SOMiMS - Topographic Mapping in the Model Space

Xinyue Chen[1(✉)], Yuan Shen[2], Eder Zavala[3], Krasimira Tsaneva-Atanasova[4], Thomas Upton[5], Georgina Russell[5], and Peter Tino[1]

[1] School of Computer Science, University of Birmingham, Birmingham B15 2TT, UK
{xyc588,p.tino}@cs.bham.ac.uk
[2] Nottingham Trent University, Nottingham NG1 4FQ, UK
Yuan.Shen@ntu.ac.uk
[3] Centre for Systems Modelling and Quantitative Biomedicine (SMQB), University of Birmingham, Edgbaston B15 2TT, UK
[4] University of Exeter, Exeter EX4 4PY, UK
[5] University of Bristol, Bristol BS8 1TH, UK

Abstract. Learning in the model space (LiMS) represents each observational unit (e.g. sparse and irregular time series) with a suitable model of it (point estimate), or a full posterior distribution over models. LiMS approaches take the mechanistic information of how the data is generated into account, thus enhancing the transparency and interpretability of the machine learning tools employed. In this paper we develop a novel topographic mapping in the model space and compare it with an extension of the Generative Topographic Mapping (GTM) to the model space. We demonstrate these two methods on a dataset of measurements taken on subjects in an adrenal steroid hormone deficiency study.

Keywords: LiMS · Topographic mapping · Sparse · Irregular time series

1 Introduction

Topographic visualisation techniques have been established as an important tool in data analysis and data mining, e.g. Self-Organising Map (SOM) [7,8] and its probabilistic reformulation - Generative Topographic Mapping (GTM) [3,4]. However, most of these methods were designed to operate in a vectorial data space. Also, there has been an increasing interest in formulating SOM and GTM in the model space to deal with data with more complex structures, e.g. [5,10,14]. The approach in [10] establishes the Self-Organising mixture autoregressive (SOMAR) model, in which components in the construction of the topological mixture model are AR models to model foreign exchange (FX) rates. For visualising sets of symbolic sequences, [14] attempts to extend GTM in the model space based on a constrained mixture of discrete hidden Markov models. As an extension of [14], the work in [5] formulates GTM by extending it to the space of Hidden Tree Models for tree-structured data.

H. Yin et al. (Eds.): IDEAL 2021, LNCS 13113, pp. 502–510, 2021.
https://doi.org/10.1007/978-3-030-91608-4_50

In this work we are interested in using Learning in the Model Space (LiMS) approaches to deal with potentially sparsely sampled and noisy time series. We offer a non-GTM learning method for SOM formulated directly in the model space termed *SOM in model space* (SOMiMS), which takes advantage of the probabilistic model formulation of our base inferential models. Keeping the SOM philosophy translated into the model space, we can retain control over the neighbourhood-shrinking rate and make the components move directly in the direction of gradients with respect to the model parameters of the inferential model class. We also extend the GTM to the model space. In general, GTM offers a clean formulation, but direct manipulation of dynamic neighbourhood size is not possible and prototype movements are controlled only implicitly through parameters of the embedding kernel regression function. We demonstrate the two methods on a real dataset of measurements taken on subjects in an adrenal steroid hormone deficiency study.

The rest of the paper is organised as follows: Sect. 2 briefly introduces the SOMiMS and extended GTM models. Section 3 presents the base inferential model and describes the real data set we use. Section 4 provides experimental results. We conclude the paper with a brief summary of key elements in Sect. 5.

2 Topographic Mapping of Time Series in the Model Space

Consider a set of time series $\mathcal{Y} = \{Y^{(1)}, Y^{(2)}, ..., Y^{(N)}\}$, $n = 1 : N$. The n-th time series will be denoted by $Y^{(n)} = \{Y_t^{(n)}\}_{t=1:T_n}$, where T_n is the length of n-th time series. The individual time series can be of different length, but we assume that there is a unique time grid where the observations are allowed to be taken. Generalization to time grids specific for each individual time series is relatively straightforward, but beyond the scope of this paper.

In our LiMS approach, each time series is considered as a set of partial observations of some underlying mechanistic model parametrised via $\vec{\theta} \in \mathbb{R}^d$ [12]. Mathematically, this parametric mechanistic model will be formulated as a multivariate Ordinary Differential Equation (ODE).

The topographic mapping of a vectorial data set is given by a (usually) nonlinear mapping from input vector space to a low-dimensional topographic mapping space (usually two dimensional) [8,15]. Topographic mappings we are interested in will operate in the model space rather than the original signal space. Each node of a topographic map corresponds to an instance from the inferential model class. The aim is to represent each time series as an individual projection on the topographic map.

2.1 SOMiMS

We will assume a SOM structure with $k \times k$ nodes. Each node i will be assigned an inferential model representative parameterized by a parameter vector $\vec{\theta}_i$. Considering the n-th time series $Y^{(n)}$, the log likelihood of i-th node is

$$\mathcal{L}(Y^{(n)}|\vec{\theta}_i, \Sigma) = \ln \prod_{t=1}^{T_n} p(Y_t^{(n)}|\vec{\theta}_i, \Sigma) = \sum_{t=1}^{T_n} \ln p(Y_t^{(n)}|\vec{\theta}_i, \Sigma),$$

where Σ collects parameters of the observational noise. Since each time series may have different length, we will operate with log likelihood per observation,

$$\mathcal{Q}(Y^{(n)}|\vec{\theta}_i, \Sigma) = \frac{1}{T_n}\mathcal{L}(Y^{(n)}|m_{\vec{\theta}_i}, \Sigma) = \frac{1}{T_n}\sum_{t=1}^{T_n} \ln p(Y_t^{(n)}|\vec{\theta}_i, \Sigma).$$

The "quality measure" of the i-th node, given the time series $Y^{(n)}$, is obtained by renormalization through all nodes,

$$\overline{\mathcal{Q}}(Y^{(n)}|\vec{\theta}_i, \Sigma) = \frac{\mathcal{Q}(Y^{(n)}|\vec{\theta}_i, \Sigma)}{\sum_a \mathcal{Q}(Y^{(n)}|\vec{\theta}_a, \Sigma)} = \frac{-\mathcal{Q}(Y^{(n)}|\vec{\theta}_i, \Sigma)}{\sum_a -\mathcal{Q}(Y^{(n)}|\vec{\theta}_a, \Sigma)}.$$

Note that $-\mathcal{Q}(Y^{(n)}|\vec{\theta}_i, \Sigma)$ can be thought of as the information (per observation) the node i contains about the time series $Y^{(n)}$. The quality measure $\overline{\mathcal{Q}}(Y^{(n)}|\vec{\theta}_i, \Sigma)$ is then the normalized information the node i holds on $Y^{(n)}$, renormalized in the competition across all the nodes.

Adopting a Gaussian observational noise model, we have:

$$p(Y_t^{(n)}|\vec{\theta}_i, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}}\left\{\exp(-\frac{1}{2}(Y_t^{(n)} - X_{i,t})^{\mathbf{T}}\Sigma^{-1}(Y_t^{(n)} - X_{i,t})\right\},$$

where $X_{i,t}$ is the (noiseless) observational vector at time t obtained from the inferential model parametrized with $\vec{\theta}_i$. Here we assume a homoscedastic process with a fixed covariance Σ. Again, generalization to time varying noise model is relatively straightforward, but out of the scope of the present paper.

During the training phase, in each iteration we randomly pick (with replacement) a time series $Y^{(n)}$ from the data set $\mathcal{Y}$. In each iteration, rather than updating the winner node (the node with maximum quality $\mathcal{Q}$) and its neighborhood as in the classic topographic mapping [8,15], we will update each node and its neighborhood according to the normalized quality measure $\overline{\mathcal{Q}}(Y^{(n)}|\vec{\theta}_i, \Sigma)$. This is needed, since especially for sparsely observed and/or noisy data, several prototypical node models can be likely and committing to a single winner node may bias the final topographic map. All nodes are considered in turn. For the i-th node, nodes c its neighbourhood are updated as

$$\vec{\theta}_c(l+1) = \vec{\theta}_c(l) \ + \ \overline{\mathcal{Q}}(Y^{(n)}|\vec{\theta}_i, \Sigma) \cdot h_{(c,i)}(l) \cdot \eta(l) \cdot \nabla_{\vec{\theta}_c} Q(Y^{(n)}|\vec{\theta}_c(l)),$$

where $h_{(c,i)}(l)$ is the neighborhood function and $\eta(l)$ is the learning rate, both monotonically decreasing in algorithmic time steps l:

$$\eta(l) = \eta(0) \cdot \exp\left(-\frac{l}{\tau}\right) \tag{1}$$

$$h_{(c,i)}(l) = \exp\left\{-\frac{||c-i||^2}{2(\alpha(l))^2}\right\} \tag{2}$$

$$\alpha(l) = \alpha(0) \cdot \exp\left(-\frac{l}{\tau}\right) \tag{3}$$

Here τ is a time scale parameter and l is the current iteration index [9]. $\eta(0)$ is the initial learning rate in the power series learning rate function [13]; $\alpha(0)$ is the initial neighborhood size.

Note two crucial aspects of the SOMiMS methodology: (1) for each time series $Y^{(n)}$ we perform a double scan through the grid nodes, the outer scan through the pivotal nodes $\vec{\theta}_i$ with the inner scan through their neighbours $\vec{\theta}_c$; (2) the node updates are performed in the model space in the directions improving the node likelihoods, given $Y^{(n)}$, i.e. directions given by the gradient $\nabla_{\vec{\theta}_c} Q(Y^{(n)}|\vec{\theta}_c)$.

To visualize the time series data, we embed the $k \times k$ node grid in a square, e.g. $[-1,1]^2$, resulting in the embedded grid of points $\mathbf{g}_i \in [-1,1]^2$. The n-th time series $Y^{(n)}$ is then visualized (GTM-style) in $[-1,1]^2$ as the mean of the posterior distribution over the grid points [4,14],

$$Proj(Y^{(n)}) = \sum_{i=1}^{J} P(\mathbf{g}_i|Y^{(n)}, \vec{\theta}_i, \Sigma) \cdot \mathbf{g}_i,$$

where (imposing a uniform prior distribution over the grid),

$$P(\mathbf{g}_i|Y^{(n)}, \vec{\theta}_i, \Sigma) = \frac{p(Y^{(n)}|\vec{\theta}_i, \Sigma)}{\sum_{j=1}^{J} p(Y^{(n)}|\vec{\theta}_j, \Sigma)}.$$

2.2 Generative Topographic Mapping in the Model Space

As an alternative to SOMiMS, we also extend the Generative Topographic Mapping (GTM) [4] to the model space along the lines of [14] and [5]. Consider a 2-dimensional latent space $\mathcal{H}$, e.g. $\mathcal{H} = [-1,1]^2$. The aim is to represent each time series using this latent space through imposing a uniform prior over regular grid $\{\mathbf{g}_i\}_{i=1}^{J}$ of J points $\mathbf{g}_i \in \mathcal{H}$ covering the latent space. One imposes a function $\ell(\mathbf{g}; W)$ parametrized by W that maps the latent space into the model space:

$$\ell(\mathbf{g} : W) = W\phi(\mathbf{g}),$$

where W is a $d \times M$ matrix of parameters that governs the mapping $\ell(\mathbf{g}; W)$ and $\phi(\mathbf{g})$ contains M fixed basis functions $\phi_m(\mathbf{g}) : \mathcal{H} \to \mathbb{R}$. Note that $\ell(\mathbf{g}_i : W)$ now plays the role of the i-th prototypical model setting $\vec{\theta}_i$ in SOMiMS.

Given the n-th $Y^{(n)}$ time series in $\mathcal{Y}$ of length T_n, we can again calculate its probability, given the forward ODE model parametrized by $\ell(\mathbf{g}_i : W)$ and observational noise model parametrized by Σ:

$$p(Y^{(n)}|\mathbf{g}_i, W, \Sigma) = \prod_{t=1}^{T_n} p(Y_t^{(n)}|\ell(\mathbf{g}_i : W), \Sigma).$$

Since GTM is a flat mixture model of the latent grid, we have for the data log likelihood:

$$\mathcal{L} = \sum_{n=1}^{N} \ln \left\{ \frac{1}{J} \sum_{i=1}^{J} p(Y^{(n)}|\mathbf{g}_i, W, \Sigma) \right\}. \tag{4}$$

Expectation Maximization (EM) algorithm is used to obtain W by maximizing $\mathcal{L}$. The 'responsibilities' of grid points $\mathbf{g}_i, i = 1 : J$, for time series $Y^{(n)}$ are calculated in the E-step as

$$R_{in} = p(\mathbf{g}_i|Y^{(n)}, W, \Sigma) = \frac{p(Y^{(n)}|\mathbf{g}_i, W, \Sigma)}{\sum_j p(Y^{(n)}|\mathbf{g}_j, W, \Sigma)}.$$

The expected complete-data log likelihood then takes the form

$$\langle \mathcal{L}_{comp} \rangle = \sum_{n=1}^{N} \sum_{j=1}^{J} R_{in} \ln\{p(Y^{(n)}|\mathbf{g}_i, W, \Sigma)\}.$$

The M-step consists of maximizing $\langle \mathcal{L}_{comp} \rangle$ with respect to W.

After training, each time series $Y^{(n)}$ can be visualized in the latent space $\mathcal{H}$ as the mean of the posterior distribution over the latent grid points [4,14],

$$Proj(Y^{(n)}) = \sum_{i=1}^{J} R_{in} \cdot \mathbf{g}_i.$$

3 Biomedical Background, Mechanistic Model, and the Data

Major adrenal steroid hormones are synthesized in different areas of the adrenal cortex[2]. We are particularly interested in the glucocorticoid and mineralocorticoid pathways. An appreciation of these pathways helps to understand the different forms of congenital adrenal hyperplasia (CAH) and isolated hypoaldosteronism characterized by defects in functionality of enzymes involved in adrenal steroid hormone synthesis[1].

The real data set we used in the experiments includes three conditions: Healthy control, Cushing's and Primary Aldosteronism. Cushing's usually results from the excessive production of Cortisol. Primary Aldosteronism is corresponding to the Aldosterone excess. The dataset contains subject-specific multivariate time series (of Corticosterone, Aldosterone, Cortisol, and Cortisone) obtained from 60 subjects covering the three conditions - Control (30), Cushing's (15) and Primary Aldosteronism (15). Each time series was sampled every 20 min within 24 h. However, there are some missing values due to certain operational problems. Thus, the length of time series may vary.

Below we introduce the mechanistic model representing the adrenal steroid hormone biosynthesis pathway that will be used to represent the observed data

Table 1. Model parameters.

Parameter	Description	Parameter	Description
k_C	Corticosterone synthesis rate	k_A	Aldosterone synthesis rate
k_F	Cortisol synthesis rate	k_E	Cortisone synthesis rate
k_b	Cortisone to Cortisol conversion rate	γ_C	Corticosterone degradation
γ_A	Aldosterone degradation rate	γ_F	Cortisol degradation rate
γ_E	Cortisone degradation rate	α_c	Amplitude of circadian drive
T_s^c	Phase shift of circadian drive	σ	Asymmetry of circadian drive
β	Offset of circadian drive	α_u	Amplitude of ultradian drive
T_s^u	Phase shift of ultradian drive	n_p	Number of ultradian pulses

in the LiMS framework. In the first instance, four hormones (Corticosterone (C), Aldosterone (A), Cortisol (F), Cortisone (E)) are modeled through a system of ODEs. The joint model reads:

$$\frac{d}{dt}C = k_C\varphi_c(t) - k_A C - \gamma_C C \qquad \frac{d}{dt}A = k_A C\varphi_u(t) - \gamma_A A$$

$$\frac{d}{dt}F = k_F\varphi_c(t) - k_E F + k_b E - \gamma_F F \qquad \frac{d}{dt}E = k_E F - k_b E - \gamma_E E,$$

where $\varphi_c(t)$ and $\varphi_u(t)$ are periodic circadian and ultradian drives specified by

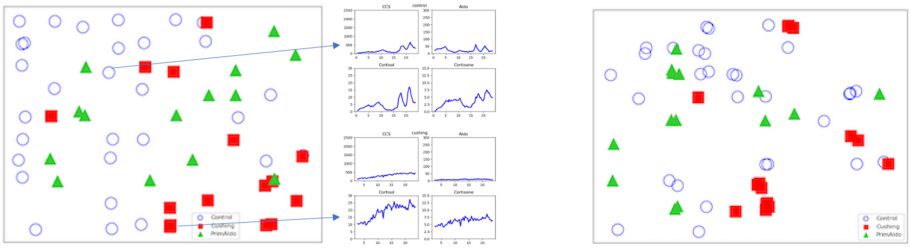
$$\varphi_c(t) = \alpha_c \sin(2\pi(t + T_s^c) + \sigma * \sin(2\pi(t + T_s^c))) + \beta$$

$$\varphi_u(t) = 1 + \alpha_u \sin(2\pi(t + T_s^u)n_p).$$

All sixteen parameters used in models and their descriptions are listed in Table 1. Three drive parameters were fixed to $\alpha_c = 1$, $\alpha_u = 1$, $T_s^u = 0.5$, leaving thirteen free parameters.

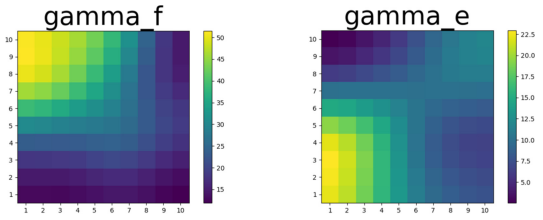
4 Experiments and Results

In this section we present results of applying the SOMiMS and extended GTM methodologies on the real adrenal steroid dataset. We used a 10×10 grid and the models were initialized by applying the classic Self-Organising Map [8] in the signal space on all 60 subjects. The missing values were imputed using Gaussian Process model [11]. After training, most grid points of the classical SOM contain in their Voronoi compartments one or several time series. For grid points with no time series assigned, we used time series of their closest neighbours on the grid. Each grid point was then transformed to the model space by calculating the maximum likelihood parameter estimate obtained on the time series assigned to it. The 10×10 classic map thus became a map in the model space, each grid point corresponding to a setting of the 13-dimensional parameter vector.



(a) SOMiMS (b) Extended GTM

Fig. 1. Topographic visualization of the data obtained by the SOMiMS (a) and Extended GTM (b) models. The Control, PrimaryAldo and Cushing's conditions are marked as blue circles, green triangles and red squares, respectively. (Color figure online)



(a) (b)

Fig. 2. Parameter heat maps of γ_F (a) and γ_E (b) for the SOMiMS model.

The classic SOM (initialization stage for SOMiMS and extended GTM) was trained for 300 epochs with initial learning rate and initial neighbourhood size equal to 0.2 and 6, respectively. For SOMiMS, the initial learning rate and neighbourhood size were set to 0.1 and 2, respectively. This accounts for the fact that some very rough initial topographic organisation was already achieved in the classic SOM. SOMiMS was trained for 200 epochs. In the extended GTM, we employed $M = 4 \times 4 = 16$ basis functions ϕ_m and one additional constant basis function as the bias term. Basis functions were radial Gaussian functions with the same width $\sigma = 1$. The likelihood leveled up after 80–100 E-M cycles.

Topographic maps obatained by SOMiMS and extended GTM are shown in Fig. 1. The models were trained in a completely unsupervised manner, i.e. markers on the data projections signifying their corresponding conditions were not used during the training in any way. Overall, both topographic maps constructed in the model space show a good degree of separation of the conditions, noting that this is a noisy data set measured on real subjects. Both maps also show a tendency of sub-grouping the Cushing's cohort into at least two sub-populations. The signal plots to the right of the SOMiMS map illustrate the steroid time series corresponding to the selected projections (subjects).

Table 2. Confusion matrix

		Actual value					
		SOMiMS			Extended GTM		
Predicted value		Control	Cushing's	Primary Aldosteronism	Control	Cushing's	Primary Aldosteronism
	Control	0.80	0.10	0.10	0.68	0.27	0.05
	Cushing's	0.13	0.67	0.20	0.13	0.80	0.07
	Primary Aldosteronism	0.21	0.29	0.50	0.14	0.22	0.64

A detailed bio-medical analysis of the visualization plots is beyond the scope of this paper. We nevertheless stress that one of the primary advantages of topographic maps in the model space is the opportunity to readily interpret the topographic data organization from the mechanistic point of view of the underlying processes that generated the data. To that end, one can create parameter plots where the values learnt for each individual mechanistic model parameter across the prototypes on the grid are shown as heat maps. As an example, Fig. 2 presents parameter heat maps for Cortisol and Cortisone degradation rates, γ_F and γ_E, respectively. The two parameters have low values in the regions of the SOMiMS topographic map containing Cushing's projections. It is clear that the Cortisol excess associated with the Cushing's condition is partially caused by reduced degradation rates of Cortisol and Cortisone (which is positively coupled to Cortisol through k_b).

To quantify the amount of separation of the different conditions on the visualization plot, we also performed K-nearest neighbor (KNN) classification [6] on the map projections. Based on the cross-validated hyper-parameter tuning, we picked $K = 3$. Table 2 presents KNN confusion matrices for SOMiMS and Extended GTM projections. Thanks to the possibility of explicit control over the topographic map formation offered by SOMiMS (neighbourhood function and its shrinkage), the projections on the SOMiMS map are much more spread than those of the Extended GTM. Obviously, topographic organization does not correspond directly to the classification performance. After all, this is an unsupervised learning scenario. Such an analysis does, however, demonstrate that a full formation of a topographic map may disrupt cases of multiple projections in a very close neighbourhood of the visualisation space - a scenario that could yield good distance-based classifications, but is not preferable from the visualisation point of view.

5 Conclusion

We have presented a new learning method for SOM formulated directly in the model space, termed SOM in model space (SOMiMS), together with an extended GTM formulation in the model space for visualizing sets of sparse time series. We illustrated the methodologies on a real data set of measurements on subjects with different steroid hormone biosynthesis conditions. To that end, we formed a

parameterized mechanistic inferential model in the form of coupled ordinary differential equations and demostrated how the topographic maps could be formed in the space of such inferential models, given the data.

Compared to the traditional approaches working in the signal space, SOMiMS and extended GTM are not only naturally able to deal with sparse time series, but also capable of taking the mechanistic information into account, creating scientifically interpretable readily data visualisations.

References

1. Arlt, W., et al.: Steroid metabolome analysis reveals prevalent glucocorticoid excess in primary aldosteronism. JCI Insight **2**(8), e93136 (2017)
2. Arlt, W., Stewart, P.M.: Adrenal corticosteroid biosynthesis, metabolism, and action. Endocrinol. Metab. Clin **34**(2), 293–313 (2005)
3. Bishop, C.M., Svensén, M., Williams, C.K.: Developments of the generative topographic mapping. Neurocomputing **21**(1–3), 203–224 (1998)
4. Bishop, C.M., Svensén, M., Williams, C.K.: GTM: the generative topographic mapping. Neural Comput. **10**(1), 215–234 (1998)
5. Gianniotis, N., Tino, P.: Visualization of tree-structured data through generative topographic mapping. IEEE Trans. Neural Netw. **19**(8), 1468–1493 (2008)
6. Keller, J.M., Gray, M.R., Givens, J.A.: A fuzzy k-nearest neighbor algorithm. IEEE Trans. Syst. Man Cybern. **4**, 580–585 (1985)
7. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biol. Cybern. **43**(1), 59–69 (1982)
8. Kohonen, T.: Essentials of the self-organizing map. Neural Netw. **37**, 52–65 (2013)
9. Natita, W., Wiboonsak, W., Dusadee, S.: Appropriate learning rate and neighborhood function of self-organizing map (som) for specific humidity pattern classification over southern thailand. Int. J. Model. Optim. **6**(1), 61 (2016)
10. Ni, H., Yin, H.: A self-organising mixture autoregressive network for fx time series modelling and prediction. Neurocomputing **72**(16–18), 3529–3537 (2009)
11. Rasmussen, C.E., Williams, C.: Gaussian Processes for Machine Learning, vol. 32, p. 68. The Mit Press, Cambridge (2006)
12. Shen, Y., Tino, P., Tsaneva-Atanasova, K.: Classification framework for partially observed dynamical systems. Phys. Rev. E **95**(4), 043303 (2017)
13. Stefanovič, P., Kurasova, O.: Visual analysis of self-organizing maps. Nonlinear Anal. Model. Control **16**(4), 488–504 (2011)
14. Tino, P., Kabán, A., Sun, Y.: A generative probabilistic approach to visualizing sets of symbolic sequences. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 701–706 (2004)
15. Torma, M.: Kohonen self-organizing feature map and its use in clustering. In: ISPRS Commission III Symposium: Spatial Information from Digital Photogrammetry and Computer Vision, vol. 2357, pp. 830–835. International Society for Optics and Photonics (1994)