Characterizing the spatial variability of groundwater quality using the entropy theory: II. Case study from Gaza Strip

Y. Mogheir,1,2 J. L. M. P. de Lima3* and V. P. Singh4

1 Department of Civil Engineering, Faculty of Science and Technology, Campus 2, University of Coimbra, 3030-290 Coimbra, Portugal
2 Palestinian Water Authority, Building No. 130/61, League of Arab States St. Tel El Hawa, Gaza, Palestine
3 Department of Civil Engineering, Institute of Marine Research, Coimbra Interdisciplinary Centre, Faculty of Science and Technology, Campus 2, University of Coimbra, 3030-290 Coimbra, Portugal
4 Department of Civil and Environmental Engineering, Louisiana State University, Baton Rouge, LA 70803-6405, USA

Abstract:
This paper, the second in the series, uses the entropy theory to describe the spatial variability of groundwater quality data sets. The application of the entropy theory is illustrated using the chloride observations obtained from a network of groundwater quality monitoring wells in the Gaza Strip, Palestine. The application involves calculating information measures, such as transinformation, the information transfer index and the correlation coefficient. These measures are calculated using a discrete approach, in which contingency tables are used. An exponential decay fitting approach was applied to the discrete models. The analysis shows that transinformation, as a function of distance, can be represented by the exponential decay curve. It also indicates that, for the data used in this study, the transinformation model is superior to the correlation model for characterizing the spatial variability. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS correlation; entropy; information; spatial variability; Gaza Strip; Palestine

INTRODUCTION
Entropy theory (information theory) came to be viewed as a statistical concept at the beginning of the twentieth century. About 50 years later, it found its way into engineering and mathematics, notably through the work of Shannon in communication engineering. Shannon (1948) used entropy as a measure of uncertainty in the mind of someone receiving a message that contains noise. Later, in 1957, Jaynes made use of Shannon’s entropy metric to formulate the maximum entropy principle that formed a basis for estimation and inference problems (Golan et al., 1997). In 1972 Amorocho and Esplidora were the first to apply the entropy concept to hydrological modelling (Singh, 1997). Since then, there has been a great variety of entropy applications in hydrology and water resources management (e.g. Rajagopal et al., 1987; Singh and Rajagopal, 1987; Singh, 1998; Harmancioglu et al., 1999). Entropy theory can be used in modelling and decision-making in environmental and water resources, especially in developing countries (Singh, 2000).

Entropy theory also has been applied to assess and evaluate monitoring networks with respect to: water quality (Harmancioglu et al., 1994, Ozkul et al., 2000), rainfall (Krastanovic and Singh, 1992) and groundwater (Bueso et al., 1999; Mogheir and Singh, 2002). Most of these applications involve applying entropy theory to the evaluation, assessment and design of monitoring networks, and they used an analytical approach with a presumed knowledge of the probability distributions of the random variables involved. In the first paper of this series, Mogheir et al. (2004) adopted discrete and analytical approaches using a synthetic...
data set, where the data were spatially correlated and fitted the normal distribution function. Under these conditions, it was found that there was a reasonable agreement between discrete and analytical approaches for developing the transinformation model (T model), and it was shown that the T model also could be used instead of the correlation model (C model) to characterize the spatial variability.

In this paper, a different set of data is used. The set of data includes groundwater quality from the Gaza Strip monitoring network (chloride data). For these data, the spatial correlation is low and the normal distribution function does not fit. The objective of this paper is to:

1. use a discrete approach (contingency table) for calculating information measures, such as transinformation (T), information transfer index (ITI) and correlation coefficients.
2. apply an exponential decay fitting approach to the discrete T model and C model;
3. use the T model and C model to describe the spatial variability of the Gaza Strip data set.

GAZA STRIP GROUNDWATER QUALITY DATA

The set of data used in the analysis is part of groundwater quality data from the Gaza Strip, Palestine. The data were selected from the groundwater quality data monitored in the middle part of the Gaza Strip. This part of the Gaza Strip is the area with the most serious problems of seawater intrusion. More than 150 wells are used to monitor the groundwater quality in this area. In this study, 26 monitoring wells that monitor chloride were selected. Each well has 52 chloride data measured between 1972 and 1997. Chloride is measured twice per year: in winter and summer. The winter cycle is considered to be taken in April and May whereas the summer cycle is in October and November. The locations of these 26 wells in the middle part of the Gaza Strip are shown in Figure 1. The chloride time-series of the 26 wells are presented in Table I. In the table, \( \bar{x} \) is the mean and \( S_x \) is the standard deviation of the chloride data. The spatial variation of the mean of the chloride time-series in each well is presented in Figure 2. The contour lines were drawn using the kriging technique, which is an option in the Surfer-7 mapping program (Golden Software, 1999). Additionally, the chloride time-series of some of these wells are plotted in Figure 3. The groundwater data in the Gaza Strip (quality and water level) were summarized and presented by the Palestinian Water Authority (PWA, 2000). These data were also used in the modelling of the Gaza Strip aquifer by Metcalf and Eddy (2000).

METHOD

The method used in this study follows that presented in Mogheir et al. (2004). A contingency table is used for the discrete approach. The discrete models’ results were smoothed using the moving average method. For convenience, the base \( e \) and the unit \( \text{nats} \) were used for computing numerical results.

This study differs from Mogheir et al. (2004) mainly in the analytical approach. As the Gaza Strip data, which were used in this study, do not follow the Gaussian distribution function, and their spatial correlations are low, an exponential decay curve is fitted to the discrete models and to the smoothed discrete models (exponential decay fitting approach).

Harmancioglu et al. (1999) investigated the fitting of a semi-exponential curve to the discrete T model. The analysis of the synthetics data (Mogheir et al., 2004) and the shape of the discrete T model, smoothed by the moving average method, of the chloride data set signified that the exponential decay curve could be the best representation of the discrete T model, and could be presented as (e.g. Motulsky, 1999)

\[
T(d) = G e^{(-Kd)} + Q
\]

where the exponential decay curve starts with \( T_0 = G + Q \) at distance \( (d) = 0 \); and the curve decays to reach \( Q \) value with a constant rate \( K \). The units of \( G \) and \( Q \) are expressed in the same way as the \( T \) unit (nats),

Copyright © 2004 John Wiley & Sons, Ltd. 

Figure 1. Location of selected groundwater quality monitoring wells in the middle part of the Gaza Strip, Palestine.
Table I. Chloride (mg/L) time-series of the 26 monitoring wells as presented in Figures 1 and 2

<table>
<thead>
<tr>
<th>Date</th>
<th>Monitoring well</th>
</tr>
</thead>
<tbody>
<tr>
<td>01-05-72</td>
<td>294 378 805 581 595 1064 798 844 343 742 700 525 574 725 427 441 399 816 707 875 427 637</td>
</tr>
<tr>
<td>28-10-72</td>
<td>322 385 840 539 616 1008 406 560 532 515 790 679 329 742 658 469 644 686 574 504 483 693 700 882 413 609</td>
</tr>
<tr>
<td>26-04-73</td>
<td>462 378 826 623 700 1001 420 476 539 462 798 721 322 756 672 518 707 756 427 623 511 714 728 910 483 588</td>
</tr>
<tr>
<td>23-10-73</td>
<td>371 385 840 616 700 1022 399 483 532 525 784 805 315 749 714 476 686 847 574 665 546 742 812 924 497 609</td>
</tr>
<tr>
<td>21-04-74</td>
<td>322 378 910 623 700 1015 420 483 546 525 798 693 336 756 672 462 707 756 427 609 532 726 756 889 483 588</td>
</tr>
<tr>
<td>10-04-77</td>
<td>462 406 840 623 770 1015 406 560 518 525 791 750 315 749 714 448 686 847 518 595 518 742 812 882 497 609</td>
</tr>
<tr>
<td>13-10-75</td>
<td>371 399 945 644 770 1008 413 560 532 532 812 679 427 721 714 511 714 854 574 616 546 735 928 924 518 595</td>
</tr>
<tr>
<td>01-04-76</td>
<td>385 420 882 644 763 1043 434 476 553 559 756 721 399 756 742 518 735 847 560 623 532 805 728 861 511 588</td>
</tr>
<tr>
<td>07-10-76</td>
<td>399 432 882 644 721 1016 455 476 560 567 861 805 420 770 735 476 725 931 574 665 538 791 812 945 553 602</td>
</tr>
<tr>
<td>05-04-77</td>
<td>350 434 924 686 791 1029 406 504 553 539 539 791 717 420 721 866 462 678 854 504 609 567 749 700 952 584 574</td>
</tr>
<tr>
<td>02-10-77</td>
<td>399 392 791 616 707 903 448 441 511 497 798 756 357 707 679 455 602 896 564 616 574 756 686 840 497 574</td>
</tr>
<tr>
<td>31-03-78</td>
<td>371 406 805 525 791 917 399 417 504 525 812 735 371 714 742 476 735 924 581 644 588 742 700 842 504 644</td>
</tr>
<tr>
<td>22-09-79</td>
<td>399 385 840 609 749 949 444 455 518 546 812 728 378 735 665 434 602 1015 588 693 665 756 924 875 525 610</td>
</tr>
<tr>
<td>20-03-80</td>
<td>364 406 868 816 693 924 441 417 532 555 749 721 392 763 700 441 623 1008 581 707 679 931 735 861 532 679</td>
</tr>
<tr>
<td>16-09-80</td>
<td>406 406 868 651 742 949 441 483 602 560 798 749 385 756 700 399 630 1043 756 742 665 854 777 968 546 497</td>
</tr>
<tr>
<td>15-03-81</td>
<td>364 430 882 688 693 924 469 490 502 553 749 756 392 763 665 420 630 1008 581 707 700 966 735 861 518 490</td>
</tr>
<tr>
<td>11-09-81</td>
<td>399 420 890 651 756 1036 516 518 602 560 861 861 434 840 784 420 658 1002 574 742 693 854 777 854 525 497</td>
</tr>
<tr>
<td>10-03-82</td>
<td>462 504 959 801 777 833 924 518 490 623 637 861 840 448 861 791 420 693 1169 581 707 705 1071 910 952 602 490</td>
</tr>
<tr>
<td>06-09-82</td>
<td>476 450 1001 756 861 1071 532 483 644 651 861 861 441 840 784 399 665 1162 585 882 784 1015 1093 939 630 553</td>
</tr>
<tr>
<td>04-06-83</td>
<td>455 504 1001 756 840 924 518 490 651 658 819 840 448 1008 840 441 707 1190 581 924 847 1134 840 945 630 546</td>
</tr>
<tr>
<td>01-09-83</td>
<td>497 511 1029 812 861 1057 525 483 658 658 791 882 455 889 784 420 707 1155 756 902 819 1085 910 945 644 595</td>
</tr>
<tr>
<td>28-02-84</td>
<td>490 504 1064 791 875 1057 546 602 672 651 889 854 497 861 826 434 735 1071 791 896 840 1169 910 980 651 623</td>
</tr>
</tbody>
</table>
whereas $K$ is expressed in the inverse unit used by $d$ (1/m). Note that Equation (1) was also used to represent the analytical lognormal $T$, ITI and correlation models.

The fitting of the exponential decay curve to the discrete models was performed using the least-square fitting procedure with the GRAPHPAD PRISM statistical software (Motulsky, 1999). The coefficient of determination was used to quantify the goodness of fit between the exponential decay curve and discrete models. The coefficient of determination ($R^2$) was computed as (e.g. Motulsky, 1999)

$$R^2 = 1 - \frac{SS_{reg}}{SS_{tot}}$$

where $SS_{reg}$ is the sum of the squares of the residuals between the discrete model and the best-fit exponential decay curve, and $SS_{tot}$ is the sum of the squares of the residuals between the discrete model and the horizontal line through the mean.

As in Mogheir et al. (2004), the $T$ model and $C$ model were compared to characterize the spatial variability of the Gaza Strip data set.

**COMPARISON OF DISCRETE AND EXPONENTIAL DECAY FITTING APPROACHES**

**Correlation model ($C$ model)**

The discrete $C$ model is obtained by computing the correlation values using the discrete approach and the distance between wells. The discrete $C$-Model data is smoothed by using the moving average method.
Figure 3. Chloride time-series of 12 monitoring wells (H-10, H-11, H-20, H-28, H-35, H-45, H-48, H-50, H-41, H-61, H-15 and H-4) for the period winter 1972 to summer 1997, used in the analyses. In the graph $\tau$ is the mean of the chloride time-series and $S_\tau$ is the standard deviation.

(DCM$_{MA}$). The exponential decay fitting approach is applied to the discrete C model and DCM$_{MA}$. A summary of the best-fit equations of the exponential decay curve to the discrete T, lognormal T, ITI, correlation models and $R^2$ values for each model is presented in Table II.

The discrete C model (DCM), the C model smoothed by the moving average method (DCM$_{MA}$) and the exponential decay of the discrete C model (DCM$_{ED}$) are plotted in Figure 4. This figure and Table II show that DCM$_{ED}$ does not fit the discrete C model well, as $R^2 = 0.07$, which is very low. The coefficient $R^2$ is increased by applying the exponential decay fitting approach to the DCM$_{MA}$($R^2 = 0.22$). Nevertheless, for both the DCM$_{MA}$ and discrete C models the coefficient $R^2$ is quite small. Therefore, the exponential decay curve, which was selected to present the discrete C model, does not infer the spatial variability of the chloride data adequately.

<table>
<thead>
<tr>
<th>Model type</th>
<th>Fitting equation</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete C model</td>
<td>$r(d) = 0.43 e^{-0.0033 d} + 0.53$</td>
<td>0.07</td>
</tr>
<tr>
<td>Discrete T model</td>
<td>$T(d) = 0.29 e^{-0.0087 d} + 0.90$</td>
<td>0.33</td>
</tr>
<tr>
<td>Lognormal discrete T model</td>
<td>$T(d) = 0.90 e^{-0.0102 d} + 0.59$</td>
<td>0.43</td>
</tr>
<tr>
<td>Discrete ITI model</td>
<td>ITI($d$) = $0.39 e^{-0.0359 d} + 0.61$</td>
<td>0.57</td>
</tr>
</tbody>
</table>
Transinformation model (T model)

The discrete T model was obtained in the same way, by computing the T values using the discrete approach and the distance between wells. The discrete T model data were smoothed using the moving average method (DTMMA). The exponential decay fitting approach is applied to the discrete T model and DTMMA. For the discrete T model, the $R^2$ coefficient is 0.33, which is smaller than that for DTMMA ($R^2 = 0.71$). This indicates that the exponential decay curve fits the DTMMA much better than does the discrete T model. The discrete T model, the DTMMA and exponential decay of the discrete T model (DTMED) are plotted in Figure 5.

T-model using logarithmic chloride data

As the normal distribution did not fit the chloride data well, the lognormal distribution was assumed. The chloride logarithmic data from the Gaza Strip monitoring wells are used to compare the discrete and exponential decay fitting approaches in obtaining the T values. The logarithmically transformed chloride data are used to check the fitting of the normal function by constructing the histogram and plotting the probability diagram. The chi-square test was used to assess the adjustment of the lognormal distribution to the empirical data.

After fitting the lognormal function of the chloride data from the Gaza monitoring wells, the lognormal discrete T model (lognormal DTM) is obtained by computing the T values of the logarithm of the chloride data, using the discrete approach and the distance between wells. The lognormal discrete T model is smoothed.
using the moving average method (lognormal DTMA). The exponential decay fitting approach is applied to the lognormal discrete T model and to the lognormal DTMA. Figure 6 illustrates the lognormal discrete T model, the lognormal DTMA and the exponential decay of the lognormal discrete T model (lognormal DTMED). From Figure 6, it can be seen that, for the lognormal DTMA, \( R^2 \) is 0.68 which is greater than the \( R^2 \) value obtained by the lognormal discrete T model (\( R^2 = 0.43 \)). As for all the T models, the exponential decay curve fits better to the DTMA than that to the discrete T model. The lognormal DTMED is compared with the DTMED. As shown in Figure 6, the minimum value of the transinformation in the lognormal DTMED is 0.3 nats less than that found in the DTMED. Additionally, the initial value of the transinformation in the lognormal DTMED is 0.3 nats greater than that in the DTMED. This indicates that the T model is sensitive to the type of distribution of the data, whether its normal or lognormal.

CHARACTERIZATION OF SPATIAL VARIABILITY

When comparing the correlation model (C model) and the transinformation model (T model), to characterize the spatial variability of the chloride data, Figure 4 shows that the discrete C model is highly scattered and the exponential decay curve does not fit to the discrete C model well. This is also found where \( R^2 = 0.07 \) and 0.22 for the discrete C model and DCMMA, respectively. On the other hand, Figure 5 shows that the exponential decay curve fits to the DTMA better than it does to the discrete T model, as \( R^2 = 0.33 \) and 0.71 for the discrete T model and DTMMA, respectively. Furthermore, the \( R^2 \) values are greater if the logarithmically transformed chloride data are used.

As the ITI and correlation models have the same range from 0 to 1, they are compared in Figure 7, which demonstrates that there is less scatter in the discrete ITI model, which is smoothed by the moving average method (DITIMA), than there is in the DCMMA. The \( R^2 \) value for DITIMA is 0.79, which is greater than that for the DCMMA (\( R^2 = 0.22 \)). These values suggest that the exponential decay curve is representing the ITI model much better than it represents the C model. As a result, it can be inferred from Figures 5–7 that the T model and ITI model represent the dependency between wells better than the discrete C model.

In the above analysis, the dependency is described by an exponential decay model, which is relevant to the T model because the T value is maximized at a distance equal to zero. The maximum T value equals the average of the marginal entropies of the 26 wells. There is a sharp drop in the T value when the distance is around 500 m. With a further increase in the distance, T becomes essentially constant. Therefore, what is significant for the spatial assessment and redesign of monitoring wells is selecting the distance at which T has a minimum steady value. The prescribed 500 m value may be adopted as the recommended distance between wells. This distance can be utilized in the assessment stage under the following conditions.
1. If the distances between wells are less than the recommended distance, then there is available transinformation (redundant information) between wells.
2. If the distances between the existing wells are greater than the recommended distance, then the transinformation between wells is less than the minimum transinformation value (not enough information).
3. The adequate information that can be available between wells is found only where the distances between wells equal the recommended distance and the transinformation is minimum.

These arguments afford efficient criteria to assess and redesign the existing wells according to that recommended distance and minimum redundant information between wells. Consequently, the number of wells can be extended or reduced.

It is also useful for redesigning groundwater quality monitoring networks, and developing an analytical equation to relate T and distance. This equation can form an exponential decay curve, as in the synthetic data (Mogheir et al., 2004) and the chloride data example, or any other type of curve. The monitoring network redesign procedure also might need to look at the variations of the value of T and the shape of the T model by changing the number of wells and the size of the time-series used for constructing the T model.

CONCLUSIONS

This article has presented a comparison between the discrete and exponential decay fitting approaches, using a groundwater quality data set from the Gaza Strip (chloride data). The following conclusions can be drawn.

1. The exponential decay fitting approach shows that the exponential decay curve does not fit to the discrete correlation model well.
2. The exponential decay curve fits to the discrete T model, the lognormal discrete T model and the discrete ITI model much better than does to the discrete correlation model.
3. The characteristics of the exponential decay of the lognormal discrete T model, such as the minimum T and initial T, differ from those of the exponential decay of the discrete T model.
4. The discrete T and ITI models are superior to the discrete correlation model for characterizing the spatial variability by means of an exponential decay model.

The exponential decay T model can be used to evaluate a groundwater monitoring network. Furthermore, the T model can be used to redesign the monitoring network by either increasing or decreasing the number of wells. The assessment and redesigning of a groundwater quality monitoring network, using the sensitivity
of the T model to the number of monitoring wells and the size of time-series, are part of an ongoing study by the first author.

ACKNOWLEDGEMENTS

The authors gratefully thank the Foundation for Science and Technology of the Portuguese Ministry of Science and Technology for sponsoring the fellowship (Reference: SFRH/BD/6089/2001). The fellowship was provided for the first author’s PhD programme entitled ‘Quantification of Information for Groundwater Quality Networks’. The programme is being undertaken at the Department of Civil Engineering, Faculty of Science and Technology, University of Coimbra, Coimbra, Portugal.

APPENDIX

List of symbols and abbreviations

Symbols.

- \( G \) the value of transinformation where distance equals 0 deducted from \( Q \) (NATS)
- \( K \) the transinformation decay rate (1/m)
- \( \text{ITI}(d) \) information transfer index as a function of distance (NATS)
- \( Q \) the end value of transinformation at which the distance is maximum (NATS)
- \( R^2 \) coefficient of determination
- \( r(d) \) correlation as a function of distance.
- \( S_{\text{reg}} \) sum of the squares of the residuals between the discrete model and the best fit curve (analytical model)
- \( S_{\text{tot}} \) sum of squares of the residuals between the discrete model and the horizontal line through the mean
- \( S_x \) sample standard deviation of variable \( x \)
- \( T(d) \) transinformation as a function of distance (NATS)
- \( \bar{x} \) sample mean of variable \( x \)

Abbreviations.

- C model correlation model
- DCM discrete correlation model
- DCM\(_{\text{ED}}\) exponential decay of the discrete C model
- DCM\(_{\text{MA}}\) smoothed discrete correlation model by the moving average method
- DITIM discrete ITI model
- DITIM\(_{\text{ED}}\) exponential decay of the discrete ITI model
- DITIM\(_{\text{MA}}\) smoothed discrete ITI model by moving average method.
- DTM discrete transinformation model
- DTM\(_{\text{ED}}\) exponential decay of transinformation model
- DTM\(_{\text{MA}}\) smoothed discrete transinformation model by the moving average method
- ITI model information transfer index model
- Lognormal DTM\(_{\text{ED}}\) exponential decay of lognormal discrete transinformation model
- Lognormal DTM lognormal discrete transinformation model
- Lognormal DTM\(_{\text{MA}}\) smoothed lognormal discrete transinformation model by the moving average method
- T model transinformation model

REFERENCES


