NODE GROUPING FOR CONSUMER DEMAND ESTIMATION USING A SELF-ORGANIZING MAP

S.M. Masud Rana\textsuperscript{1}, Dominic L. Boccelli\textsuperscript{1}, Angela Marchi\textsuperscript{2}, and Graeme C. Dandy\textsuperscript{2}

\textsuperscript{1}Dept. of Chemical and Environmental Engineering, University of Cincinnati, Cincinnati, OH 45221, USA
\textsuperscript{2}School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, 5005, Australia
\textsuperscript{*} ranasd@mail.uc.edu

ABSTRACT

Real-time demand estimation is crucial for real-time management of a drinking water system (DWS), which includes minimization of operating cost, emergency response, water quality maintenance, etc. On the other hand, real-time demand estimation from a limited number of measurements is not possible without clustering consumer nodes assuming that the consumer nodes within a cluster have the same (or similar) temporal demand patterns. However, clustering nodes within a DWS model without consideration of the spatial distribution of measurement locations may make the demands unobservable within the estimation process. A machine learning based clustering algorithm, called the self-organizing map (SOM), is presented that clusters the nodes of a DWS model based on the sensitivities of the measurement locations to the nodal demands. The algorithm was applied to the Net3 network, an example network distributed with EPANET. The performance of the SOM based clustering was evaluated using data generated with a hypothetical cluster scenario representing unknown actual consumer distributions. Demand multipliers for the SOM clusters were estimated with a Bayesian approach. The SOM clusters showed good observability properties, and resulted in good representation between the simulated and synthetically generated flow measurements. However, the simulated flow measurements at unmonitored locations varied widely, which is an important consideration if adequate transport characteristics are of interest.

Keywords: Self-Organizing Map, Water Distribution Network, Demand

1 BACKGROUND

Demand driven modeling of a drinking water system (DWS) model, such as modeling using EPANET \textsuperscript{1}, is a popular modeling method for the design and management of DWSs in the USA and other developed countries. For adequately pressurized DWSs, calibrated demand driven models can provide good approximations of the network hydraulics that are essential for ensuring system reliability (i.e., adequate supply, pressure, and water quality.) On the other hand, consumer demands are one of the largest sources of uncertainty in real-time modeling of a DWS. Traditionally, long term (i.e., monthly or annually) averaged consumer demands are estimated from billing data and short term (minutes to hours) temporal variations (called demand patterns) are assumed fixed and estimated indirectly using measurements from different locations in the network \textsuperscript{2}. DWS modeling with deterministic demand patterns has provided many benefits to DWS managers, such as, long term planning and water quality management, but such demand patterns often fail to represent the daily variability in

1. EPANET: https://www.epa.gov/water-enforcement/epanet-software

...
consumer demands and hence the network hydraulics. On the other hand, real-time operations, such as daily operational planning, emergency event management, and demand forecasting, require real-time demand information \cite{3, 4, 5}. With widespread implementation of supervisory control and data acquisition (SCADA) systems, real-time demand estimation has become a possibility. Flow and pressure measurements are continuously monitored at key locations (such as, tank inlets and pump outlets) and can be used to estimate real-time consumer demands. As well as SCADA information, a vast amount of spatial information about consumers, such as, land-use, socioeconomic data, etc., are also available, which can play important role in demand estimation.

In general, demand estimation using the observed information will require some degree of clustering nodes. However, the number of unknown demand multiplier must be equal to or less than the number of available measurement locations, which will make the estimation problem even- or over-determined. The assumption inherent in such clustering is that the consumers within a cluster have almost perfectly correlated temporal demand patterns. On the other hand, observability of clustered demands is strongly linked to the spatial distribution of the measurement locations \cite{6, 7} in relation to the clusters. When the parameters are unobservable, the parameter estimation problem may become ill-conditioned or provide non-unique solutions. Clustering of consumer nodes in DWS models has been an overlooked problem, and although several clustering methods exists in the literature (e.g., \cite{8, 9, 10}) for different objectives, very few have considered demand observability of the clustered nodes as an objective (a notable exception is \cite{11}.) As such, the aforementioned clustering methods in most cases are not useful for demand estimation purposes.

The objective of this study is to present a clustering method using measurement sensitivities to demand perturbations through a self-organizing map (SOM). The SOM is an unsupervised machine learning method that is capable of handling numeric and non-numeric information, such that non-numeric spatial information about consumers can be utilized in addition to the measurement sensitivities. The SOM projects high dimensional data to a lower dimensional space in an orderly fashion \cite{12}. The ordered projection, e.g., in two dimensions, can then be used to visualize the stratification or grouping inherent in the data. A large number of disciplines have used SOM for unsupervised clustering, for example, genetic research, financial research, climate change, and data division \cite{13, 14, 15}. The SOM based method was applied to the example network Net3, which is distributed with EPANET2.0, to cluster consumer nodes using measurement sensitivity information. The aggregated demands of the clustered nodes were estimated using a Bayesian parameter estimation method and showed a high level of observability, i.e., the demand uncertainties were well bounded and tight.

2 METHODOLOGY

The SOM based clustering method was applied to Net3, which has 97 nodes and 117 pipes. Synthetic measurements from pre-specified locations were utilized to estimate distributions of clustered demands using a Markov chain Monte Carlo (MCMC) based algorithm, which is a Bayesian parameter inference method. The impacts of demand nodes on measurement locations are equivalent to the measurement sensitivities to demand perturbations given by the Jacobian in a non-linear least squares formulation for demand estimation. These impacts are calculated by the hydraulic solvers during DWS simulations, and can be easily obtained from popular demand driven DWS solvers, such as EPANET \cite{1}. Recently, excellent reviews associated with the mathematical formulation of the
Jacobian, which can be used for SOM based clustering, have been performed [7, 16].

2.1 The SOM Algorithm and Node Clustering

A typical two-dimensional (2D) SOM map contains a 2D grid of “neurons,” where each neuron represents a certain region of the data space. The distribution of the neurons in the 2D grid of a trained SOM map will reflect the different strata in the training data. For example, if a particular input dataset has three distinct groups, the neurons of a 2D SOM map for that dataset will arrange themselves in three regions. The following steps generally describe the procedure for the development of a 2D SOM map for a particular dataset: (1) Specify the SOM grid size and initialize the vectors representing the SOM neurons using training data points; (2) Compare each data point with the neurons to find the neuron closest to that data point, called the “winner” neuron; and (3) Update the vectors of the winner neuron and the neurons that are within a certain radius of the winner neuron on the 2D grid so as to make the vectors come closer to that data point. Steps 2 and 3 are repeated for all data points and for multiple epochs, and as training progresses the radius of influence of the winner neurons is reduced. The process of updating multiple neurons with a single input is key to the SOM algorithm, which has a smoothing effect on the output plane that results in an ordered plane reflecting different strata of the input space. Figure 1 shows a hypothetical 4 × 6 SOM map that, after training, has revealed three regions of its training dataset represented by three groups of neurons. Although, crisp boundaries are shown in Figure 1 for convenience, SOM neurons, in general, will vary across the grid smoothly. Thus, the size of a SOM grid impacts the resolution of the stratification of the data space – a small grid may aggregate multiple strata, while a large grid would result in very gradual neuron variation, which might make the formation of neuron groups difficult. A more general discussion and mathematical descriptions of the SOM algorithm can be found in [12, 17].

After successful training of the SOM map, the consumer nodes can be classified by the indexes of the neurons corresponding to each of the data points. Since the size of the SOM grid would likely be greater than the number of desired clusters, a second level of clustering of the trained SOM map is often required. For example, in Figure 1 if only two clusters were desired, and assuming vectors of the group 2 neurons were closer to those of group 3, neuron groups 2 and 3 would be combined. Multiple studies provide automated methods for grouping the output layer of SOM maps (e.g., [18]). In this study, the output layer was grouped visually since the output layer was crisp enough for the small Net3 network. The SOM-Toolbox software package [19] for Matlab (MathWorks, Inc., Natick, Massachusetts) was used to generate the SOM based clusters in this paper.
2.2 Demand Estimation

To assess the performance of demand observability of the SOM based clusters, a scenario with three arbitrary actual clusters was created and simulated flow measurements were assumed to be collected from 5 locations throughout the network. The left panel of Figure 2 shows the actual cluster scenario with three different consumer groups along with the locations of the flow monitoring points. Flow measurements have already been shown by several authors to be more important than pressure measurements for demand estimation purposes, hence in this paper only flow sensitivities have been used for clustering purposes. The flow measurement locations correspond to the three tank inlets/outlets (link id: 20, 40, 50), a pipe (link id: 223), and a pump station (link id: 60). Since tank inlet/outlet and pump flows are typically measured in real networks, the chosen measurement locations were deemed to be representative of actual practice. Link id: 223 was chosen so as to obtain additional information about the north-west part of the network. Assuming the number of actual consumer groups is unknown, the objective was to generate four clusters such that the demand estimation problem would be overdetermined with the measurements from the five locations. Normally distributed measurement errors with standard deviation equal to 2% of the magnitude of the measurements were added to simulate flow measurement errors. The network model was assumed to be well calibrated with no other sources of error (e.g., tank levels, pipe roughness, etc.) considered. The network was then clustered using the measurement sensitivities at the specified measurement locations with the SOM algorithm. Demand multiplier distributions for a single time step for the clusters were then estimated using the MCMC algorithm. During the demand estimation procedure and clustering, the actual clusters were assumed to be unknown, i.e., the actual clusters were only used for synthetic measurement generation.

3 RESULTS

A $4 \times 5$ SOM grid was chosen to be trained with the sensitivity information corresponding to the 5 measurement locations of the Net3 network. Four regions of neurons could be easily discerned from the output layer and the corresponding node clustering is shown in the right panel of Figure 2. The SOM based clusters are different from the actual cluster scenario, hence the demand multipliers estimated for these four clusters using the observed flow measurements represent an abstract average of the actual demand multipliers of the nodes.
Table 1: Simulated flows at the measurement locations for the actual cluster scenario and the SOM based clustering. Measurements are assumed to be the summation of the simulated actual flows and errors.

<table>
<thead>
<tr>
<th>Pipe id</th>
<th>Actual Flow (GPM)</th>
<th>Simulated Error (GPM)</th>
<th>Synthetic Measurements (GPM)</th>
<th>SOM Flow (GPM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>1287.90</td>
<td>34.29</td>
<td>1322.20</td>
<td>1322.23</td>
</tr>
<tr>
<td>40</td>
<td>796.49</td>
<td>11.39</td>
<td>807.88</td>
<td>807.88</td>
</tr>
<tr>
<td>50</td>
<td>1196.91</td>
<td>-36.99</td>
<td>1159.91</td>
<td>1159.91</td>
</tr>
<tr>
<td>223</td>
<td>-1261.80</td>
<td>-0.21</td>
<td>-1261.59</td>
<td>-1261.58</td>
</tr>
<tr>
<td>60</td>
<td>13276.68</td>
<td>164.98</td>
<td>13441.66</td>
<td>13275.74</td>
</tr>
</tbody>
</table>

Table 1 shows the synthetic flow measurements generated with the actual cluster scenario with multipliers 1.34, 1.71 and 1.10 for clusters one through three, respectively. The expected demand multipliers of the four SOM based clusters, estimated using the measurements from the actual cluster scenario, were 1.29, 1.70, 1.59, and 1.47, respectively. Table 1 also shows the flows obtained with the means of the multiplier distributions for the SOM based clusters. A good match at the measurement locations was observed with a root mean squared error (RMSE) of 74.2 GPM, representing an average error of about 0.25%. While simulated flows for the actual clusters and the SOM based clusters are very similar at the measurement locations, flows at the unmonitored locations varied widely. The deviation of the underlying hydraulics of the entire network from the actual can be assessed by calculating the RMSE of flow considering all pipes in the network, a metric that obviously cannot be calculated for a real network, unless, of course, the flows in all pipes are actually monitored. Figure 3 shows the histogram and empirical cumulative distribution of the flow errors in all pipes. The total RMSE for all pipes was about 1800 GPM. Two other orientations of actual clusters with 4 and 5 total clusters, respectively, were also used to assess the SOM based clustering (results not shown) and very similar results were obtained i.e., flows at the measurement locations matched well with the synthetically derived actual flows, while flows at the unmonitored locations varied widely.

The distribution of the demand multipliers of the SOM based clusters, estimated using the synthetic measurements generated with the actual cluster scenario, are shown in Figure 4. As expected the demand multipliers were tightly bounded and represented some degree of averaging of the actual flows.
nodal demands. The demand multiplier distributions estimated using synthetic data generated with the other two actual cluster scenarios were also tightly bounded and smooth, indicating good observability of the SOM clusters with respect to the measurement locations.

4 SUMMARY

A new method of clustering drinking water distribution system nodes using the SOM algorithm was presented. Sensitivity of different nodes, as measured by the sensitivity of measurements to changes in nodal demands, were used as the criteria to cluster nodes. After training, the SOM stratified the sensitivity information and multiple nodal groups were formed. From the stratified sensitivity map, the desired number of clusters could be obtained. The clusters identified with the SOM were used to approximate the underlying hydraulics of some arbitrarily generated actual clusters using an MCMC-based parameter estimation procedure. The modeled flow using the estimated multipliers for the SOM based clusters matched the data obtained at the monitored locations very well. On the other hand, the approximation of the flows in the non-monitored pipes was relatively poor. This study highlights the difficulty of clustering a network with the objective of matching monitored data alone. Spatial information about consumers and land use that can be easily obtained from zoning maps or satellite imagery may also have a valuable role in clustering a network to improve the approximation of underlying hydraulics. The SOM, being capable of handling non-numeric information, can also utilize such spatial information along with sensitivity information to produce clusters that will not only ensure demand observability but also will improve the modeling of the actual network hydraulics.

References


