8-1-2014

Evaluating North Sea Water Level Monitoring Network Considering Uncertain Information Theory Quantities

Leonardo Alfonso

Elena Ridolfi

Sandra Gaytan

Francesco Napolitano

Fabio Russo

EVALUATING NORTH SEA WATER LEVEL MONITORING NETWORK CONSIDERING UNCERTAIN INFORMATION THEORY QUANTITIES

LEONARDO ALFONSO (1), ELENA RIDOLFI (2,3), SANDRA GAITAN-AGUILAR (4), FRANCESCO NAPOLITANO (2), FRANCESCO RUSSO (2)
(1): Hydroinformatics Chair Group, UNESCO-IHE, Westvest 7, Delft 2611AX, The Netherlands
(2): Dipartimento di Ingegneria Civile, Edile e Ambientale, Sapienza Università di Roma, Rome 00184, Italy
(3): H2CU-Honors Center of Italian Universities, Sapienza Università di Roma, Rome 00184, Italy
(4): Deltares, Rotterdamseweg185, Delft 2629 HD, The Netherlands

Although Information Theory concepts have been successfully applied in hydrology and other fields to quantify the amount of information contained in singular variables and shared by multiple variables, the valuation of these quantities is sensible to different parameters. In particular, the bin size of histograms and the Pearson correlation coefficient to estimate probabilistic measures such as Joint Entropy and Total Correlation are of interest when evaluating a monitoring network. This work extends the ensemble entropy method developed by the authors to take into consideration the uncertainty coming from these parameters in the assessment of the North Sea's water level network for large number of sensors. The main idea is to represent entropy of random variables through their probability distribution, instead of considering entropy as a deterministic value. The method considers solving multiple scenarios of Multi-Objective Optimization in which information content (Joint Entropy) of a set of stations is maximized and redundancy (Total Correlation) is minimized. These scenarios are generated with parameter sampling methods such as the Latin Hypercube. Results include probabilistic Pareto fronts generated by parameter sampling, which provided additional criteria on the selection of the final set of monitoring points and the elimination of redundant/non-informative points.

INTRODUCTION

In hydrology and water resources data collection is an important task. Among other problems, both the lack and the imprecision of data series may cause inaccuracy in hydraulic modelling. Therefore, methods to design networks of sensors are of great relevance to collect reliable data series. Many authors have dealt with the issue of determining the best sensor layout on the base of statistical analysis as regression techniques [1] and cross correlation reduction [2].

Information Theory has been used for monitoring network design and assessment of various water systems [3-6]. Recently, the issue of determining the best sensors location has been faced using the two Information theory quantities of joint information (JH) and total correlation (C). The optimal sensors location is defined in such a way that their measurements cover as much as possible the area and that the redundancy of the data is minimum. The two requirements are satisfied through the contemporary maximization of JH (i.e. a measure of information content) and minimization of C (i.e. a measure of information redundancy). The methodology has been applied to analyse pair-wise station, to evaluate the distribution of water level monitors in polders and in river systems [7,8]. Since entropy quantities involve the determination of the marginal and joint probabilities of the considered random vectors (RVs), their determination is of crucial importance. In a previous work [9] investigate the sensitivity of Information Theory quantities to the assumptions made to calculate probabilities as the bin size of histograms. The authors introduced the concept of ensemble entropy that is given by the envelope of possible entropy values. The aim of this work is to evaluate the uncertainty of the entropy–based approach and most of all its effect on the sensors network design..

ENTROPY-BASED METHODOLOGY

In recent years the issue of monitoring sensors location has been investigated through the entropy concept [e.g. 7,8]. The main idea is to determine which sensors layout and number is the most representative of the network. The optimized sensors have to provide the highest amount of information about the variable of interest and at the same time, the information should be essential, without repetition. In terms of Information Theory quantities it means that sensors location should fulfill the following objectives:

$$\begin{aligned} &Max(JH) = Max\left[H(X_1, X_2, ...X_M)\right] \\ &Min(C) = Min\left[C(X_1, X_2, ...X_M)\right] \end{aligned} \tag{1}$$

JH is the joint entropy of the M discrete RVs:

$$JH = H(X_1, X_2, ...X_M) = -\sum_{i_1=1}^{n_1} ... \sum_{i_M=1}^{n_M} p_{i_1,...,i_M} \log_2(p_{i_1,...,i_M}) \tag{2}$$

where $pi_1...i_M$ is the joint probability of the M variables. C is the total correlation:

$$C(X_1, X_2, ...X_M) = \sum_{i=1}^{M} H(X_i) - H(X_1, X_2, ..X_M), \tag{3}$$

H(X*i*) is the marginal entropy of the discrete RV X:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2\left[p(x_i)\right] \quad (4)$$

where p*x* is the probability that X_i equals the outcome x, $P(X_i = x)$.
The problem of determining the best monitors location is, then, posed as a multi-objective (MO) problem, where the two objectives are presented in Eq.(1) and it is solved through the Non-dominated Sorting Genetic Algorithm (NSGA-II; [10]). The algorithm searches for the set of solutions that simultaneously optimize the two independent objective functions. The solution of the MO problem is the set Yi (i = 1,2,...,M) chosen among all sets Xi (i = 1,2,...,M) that is then, plotted on a Pareto front of quasi-optimal solutions.

THE ENSEMBLE ENTROPY CONCEPT

The concept of ensemble entropy was presented in a previous work [9] in order to investigate the sensitivity of Information Theory quantities to the assumptions made to calculate probabilities as the bin size of histograms. Several authors have dealt also with this issue, studying the differences raised when using different bin size and defining the bin size as a function of the RV length.
Specifically, for discrete type RVs we investigate the quantization method suggested in [7], which converts an analog (i.e., continuous) sign into a discrete pulse using mathematical floor function. For continuous type RVs we investigate the effect of choice of the copula function to describe the correlation structures among variables. Both procedures lead to estimate JH and C not as punctual values, but as an envelope of values: the so-called ensemble quantities.

Discrete type RVs

The conversion of an analog value x to a quantized value x_q, rounded to the nearest multiple of a, is given by:

$$x_q = \gamma a\left[\frac{2x + a}{2a}\right] \quad (5)$$

where a is the quotient of the difference between the maximum and the minimum of the time series and the bin-size used in frequency analysis, γ is a numeric parameter to normalise data series deriving from multiple sources. Therefore, RVs are transformed through Eq.(5) so that

series from different sensors are normalised for fair comparison. Probabilities are computed evaluating the frequency of occurrence of each normalised value.
This work investigates how the assumptions of parameters γ and a affect results of sensors network design. Details on the procedure can be found in [9], and it is summarized as follows: 1) assume a value for parameters γ and a; 2) normalise data series of each sensor following Eq. (5); 3) solve the MO problem as in Eq. (1), obtaining the Pareto front of quasi-optimal vector Yi of sensor locations; 4) assume S different sample combinations for parameters γ and a; 5) normalise quasi-optimal Yi data series using the S sample combinations and obtain the transformed time series Dij*; 6) evaluate JH and C of the time series Dij* through Eq.(2) and Eq.(3), respectively; 7) obtain the Pareto of two-dimensional distributions of JH and C; 8) find the optimal combination of sensors location from the analysis of the last Pareto front.

Continuous type RVs

This section presents the analysis made through copula functions to determine JH and C envelopes. The procedure follows the one presented previously, but the JH and C values are computed using copula functions of the Archimedean family: Gumbel, Clayton and Frank [10]. The copula function [11] is a bivariate function $C:[0,1]x[0,1] \rightarrow [0,1]$ with properties, in which $C(u,v)$ is a non-decreasing function for each u, v. Also, for every $u, v \in [0,1]$, $C(u,0)=C(0,v)=0, C(u,1)=u, C(1,v)=v$ and for every $u_1, v_1, u_2, v_2 \in [0,1]$ so that $u_1 \leq u_2$ and $v_1 \leq v_2$. In addition, $C(u_2,v_2)-C(u_1,v_2)-C(u_2,v_1)+C(u_1,v_1) \geq 0$. If u and v are independent, the copula is $C(u,v)=u \times v$, where $u \times v$ is the copula product $\Pi(u,v)$. The joint cumulative distribution function of two RVs X and Y (i.e, $F(X,Y)$), with marginal distribution $F(X)$ and $G(Y)$, is defined as:

$$F(X,Y)=C[F(X),G(Y)] \quad (6)$$

If the marginal functions are continuous, C is unique; otherwise C is determined by Ran F_i x Ran G_i, where *Ran* is the range of the marginals. By definition, the copula is a distribution with marginals U(0,1), so that it is necessary to transform a RV X_i i=1, 2 with a given continuous distribution $F(X_i)$ in a RV with distribution U(0,1). This transformation is made through the Probability Inverse Transformation (PIT), so that $U_i=F(X_i)$. Thus Eq. (6) becomes $F(X,Y)=P[X \leq x, Y \leq y]=C(u,v)=C(F(x),G(y)$. Eq. (6) is the key to the Sklar theorem. Under the assumption that the two marginal distributions are univariate, the joint probability density function is given by:

$$f(x,y)=c(F(x),G(y))\, f(x)\, g(y) \quad (7)$$

where c is the density function of C, and it is:

$$c(u,v) = \frac{d^2 C(u,v)}{dudv} \tag{8}$$

where u and v are two dependent cumulative distribution functions, $F(X)$ and $G(Y)$. A generic Archimedean 2-copula is written as:

$$C(u,v) = \varphi^{-1}\left(\varphi(u) + \varphi(v)\right) \tag{9}$$

where $\boldsymbol{\varphi}$ is the function that generates the copula. It is a strictly decreasing function from [0,1] to [0, ∞): $\boldsymbol{\varphi}(0)=\infty$ and $\boldsymbol{\varphi}(1)=0$, and its inverse is completely monotone on [0, ∞).

CASE STUDY

Figure 1. North Sea and the Netherlands delta and location of one of the most robust sensors set found (blue dots).

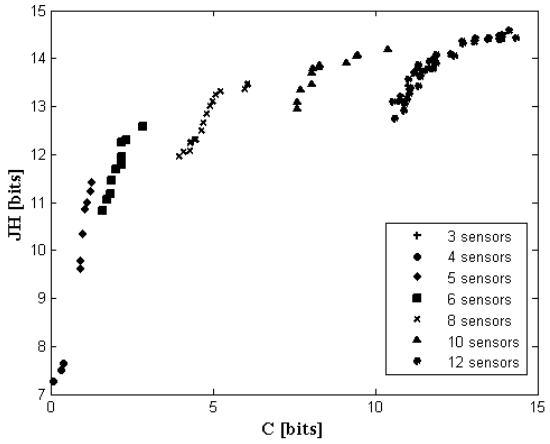

Figure 2. Pareto front of quasi-optimal solutions considering several number of monitoring sensors.

The case study is the North Sea area, where a network of 47 water level sensors is deployed as represented in Figure 1. Specifically, the network monitors territories below the sea level and the delta of The Netherlands, because it is the location of one of the most important ship transit port in the world, the harbor of Rotterdam. However, as the maintenance of the sensors network is expensive, it is necessary to optimize their number. To this end, the data series used in the MO problem are the water stage values gauged by sensors, collected with 10 minutes resolution time and covering the period from 2007 to 2008.

RESULTS AND DISCUSSION

The aim of this study is to investigate the effects of the assumptions made to determine the probabilities of variables on the Information Theory quantities. Following the procedure previously presented, values of γ and a are hypothesized and the corresponding Pareto front of quasi-optimal solutions is plotted, Figure 2.

The optimal number of sensors is 8, because corresponding set have a high value of JH and simultaneously a restrained C value. Indeed, considering more sensors, JH slightly increases, while C consistently raised. This means that it not worthy adding a relevant number of sensors because the gain in terms of information (i.e. JH) is not as high as the loss due to the redundancy (i.e. C). Quasi-optimal sets of 8 and 10 sensors are then analysed to evaluate their ensemble entropy. As previously explained quasi-optimal sets are normalised using several values of γ and a parameters and corresponding JH and C values are computed. The resulting Pareto fronts for 8 and 10 sensors are presented in Figure 3 (a) and (b) respectively.

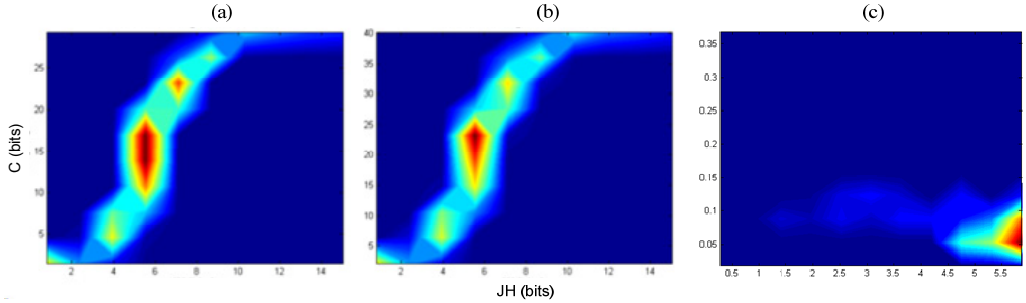

Figure 3. Pareto envelope plots for 8 sensors (a) and 10 sensors (b), built considering different bin size values. The envelop for copulas for two variables is presented in (c)

The dark red area, Figure 3, corresponds to the most robust sensor networks (i.e., a variation in entropy parameters would not greatly change their location). Indeed, in this area, although γ and a change, JH and C values are the same. One of these robust configurations of sensors network is presented in Figure 1. As we can see the majority of the sensors are located in the lower part of the delta. This result is in agreement with the one presented by [9], where an optimal network of three sensors was deployed in the same area case of study. This means that the most sensitive area to water level changes is the one located in the lower area, where the presence of the Rotterdam port makes delicate the water level monitoring issue. Envelope values are also presented for the copula functions case, Figure 3(c). As previously mentioned, these results can be extended to a multidimensional case. In this case, using different copula functions (i.e. a different correlation structure) values of JH and C do not change consistently.

CONCLUSIONS

In this work a methodology to estimate the uncertainty related to the Information Theory quantities of joint entropy and total correlation is presented. The issue is set in the framework of a water level monitoring sensors design: monitoring sensors are optimized maximizing their JH and minimizing their C. The multi-objective problem is solved using a genetic algorithm and the resulting Pareto front is plotted. The main idea is to consider entropy as a probability distribution, rather than a deterministic value. Therefore, when dealing with the issue of choosing the best set of sensors, the best solution is the one with the most probable values of JH and C in the Pareto envelope.

The methodology is applied in the case of both the discrete and continuous random vectors. For the former, discrete probability values are estimated, for the latter the joint probability distributions functions are evaluated through the use of copula functions. It can be concluded that the entropy uncertainty highly influences the deployment of the sensors over the area, therefore, the estimation of all possible entropy values scenarios and the determination of the most robust values make the network deployment more objective.

REFERENCES

[1] *Moss M.E. and Karlinger M.R., "Surface water network design by regression analysis simulation", Water Resources Research, Vol. 10, (1974).*

[2] *Bonaccorso B., Cancelliere A. and Rossi G., "Network design for drought monitoring by geostatistical techniques", European Water, Vol. 8, (2003).*

[3] *Fiering M.B., "An optimization scheme for gaging", Water Resources Research, Vol. 1, (1965), pp. 463-470.*

[4] *Harmancioglu, N.B., Water quality monitoring network design, Kluwer Academic Publishers, (1999).*

[5] *Caselton W.F. and Husain T., "Hydrologic networks: Information transmission", Journal of the Water Resources Planning and Management Division, Vol. 106, (1980), pp 503-520.*

[6] *Mogheir Y. and Singh V.P., "Application of information theory to groundwater quality monitoring networks", Water Resources Management, Vol. 16, (2002), pp 37-49.*

[7] *Alfonso L., Lobbrecht A. and Price R., "Information theory-based approach for location of monitoring water level gauges in polders", Water Resour. Res., Vol. 46, (2010), W03528.*

[8] *Ridolfi E., Alfonso L., Di Baldassarre G., Dottori F., Russo F. and Napolitano F., "An entropy approach for the optimization of cross-section spacing for river modelling", Hydrological Sciences Journal, Vol. 59, No. 1, (2013), 126-137.*

[9] *Alfonso L., Ridolfi E., Gaytan-Aguilar S., Napolitano F. and Russo F., "Ensemble entropy for monitoring network design", Entropy, Vol. 16, (2014), pp. 1365-1375.*

[10] *Deb K., Pratap A., Agarwal S. and Meyarivan T., "A fast and elitist multiobjective genetic algorithm: NSGA-II", IEEE Trans. Evol. Comput., Vol. 6, (2002), pp. 182–197.*

[11] *Grimaldi S. and Serinaldi F., "Asymmetric copula in multivariate flood frequency analysis", Advances in Water Resources, Vol. 29, (2006), pp. 1155–1167.*

[12] *Nelsen RB., "An introduction to copulas. Lecture notes in statistics",. New York: Springer-Verlag, Vol. 139, (1999).*