensional node vectors are clustered in data space, and the cluster memberships are projected onto the low-dimensional map for visualisation, with the subsets of nodes grouped together by colour or outlines. All the data matched to any of the member nodes are now members of each cluster (Vesanto & Alhoniemi, 2000). As might be expected, the weights of the most influential variables will change rapidly at the borders of the clusters (Kaski et al., 1998). Toth (2009) finds second-level clustering of a larger SOM to be more suitable for preserving the distinctive features of the classes when compared with using a smaller initial map size.

Some common methods of second level clustering include visually assessing the distance in data space between nodes with the U-matrix, similarity colouring, or the number of 'hits' in each region; another SOM; partitive clustering such as k-means (as in Skupin, 2004; Wang et al., 2013); and hierarchical agglomerative clustering (as in Wang & Feng, 2011; and Guo et al., 2006). The remainder of this section will describe these techniques.

Cluster validity measures

If attempting to have each node of the map represent a cluster of the data, it is necessary to estimate the number of clusters present in the data set before the SOMs analysis begins. There are various methods available for this, which can also be used to determine k in k-means clustering.

The Davies-Bouldin index (Davies & Bouldin, 1979) is a popular measure for determining the number of clusters present in a data set (as in Gamble & Babbar-Sebens, 2012; Garcia & Gonzalez, 2004) and used in kmeans_clusters.m of the MATLAB SOM Toolbox (Vesanto et al., 2000). It measures the ratio of within-cluster to between-cluster distances. A small ratio indicates compact, well-separated clusters.

The Dunn index (Dunn, 1973) and Silhouette coefficient (Rousseeuw, 1987) are also used to determine the number of clusters (as in Sarlin & Yao, 2013). Both are also concerned with the ratio of inter-cluster to intra-cluster distances. Silhouette clustering determines how appropriately the data has been clustered by ranking data points as: well clustered (1); would be better in neighbouring cluster (-1); or on the border of two natural clusters (0), and taking the average over the entire dataset. Clusters with narrower silhouettes than the rest will appear if there are too many or too few clusters.

If the data items are in k compact, separated clusters, a function based on within- or between-cluster distances could be expected to decrease rapidly as the number of nodes increases, until the number of nodes equals k (Flexer, 1999). The 'kink' in the curve of cluster dissimilarity as a function of the number of clusters, where the curve begins to decrease less rapidly, may indicate the number of clusters in the data (Hastie et al., 2009).

8.5.3.1 Visualising the space between clusters

Unified distance matrix (U-matrix)

The U-matrix (Ultsch, 2003) visualises distances between regions of the data space represented by each node. By computing high-dimensional similarities between neighbouring nodes (how close they are in data space), the U-matrix determines cluster boundaries based on large dissimilarities (Ultsch, 2003). This is also known as the 'degree of distortion', that is the change in relative distance between the high-dimensional locations of the nodes in data space and the low-dimensional map representation (Agarwal & Skupin, 2008). The distances are indicated on the U-matrix map by colour, with differing colours indicating the boundaries between clusters. It can be seen on the U-matrix in Figure 46(a) that distinct clusters (blue) exist, separated by lighter coloured boundary areas.

Similarity colouring

Similarity colouring involves spreading a two-dimensional colourmap over the principal component projection of the nodes, thereby colouring similar nodes similar colours. Further apart (more different) nodes become coloured with colours that are perceived as more distinct (Kaski et al., 2000). The similarity colouring and 'empty' map nodes in Figure 46(b) reveal the same cluster structure as the U-matrix. On this plot, the clusters evident in the lower portion of the map are outlined.

Hits

Plotting the number of data items matched to each node, known as the 'hits', on the output map may allow cluster structure in the data to become visible. High intensities of data might become evident on clearly separated regions of the map. Nodes with zero (or relatively low) hits delineate the cluster borders (Zhang & Li, 1993; Vesanto & Alhoniemi, 2000).

Using a surface plot to record the hits will produce raised regions of the map where the clusters exist (as in Gopakumar et al., 2005). Linearly scaling the size of the output map nodes in proportion to the number of hits each receives provides another method of visually indicating cluster structure. Figure 46(c) shows the use of hits to confirm the cluster structure.

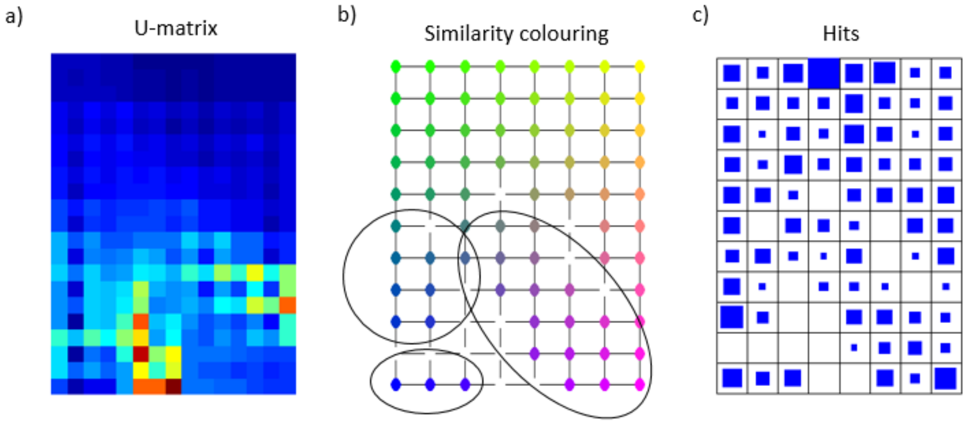


Figure 46: U-matrix, similarity colouring, and hits. The cluster structure evident on all three plots is delineated on plot b. a) Colouring is based on the distance between neighbouring nodes in data space. Dark blue indicates smaller distances and red indicates greater distances, therefore the yellow, orange and red regions indicate boundaries between clusters. b) Similarity colouring visually indicates similarities between nodes, with empty (white) nodes forming boundaries between clusters. c) The density structure of the data is shown by sizing map nodes based on the number of data items matched to them.

K-means clustering

K-means clustering is a popular method for second-level clustering as the topological preservation of the data has already been captured in the first SOM (as in Garcia & Gonzalez, 2004). K-means clustering produces good results if the clusters are compact, hyper-spherical and well-separated (Garcia & Gonzalez, 2004). Determining the optimal 'k' value, or number of clusters to extract, can be done with the methods in Section 0.

SOM for second level clustering

A SOM can be used for second-level clustering (as in Clark et al., 2014). In this method, the node vectors of the first SOM become the input vectors for the second SOM. An example is shown in Figure 47 in which a 5x2 SOM is used to cluster the output of an 18x18 SOM, reducing the number of clusters from 324 to 10. Results produced are similar to k-means clustering with the added benefits of maintaining an order to the clusters and allowing presentation in the familiar SOM output.

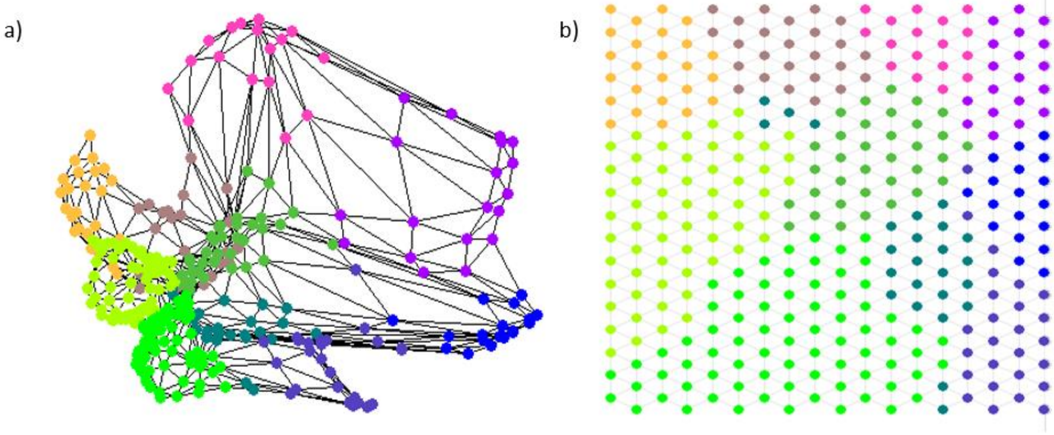


Figure 47: Second-level clustering: an 18x18 SOM is clustered into 10 clusters via a 5x2 SOM, grouping similar nodes together. Each colour represents membership in one of the 10 second-level clusters. a) the map nodes are shown in a two-dimensional principal component projection of the high-dimensional space; b) the nodes are displayed in map space.

Hierarchical clustering

Hierarchical clustering (eg. Ward, 1963) can be used to determine many levels of progressively larger clusters on the map. An advantage of hierarchical clustering over k-means is that many nested levels of clusters can be shown simultaneously on one output map (as in Wang & Feng, 2011). In Skupin (2004), five levels of clustering are shown on a single map.

Hierarchical clustering can also be performed based on an associated variable (as in Sarlin & Marghescu, 2011). Another form of agglomerative hierarchical clustering is 'neuron label clustering' used by Skupin et al. (2013) in which neighbouring clusters are merged if they share the top-ranked label term (most influential dimension). This can be repeated for the second-ranked label terms (etc.) to get many separate cluster layers.

8.5.4 Trend visualisation, infilling missing data, prediction, incorporating new data

An indication of temporal changes (trends) in data sets can be gained through SOMS visualisations using any of the following methods: a SOM is trained with all the available input data and changes over time in the data mapping to each node (or second-level cluster) reveal temporal trends in the data structure (as in Wang & Feng, 2011); all the data matching one map node could be used as input for a local prediction model (Vesanto, 1997); trajectories, or lines, connecting the BMUs of consecutive data points in a time series may be used to indicate trends (as described in Principe et al., 1998; see Schreck et al., 2009, for more cluster analysis of trajectory data.); consecutive sections of the input data (ie. years or decades) can be plotted on to the map to visualise changes (as in Wang, 2015); or, temporal patterns may be extracted based on trends of the cluster centroids when a separate map is created with data from each time step (as in Sarlin, 2012).

To use SOMs for infilling missing data, the best matching map node can be found based on the available variables of the data item, and then the value of the missing variable adopted from the node vector (as in Mwale et al., 2012).

A similar method of value adoption based on established intervariable relationships is used for prediction. The best matching map node can be found based on a set of easily predicted variables, and then the value of the unknown variable adopted from the node vector (as in Steynor et al., 2009).

The incorporation of new data onto the trained and clustered map reveals how it relates to the other data items based on the established relationships and clusters. New data can be added to the input data set during Step 6 (placing the data on the map). It will be placed into the established clusters and can be compared to the other cluster members (as in Dejean et al., 2011).

8.6 Conclusion

This paper leads researchers through the creation and interpretation of a SOM relevant to a specific data set, providing a practical guide to understanding meaningful parameter choices and interpreting SOM results for the extraction of interesting information from large sets of data. Information and guidance on the SOM method has been consolidated in this cohesive document to aid readers interested in using SOMs for data-driven exploratory analysis of multivariate, nonlinear data sets.

Though SOMs are widely used and increasing in popularity for environmental applications, uncertainty remains in the SOMs method due to the two inherent competing goals (approximating the data with the map nodes and preserving the topology of the data set) which negate the possibility of specifying a single objective function that the map aims to optimise. This gives rise to complications in parameter specification, as different choices will improve some aspects of the results and possibly inhibit others. Parameter choices impact the formation of the output SOMs, with different maps resulting from the use of distinct parameter sets.

It is therefore important for the analyst to appreciate and understand the parameter options available. A sole reliance on software default parameters may result in maps that are not the most suitable size and shape to represent a particular data set, or the level of smoothing and generalisation provided on the output map may not suit the specific purposes of the analysis. Interpretations of the data based on such maps may reveal less information about the data set than there is potential to uncover, for example with respect to the distributions of individual variables and the intervariable relationships. Analysts must therefore make an informed choice between the options based on the relative benefits of each choice, until an automated method is incorporated into SOMs for this multi-objective parameter selection.

8.7 APPENDIX 1 – MATHEMATICAL FORMULATION OF THE SOM

The mathematical formulation of the SOM training mechanism is detailed here. The calculation of the BMU (finding the nearest map node for each data point), the application of the neighbourhood function (smoothing kernel), the process of updating the location of each node at each iteration and the traditionally used quantisation error and distortion measure are discussed. The matrix calculations used in the MATLAB SOM Toolbox (Vesanto et al., 2000) for computing the BMU, quantisation error and distortion measure, and implementing the neighbourhood function and updating rule are outlined, including the use of the 'mask' for weighting the influence of each variable. This investigation provides the basis for examining possibilities for mathematical modifications to the SOM, and to explore how such modifications might be implemented in MATLAB code.

BMU search and quantization error: The search for BMUs and the calculation of the quantization error require the recurring computation of Euclidean distance between each data point and each map unit at each iteration. To do so, the first dimension of all data points is compared with the first dimension of each map unit, then the second dimension and so on. The mask is incorporated to weight each dimension, usually by a 1 or 0.

The Euclidean distance calculation between high-dimensional data vectors (x_i) and map vectors (m_j): $squared\ distance = (m_j - x_i)^2 = (m_j)^2 + (x_i)^2 - 2(m_j) * (x_i)$ is implemented below in matrix terms where **Map** is a matrix of M d-dimensional map vectors $(m_1, \dots, m_M)$, **D** is a matrix of N d-dimensional data points $(d_1, \dots, d_N)$, and mask is a one-dimensional vector of length d:

$$\begin{array}{llclcl}
\text{Dist =} & \mathrm{Map.}^2 * \mathrm{mask} * \mathrm{ones}(1,\mathrm{N}) & + & \mathrm{ones}(\mathrm{M},1) * \mathrm{mask}' * \mathrm{D}'.^2 & \text{-} & 2\mathrm{Map} * \mathrm{diag}(\mathrm{mask}) * \mathrm{D}' \\
\text{size: [M,N] =} & \text{[M,d] *[d,1]* [1,N]} & + & \text{[M,1]*[1,d]*[d,N]} & \text{-} & \text{2 [M,d]* [d,d]*[d,N]}
\end{array}$$

$$\begin{bmatrix} Dist_{m_1,x_1} & \cdots & Dist_{m_1,x_N} \\ \vdots & \ddots & \vdots \\ Dist_{m_M,x_1} & \cdots & Dist_{m_M,x_N} \end{bmatrix} = \begin{bmatrix} {m_{1,1}}^2 & \cdots & {m_{1,d}}^2 \\ \vdots & \ddots & \vdots \\ {m_{M,1}}^2 & \cdots & {m_{M,d}}^2 \end{bmatrix} * \begin{bmatrix} mask_1 \\ \vdots \\ mask_d \end{bmatrix} * [1_1 \quad \dots \quad 1_N]$$

$$+ \begin{bmatrix} 1_1 \\ \vdots \\ 1_M \end{bmatrix} * [mask_1 \quad \dots \quad mask_d] * \begin{vmatrix} {d_{1,1}}^2 & \cdots & {d_{N,1}}^2 \\ {d_{1,d}}^2 & \cdots & {d_{N,d}}^2 \end{vmatrix} - 2 * \begin{bmatrix} m_{1,1} & \dots & m_{1,d} \\ \vdots & \ddots & \vdots \\ m_{M,1} & \cdots & m_{M,d} \end{bmatrix}$$

$$* \begin{bmatrix} mask_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & mask_d \end{bmatrix} * \begin{bmatrix} d_{1,1} & \cdots & d_{N,1} \\ d_{1,d} & \cdots & d_{N,d} \end{bmatrix}$$

The calculated distance between each map node and each data point is stored in matrix **Dist** (of size M*N), with a column for each data point and a row for each map node. Each 'distance' is essentially the sum of squared distances calculated separately in each dimension, with mask=0 dimensions ignored.

The BMU search seeks the minimum value in each column of the **Dist** matrix. This is the minimum distance to any map node from that data point, and becomes recorded as the squared quantization error for the data point. The square roots of the minimum distances in each column are averaged to give the mean quantisation error for the map.

$$QE = \frac{1}{N}\sum_{i}\sqrt{\min\bigl(Dist(:,i)\bigr)}$$

Distortion measure and updating rule: The distortion measure is described as a 'statistical measure of within-group variation augmented to take account of the neighbourhood function which is an essential distinguishing characteristic of the SOM' (Curry & Morgan, 2004). Vesanto (2003) decomposes the distortion measure into aspects of quantization, neighbourhood bias (comparing the mean of the data points to the mean of the map units, linking quantisation and topology together) and topological quality for each map unit.

The SOM Toolbox calculates the total distortion per map unit by finding the distance between each data point and each map unit with the neighbourhood kernel and mask applied (Vesanto et al., 2000). (This differs slightly to the calculations of BMU with the addition of matrix H as the neighbourhood kernel.)

In matrix terms, for each data point, x_i, row i of the transposed **Dist** matrix is multiplied by the column of the neighbourhood matrix, H, corresponding to the BMU of x_i. The amount of distortion influenced by each data point is therefore determined as:

$$\text{distortion per data point} = \mathrm{H}\bigl(:, bmu_{x_i}\bigr)' * \mathrm{Dist}$$

$$= \begin{bmatrix} h_{1,bmu_{x_i}} & \cdots & h_{M,bmu_{x_i}} \end{bmatrix} * \begin{bmatrix} Dist_{m_1,x_1} & \cdots & Dist_{m_1,x_N} \\ \vdots & \ddots & \vdots \\ Dist_{m_M,x_1} & \cdots & Dist_{m_M,x_N} \end{bmatrix}$$

The distortion can be reported as either the distortion of each map unit due to all the input samples, or the average distortion over all map units. The amount of distortion experienced by each map unit, dmu, is given by the sum of distortion per data point over all data points mapping to each unit. The average distortion for the map is given by:

$$\text{average distortion} = \frac{\sum_m dmu}{M}$$

The updating of each map vector at each time step, t, proceeds as:

$$m_i(t+1) = \frac{\sum_{j=1}^{m} h_{ij}(t)s_j(t)}{\sum_{j=1}^{m} n_{v_j}h_{ij}(t)}$$

where n_{v_j} is the number of items mapping to node j (in Voronoi set, v_j) and $s_j(t) = \sum_{i=1}^{n_{v_j}} x_i$ is the sum of vectors in v_j. For example, the updating of map vector 5 proceeds as:

$$m_5(t+1) = \frac{h_{5,1}(t)s_1(t) + h_{5,2}(t)s_2(t) + \cdots + h_{5,M}(t)s_M(t)}{n_{v_1}h_{5,1}(t) + n_{v_2}h_{5,2}(t) + \cdots + n_{v_M}h_{5,M}(t)}$$

8.8 Appendix 2 - Objective function search

The relationship between SOMs, objective functions and probabilistic methods were investigated in detail to generate ideas for the potential improvement of SOMs implementation through the optimization of an objective function. This section of the literature review is split into two subsections: attempts to define the objective function that is minimized by the SOM, and alteration of the SOM itself to fit it into a probabilistic framework. Following is a summary of the literature on these topics and some basic experimental modelling of the techniques.

8.8.1 Defining an objective function for SOMs

Objective functions in general

The search for an objective function was motivated by the notion that map parameters may be selected through a quantifiable assessment of what is gained or lost through different map setups. An objective function provides a real number output from the solution of a function using a given set of parameters. In terms of regression or classification with supervised or unsupervised learning, an objective function aids parameter selection through being minimized over a set of training data, leading to the choice of the parameter set producing the smallest result.

The objective function, in general, quantifies the difference between the estimated values from the model, $x_i(calc)$ and the true values, $x_i(obs)$, where w_i is the weight of observation i, and N is the number of observations: $objective\ function, Q_i = \sum_{i=1}^{N} w_i[x_i(obs) - x_i(calc)]^2$. A continuously differentiable objective function can be minimized using the gradient descent method, where at each iteration the value moves in the direction of the negative gradient of the objective function, multiplied by the value of some learning parameter, $\propto$, as follows, where $Q_i(a)$ is the objective function, and a is the parameter to be optimised: $b = a - \propto \sum_{i=1}^{N} \nabla Q_i(a)$.

Why the SOM doesn't minimize a single objective function

The issue of objective functions has been the source of much discussion in SOMs literature. The theory of objective functions as described above would appear to correspond easily to the SOM approach, as Varsta (2001) indicates the target of the SOM is to minimize the sum of weighted errors between the input and the map vectors. However, it has been proven that the SOM does not follow a gradient descent of any single objective function (Erwin, 1992). This is due to the opposing aims of quantization (data reduction) and the topology-preserving projection (visualisation) of the SOM. Instead, a set of energy functions must be used (one for each node) and independently minimized using stochastic gradient descent, which becomes complicated for multi-dimensional data. Erwin (1992) states that 'it is easy to propose a cost function which should be minimized by the ordered map, but much more difficult to find an energy function on which a gradient descent can be guaranteed to lead from any disordered map to a map minimizing the cost function.' Kohonen (1990) indicated that no solution for the optimal placement of map nodes is possible, and placement must be determined by iterative approximation techniques. He later states (Kohonen, 2001) that there is not any theoretical reason for which the basic SOM should ensue from any objective function. Yin (2008) lists six attempts at a proof of convergence and ordering for multidimensional SOM systems and yet a full proof remains unattained. Yin (2008) provides a full review of the issues surrounding the objective function of the SOM with a good description of the recent literature.

Distortion measure as objective function

There is discussion in the literature about whether the distortion measure can serve as an objective function for the SOM. The SOM has been described as the 'set of nodes that globally minimizes the average expected distortion measure' (Kohonen, 2001), and the distortion measure has been referred to as the 'energy function' of the SOM (website 9). Rynkiewicz (2006) states the distortion measure is often used as a criterion for assessing the quality of the SOM as it overcomes the absence of a cost function, though no papers actually doing so have been found to support this statement.

The original (sequential, Kohonen, 1990) updating rule for SOMs is a Robbins-Munro stochastic approximation (a method of approximate optimization) of the distortion measure (Kohonen, 2001). The Robbins-Munro method approximates the gradient of the average expected distortion measure by the gradient of the distortion measure with the input samples. This leads to the basic SOM training algorithm, though 'the convergence limit of the Robbins-Munro stochastic approximation does not necessarily represent the exact minimum of the average expected distortion measure' (Kohonen, 2001). In the batch algorithm (Kohonen, 2013), the approximate gradient is evaluated for the entire input set and the weights are updated to the global optimum giving the current partitioning of the data. (Varsta, 2001).

If an infinite set of input samples (X) were available, the average expected distortion measure would be (Kohonen, 2001):

$$\text{average expected distortion measure} = \int \sum_{j=1}^{M} \mathrm{h_{ij}} \left\| \mathrm{x_i} - \mathrm{m_j} \right\|^2 p(X)dx$$

where p(X) is the probability density function of x. The average expected distortion measure returns a scalar value and can be considered an objective function for a continuous distribution (Kohonen, 2001).

When the probability density function isn't known, though, an approximation must be made with available samples of x. In this case, the distortion measure of the SOM is not continuously differentiable, and therefore it cannot be minimized exactly with the gradient descent method. The function is not differentiable at the borders of the Voronoi regions (areas containing groups of data items mapping to a single map unit), due to the fact that the input space at the boundaries has exactly the same distance to two separate map units (Yin, 2008). Kohonen (2001) states that the distortion measure 'is not continuously differentiable, and c (the index of the BMU) changes abruptly when crossing a border in input space'.

Distortion measure vs updating rule

There is a difference between the distortion measure and the updating rule of the SOM. The updating rule of the SOM has been arrived at by ignoring the discontinuities at the boundaries of the Voronoi regions (Varsta, 2001). The distortion measure considers distances between the nodes and the data points, whereas the updating rule only uses the centroids of the sets of data points and isn't concerned with current node locations beyond using them to find sets of data points to determine the next node locations.

The derivation of an updating rule that would minimize the average expected distortion measure runs into difficulty as the Voronoi tessellation is only piecewise continuous, and therefore the function is only piecewise differentiable with respect to the weights (Varsta, 2001).

The distortion measure is proportional to neighbourhood size. As the neighbourhood size decreases with each training iteration, the distortion measure also decreases. Kohonen (2001) gives a detailed description of the map that would result from optimizing the distortion measure *with a constant neighbourhood function,* though the exact optimization of the average expected distortion measure is still an unsolved issue. This optimization attempt assumes the neighbourhood function to be constant, even though an important property of the SOM is that it decreases in size as training progresses (Kohonen, 2001). Therefore, the SOM only minimizes the distortion measure if the neighbourhood kernel is constant (Vesanto, 2000), and this is a stipulation which does not hold in the fundamental SOM method. The learning rule that would follow directly from the distortion measure would therefore be different to the actual SOM training rule, and so the SOM is only an approximate minimization of the distortion measure (website 9).

8.8.2 Altering the SOM structure and probabilistic alternatives

Having established that the traditional SOM output map cannot be arrived at through the optimization of any objective function, researchers have attempted to alter the SOM method itself to encourage it to follow an objective function.

These attempts have contributed a number of alternatives to the SOM. Three new algorithms for topographic mappings are offered by Graepel et al. (1998) including STVQ and 2 generalizations. Each is based on the minimization of a cost function: STVQ (soft topographic vector quantization) also uses a stable neighbourhood radius making it possible to use a fixed neighbourhood function to encode desired neighbourhood relations between nodes; STMK (kernel based soft topographic mapping) is a generalization of STVQ that introduces new distance measures in input space based on kernel functions, which equates to performing STVQ in high dimensional feature space, revealing structure in data that is not revealed by STVQ in Euclidean space. STMP (soft topographic mapping for proximity data) is also a generalization of STVQ, for data given in terms of pairwise proximities. Heskes (1999) slightly changes the definition of the winning unit, to enable SOMs to perform stochastic gradient descent on an energy function. Kostiainen & Lampinen (2002) derive a probability density model for which the converged state of the traditional SOM training algorithm gives the maximum likelihood estimate, based on isotropic Gaussian components.

For the most part, these methods which alter the fundamental structure of the SOM algorithm to force it to follow an objective function are not being found outside theoretical literature. They are not being adopted for research applications, and researchers appear to be reluctant to vary from the widely used traditional SOMs method and are continuing to base their applications on the original heuristic version (Olier & Vellido, 2006). One notable exception, the generative topographic mapping (GTM, Bishop et al., 1997a), with over 1400 cites is by far the most popular probabilistic alternative to the SOM. The GTM consists of a constrained mixture model of Gaussians, allowing model parameters to be determined by the maximum likelihood method. A brief study of the GTM, and comparison with the SOM, is provided here. Whilst the SOM represents a data set by a discrete set of reference vectors, the GTM uses a continuous manifold. The SOM assigns each data point to a single reference vector, and the GTM distributes responsibility over a number of components. The smoothness of the SOM is determined by the choice of neighbourhood function, and the smoothness of the GTM is controlled directly by the basis function parameters.

The GTM is a nonlinear, probabilistic visualisation and clustering model which is considered a probabilistic reformulation of the SOM. GTM was created to overcome the absence of a SOM objective function and the lack of a theoretical basis for parameter choices.

The GTM belongs to the same family of 'unsupervised methods' as the SOM, for which visualisation is a key aspect. High dimensional data is mapped to two-dimensional space, preserving topology and clusters in the original data. The GTM directly computes the topological relationships between grid nodes to define a system similar to the SOM (Kohonen, 2013). Contrary to the SOM method, the GTM constructs a mapping from latent (low-dimensional) space into data space (rather than data space to low-dimensional space as with the SOM) and this mapping is then inverted into latent space for visualisation (Bishop et al., 1997a). While the SOM involves hard assignments of data to nodes, vectors in the GTM involve soft assignments weighted by posterior probabilities. This is analogous to the distinction between the k-means clustering algorithm and using the expectation-maximisation (EM) algorithm to fit a Gaussian mixture model (Bishop et al., 1997a).

8.9 References

Abrahart, R. J., et al. (2012). "Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting." Progress in Physical Geography **36**(4): 480-513.

Adeloye, A. J. and R. Rustum (2012). "Self-organizing map rainfall-runoff multivariate modelling for runoff reconstruction in inadequately gauged basins." Hydrology Research **43**(5): 603-617.

Agarwal, P. and A. Skupin (2008). Self-organizing maps: Applications in geographic information science, John Wiley & Sons.

Astudillo, C. A. and B. J. Oommen (2014). "Topology-oriented self-organizing maps: a survey." Pattern Analysis and Applications **17**(2): 223-248.

Bache, K. and M. Lichman (2013). "UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2013." URL: http://archive. ics. uci. edu/ml.

Barbariol, F., et al. (2016). "Wave extreme characterization using self-organizing maps." Ocean Science **12**(2): 403-415.

Barreto, G. (2007). "Time series prediction with the self-organizing map: A review." Perspectives of neural-symbolic integration: 135-158.

Cereghino, R. and Y. S. Park (2009). "Review of the Self-Organizing Map (SOM) approach in water resources: Commentary." Environmental Modelling & Software **24**(8): 945-947.

Chang, F. J., et al. (2007). "Enforced self-organizing map neural networks for river flood forecasting." Hydrological Processes **21**(6): 741-749.

Choi, B. Y., et al. (2014). "Hydrogeochemical interpretation of South Korean groundwater monitoring data using Self-Organizing Maps." Journal of Geochemical Exploration **137**: 73-84.

Clark, S., et al. (2014). "Increasing dependence on foreign water resources? An assessment of trends in global virtual water flows using a self-organizing time map." Ecological Informatics **26**: 192-202.

Clark, S., et al. (2016). "A dimension range representation (DRR) measure for self-organizing maps." Pattern recognition **53**: 276-286.

Clark, S., et al. (2017). "Nonlinear manifold representation in natural systems: The SOMersault." Environmental Modelling & Software **89**: 61-76.

Cottrell, M., et al. (2016). Theoretical and applied aspects of the self-organizing maps. Advances in Self-Organizing Maps and Learning Vector Quantization, Springer**:** 3-26.

da Anunciacao, Y. M. T., et al. (2014). "Observed summer weather regimes and associated extreme precipitation over Distrito Federal, west-central Brazil." Environmental Earth Sciences **72**(12): 4835-4848.

Davies, D. L. and D. W. Bouldin (1979). "A cluster separation measure." Pattern Analysis and Machine Intelligence, IEEE Transactions on(2): 224-227.

Deboeck, G. (1998). Best practices in data mining using self-organizing maps. Visual Explorations in Finance, Springer**:** 203-229.

Dejean, A., et al. (2011). "Climate change impact on Neotropical social wasps." PloS one **6**(11): e27004.

Dunn, J. C. (1973). "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters."

DuVivier, A. K., et al. (2016). "Winter Atmospheric Buoyancy Forcing and Oceanic Response during Strong Wind Events around Southeastern Greenland in the Regional Arctic System Model (RASM) for 1990-2010." Journal of Climate **29**(3): 975-994.

Eckardt, F. D., et al. (2013). "The nature of moisture at Gobabeb, in the central Namib Desert." Journal of Arid Environments **93**: 7-19.

Edwards, T. W. D., et al. (2017). "Seasonal variability in Northern Hemisphere atmospheric circulation during the Medieval Climate Anomaly and the Little Ice Age." Quaternary Science Reviews **165**: 102-110.

Ellis, K., et al. (2014). "Use of on-line water quality monitoring data to predict bacteriological failures." Procedia Engineering **70**: 612-621.

Erwin, E., et al. (1992). "Self-organizing maps: ordering, convergence properties and energy functions." Biological cybernetics **67**(1): 47-55.

Falcieri, F. M., et al. (2014). "Po River plume pattern variability investigated from model data." Continental Shelf Research **87**: 84-95

Flexer, A. (1999). On the use of self-organizing maps for clustering and visualization. Principles of Data Mining and Knowledge Discovery, Springer**:** 80-88.

Folguera, L., et al. (2015). "Self-organizing maps for imputation of missing data in incomplete data matrices." Chemometrics and Intelligent Laboratory Systems **143**: 146-151

Fyfe, C. (2008). Topographic maps for clustering and data visualization. Computational Intelligence: A Compendium, Springer**:** 111-153.

Gamble, A. and M. Babbar-Sebens (2012). "On the use of multivariate statistical methods for combining in-stream monitoring data and spatial analysis to characterize water quality conditions in the White River Basin, Indiana, USA." Environmental Monitoring and Assessment **184**(2): 845-875.

García, H. L. and I. M. González (2004). "Self-organizing map and clustering for wastewater treatment monitoring." Engineering Applications of Artificial Intelligence **17**(3): 215-225.

Gopakumar, R., et al. (2007). "Hydrologic data exploration and river flow forecasting of a humid tropical river basin using artificial neural networks." Water Resources Management **21**(11): 1915-1940.

Guo, D., et al. (2006). "A visualization system for space-time and multivariate patterns (vis-stamp)." Visualization and Computer Graphics, IEEE Transactions on **12**(6): 1461-1474.

Hammer, B., et al. (2005). Self organizing maps for time series. Proceedings of WSOM.

Hastie, T., et al. (2009). The Elements of Statistical Learning, New York: Springer.

Hewitson, B. and R. Crane (2002). "Self-organizing maps: applications to synoptic climatology." Climate Research **22**(1): 13-26.

Hong, D. G., et al. (2016). "Limnological assessment of the meteo-hydrological and physicochemical factors for summer cyanobacterial blooms in a regulated river system." Annales De Limnologie-International Journal of Limnology **52**: 123-136.

Hsu, K. l., et al. (2002). "Self-organizing linear output map (SOLO): An artificial neural network suitable for hydrologic modeling and analysis." Water Resources Research **38**(12): 38-31-38-17.

Huang, Y., et al. (2013). "Discriminant diffusion maps analysis: A robust manifold learner for dimensionality reduction and its applications in machine condition monitoring and fault diagnosis." Mechanical Systems and Signal Processing **34**(1): 277-297.

Jutagate, T., et al. (2016). "Spatio-temporal variations in abundance and assemblage patterns of fish larvae and their relationships to environmental variables in Sirindhron Reservoir of the Lower Mekong Basin, Thailand." Indian Journal of Fisheries **63**(3): 11-23

Kalteh, A. M., et al. (2008). "Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application." Environmental Modelling & Software **23**(7): 835-845.

Kaski, S. and T. Kohonen (1996). Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world. Neural networks in financial engineering. Proceedings of the third international conference on neural networks in the capital markets, Citeseer.

Kaski, S. and K. Lagus (1996). Comparing self-organizing maps. Artificial Neural Networks—ICANN 96, Springer**:** 809-814.

Kaski, S., et al. (1998). Methods for interpreting a self-organized map in data analysis. In Proc. 6th European Symposium on Artificial Neural Networks (ESANN98). D-Facto, Brugfes, Citeseer.

Kaski, S., et al. (2000). "Coloring that reveals cluster structures in multivariate data." Australian Journal of Intelligent Information Processing Systems **6**(2): 82-88.

Kim, J. Y., et al. (2016). "Application of multivariate analysis to determine spatial and temporal changes in water quality after new channel construction in the Chilika Lagoon." Ecological Engineering **90**: 314-319.

Kingston, G. B., et al. (2005). "Bayesian training of artificial neural networks used for water resources modeling." Water Resources Research **41**(12).

Kiviluoto, K. (1996). Topology preservation in self-organizing maps. IEEE International Conference on Neural Networks.

Kohonen, T. (1982). "Self-organized formation of topologically correct feature maps." Biological cybernetics **43**(1): 59-69.

Kohonen, T. (1990). "The self-organizing map." Proceedings of the IEEE **78**(9): 1464-1480.

Kohonen, T. (1993). Things you haven't heard about the Self-Organizing Map. Neural Networks, 1993., IEEE International Conference on, IEEE.

Kohonen, T. (1995). "Self-Organizing Maps." Springer series in information sciences **30**.

Kohonen, T. (1998). "The self-organizing map." Neurocomputing **21**(1): 1-6.

Kohonen, T. (2001). Self-organizing maps, Springer.

Kohonen, T. (2013). "Essentials of the self-organizing map." Neural Networks **37**: 52-65.

Kohonen, T. and T. Honkela (2007). "Kohonen network." Scholarpedia **2**(1): 1568.

Lek, S. and J.-F. Guégan (1999). "Artificial neural networks as a tool in ecological modelling, an introduction." Ecological Modelling **120**(2): 65-73.

Ley, R., et al. (2011). "Catchment classification by runoff behaviour with self-organizing maps (SOM)." Hydrology and Earth System Sciences **15**(9): 2947.

Li, W., et al. (2015). "Spatiotemporal Classification Analysis of Long-Term Environmental Monitoring Data in the Northern Part of Lake Taihu, China by Using a Self-Organizing Map." Journal of Environmental Informatics **26**(1): 71-79.

Lin, G.-F. and L.-H. Chen (2006). "Identification of homogeneous regions for regional frequency analysis using the self-organizing map." Journal of Hydrology **324**(1): 1-9.

Liu, W. B., et al. (2016). "Large-scale circulation classification and its links to observed precipitation in the eastern and central Tibetan Plateau." Climate dynamics **46**(11-12): 3481-3497.

Liu, Y. and R. H. Weisberg (2011). "A review of Self-Organizing Map applications in meteorology and oceanography." Self-Organizing Maps-Applications and Novel Algorithm: 253-272.

Liu, Y., et al. (2006). "Performance evaluation of the self-organizing map for feature extraction." Journal of Geophysical Research: Oceans (1978–2012) **111**(C5).

Lukacs, B. A., et al. (2015). "Rainfall fluctuations and vegetation patterns in alkali grasslands - using self-organizing maps to visualise vegetation dynamics." Tuexenia(35): 381-397.

Lynch, A. H., et al. (2016). "Linkages between Arctic summer circulation regimes and regional sea ice anomalies." Journal of Geophysical Research-Atmospheres **121**(13): 7868-7880.

Matic, F., et al. (2017). "Oscillating Adriatic temperature and salinity regimes mapped using the Self-Organizing Maps method." Continental Shelf Research **132**: 11-18.

Mattingly, K. S., et al. (2016). "Increasing water vapor transport to the Greenland Ice Sheet revealed using self-organizing maps." Geophysical Research Letters **43**(17): 9250-9258.

Mihanovic, H., et al. (2015). "Mapping of decadal middle Adriatic oceanographic variability and its relation to the BiOS regime." Journal of Geophysical Research-Oceans **120**(8): 5615-5630.

Morioka, Y., et al. (2010). "Climate variability in the southern Indian Ocean as revealed by self-organizing maps." Climate dynamics **35**(6): 1059-1072.

Mothe, J., et al. (2006). "Combining mining and visualization tools to discover the geographic structure of a domain." Computers, environment and urban systems **30**(4): 460-484.

Mwale, F., et al. (2012). "Infilling of missing rainfall and streamflow data in the Shire River basin, Malawi–A self organizing map approach." Physics and Chemistry of the Earth, Parts A/B/C **50**: 34-43.

Newton, B. W., et al. (2014). "Evaluating the distribution of water resources in western Canada using synoptic climatology and selected teleconnections. Part 2: summer season." Hydrological Processes **28**(14): 4235-4249.

Nguyen, T. T., et al. (2015). "Identification of spatio-seasonal hydrogeochemical characteristics of the unconfined groundwater in the Red River Delta, Vietnam." Applied Geochemistry **63**: 10-21.

Nkiaka, E., et al. (2016). "Using self-organizing maps to infill missing data in hydro-meteorological time series from the Logone catchment, Lake Chad basin." Environmental Monitoring and Assessment **188**(7).

Olkowska, E., et al. (2014). "Assessment of the water quality of Klodnica River catchment using self-organizing maps." Science of the Total Environment **476**: 477-484.

Ota, K., et al. (2011). "Asymmetric neighborhood functions accelerate ordering process of self-organizing maps." Physical Review E **83**(2): 021903.

Pampalk, E. (2001). "Limitations of the SOM and the GTM."

Park, Y. S., et al. (2014). "Characterizing effects of landscape and morphometric factors on water quality of reservoirs using a self-organizing map." Environmental Modelling & Software **55**: 214-221.

Pena, M., et al. (2008). Topology-preserving mappings for data visualisation. Principal Manifolds for Data Visualization and Dimension Reduction, Springer**:** 131-150.

Principe, J. C., et al. (1998). "Local dynamic modeling with self-organizing maps and applications to nonlinear system identification and control." Proceedings of the IEEE **86**(11): 2240-2258.

Ramos, E., et al. (2016). "An ecological classification of rocky shores at a regional scale: a predictive tool for management of conservation values." Marine Ecology-an Evolutionary Perspective **37**(2): 311-328.

Reusch, D. B., et al. (2007). "North Atlantic climate variability from a self-organizing map perspective." Journal of Geophysical Research: Atmospheres **112**(D2).

Rivera, D., et al. (2015). "Exploring soil databases: a self-organizing map approach." Soil Use and Management **31**(1): 121-131.

Rodriguez-Alarcon, R. and S. Lozano (2017). "SOM-Based Decision Support System for Reservoir Operation Management." Journal of Hydrologic Engineering **22**(7).

Rousseeuw, P. J. (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." Journal of computational and applied mathematics **20**: 53-65.

Rustum, R. and A. J. Adeloye (2007). "Replacing outliers and missing values from activated sludge data using Kohonen self-organizing map." Journal of Environmental Engineering **133**(9): 909-916.

Sarlin, P. (2012). "Self-organizing time map: An abstraction of temporal multivariate patterns." Neurocomputing.

Sarlin, P. and D. Marghescu (2011). "Visual predictions of currency crises using self-organizing maps." Intelligent Systems in Accounting, Finance and Management **18**(1): 15-38.

Sarlin, P. and Z. Y. Yao (2013). "Clustering of the Self-Organizing Time Map." Neurocomputing **121**: 317-327.

Schreck, T., et al. (2009). "Visual cluster analysis of trajectory data with interactive kohonen maps." Information Visualization **8**(1): 14-29.

Shanmuganathan, S., et al. (2006). "Self-organizing map methods in integrated modelling of environmental and economic systems." Environmental Modelling & Software **21**(9): 1247-1256.

Sharif, S. M., et al. (2015). Characterization of water quality conditions in the Klang River Basin, Malaysia using self organizing map and K-means algorithm. Environmental Forensics 2015. A. Z. Aris. **30:** 73-78.

Simon, G., et al. (2005). "Time series forecasting: Obtaining long term trends with self-organizing maps." Pattern Recognition Letters **26**(12): 1795-1808.

Skupin, A. (2004). "A picture from a thousand words." Computing in Science & Engineering **6**(5): 84-88.

Skupin, A., et al. (2013). "Visualizing the topical structure of the medical sciences: a self-organizing map approach." PloS one **8**(3): e58779.

Skupin, A. and R. Hagelman (2005). "Visualizing demographic trajectories with self-organizing maps." GeoInformatica **9**(2): 159-179.

Steynor, A., et al. (2009). "Projected future runoff of the Breede River under climate change." Water SA **35**(4): 433-440.

Swales, D., et al. (2016). "Examining moisture pathways and extreme precipitation in the US Intermountain West using self-organizing maps." Geophysical Research Letters **43**(4): 1727-1735.

Takala, M., et al. (2008). "Detecting the onset of snow-melt using SSM/I data and the self-organizing map." International journal of remote sensing **29**(3): 755-766.

Tiwari, M. K., et al. (2013). "Improving reliability of river flow forecasting using neural networks, wavelets and self-organizing maps." Journal of Hydroinformatics **15**(2): 486-502.

Toth, E. (2009). "Classification of hydro-meteorological conditions and multiple artificial neural networks for streamflow forecasting." Hydrology and Earth System Sciences **13**(9): 1555-1566.

Toth, E. (2013). "Catchment classification based on characterisation of streamflow and precipitation time series." Hydrology and Earth System Sciences **17**(3): 1149.

Tsai, W. P., et al. (2016). "Exploring the ecological response of fish to flow regime by soft computing techniques." Ecological Engineering **87**: 9-19.

Ultsch, A. (2003). "U-matrix: a tool to visualize clusters in high dimensional data." Marburg: Fachbereich Mathematik und Informatik.

Václavík, T., et al. (2013). "Mapping global land system archetypes." Global Environmental Change **23**(6): 1637-1647.

Van Laerhoven, K. (2001). "Combining the self-organizing map and k-means clustering for on-line classification of sensor data." Artificial Neural Networks—ICANN 2001: 464-469.

Vereecken, H., et al. (2016). "On the role of patterns in understanding the functioning of soil-vegetation-atmosphere systems." Journal of Hydrology **542**: 63-86.

Vesanto, J. (1997). Using the SOM and local models in time-series prediction. Proc. Workshop on Self-Organizing Maps 1997.

Vesanto, J. (1999). "SOM-based data visualization methods." Intelligent data analysis **3**(2): 111-126.

Vesanto, J. (2000). Neural network tool for data mining: SOM toolbox. Proceedings of symposium on tool environments and development methods for intelligent systems (TOOLMET2000).

Vesanto, J. and E. Alhoniemi (2000). "Clustering of the self-organizing map." Ieee Transactions on Neural Networks **11**(3): 586-600.

Vesanto, J., et al. (1999). Self-organizing map in MATLAB: the SOM toolbox. Proceedings of the MATLAB DSP conference.

Vesanto, J., et al. (2000). "SOM toolbox for MATLAB 5." Helsinki University of Technology, Finland.

Vesanto, J., et al. (2003). On the decomposition of the self-organizing map distortion measure. Proceedings of the workshop on self-organizing maps (WSOM'03).

Wang, N., et al. (2013). "Visualizing gridded time series data with self organizing maps: an application to multi-year snow dynamics in the Northern Hemisphere." Computers, environment and urban systems **39**: 107-120.

Wang, X., et al. (2006). "Characteristic-based clustering for time series data." Data Mining and Knowledge Discovery **13**(3): 335-364.

Wang, Y. B., et al. (2015). "Differentiating the Spatiotemporal Distribution of Natural and Anthropogenic Processes on River Water-Quality Variation Using a Self-Organizing Map With Factor Analysis." Archives of Environmental Contamination and Toxicology **69**(2): 254-263.

Wang, Y.-C. and C.-C. Feng (2011). "Patterns and trends in land-use land-cover change research explored using self-organizing map." International journal of remote sensing **32**(13): 3765-3790.

Ward, J. H. J. (1963). "Hierarchical grouping to optimize an objective function." Journal of the American Statistical Association **58**: 236-244.

Wehrens, R. and L. M. Buydens (2007). "Self-and super-organizing maps in R: the Kohonen package." J Stat Softw **21**(5): 1-19.

Ye, C., et al. (2015). "Advancing Analysis of Spatio-Temporal Variations of Soil Nutrients in the Water Level Fluctuation Zone of China's Three Gorges Reservoir Using Self-Organizing Map." PloS one **10**(3).

Yin, H. (2008). Learning nonlinear principal manifolds by self-organizing maps. Principal manifolds for data visualization and dimension reduction, Springer**:** 68-95.

Yin, H. (2008). "On multidimensional scaling and the embedding of self-organizing maps." Neural Networks **21**(2): 160-169.

Yin, H. (2008). The self-organizing maps: Background, theories, extensions and applications. Computational intelligence: a compendium, Springer**:** 715-762.

Zhang, X. and Y. Li (1993). Self-organizing map as a new method for clustering and data analysis. Neural Networks, 1993. IJCNN'93-Nagoya. Proceedings of 1993 International Joint Conference on, IEEE.

Websites:

[1] Scopus, refined by ‘earth and planetary sciences’ and ‘environmental sciences’, accessed June 2017

[2] Google Scholar, accessed June 2017

[3] Environment Canada’s HYDAT database (National Water Data Archive): https://ec.gc.ca/rhc-wsc/default.asp?lang=En&n=9018B5EC-1, accessed July 2017

[4] http://www.cis.hut.fi/somtoolbox/, accessed July 2017

9 CONCLUSION

This thesis extends current SOMs theory through a closely-tied set of advancements focused on improving the extraction and interpretation of useful information from nonlinear and spatiotemporal data sets. Data encountered in water-related sciences often result from high-dimensional, frequent measurements of systems with nonlinear and spatiotemporal aspects. The developments presented here have been aimed at improving the summarisation, sorting and visualisation of this data, to increase insight into the interrelationships of system components.

The new methods are suited to environmental data with missing values and data structures that change over time. They have been demonstrated on current water-related issues with complex hydrological-human relationships, revealing the enhanced pattern extraction and clustering capabilities. Though this project is motivated by and based on hydrologic and water resource applications, the approaches and extensions to the SOM introduced here could be applied to the exploratory analysis of a wide variety of data sets from any field.

A summary of the new methods introduced in Papers 1-4 is provided here:

- In Paper 1, data items have been clustered in terms of their similarities based on the nonlinear relationship between the variables. Temporal dynamics over the timeline of the study were investigated, revealing a global trend of the data set and groups of data items with similar temporal movements through the map. A single visual output is produced to convey data items with similar and diverging trends. The large spatiotemporal data set has been reduced into a specified number of representative vectors which are ordered based on similarity, leading to the possibility of further quantitative analysis of the trends in the datasets without having to manipulate vast amounts of individual observations or explicitly define the nonlinear relationships.
- In Paper 2, the 'dimension range representation' measure is introduced to quantify how well a map represents each dimension of a data set, indicating if any dimensions of input data become under-represented in the creation of the map. It has been shown that though a data item may more closely match its allocated map vector in certain dimensions, the map vector will be interpreted as being representative of the data in all dimensions. This new measure, used in conjunction with existing quality measures, can aid the choice of number and configuration of map nodes.
- In Paper 3, a method is introduced to improve the extraction of underlying nonlinear relationships from complex high-dimensional measurements. The SOM framework is expanded to enable the characterisation of highly nonlinear manifolds, transferring the global ordering process into low-dimensional space by focusing the first approximation of map node locations along the geodesic surface, and applying a restricted neighbourhood kernel when refining the map node locations to limit the influence of each data item to nodes directly around it.
- In Paper 4, dimension reduction and clustering are performed on maps in a series, created with different configurations and variables. Each map indicates the

predominant characteristics of the data at that time step. Temporal patterns are extracted by identifying the relationship of each data item to these predominant characteristics of each map.

Through the series of papers, the gaps in the general knowledge base of self-organizing maps identified in Section 2.2 have been progressively addressed, as follows:

1. **Spatiotemporal clustering.** The issue of presenting spatiotemporal analyses within a single visualisation, rather than a series of maps requiring subjective user interpretation, was a focus of Paper 1. The result was a method that is able to trace the individual data items as they move through the evolving global cluster structure of the data over the time period of the study. This effectively reduces the spatiotemporal results to a single visualisation representing the changing cluster structure of the data as well as the changing relationships between data items. This method, however, is still lacking the freedom to represent each timestep of data with maps of different configurations, even though the data structure may differ at each timestep.
   Spatiotemporal clustering is revisited in Paper 4, which expands temporal SOMs methodology to allow the two-dimensional structure of the SOM to shift in time, representing a possible shifting temporal structure of the data distribution. Whilst the study leading to Paper 1 focused on individual trends of data items through an evolving global cluster structure, the study leading to Paper 4 was concerned with individual trends of data items related to the changing structure of the data set. Two-dimensional maps were used in this study, as a three-dimensional representation of the overall data set (the two-dimensional maps vs time) was not feasible due to visualisation issues. However, this is an issue that would benefit from further attention in future as three dimensions would allow increased insight into the system patterns.
2. **Parameter selection.** The issue of deliberately choosing an appropriate map structure to best represent a particular data set was addressed in Paper 2. This is accomplished by monitoring the dimension-specific intra-cluster range of data assigned to each map node, ensuring the choice of map configuration that allows all dimensions to be represented relatively equally in the resulting clustering and visualization. This concept of providing an unbiased coverage of each data dimension is combined into the spatiotemporal analysis of Paper 4. In future work, an automated system of selecting optimal map configuration with less user input could be attempted to make the method more reproducible.
3. **Nonlinear manifold representation**. The summarising of patterns and clusters within data containing nonlinear manifolds was addressed in Paper 3, with the introduction of the SOMersault algorithm. This new algorithm effectively unrolls the initial map in alignment with any low-dimensional nonlinear manifold within high-dimensional data measurements, performs map finetuning in high-dimensional space ensuring accurate pattern extraction, and restricts clustering to regions along the geodesic surface to ensure cluster members share similarities

in consideration of the nonlinear manifold. Future work may include an improved method for mapping the prototype vectors back and forth between the low and high dimensional spaces.

4. **Objective function minimisation.** A detailed investigation into the search for an objective function is included in Section 8.8. This study was conducted with the ambition to apply an objective function to parameter selection, however, it was found that the SOM does not result from the optimisation of any objective function. The stochastic optimization of the distortion measure only approximately represents the convergence of the SOM and cannot be used to determine map structure as it decreases with decreasing map size. The literature provides some attempts to create probabilistic alternatives to the SOM, such as the generative topographic mapping, though researchers in general are continuing to use the traditional SOM rather than adopting the new methods. This is perhaps due to the statistical complexity of these alternatives compared to the intuitive nature of the traditional SOM method. This line of investigation was therefore not pursued further as little impact on practical applications seemed probable. This decision was reinforced by the fact that this line of research has lost momentum in the literature since 2001, with all references since that date merely stating the lack of an objective function for SOMs.
5. **Crossover of theory into applied research.** The accessibility of SOMs theory (both traditional and innovative) to non-statistical users was addressed in Paper 5 with the production of a practical implementation guide. Background theory, current best practice, basic and innovative examples, and a step-by-step guide are included to ensure researchers, engineers and scientists interested in using SOMs to explore their high-dimensional, nonlinear data sets have a resource to do so. Though this paper does not transfer all historical SOMs technical innovations into the applied realm, it will raise consciousness that capabilities of the SOM for representing specific data sets exist beyond those realised with default applications. In addition, the new techniques presented in Papers 1-4 are demonstrated on real-world applications providing examples that can be followed.

Though this thesis is primarily aimed at expanding the general knowledge base of SOMs theory, the applications contained within it have also contributed to increased knowledge of global and regional water resource relationships through providing:

- An understanding of the role of 172 countries in the global exchange of virtual water and the transformation of individual countries' dependencies on foreign water resources within the context of the shifting virtual water market.
- Analysis of the relative conditions of 142 countries with respect to Millennium Development Goal 7c, aimed at reducing the proportion of people without sustainable access to safe water and basic sanitation. Clusters of countries sharing similar states of urban and rural water and sanitation development were identified.

- The integration of satellite-based water storage measurements with basin-level water scarcity calculations, identifying the disparity between the quantity of all forms of water present in a basin and anthropogenically-induced shortages in availability.
- The depiction of projected trends and comparisons between changing river flood impacts on population and property induced by anticipated urbanization and climate change in 98 cities.

Addressing the need for increased integration of SOMs theoretical knowledge into applied research, science and engineering is an important component of this thesis. SOM users in the environmental field appear to require some encouragement to transfer theoretical innovations into applied research. The applications presented in these papers may be used as examples for implementation of these new methods within applied fields of research and engineering, and it is hoped that Paper 5, in particular, will provide enough information to lead the interested researcher through a deliberate application of SOM techniques.

Opportunities for future work have been considered during the preparation of the thesis. Papers 1-4, taken together, lead towards a variety of possible next steps for the expansion of spatiotemporal, nonlinear SOMs. The methods could be expanded into three output dimensions, ideally including a shifting structure of the SOM that accounts for temporal changes in the data structure based on ensuring appropriate coverage of each time step of the data by the corresponding portion of the SOM. This would require specialised visualizations to allow intuitive interpretation of the temporal flow. A projection of visible cluster trends into the next time step could also be contemplated. Furthermore, any possible methods for reducing the requirement for subjective user input in the setup and interpretation phases of the SOM process will benefit the method by leading to a more reproducible product.

10 REFERENCES

Abe, H., & Tsumoto, S. (2011). Evaluating a temporal pattern detection method for finding research keys in bibliographical data. *Transactions on rough sets XIV*, 1-17.

Abrahart, R. J., Anctil, F., Coulibaly, P., Dawson, C. W., Mount, N. J., See, L. M., Wilby, R. L. (2012). Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. *Progress in Physical Geography, 36*(4), 480-513.

Adeloye, A. J., & Rustum, R. (2012). Self-organizing map rainfall-runoff multivariate modelling for runoff reconstruction in inadequately gauged basins. *Hydrology Research, 43*(5), 603-617.

Alahakoon, D., Halgamuge, S. K., & Srinivasan, B. (2000). Dynamic self-organizing maps with controlled growth for knowledge discovery. *Neural Networks, IEEE Transactions on, 11*(3), 601-614.

Alsdorf, D. E., Rodriguez, E., & Lettenmaier, D. P. (2007). Measuring surface water from space. *Reviews of Geophysics, 45*(2).

Andrade, A. O., Nasuto, S., Kyberd, P., & Sweeney-Reed, C. M. (2005). Generative topographic mapping applied to clustering and visualization of motor unit action potentials. *Biosystems, 82*(3), 273-284. doi:10.1016/j.biosystems.2005.09.004

Arroyo, J., González-Rivera, G., Maté, C., & San Roque, A. M. (2011). Smoothing methods for histogram-valued time series: an application to value-at-risk. *Statistical Analysis and Data Mining: the ASA Data Science Journal, 4*(2), 216-228.

Bache, K., & Lichman, M. (2013). UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2013. *URL: http://archive. ics. uci. edu/ml*.

Barreto, G. (2007). Time series prediction with the self-organizing map: A review. *Perspectives of neural-symbolic integration*, 135-158.

Bartkowiak, A. (2004). Visualizing large data by the SOM and GTM methods—what are we obtaining? *Intelligent Information Processing and Web Mining* (pp. 399-403): Springer.

Bierman, P., Lewis, M., Ostendorf, B., & Tanner, J. (2011). A review of methods for analysing spatial and temporal patterns in coastal water quality. *Ecological Indicators, 11*(1), 103-114.

Billard, L. (2011). Brief overview of symbolic data and analytic issues. *Statistical Analysis and Data Mining: the ASA Data Science Journal, 4*(2), 149-156.

Bishop, C. M., Hinton, G. E., Strachan, I. G. D., & Inst Elect Engineers; Inst Elect, E. (1997). GTM through time *Fifth International Conference on Artificial Neural Networks* (pp. 111-116).

Bishop, C. M., Svensen, M., & Williams, C. K. I. (1997). GTM: A principled alternative to the self-organizing map. In M. C. Mozer, M. I. Jordan, & T. Petsche (Eds.), *Advances in Neural Information Processing Systems 9: Proceedings of the 1996 Conference* (Vol. 9, pp. 354-360).

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual review of information science and technology, 37*(1), 179-255.

Chan, A., & Pampalk, E. (2002). *Growing hierarchical self organizing map (ghsom) toolbox: visualisations and enhancements.* Paper presented at the Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on.

Chapman, C., & Charantonis, A. A. (2017). Reconstruction of Subsurface Velocities From Satellite Observations Using Iterative Self-Organizing Maps. *Ieee Geoscience and Remote Sensing Letters, 14*(5), 617-620. doi:10.1109/lgrs.2017.2665603

Chen, S., Amid, D., Shir, O. M., Limonad, L., Boaz, D., Anaby-Tavor, A., & Schreck, T. (2013). *Self-organizing maps for multi-objective Pareto frontiers.* Paper presented at the Visualization Symposium (PacificVis), 2013 IEEE Pacific.

Cheng, Y. (1997). Convergence and ordering of Kohonen's batch map. *Neural computation, 9*(8), 1667-1676.

Chiu, G., & Lehmann, E. (2011). *Bayesian hierarchical modelling: incorporating spatial information in water resources assessment and accounting.* Paper presented at the International Congress on Modelling and Simulation (MODSIM).

Clark, S., Sisson, S. A., & Sharma, A. (2016). A dimension range representation (DRR) measure for self-organizing maps. *Pattern recognition, 53*, 276-286.

Clark, S., Sisson, S. A., & Sharma, A. (2017). Nonlinear manifold representation in natural systems: The SOMersault. *Environmental Modelling & Software, 89*, 61-76.

Cornford, D., Schroeder, M., & Nabney, I. T. (2009). Data visualisation and exploration with prior knowledge.

Cottrell M, Olteanu M, Rossi F, Villa-Vialaneix N. (2016). Theoretical and applied aspects of the self-organizing maps. Advances in self-organizing maps and learning vector quantization: Springer. p. 3-26.

Curry, B., & Morgan, P. H. (2004). Evaluating Kohonen's learning rule: An approach through genetic algorithms. *European Journal of Operational Research, 154*(1), 191-205.

Decharme, B., Douville, H., Prigent, C., Papa, F., & Aires, F. (2008). A new river flooding scheme for global climate applications: Off-line evaluation over South America. *Journal of Geophysical Research: Atmospheres, 113*(D11).

Dejean, A., Cereghino, R., Carpenter, J. M., Corbara, B., Herault, B., Rossi, V., . . . Bonal, D. (2011). Climate change impact on Neotropical social wasps. *PloS one, 6*(11), e27004.

Demartines, P., & Hérault, J. (1997). Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *Neural Networks, IEEE Transactions on, 8*(1), 148-154.

Dorigo, W., Jeu, R., Chung, D., Parinussa, R., Liu, Y., Wagner, W., & Fernández-Prieto, D. (2012). Evaluating global trends (1988–2010) in harmonized multi-satellite surface soil moisture. *Geophysical Research Letters, 39*(18).

Eckardt, F. D., Soderberg, K., Coop, L. J., Muller, A. A., Vickery, K. J., Grandin, R. D., . . . Henschel, J. (2013). The nature of moisture at Gobabeb, in the central Namib Desert. *Journal of Arid Environments, 93*, 7-19. doi:10.1016/j.jaridenv.2012.01.011

Erwin, E., Obermayer, K., & Schulten, K. (1992). Self-organizing maps: ordering, convergence properties and energy functions. *Biological cybernetics, 67*(1), 47-55.

Fahimi, F., Yaseen, Z. M., & El-shafie, A. (2017). Application of soft computing based hybrid models in hydrological variables modeling: a comprehensive review. *Theoretical and Applied Climatology, 128*(3-4), 875-903. doi:10.1007/s00704-016-1735-8

Fekete, B. M., Vörösmarty, C. J., & Grabs, W. (2002). High-resolution fields of global runoff combining observed river discharge and simulated water balances. *Global Biogeochemical Cycles, 16*(3).

Fenn, D. J., Porter, M. A., Mucha, P. J., McDonald, M., Williams, S., Johnson, N. F., & Jones, N. S. (2012). Dynamical clustering of exchange rates. *Quantitative Finance, 12*(10), 1493-1520.

Fritzke, B. (1994). Growing cell structures—a self-organizing network for unsupervised and supervised learning. *Neural Networks, 7*(9), 1441-1460.

Fung, G., Yu, J., & Lam, W. (2002). News sensitive stock trend prediction. *Advances in knowledge discovery and data mining*, 481-493.

Fyfe, C. (2008). Topographic maps for clustering and data visualization *Computational Intelligence: A Compendium* (pp. 111-153): Springer.

Gamble, A., & Babbar-Sebens, M. (2012). On the use of multivariate statistical methods for combining in-stream monitoring data and spatial analysis to characterize water quality conditions in the White River Basin, Indiana, USA. *Environmental Monitoring and Assessment, 184*(2), 845-875.

Gao, P., Geissen, V., Ritsema, C., Mu, X., & Wang, F. (2013). Impact of climate change and anthropogenic activities on stream flow and sediment discharge in the Wei River basin, China. *Hydrology and Earth System Sciences Discussions, 17*, 961-972.

Goodhill, G. J., & Sejnowski, T. J. (1997). A unifying objective function for topographic mappings. *Neural computation, 9*(6), 1291-1303.

Gopakumar, R., Takara, K., & James, E. (2007). Hydrologic data exploration and river flow forecasting of a humid tropical river basin using artificial neural networks. *Water Resources Management, 21*(11), 1915-1940.

Gorban, A. N., & Zinovyev, A. Y. (2008). Elastic maps and nets for approximating principal manifolds and their application to microarray data visualization *Principal manifolds for data visualization and dimension reduction* (pp. 96-130): Springer.

Graepel, T., Burger, M., & Obermayer, K. (1998). Self-organizing maps: generalizations and new optimization techniques. *Neurocomputing, 21*(1), 173-190.

Guo, X., Wang, H., & Glass, D. H. (2011). *The impact of learning parameters on Bayesian self-organizing maps: An empirical study.* Paper presented at the Natural Computation (ICNC), 2011 Seventh International Conference on.

Guo, X.-L., Wang, H.-Y., & Glass, D. H. (2012). *A growing Bayesian self-organizing map for data clustering.* Paper presented at the Machine Learning and Cybernetics (ICMLC), 2012 International Conference on.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: New York: Springer.

He, H., Chen, J., Jin, H., & Chen, S.-H. (2007). Trading strategies based on K-means clustering and regression models. *Computational Intelligence in Economics and Finance, 2*, 123-134.

Helsel, DR, & Hirsch, RM. (2002). Statistical methods in water resources. Techniques of water resources investigations, book 4, chapter A3. U.S. Geological Survey.

Heskes, T. (1999). Energy functions for self-organizing maps. *Kohonen maps*, 303-316.

Hewitson, B., & Crane, R. (2002). Self-organizing maps: applications to synoptic climatology. *Climate Research, 22*(1), 13-26.

Houborg, R., Rodell, M., Li, B., Reichle, R., & Zaitchik, B. F. (2012). Drought indicators based on model-assimilated Gravity Recovery and Climate Experiment (GRACE) terrestrial water storage observations. *Water Resources Research, 48*(7).

Hsu, K. l., Gupta, H. V., Gao, X., Sorooshian, S., & Imam, B. (2002). Self-organizing linear output map (SOLO): An artificial neural network suitable for hydrologic modeling and analysis. *Water Resources Research, 38*(12), 38-31-38-17.

Jongman, B., Ward, P. J., & Aerts, J. C. (2012). Global exposure to river and coastal flooding: Long term trends and changes. *Global Environmental Change, 22*(4), 823-835.

Kaski, S. (1997). PhD thesis: Data exploration using self-organizing maps. Helsinki University of Technology.

Kingston, G. B., Lambert, M. F., & Maier, H. R. (2005). Bayesian training of artificial neural networks used for water resources modeling. *Water Resources Research, 41*(12).

Kiviluoto, K., & Oja, E. (1998). S-Map: A network with a simple self-organization algorithm for generative topographic mappings. In M. I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in Neural Information Processing Systems 10* (Vol. 10, pp. 549-555).

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE, 78*(9), 1464-1480.

Kohonen, T. (1998). The self-organizing map. *Neurocomputing, 21*(1), 1-6.

Kohonen, T. (2001). *Self-organizing maps* (Vol. 30): Springer.

Kohonen, T. (2008). Data management by self-organizing maps. *Computational intelligence: research frontiers*, 309-332.

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Networks, 37*, 52-65. doi:10.1016/j.neunet.2012.09.018

Kostiainen, T., & Lampinen, J. (2002). On the generative probability density model in the self-organizing map. *Neurocomputing, 48*(1), 217-228.

Kurasova, O., Petkus, T., & Filatovas, E. (2013). Visualization of Pareto Front Points when Solving Multi-objective Optimization Problems. *Information Technology and Control, 42*(4), 353-361.

Lee, M. R., & Chen, T. T. (2012). Revealing research themes and trends in knowledge management: From 1995 to 2010. *Knowledge-Based Systems, 28*, 47-58.

Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition, 38*(11), 1857-1874.

Liu, Y., & Weisberg, R. H. (2011). A review of Self-Organizing Map applications in meteorology and oceanography. *Self-Organizing Maps-Applications and Novel Algorithm*, 253-272.

Liu, Y., Weisberg, R. H., & Mooers, C. N. (2006). Performance evaluation of the self-organizing map for feature extraction. *Journal of Geophysical Research: Oceans (1978–2012), 111*(C5).

Luttrell, S. P. (1994). A Bayesian analysis of self-organizing maps. *Neural computation, 6*(5), 767-794.

Marsland, S., Shapiro, J., & Nehmzow, U. (2002). A self-organizing network that grows when required. *Neural Networks, 15*(8), 1041-1058.

Mendoza, M., Bocco, G., & Bravo, M. (2002). Spatial prediction in hydrology: status and implications in the estimation of hydrological processes for applied research. *Progress in Physical Geography, 26*(3), 319-338.

Mothe, J., Chrisment, C., Dkaki, T., Dousset, B., & Karouach, S. (2006). Combining mining and visualization tools to discover the geographic structure of a domain. *Computers, environment and urban systems, 30*(4), 460-484.

Noirhomme-Fraiture, M., & Brito, P. (2011). Far beyond the classical data models: symbolic data analysis. *Statistical Analysis and Data Mining: the ASA Data Science Journal, 4*(2), 157-170.

Obayashi, S., & Sasaki, D. (2003). *Visualization and data mining of Pareto solutions using self-organizing map.* Paper presented at the Evolutionary multi-criterion optimization.

Okamoto, T., Hanaoka, Y., Aiyoshi, E., & Kobayashi, Y. (2014). Optimal Design of Buffer Material in the Geological Disposal of Radioactive Wastes Using the Satisficing Trade-off Method and a Self-Organizing Map. *Electrical Engineering in Japan, 187*(2), 17-32.

Olier, I., Vellido, A., & Ieee. (2006). Capturing the dynamics of multivariate time series through visualization using generative topographic mapping through time.

Pampalk, E. (2001). Limitations of the SOM and the GTM.

Pampalk, E., Rauber, A., & Merkl, D. (2002). Using smoothed data histograms for cluster visualization in Self-Organizing Maps. In J. R. Dorronsoro (Ed.), *Artificial Neural Networks - Icann 2002* (Vol. 2415, pp. 871-876).

Pan, M., Wood, E. F., Wójcik, R., & McCabe, M. F. (2008). Estimation of regional terrestrial water cycle using multi-sensor remote sensing observations and data assimilation. *Remote Sensing of Environment, 112*(4), 1282-1294.

Papa, F., Prigent, C., & Rossow, W. (2008). Monitoring flood and discharge variations in the large Siberian rivers from a multi-satellite technique. *Surveys in Geophysics, 29*(4), 297-317.

Postel, S. L., Daily, G. C., & Ehrlich, P. R. (1996). Human appropriation of renewable fresh water. *Science, 271*(5250), 785-788. doi:10.1126/science.271.5250.785

Powell, N., Foo, S. Y., & Weatherspoon, M. (2008). *Supervised and unsupervised methods for stock trend forecasting.* Paper presented at the System Theory, 2008. SSST 2008. 40th Southeastern Symposium on.

Reichle, R. H. (2008). Data assimilation methods in the Earth sciences. *Advances in water resources, 31*(11), 1411-1418.

Reusch, D. B., Alley, R. B., & Hewitson, B. C. (2007). North Atlantic climate variability from a self-organizing map perspective. *Journal of Geophysical Research: Atmospheres, 112*(D2).

Ruiz-Medina, M. (2012). New challenges in spatial and spatiotemporal functional statistics for high-dimensional data. *Spatial Statistics, 1*, 82-91.

Rynkiewicz, J. (2006). Self-organizing map algorithm and distortion measure. *Neural Networks, 19*(6-7), 830-837. doi:10.1016/j.neunet.2006.05.016

Salhi, M. S., Arous, N., & Ellouze, N. (2009). Principal temporal extensions of SOM: Overview. *International Journal of Signal Processing, Image Processing and Pattern Recognition, 2*(4), 61-84.

Sarlin, P. (2012). Self-organizing time map: An abstraction of temporal multivariate patterns. *Neurocomputing*.

Sarlin, P. (2013). *A self-organizing time map for time-to-event data.* Paper presented at the Computational Intelligence and Data Mining (CIDM), 2013 IEEE Symposium on.

Schroeder, M., Cornford, D., Farrimond, P., & Cornford, C. (2008). Addressing missing data in geochemistry: A non-linear approach. *Organic Geochemistry, 39*(8), 1162-1169. doi:10.1016/j.orggeochem.2008.02.016

Segev, A., & Kantola, J. (2012). Identification of trends from patents using self-organizing maps. *Expert Systems with Applications, 39*(18), 13235-13242.

Shanmuganathan, S., Sallis, P., & Buckeridge, J. (2006). Self-organizing map methods in integrated modelling of environmental and economic systems. *Environmental Modelling & Software, 21*(9), 1247-1256. doi:10.1016/j.envsoft.2005.04.011

Skupin, A. (2004). A picture from a thousand words. *Computing in Science & Engineering, 6*(5), 84-88.

Skupin, A., Biberstine, J. R., & Börner, K. (2013). Visualizing the topical structure of the medical sciences: a self-organizing map approach. *PloS one, 8*(3), e58779.

Smith, L. C. (1997). Satellite remote sensing of river inundation area, stage, and discharge: A review. *Hydrological Processes, 11*(10), 1427-1439.

Smith, L. I. (2002). A tutorial on principal components analysis. *Cornell University, USA, 51*(52), 65.

Sofia, G., Roder, G., Dalla Fontana, G., & Tarolli, P. (2017). Flood dynamics in urbanised landscapes: 100 years of climate and humans' interaction. *Scientific reports, 7*.

Steynor, A., Hewitson, B., & Tadross, M. (2009). Projected future runoff of the Breede River under climate change. *Water SA, 35*(4), 433-440.

Sturn, A., Quackenbush, J., & Trajanoski, Z. (2002). Genesis: cluster analysis of microarray data. *Bioinformatics, 18*(1), 207-208.

Svensen, J. (1998). PhD thesis: Generative Topographic Mapping. Aston University.

Tay, F. E. H., & Cao, L. J. (2001). Improved financial time series forecasting by combining support vector machines with self-organizing feature map. *Intelligent data analysis, 5*(4), 339-354.

Tino, P., & Nabney, I. (2002). Hierarchical GTM: Constructing localized nonlinear projection manifolds in a principled way. *Ieee Transactions on Pattern Analysis and Machine Intelligence, 24*(5), 639-656. doi:10.1109/34.1000238

Tiwari, M. K., Song, K.-Y., Chatterjee, C., & Gupta, M. M. (2013). Improving reliability of river flow forecasting using neural networks, wavelets and self-organizing maps. *Journal of Hydroinformatics, 15*(2), 486-502.

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika, 17*(4), 401-419.

Toth, E. (2009). Classification of hydro-meteorological conditions and multiple artificial neural networks for streamflow forecasting. *Hydrology and Earth System Sciences, 13*(9), 1555-1566.

Van der Maaten, L., Postma, E., & van den Herik, H. (2007). MATLAB toolbox for dimensionality reduction. *MICC, Maastricht University*.

Vanem, E., Huseby, A. B., & Natvig, B. (2012). A Bayesian hierarchical spatio-temporal model for significant wave height in the North Atlantic. *Stochastic environmental research and risk assessment, 26*(5), 609-632.

Varsta, M., Heikkonen, J., Lampinen, J., & Millán, J. D. R. (2001). Temporal Kohonen map and the recurrent self-organizing map: Analytical and experimental comparison. *Neural processing letters, 13*(3), 237-251.

Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (2000). *SOM toolbox for MATLAB 5*: Citeseer.

Vesanto, J., Sulkava, M., & Hollmén, J. (2003). *On the decomposition of the self-organizing map distortion measure.* Paper presented at the Proceedings of the workshop on self-organizing maps (WSOM'03).

Vorosmarty, C. J., Douglas, E. M., Green, P. A., & Revenga, C. (2005). Geospatial indicators of emerging water stress: An application to Africa. *Ambio, 34*(3), 230-236. doi:10.1639/0044-7447(2005)034[0230:gioews]2.0.co;2

Vorosmarty, C. J., Green, P., Salisbury, J., & Lammers, R. B. (2000). Global water resources: Vulnerability from climate change and population growth. *Science, 289*(5477), 284-288. doi:10.1126/science.289.5477.284

Wang, N., Biggs, T. W., & Skupin, A. (2013). Visualizing gridded time series data with self organizing maps: an application to multi-year snow dynamics in the Northern Hemisphere. *Computers, environment and urban systems, 39*, 107-120.

Wehrens, R., & Buydens, L. M. (2007). Self-and super-organizing maps in R: the Kohonen package. *J Stat Softw, 21*(5), 1-19.

Wikle, C. K., Berliner, L. M., & Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics, 5*(2), 117-154.

Wu, Y.-P., Wu, K.-P., & Lee, H.-M. (2012). *Stock trend prediction by sequential chart pattern via k-means and aprioriall algorithm.* Paper presented at the Technologies and Applications of Artificial Intelligence (TAAI), 2012 Conference on.

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., Meng, J. (2012). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *Journal of Geophysical Research: Atmospheres, 117*(D3).

Yang, Q., & Wu, X. (2006). 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making, 5*(04), 597-604.

Yin, H. (2005). *Self-organizing map as a natural kernel method.* Paper presented at the Neural Networks and Brain, 2005. ICNN&B'05. International Conference on.

Yin, H. (2008). The self-organizing maps: Background, theories, extensions and applications *Computational intelligence: a compendium* (pp. 715-762): Springer.

Yin, H., & Allinson, N. M. (1997). Bayesian learning for self-organizing maps. *Electronics letters, 33*(4), 304-305.

Yoshimi, M., Kuhara, T., Nishimoto, K., Miki, M., & Hiroyasu, T. (2012). Visualization of pareto solutions by spherical self-organizing map and it's acceleration on a GPU. *Journal of Software Engineering and Applications, 5*(03), 129.

Yu, L., Wang, S., & Lai, K. K. (2005). A novel nonlinear ensemble forecasting model incorporating GLAR and ANN for foreign exchange rates. *Computers & Operations Research, 32*(10), 2523-2541.

Websites:

[1] UN News centre: http://www.un.org/ apps/news/story.asp?NewsID= 56774#.WRzHPmjyvlV

[2] Scopus database: search 'self-organi*ing map*' refined by (water or hydrology or runoff or streamflow or river); accessed July 31, 2017

[3] UN Data: http://www.fao.org/statistics/databases/en/

[4] GRACE data portal: http://geoid.colorado.edu/grace/dataportal.html

[5] http://waterfootprint.org/en/resources/water-footprint-statistics/

[6] WRI's Aqueduct Global Flood Analyser tool: http://www.wri.org//resources/ maps/aqueduct-global-flood-analyzer

[7] Environment Canada's HYDAT database (National Water Data Archive): https://ec.gc.ca/rhc-wsc/default.asp?lang=En&n=9018B5EC-1