UC DAVIS DEPARTMENT OF PHYSICS

PHY 256 NATURAL COMPUTATION AND SELF ORGANIZATION

PROJECT REPORT

# USING HIDDEN MARKOV RANDOM FIELDS FOR ESTIMATING THE UNCERTAINTY IN GROUNDWATER FLOW


SERCAN CEYHAN

CIVIL & ENVIRONMENTAL ENGINEERING

HYDROLOGIC RESEARCH LAB

sercan.ceyhan@gmail.com



## ABSTRACT

The uncertainty estimation in groundwater flow models plays an important role for the planning of infrastructure projects. The uncertainty caused by geologic heterogeneity can be modeled using Monte Carlo simulations, which require multiple realizations of the geologic field. In this study a information theoretic approach is followed to model geologic regions as random fields. A Python code was written to determine the occurrence probabilities of the structures and to calculate the Entropy in the dataset. The analysis was conducted for a manually created dataset. It was found that within 16 symbols determined in the dataset, 2 symbols cover almost 87% of the occurences and 8 symbols cover almost 99%.

# INTRODUCTION

Groundwater modeling tools are used in many private, state and federal studies to help decision makers. Those studies mainly include planning and forecasting, historical reconstruction and contaminant transport problems.

Although these tools provide great guidance, we should keep in mind that they are just approximations to the complex natural systems. They hold many assumptions and simplifications that are based on engineering judgment.

Due to the assumptions and simplifications mentioned above, all models carry a certain degree of uncertainty with them. In groundwater models two major sources of uncertainty are the geological heterogeneity and the uncertainty in recharge conditions. In the latter, climate change and anthropogenic impacts play very important roles and it is widely studied by scientists in our decade. The first one on the other hand, is often overlooked as it becomes obvious in much larger scales.

The uncertainty originating from the geological heterogeneity can be modeled by using Monte Carlo simulations. In order to run the Monte Carlo simulations one needs many realizations of the uncertainty causing parameters.

In this study, we aim to develop methods to identify and model the geological heterogeneity by considering it as a spatial system and using information processing and stochastic modeling tools.

This work was conducted to find answers for two questions:

- How random is the spatial system?
- How can we generate other possible realizations of the spatial system?

This report is organized in six sections. It starts with Introduction, which is followed by the Background. In the Background section, it was aimed to provide brief information on groundwater systems. The third section is Dynamical System, which gives further information about the spatial systems and how we considered geology as a dynamical system. Fourth section, Methods, presents the methods and tools used in this study. The results are given in the fifth section and the report is concluded in the sixth section.

# BACKGROUND

## GROUNDWATER FLOW

Groundwater is the major source for clean, safe and reliable water supply. 30 to 46% of California's water is supplied from groundwater while there are some communities that are 100% groundwater dependent for agricultural and urban use.

Often mistaken with the soil water, groundwater is actually the water below the soil layer. The distance from surface to the groundwater surface can change between a couple meters to kilometers.

The groundwater moves slowly within the pores and fractures of the geological formations. The slow flowing groundwater lets us to make many valid assumptions

otherwise not possible. This simplifies the mathematical and physical equations used to solve the groundwater flow.

The groundwater flow equation is a form of diffusion equation, which is derived from a mass balance for a small REV (Representative Elementary Volume) in which the properties were assumed to be effectively constant which is an acceptable assumption in the lab-scale. Darcy's Law is used to describe the fluxes to and from the REV.

The governing equations for confined and unconfined groundwater flow is as follows:

Confined: $$S\frac{\partial h}{\partial t} = \nabla(K\ H\ \nabla h) + R$$

Unconfined: $$S_y\frac{\partial h}{\partial t} = \nabla(K\ h\ \nabla h) + R$$

where,

h: hydraulic head [L]

K: hydraulic conductivity [L/T]

S: specific storage [-]

Sy: specific yield [-]

H: depth of the confined zone [L]

R: sink/source term such as recharge or pumping per unit area [L/T]

The storage coefficients (S and Sy) and the hydraulic conductivity (K) are aquifer parameters and they are related to the geologic formation in which groundwater flows. In reality, factors other than the geology may exist that affect the aquifer parameters. However, for the sake of simplicity in this study we assume that the geologic formations control the aquifer parameters.

In regional studies groundwater masses cover much larger areas. The hydrogeologic information on the regional scale is often interpreted from the point data using geological methods. Because of the size of the scale, the heterogeneity affecting the groundwater flow cannot be represented.

# DYNAMICAL SYSTEM

The information theory and stochastic modeling tools are generally developed and used for the analysis and modeling of time series. 1-D spatial systems can also be treated the same way by using the same tools.

The geological formations were formed in multiple dimensions very slowly over a very long period of time. The rate of the process is very small so that we can say that there is no time dependency. However, formations are created next to each other, which makes them space dependent. There are transitions between formations, and similar formations often follow certain others as they were formed.

When we talk about the geology, or other spatial 2-D systems in general, things get more complicated. To begin with, as mentioned there is no time dimension in the systems we study. A good starting point is to think the system as two independent 1-D systems. However, with this assumption we lose very important information that the system harbors between the dimensions.

For this very specific reason, we have to keep the structural information between spatial dimensions together in a symbol while conducting our analysis. In this study, to detect the symbols in a dataset we chose the “Z” shaped template presented in (Feldman & Crutchfield, 2002). The “Z” shape accommodates left and right information in adjacent upper and lower rows altogether (Figure 1 - top). The set of all symbols forms our alphabet (Figure 1 - bottom).

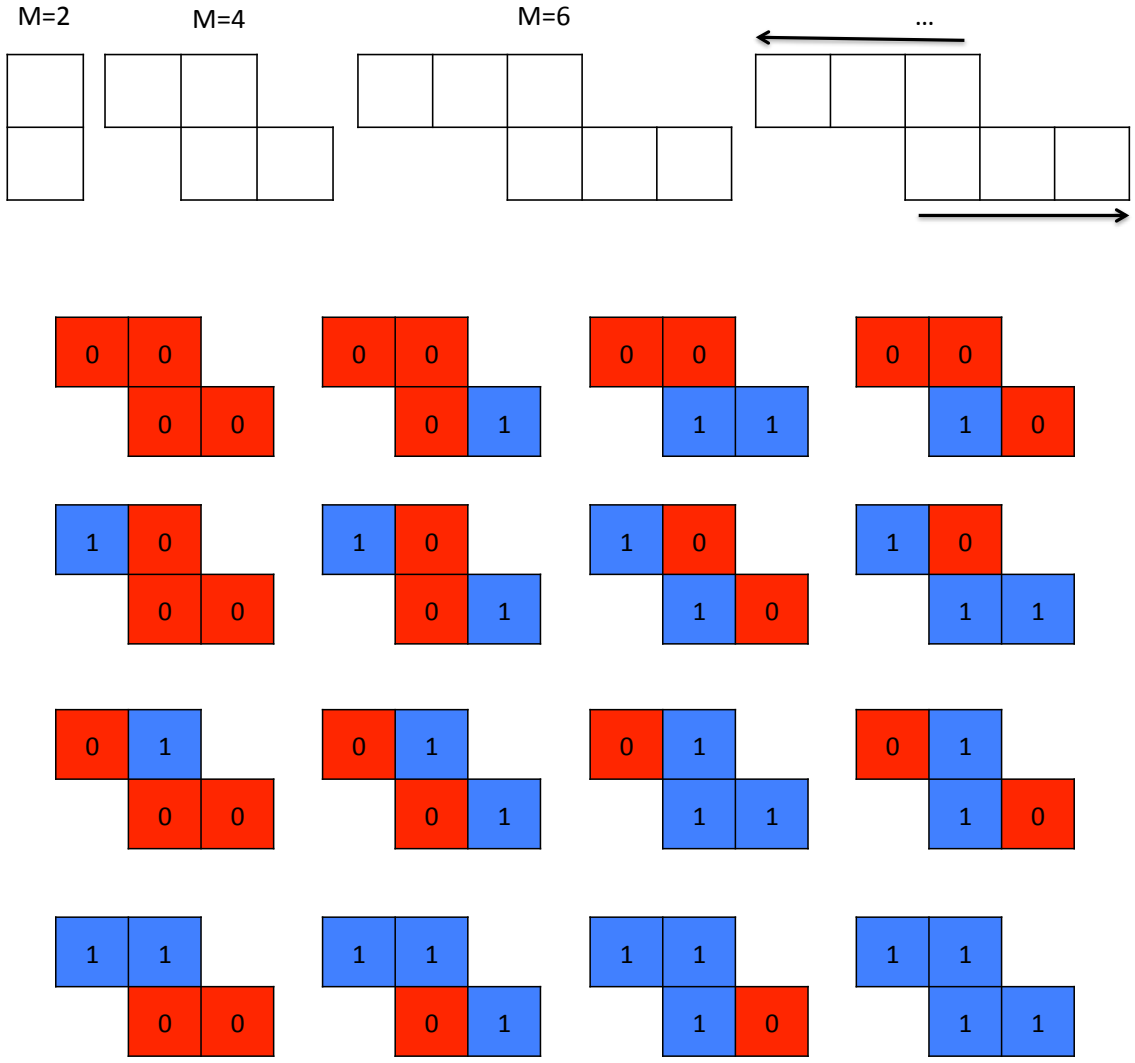



FIGURE 1. THE "Z" SHAPED TEMPLATE (TOP) AND THE ALPHABET FOR A WORD LENGTH (M) OF FOUR (BOTTOM)

Since we stored the multiple dimensional information in the symbols, we can now consider our model as a 1-D system.

# METHODS

## PROBABILITIES

Occurrence probability of each symbol is determined by ratio of the number of times each symbol was observed ($N_i$) and the total number of occurrences in the spatial dataset ($N$). The occurrence probability of symbol $i$, $p_i$, is calculated as:

$$p_i = \frac{N_i}{N}$$

## ENTROPY

Used in information theory, entropy is the number of yes/no questions we need to ask to determine which symbol was observed (Jvstone, 2014) (Crutchfield, 2014). It is calculated as:

$$H = -\sum_i p_i \log_2 p_i$$

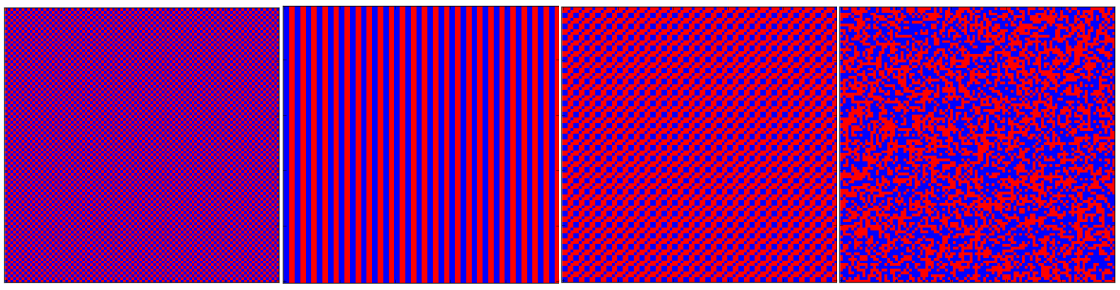
In a case with 2 classes, since a yes/no question will give us the value of a node, we can also think entropy as the number of the nodes that we need to know within the symbol to determine the rest of the symbol.

## TEST RUNS

A Python code is written to visualize and infer the dataset and to calculate entropies. Some functions are used from the package Numpy. The program code is included at the end of this report.

For test runs during the development of the code, following patterns are used as inputs. The number of classes is limited to 2 for computational ease.

- 1x1 **Checkerboard** pattern,
- **Vertical Lines** with width of 1,
- **Diagonal** pattern with 1x1 and 2x2 squares,
- Uniform **random** distribution.



The template sizes are varied between M=2 and M=16 and the entropy is recorded. For M=2 case there are $2^2 = 4$ unique possible symbols. For M=16 this number increases to

$2^{16} = 65{,}536.$

Once the code is complete and running, it was tried on a 100x100 dataset created manually. This dataset is presented below in Figure 2 and will be called as 256B dataset in the rest of the report. As you can see from the image, the dataset neither have any repeating pattern in it like the previous cases nor it is completely random. It has some structure inside it and therefore harbors information to be discovered.

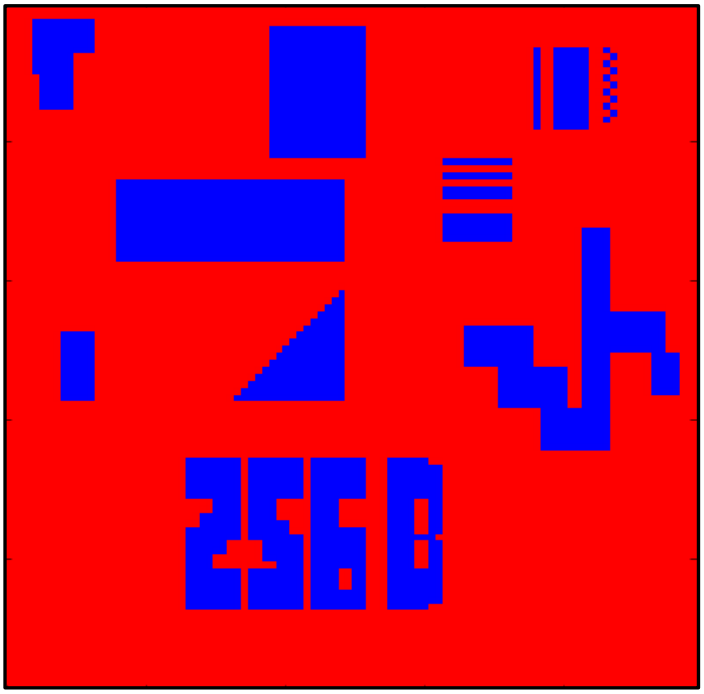



FIGURE 2. 256B DATASET

# RESULTS

## TEST RUNS

### *OCCURRENCES & PROBABILITIES*

The number of unique symbols determined in each pattern in the test runs is presented in Table 1. In the first three, depending on the pattern, we see that the number of unique symbols reach to a limit.

On the other hand, the random case never reaches to a limit as we expected. With an adequate sample size, we should see all possible symbols as unique symbols. In our computation this is what happens until M=12. After this word length our sample size becomes inadequate for the analysis.

TABLE 1. NUMBER OF UNIQUE SYMBOLS OBSERVED IN TEST RUNS FOR CHANGING WORD LENGTHS

| # OF UNIQUE SYMBOLS | Word Length | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **2** | **4** | **6** | **8** | **10** | **12** | **14** | **16** |
| **Checkerboard** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| **Vertical Lines** | 4 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |
| **Diagonal** | 4 | 14 | 22 | 28 | 28 | 28 | 28 | 28 |
| **Random** | 4 | 16 | 64 | 256 | 1024 | 4041 | 10650 | 14894 |
| **# of POSSIBLE SYMBOLS** | **4** | **16** | **64** | **256** | **1024** | **4096** | **16384** | **65536** |
| **TOTAL # of OBSERVATIONS** | **19600** | **19208** | **18816** | **18424** | **18032** | **17640** | **17248** | **16856** |

When we looked to the probabilities, the checkerboard, vertical lines and random cases output equal probabilities for each unique symbol. The diagonal pattern output different probabilities because of its different symmetry. For M=8 case, its probabilities are changing between 1/16 and 1/32.

## *ENTROPY*

Table 2 and Figure 3 presents the entropies calculated for each pattern and word length. For the checkerboard pattern it is enough for us to know only the value in one node to fill the rest. On the other hand, in the random case there is no way for us to determine the symbol if we don't know the value of every node. Once again, because of the inadequate sample size, the values diverge after M=12 for the random case.

TABLE 2. ENTROPY CALCULATED FOR TEST RUNS FOR CHANGING WORD LENGTHS

| ENTROPY | Word Length | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **2** | **4** | **6** | **8** | **10** | **12** | **14** | **16** |
| **Checkerboard** | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| **Vertical Lines** | 1.811 | 3.000 | 3.000 | 3.000 | 3.000 | 3.000 | 3.000 | 3.000 |
| **Diagonal** | 1.906 | 3.639 | 4.375 | 4.750 | 4.750 | 4.750 | 4.750 | 4.750 |
| **Random** | 2.000 | 4.000 | 6.000 | 8.000 | 10.000 | 11.800 | 13.209 | 13.800 |

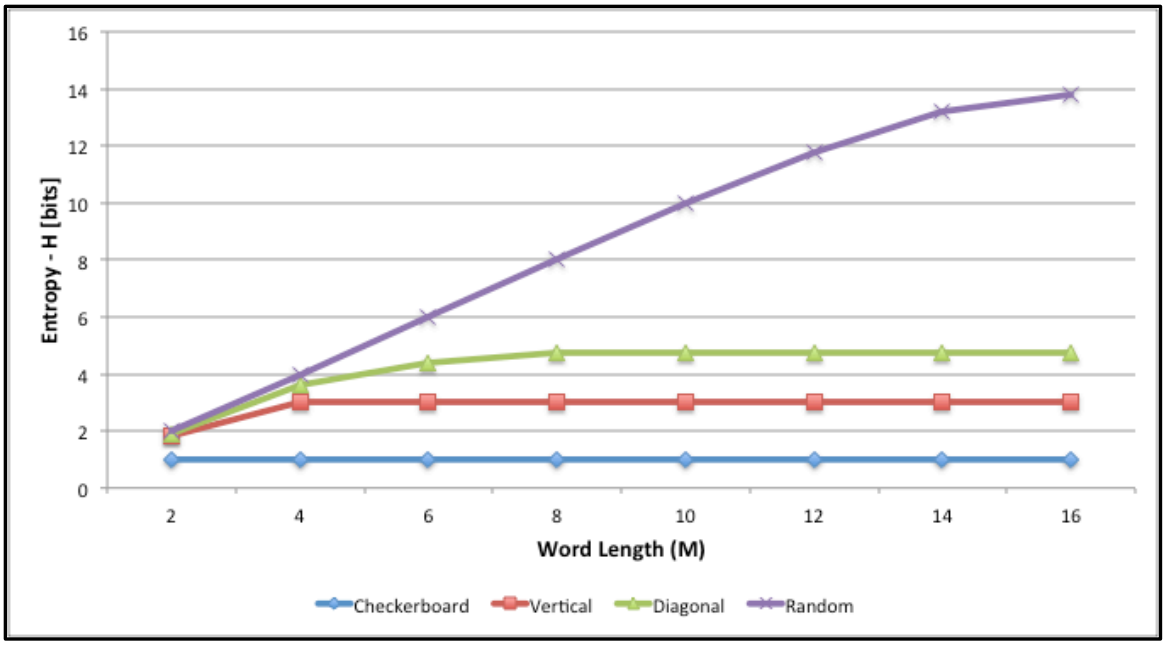



FIGURE 3. ENTROPY IN EACH PATTERN FOR CHANGING WORD LENGTHS

## 256B DATASET

The algorithm was run for 256B dataset with wordlengths between M=2 to M=10. The numbers of observed unique symbols are presented in Table 3, and the entropies are presented in Table 4 and Figure 4.

TABLE 3. NUMBER OF UNIQUE SYMBOLS OBSERVED IN TEST RUNS FOR CHANGING WORD LENGTHS

| **# OF UNIQUE SYMBOLS** | **Word Length** | | | | |
|---|---|---|---|---|---|
| | **2** | **4** | **6** | **8** | **10** |
| **256B** | 4 | 16 | 56 | 125 | 228 |
| **# of POSSIBLE SYMBOLS** | **4** | **16** | **64** | **256** | **1024** |

TABLE 4. ENTROPY CALCULATED FOR TEST RUNS FOR CHANGING WORD LENGTHS

| **ENTROPY** | **Word Length** | | | | |
|---|---|---|---|---|---|
| | **2** | **4** | **6** | **8** | **10** |
| **256B** | 1.012 | 1.565 | 2.114 | 2.635 | 3.151 |

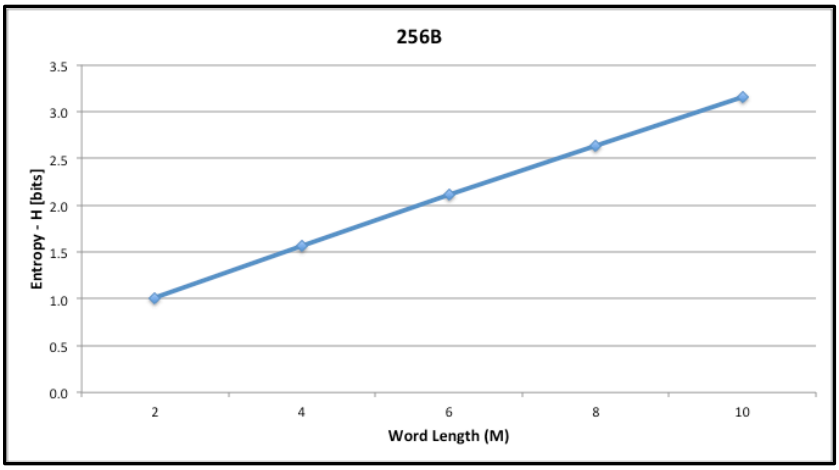



FIGURE 4. ENTROPY IN EACH PATTERN FOR CHANGING WORD LENGTHS

From above results we cannot determine a certain word length that keeps the entropy constant and it continuously increases such as the random case in the test runs. However, there is important information hidden in the occurrence probabilities of the symbols. 16 symbols determined for M=4 are given below with their occurrence probabilities (Figure 5). 2 symbols cover almost 87% of the occurrences and 8 symbols cover almost 99%. This situation significantly reduces the entropy of the system compared to the random one.

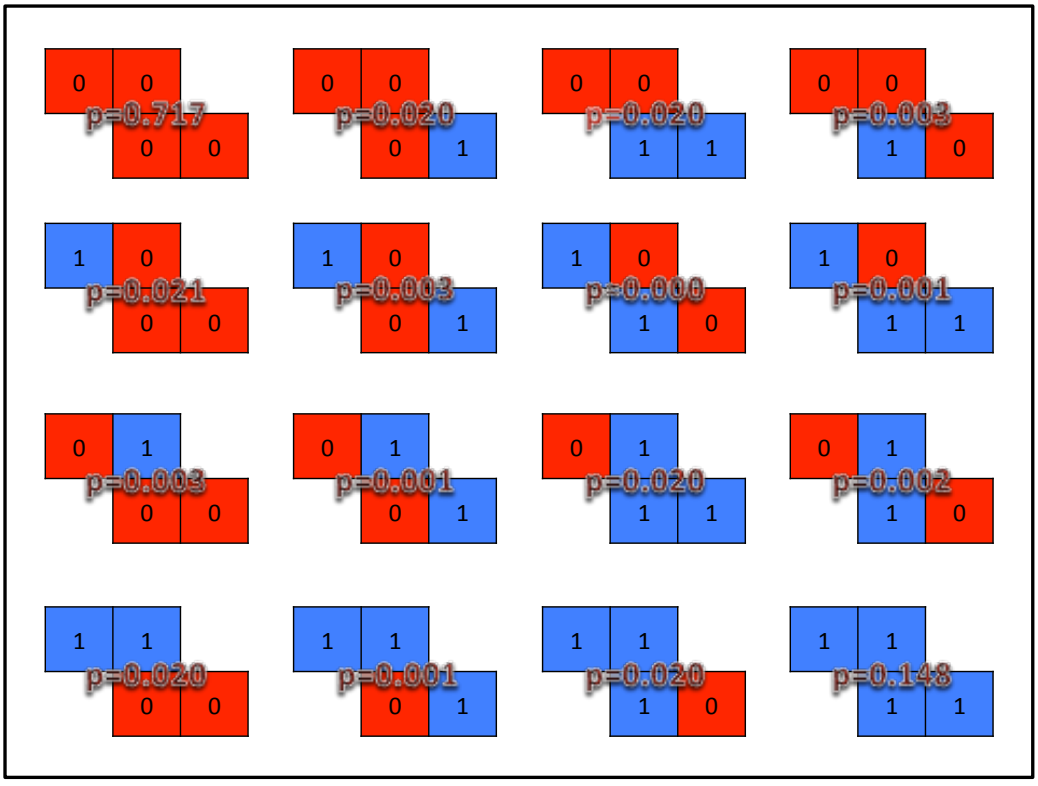



FIGURE 5. UNIQUE SYMBOLS OBSERVED IN 256B DATASET AND THEIR PROBABILITIES

# CONCLUSION

Most stochastic groundwater studies take the aquifer parameters as random constants. Considering the formation of the geologic structures, aquifer parameters should be considered as random fields rather than random constants.

In this study, an information theoretic approach is followed to model aquifer parameters as random fields. The symbols used in 1-D are redefined for 2-D structures and entropy is calculated for different datasets. In this way, the uncertainty or randomness of the data is quantified.

For the natural datasets, no word length was found after which the entropy becomes constant.

Also this method and code can be used with datasets that have more than 2 classes although the computational time will take significantly longer with longer word lengths.

The occurrence probabilities of symbols are determined during entropy calculation. They can be used for generating similar random fields to be used in Monte Carlo simulations for uncertainty estimation. The generation algorithm should be created considering the dataset.

# BIBLIOGRAPHY

Crutchfield, J. P. (2014). PHY 256 Natural Computation and Self-Organization Lecture Notes. Davis, CA, USA.

Feldman , D. P., & Crutchfield, J. P. (2002). *Structural Information in Two-Dimensional Patterns: Entropy Convergence and Excess Entropy.* Santa Fe Institute Working Paper .

Jvstone. (2014, 6 9). *Entropy (information theory)*. Retrieved 6 10, 2014, from Wikipedia: http://en.wikipedia.org/w/index.php?title=Entropy_(information_theory)&oldid=612236788