Discrete entropy theory for optimal redesigning of salinity monitoring network in San Francisco bay

Amir Boroumand and Taher Rajaee

ABSTRACT

The paper presents an entropy-based method for designing an optimum bay water salinity monitoring network in San Francisco bay (S.F. bay) considering maximum-monitoring-information and minimum-data-lost criteria. Due to cost concerns, it is necessary to design the optimal salinity monitoring network with a minimal number of sampling stations to provide reliable data. The monthly data recorded from January 1995 to December 2014 were obtained over 37 active stations located in S.F. bay and is applied in the research. Transinformation entropy in discrete mode is used to calculate the stations’ optimum distance. The discrete approach uses the frequency table to calculate transinformation measures. After calculating these measures, a transinformation-distance (T-D) curve is developed. Then, the optimum distance between salinity monitoring stations is elicited from the curve. The study shows that the S.F. bay salinity monitoring stations provide redundant information and the existing stations can be reduced to 21 with an approximate distance of 7.5 km. The coverage of the proposed monitoring network by using the optimum distance is complete and the system does not generate redundant data. The results of this research indicate that transinformation entropy is a promising method for the design of monitoring networks in bays such as those found in San Francisco bay.

Key words | optimum distance, salinity, San Francisco bay, transinformation entropy, water quality monitoring networks

INTRODUCTION

Preserving and optimal use of water resources are two main aspects of sustainable development in every country. It can be said that water management planning in every country is dependent on water resource availability. Knowing qualitative and quantitative problems in water resource monitoring systems is one of the most important steps in water resource system management and pollution reduction plans. Recent studies in the field of water quality monitoring networks have shown the need for more studies, despite capabilities and investments in this field. One of the most important problems is the difference between required data and provided data in monitoring networks. So monitoring systems should be revised and modified in several cases. High monitoring expenses necessitate optimizing monitoring systems to limit costs. Being aware of network properties is an essential step in evaluating an existing quality monitoring network. Locations of sampling stations, time frequencies, qualitative variable specifications and sampling duration should be considered in these evaluations. Several studies have been done in the field of water resource monitoring system design as described below. These studies showed that entropy theory can quantify information content and measure the monitoring network’s information.

Harmancioglu et al. (1992) developed a model based on the time and location of sampling stations and also the combination of them based on entropy theory. The results of their research showed that using this theory is very
applicable in qualitative monitoring network design. Ozkul et al. (2000) followed this study and presented a new method to evaluate and design water quality monitoring networks considering time frequency and sampling locations simultaneously.

Information entropy was used by Krstanovic & Singh (1992) to design rainfall networks in Australia. These studies on data collection systems showed the ability of the entropy concept to design optimal monitoring networks. Mogheir & Singh (2002) showed that the distance of groundwater quality monitoring stations is related to transinformation and boundary entropy. Salark & Sorman (2006) optimized and evaluated river stream monitoring networks using continuous entropy theory. Masoumi & Kerachian (2008) proposed an entropy-based approach to assess the location of salinity monitoring stations in the Tehran Aquifer. The authors used transinformation entropy to find the optimal distance among stations and showed the applicability and efficiency of the entropy in assessing the groundwater monitoring systems.

Zhang et al. (2011) and Ridolfi et al. (2011) applied information entropy for rainfall network assessment and obtained satisfactory results in this field. Ridolfi et al. (2012) located monitoring sensors in Dee River basin and then eliminated low-efficiency stations using entropy theory. Lee (2013) used entropy theory in conjunction with a genetic algorithm to determine optimal water quality monitoring points in sewer systems. The genetic algorithm was applied to select the points that maximize the total information among the collected data at multiple locations. Su & Jiing-Yun You (2014) proposed a spatial information estimation model for the analysis of precipitation gauge networks, to improve previous methods based on information theory. They employed a two-dimensional transinformation–distance relationship in conjunction with multivariate information approximation for the illustrated propose. Xu et al. (2015) used an entropy theory based on a multi-criteria method to resample the rain gauge networks.

In this study an entropy-based method of optimally redesigning salinity monitoring networks in bays using discrete entropy theory has been presented. The salinity of bay water measures the relative proportion of fresh water and seawater, which changes drastically both spatially and temporally. Salinity has a profound impact on the physical, chemical, and biological dynamics of estuaries. With regard to the importance of salinity in bays, the methodology of assessing and optimizing a bay water salinity monitoring network which takes into account the value of transinformation was applied in San Francisco bay (S.F. bay), USA. The transinformation index measures the redundant or mutual information between salinity time series to derive an optimum network. The aim of this study is to monitor system-redundant information and eliminate the data, so that using transinformation entropy can lead to the desired results.

**METHOD**

Entropy theory was used to obtain the optimum distance between monitoring stations of water salinity in the S.F. bay. It is a method of quantifying information and even controlling existing data sufficiency. Chaos and turbulence in a system can also be measured using entropy theory. Turbulence existing in a data set means many disordered variations, which can generate repeated data that are costly and redundant. Information fluctuations and their not following from a specific rule will lead to a reduction in our knowledge from the system and make uncertainties. Entropy theory provides a quantitative measure of the uncertainty or the information content of a random variable (Shannon 1948). In this theory, indices such as marginal entropy, conditional entropy and transinformation are defined for the quantification of information. The entropy indices are defined as follows for the discrete random variables \( X = \{x_i\}, i = 1, 2, \ldots, n \) and \( Y = \{y_j\}, j = 1, 2, \ldots, m \) (e.g. Mogheir & Singh 2002; Vivekanandan 2014).

**Marginal entropy**

Marginal entropy for the discrete random variable \( X \) is defined as:

\[
H(X) = - \sum_{i=1}^{n} P(x_i) \ln(P(x_i)) \tag{1}
\]
where, \( P(x_i) \) is the probability of the \( i \)th random variable \( X = \{x_i\} \), and \( n \) is the number of observations. The total of the probability values in the scope of \( X \) should be 1, i.e., \( \sum_{i=1}^{n} P(x_i) \). The marginal entropy \( H(X) \) indicates the amount of information or uncertainty that \( X \) has.

**Joint entropy**

If the variables \( X \) and \( Y \) are independent, then the joint entropy \( H(X, Y) \) is equal to the sum of their marginal entropy values and given by:

\[
H(X, Y) = H(X) + H(Y)
\]  
(2)

If the variables are stochastically dependent, then the joint entropy is less than its total entropy and is defined as follows:

\[
H(X, Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \ln[p(x_i, y_j)]
\]  
(3)

in which, \( p(x_i, y_j) \) is the joint probability between \( x_i \) and \( y_j \).

**Conditional entropy**

The conditional entropy of \( Y \) on \( X \) is defined by:

\[
H(Y|X) = H(X, Y) - H(X)
\]  
(4)

The conditional entropy value becomes zero if the value of one variable is completely determined by the value of the other variable. If the variables are independent, then \( H(Y|X) = H(Y) \).

**Transinformation entropy**

Transinformation is the form of entropy that measures the redundant or mutual information between variables. Transinformation represents the amount of information, which is common to two stochastically dependent variables \( X \) and \( Y \). The transinformation between \( X \) and \( Y \) is defined as:

\[
T(X, Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \ln \left[ \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \right]
\]  
(5)

where \( p(x_i) \) and \( p(y_j) \) are the discrete probability occurrences of \( x_i \) and \( y_j \), respectively, and \( p(x_i, y_j) \) is the joint probability of \( x_i \) and \( y_j \). For independent \( X \) and \( Y \), \( T(x, y) = 0 \).

**Entropy for a salinity monitoring network**

Since the transinformation entropy is a quantitative criterion of common or redundant data between two stations, a criterion of evaluation of the salinity monitoring network can be the value of transinformation.

Considering \( X \) and \( Y \) as a recorded time series of salinity water quality index in the stations of S.F. bay, to calculate the transinformation entropy between two stations \( [T(x, y)] \) in discrete mode, a two-dimensional frequency distribution table should be formed like Table 1.

The two-dimensional frequency distribution table records the frequency for the values that fall into each possible combination of two class intervals, and its formation and the value calculation of probabilities in Equation (5) are based on the following steps:

- Considering \( V \) class intervals for the \( X \) variable and \( U \) for \( Y \), the number of class intervals (NCI) is calculated from Equation (6) (e.g. Mogheir et al. 2003).

\[
NCI = 1 + 1.33\ln(n)
\]  
(6)

in which \( NCI \) is the class interval number and \( n \) is the size of the time series. Notice that the number of class intervals should be the same for all the salinity time series of the water quality monitoring stations.

- The joint frequencies value is shown by \( f_{ij} \) which depends on the \( i \)th row and \( j \)th column, which is equal to

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
<th>( Total )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( f_{11} )</td>
<td>( f_{12} )</td>
</tr>
<tr>
<td>2</td>
<td>( f_{21} )</td>
<td>( f_{22} )</td>
</tr>
<tr>
<td>-</td>
<td>( f_{31} )</td>
<td>-</td>
</tr>
<tr>
<td>V</td>
<td>( f_{11} )</td>
<td>( f_{22} )</td>
</tr>
<tr>
<td>Total</td>
<td>( f_1 )</td>
<td>( f_2 )</td>
</tr>
</tbody>
</table>
observations in which $x_i$ is located in the $i$th category (class interval) and $y_j$ in the $j$th.

- The marginal frequencies values, which are shown by $f_i$ and $f_j$, are equal to the summation of cell densities in each row for the $x$-variable and the summation of cell densities in each column for the $y$-variable.
- The discrete probabilities of $p(x_i), p(y_j)$ and $p(x_i, y_j)$ is calculated by dividing each cell frequency value by the total frequencies.

After the above-mentioned steps, the transinformation index in discrete mode can be calculated using Equation (5). Then, the transinformation–distance (T-D) curve will be developed. The optimum distance between salinity monitoring stations can be elicited from the curve. The obtained distance between monitoring stations will be that by which redundant data between stations are minimized. Figure 1 shows the steps of the recommended methodology in this study.

**STUDY AREAS AND DATA**

**Study area**

S.F. bay is a part of the more complex S.F. bay estuary system, which includes San Pablo Bay and Suisun Bay, the Carquinez Strait, the tidal marshes surrounding these waters, and river tributaries. The S.F. bay estuary, which consists of 480 square miles, 12 islands, and two trillion gallons of salt water, can be thought of as two separate areas: the northern, which passes south and ward from the delta through Suisun and San Pablo Bays, and the southern (also called the South Bay) which extends south-eastward toward San Jose. These two areas join in the Central Bay near the Golden Gate Bridge and flow out to the Pacific Ocean. The entire bay is relatively shallow, with narrow, deep channels near the Golden Gate.

**Data**

From a monitoring perspective, identification of bay salinity is of particular importance. Understanding the nature of salinity patterns in the estuary is also necessary for interpreting the movement of toxic substances and/or essential nutrients. The salinity regime in estuaries is controlled by a number of factors such as river discharge, coastal runoff, local precipitation–evaporation, winds, and water exchange with the ocean. Most of the variations in salinity, considering both space and time, are caused by patterns of freshwater discharge from tributary rivers and the mixing of freshwater with seawater by both tidal action and wind-driven wave action. Bay water salinity is generally defined as the salt concentration in the water. It is measured in units of PSU (Practical Salinity Unit), which is a unit based on the properties of sea water conductivity. It is equivalent to per thousand or to g/kg.

This study used salinity records consisting of data from January 1995 to December 2014 which were obtained from the United States Geological Survey (USGS) website (http://sfbay.wr.usgs.gov/access/wqdata). The data were used from all stations located in the bay (Figure 2). The sampling depth is 2 metres below the water surface and the monitoring frequency is once a month.

Figure 3 shows the spatial and temporal variation of salinity in S.F. bay in the studied period. This figure indicates the increase of the salinity in this bay from January 1995 to
December 2014. It also illustrates the minimal amount of salinity concentration near the tributary rivers because of freshwater mixing with ocean water. As shown in Figure 3 there are restricted spatial variations of salinity all over the bay.

RESULTS AND DISCUSSION

This section presents the results of the methodology explained above, for the optimization of the distances between salinity monitoring stations in San Francisco bay. Transinformation entropy provides a conceptual method to quantify the amount of information contained and shared among two or more variables. By applying the procedure detailed above, a computer program was developed and used to compute the transinformation index. The T-D curve was developed using the transinformation index in discrete mode and time series were obtained over 37 active stations located over the entire body of the bay. The optimum distance between salinity monitoring stations was extrapolated from the curve shown in Figure 4.

As shown in the figure, four lines were fitted to the data in order to obtain the equation of the T-D diagram. These lines were trended to the T-D data in a way that illustrates

Figure 2 | San Francisco bay (California) and locations of water quality sampling stations.

Figure 3 | Spatial and temporal variation of salinity in S.F. bay.
their general manner representatively. The intersection point of these two lines was reported as the optimal distance between quality monitoring stations for the qualitative salinity index. The reported distance is the minimal distance over which changes in data transfer content depending on the distance are very low and negligible, and data transfer reaches its lowest value. In other words, the line 2 slope can be ignored in comparison with the line 1 slope. This distance indeed is the optimum distance between monitoring stations for the investigated qualitative variable index, because over shorter distances excess data would be collected by the monitoring system and over greater distances the network coverage is not complete. The coverage of the suggested monitoring network by using the optimum distance was complete and the system did not lead to redundant data.

The relation of the fitted curve and derived optimal distance is shown in Table 2. This table indicates the optimum distance of 7.62 km as the salinity monitoring sampling station distance. Since the obtained optimum distance using the presented methodology in this research is the best approximation of it, the extrapolated distance was rounded to 7.5 km in order to report it in a more applicable form. Given the distance, the number of required sampling stations to monitor salinity in San Francisco bay will be 21. Considering the 37 existing active sampling stations in the bay, there are 16 excess stations. Of course, the results of this research are only applicable for the qualitative salinity index in San Francisco bay. Therefore the results may be different for other qualitative indices in other case studies.

Table 2 | Equation of the T-D diagram and derived distance

<table>
<thead>
<tr>
<th>Water quality index</th>
<th>Line 1</th>
<th>Line 2</th>
<th>D*(Km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salinity</td>
<td>$T = -0.111D + 1.398$</td>
<td>$T = -0.003D + 0.675$</td>
<td>7.62</td>
</tr>
<tr>
<td></td>
<td><strong>Optimum distance</strong></td>
<td></td>
<td>7.500</td>
</tr>
</tbody>
</table>

Figure 4 | Salinity transinformation versus distance in San Francisco bay.

The presented methodology in the paper was used to optimize salinity sampling station distance with the target of bay water quality control in S.F. bay. The visual study of temporal and spatial variations in salinity using a 3-D diagram showed that salinity increased from January 1995 to December 2014. Also it is observed that there are restricted spatial variations of salinity all over the bay and that the 37 existing stations that monitor the salinity in the bay give redundant data that are unnecessary and costly. This observation confirms the results of entropy theory in this study.

CONCLUSION

The presented methodology in the paper was used to optimize salinity sampling station distance with the target of bay water quality control in S.F. bay. The visual study of temporal and spatial variations in salinity using a 3-D diagram showed that salinity increased from January 1995 to December 2014. Also it is observed that there are restricted spatial variations of salinity all over the bay and that the 37 existing stations that monitor the salinity in the bay give redundant data that are unnecessary and costly. This observation confirms the results of entropy theory in this study.
Based on the authors’ findings, this paper is the first application of entropy-based methods for bay water monitoring network design.

The results showed that entropy theory performs promisingly in designing and updating salinity monitoring networks in bays because this theory can report the content of common data between water quality monitoring stations as quantitative values. The suggested methodology for salinity monitoring in bays not only reduces the costs of monitoring, but also provides the maximum available information to water resources managers. The results indicated that the proposed methodology can be effectively used for salinity monitoring networks in bays.

The results showed that the existing monitoring network in S.F. bay leads to excess data because of the nearness of existing stations and that approximately 21 stations can cover the salinity monitoring network completely. Although the obtained optimum distance in the study is only utilizable for salinity monitoring in S.F. bay, this methodology can be used for other qualitative indices, too. Obviously, the obtained distances for different indices may be different. Researchers can apply methods which are helpful tools to make complex decisions based on mathematics and qualitative criteria such as multi-criteria decision-making, analytic hierarchy process etc. to report the final optimum distance for the monitoring network, when they study water quality effective indices over all the bay, simultaneously. It is suggested that future studies use entropy theory in the discrete mode in order to design the sampling frequency. The proposed method in determining the sampling intervals in the depth of bays may also be used. Two lines have been fitted to the T-D data in the study such that they represent the data behavior best. Genetic algorithms, particle swarm optimization approaches etc. may be used in future research to maximize the correlation coefficients’ fitted lines.

REFERENCES


