GEOSTATISTICAL ANALYSIS OF HYDRAULIC CONDUCTIVITY FIELDS USING COPULAS

CLAUS P. HASLAUER[1], ANDRÁS BÁRDOSSY[1],
and EDWARD A. SUDICKY[2]
[1]Institute of Hydraulic Engineering, Department of Hydrology and Geohydrology, University of Stuttgart, Germany.
[2]Department of Earth and Environmental Sciences, University of Waterloo, Canada.

ABSTRACT

Hydraulic conductivity is the parameter of central importance for groundwater flow and contaminant transport. Relevant physical processes in the subsurface can not be described only by a good estimation of mean hydraulic conductivities but those processes require a sound knowledge of the underlying spatial dependence structure of hydraulic conductivity. The purpose of this contribution is to investigate real hydraulic conductivity fields using multivariate Copulas. Using Copulas offers two key advantages: (1) the possibility to investigate the spatial dependence structure of the hydraulic conductivity field independently from the marginal distributions; and (2) to be able to incorporate additional hard and soft knowledge through assumptions on the location-specific distributions. The goal of this paper is to demonstrate how these two key advantages are incorporated into spatial analysis, which is consecutively used to improve the statistical model of spatial dependence structures.

INTRODUCTION

It has been recognized that hydraulic conductivity (K) fields used in (hydro-)geologic models need to have a certain property if they should resemble nature: Spatially close points with similar quantile ranges of K need to be connected. Especially the connection of extreme values is critical for groundwater flow, because such zones act as "barriers" or "windows" [Frind et al., 2002]. Spatial fields with such properties do not exhibit a gaussian spatial dependence structure [Journel and Alabert, 1989].

Nevertheless, naturally occurring spatially distributed data has been frequently assumed to be log-normally distributed. Such a log-transformation of the original data often facilitates computing. A second assumption that is commonly made when building a model for spatial analysis is that the spatial dependence structure of spatially distributed data can be described with a normal distribution. Kriging, a typical model for spatial interpolation uses this assumption and hence all values, regardless of their magnitude, are treated equally when assigning Kriging-weights to them.

Spatial Copulas offer a method to analyze spatial dependence structures without the need of gaussian assumptions: The first assumption of a log-normal shape of the marginal distribution function is overcome by using the ranks of the hydraulic conductivities, which leads to an exactly uniform marginal distribution function, and not to a somewhat normal marginal distribution function. Secondly, the spatial dependence structure of the data-set that is being investigated can be modelled using empirical Copula densities which have properties that allow to recognize the type and magnitude of spatial dependence.

The data-set that is used for the analysis of spatial dependence structure is the same data-set used by Sudicky [1986] and by Woodbury and Sudicky [1991]. A total of 1188 K measurements were used for the analysis, taken along two perpendicular transects in the Borden aquifer in Ontario, Canada. One cross-section (called"*AA*'") is 19 m long, the other cross-section (called "*BB*'") is 6 m long, and both cross-sections are 1.75 m thick. Hydraulic conductivity was measured via permeameter tests using samples from cores that were spaced 1 m horizontally and subsampled every 0.05 m vertically.
The Borden aquifer originated as beach sand along the shore of a glacial meltwater lake that existed at the end of Pleistocene time [Cherry et al., 1996]. The aquifer has traditionally been thought of as being homogeneous, until Sudicky [1983] showed that there is strong influence of aquifer heterogeneity on dispersion even though the aquifer is relatively homogeneous in its hydraulic behaviour [Cherry et al., 1996].

METHODOLOGY

Plots of empirical Copula densities are used to evaluate if the dependence structure of the given data-set is gaussian. Supporting measures include the symmetry of an empirical Copula density as well as the rank correlation coefficient. These novel measures are compared to the more traditional measure variogram functions.

Empirical Copula Density Functions

Copulas describe the dependence structure between random variables without information on the marginal distributions. Thus they can be seen as the pure representation of the dependence between the random variables over the range of quantiles. Any multivariate distribution $F(t_1,\ldots,t_n)$ can be represented with the help of a Copula C [Sklar, 1959]:

$$F(t_1,\ldots,t_n) = C\left(F_{t_1}(t_1),\ldots,F_{t_n}(t_n)\right) \tag{1}$$

where $F_{t_i}(t_i)$ represents the i-th one-dimensional marginal distribution of the multivariate distribution.

Assuming that C is continuous then the Copula density $c(u_1,\ldots,u_n)$ can be written as

$$c(u_1,\ldots,u_n) = \frac{\partial^n C(u_1,\ldots u_n)}{\partial u_1,\ldots \partial u_n} \tag{2}$$

A key advantage of using the Copula of a multivariate distribution is that it is invariant to monotonic transformations of the marginal variables. Thus frequently applied data transformations (taking the natural logarithm, e.g.) do not influence the Copula.

An interesting and important property of a Copula is that it can express whether the corresponding dependence is different for different quantiles of the variable. For example, high values might exhibit a strong spatial dependence, low values a weak spatial dependence, and mid-range value an again different degree of dependence.

A bivariate Copula expresses a symmetric dependence with respect to the minor axis $u_2 = 1-u_1$ of the unit square, if

$$c(u_1,u_2) = c(1-u_1,1-u_2) \tag{3}$$

Such bivariate Copulas can be used to describe the spatial variability, similar to a description using variograms or covariance functions. For this purpose it is assumed that the bivariate Copula of the random variable $Z(\mathbf{x})$ corresponding to two locations $\mathbf{x}$ separated by the vector $\mathbf{h}$ does not depend on the location $\mathbf{x}$ but on $\mathbf{h}$ only. In this case, the spatial Copula C_S becomes a function of the vector $\mathbf{h}$. For any two selected quantiles $u1,u2$, C_S is given in Equation 4.

$$C_S(\mathbf{h},u_1,u_2) = P\left[F_Z(Z(\mathbf{x})) < u_1,\, F_Z(Z(\mathbf{x}+\mathbf{h})) < u_2\right] = C\left(F_Z(Z(\mathbf{x})),\, F_Z(Z(\mathbf{x}+\mathbf{h}))\right) \tag{4}$$

The assessment of the bivariate Copulas from measured data $z(x_1),\ldots,z(x_n)$ can be done by first calculating the empirical distribution function $F_n(z)$. Using this distribution function for any given vector $\mathbf{h}$, the set of pairs $S(\mathbf{h})$, consisting of distribution function values corresponding to the parameter at locations separated by the vector $\mathbf{h}$, can be calculated

$$S(\mathbf{h}) = \{F_n(z(\mathbf{x}_i)), F_n(z(\mathbf{x}_j)) \,|\, (\mathbf{x}_i - \mathbf{x}_j \approx \mathbf{h}) \text{ or } (\mathbf{x}_j - \mathbf{x}_i \approx \mathbf{h})\} \tag{5}$$

$S(\mathbf{h})$ is thus a set of points in the unit square. Note that $S(\mathbf{h})$ is by definition symmetrical regarding the major axis $u_1 = u_2$ of the unit square, namely, if $(u_1,u_2) \in S(\mathbf{h})$, then $(u_2,u_1) \in S(\mathbf{h})$.
Plots of empirical Copula densities show the spatial dependence structure for pairs of points separated by a $S(\mathbf{h})$. On each axis the standardized rank of the K at these points is plotted. Measurement points where K is very low are plotted close to the origin, and similarly points where the measured K is high are plotted far from the

origin. If the empirical Copula density is high, indicated by dark shading in this paper, then there are a lot of pairs of points separated by the given distance that have the combination of K values that is represented by the ranks on the axes.

Further details on the theory of Copulas can be found in Joe [1997] and in Nelsen [1999]. Details on using Copulas with spatially distributed data are given by Bárdossy [2006].

Supporting Measures

Three measures are used as comparison with the empirical Copula densities: (1) the traditional variogram "γ" (Equation 6), (2) the rank correlation function "Rank" (Equation 7), and (3) a measure "Sym" for the symmetry of the empirical Copula density function (Equation 8).
Each of these measures is calculated for a given magnitude and/or angle of anisotropy of the separation vector $\mathbf{h}$. The number of pairs of points for each $\mathbf{h}$ is denoted by $n(\mathbf{h})$.

$$\gamma(\mathbf{h}) = \frac{1}{2n(\mathbf{h})} \cdot \sum_{\mathbf{x_i}-\mathbf{x_j}\approx\mathbf{h}} \left(z(\mathbf{x_i}) - z(\mathbf{x_j})\right)^2 \tag{6}$$

$$\text{Rank}(\mathbf{h}) = \frac{1}{12n(\mathbf{h})} \cdot \sum_{\mathbf{x_i}-\mathbf{x_j}\approx\mathbf{h}} \left(F_n(z(\mathbf{x}_i)) - \frac{1}{2}\right) \cdot \left(F_n(z(\mathbf{x}_j)) - \frac{1}{2}\right) \tag{7}$$

$$\begin{aligned} \text{Sym}(\mathbf{h}) = {} & \frac{1}{n(\mathbf{h})} \cdot \sum_{\mathbf{x_i}-\mathbf{x_j}\approx\mathbf{h}} \left(F_n(z(\mathbf{x}_i)) - \frac{1}{2}\right)^2 \left(F_n(z(\mathbf{x}_j)) - \frac{1}{2}\right) + \\ & + \left(F_n(z(\mathbf{x}_i)) - \frac{1}{2}\right) \left(F_n(z(\mathbf{x}_j)) - \frac{1}{2}\right)^2 \end{aligned} \tag{8}$$

RESULTS

The absolute frequencies of the hydraulic conductivities obtained from Sudicky's [1986] cross-sections AA' and BB' are shown on Figure 1. Both distributions have a very similar arithmetic mean (~0.011 cm/s), a similar standard deviation s (~0.0055 cm/s), and a positive skewness g of ~ 0.8.

A key step in traditional geostatistical analysis is the construction of the empirical variograms according to Equation 6. Empirical variograms are shown for both cross-sections for half of their respective lengths in vertical direction (Figure 2a) and horizontal direction (Figure 2b).
The empirical variogram in the vertical direction shows a range of ~0.6 m and a sill of $\sim 3.0 \cdot 10^{-5}$ cm^2/s^2 for cross-section AA' symbolized with circles, and similar and maybe slightly higher values for cross-section BB'. For the horizontal direction not much can be deduced from the variogram.

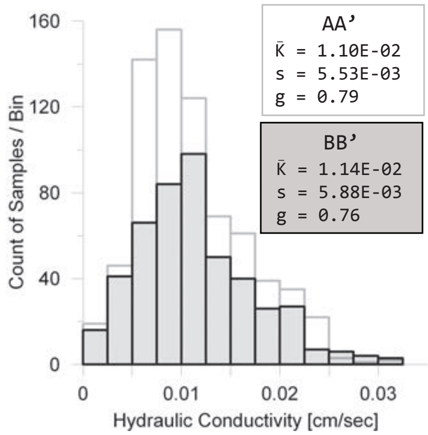
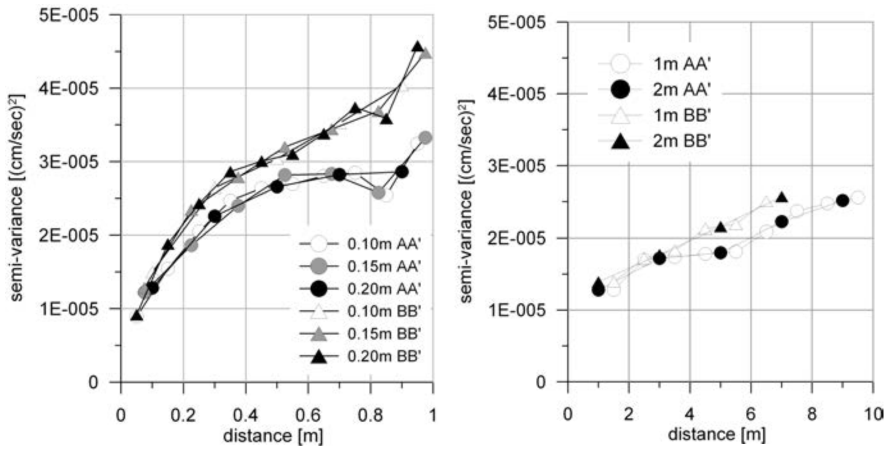


Figure 1: Absolute frequencies of hydraulic conductivities [cm/s] from the samples obtained from cross-section AA’ (count = 720) and cross-section BB’ (count = 468). Shown for each cross-section are the arithmetic averages [cm/s], the standard deviation s [cm/s], and the skewness g [-].

Geostatistical Analysis using Copulas

The same data-set that was employed for the empirical variograms (Figure 2) was used to construct empirical Copula density charts according to Equation 5 for cross-sections AA’ (Figure 3) in the vertical and horizontal directions. Rank correlation (Equation 7) and symmetry (Equation 8) functions were plotted to quantify the findings from the empirical Copula densities (Figures 4a, 4b, 4c, 4d).

There are two novel pieces of information that can be deduced from the Copula-analysis:
First, the empirical Copulas show a varying degree of dependence for different percentile values, which is a line of evidence that can not be deduced from traditional geostatistics.
Second, the empirical Copulas are not symmetric following the definition given in Equation 3. Hence the spatial dependence structure of the presented data-set does not follow a Gaussian distribution. The degree of symmetry calculated with Equation 8 confirms this finding; the calculated value differs from zero for most of the shown distances (Figures 4b and 4d).
Beyond the effective variogram range the symmetry should approach zero, indicating a flat empirical Copula density, meaning the correlation between points for that distance is zero. This is the case for the vertical symmetry at ~ 0.5 m (Figure 4d), also corresponding to the point where the rank correlation (Figure 4c) comes close to zero. For the horizontal case this point of zero symmetry is not reached (Figure 4b), and neither is a rank correlation of zero (Figure 4a). The problem of being able to determine an effective range in horizontal direction is the same as for the traditional variogram (Figure 2b).



(a) Vertical direction (b) Horizontal direction

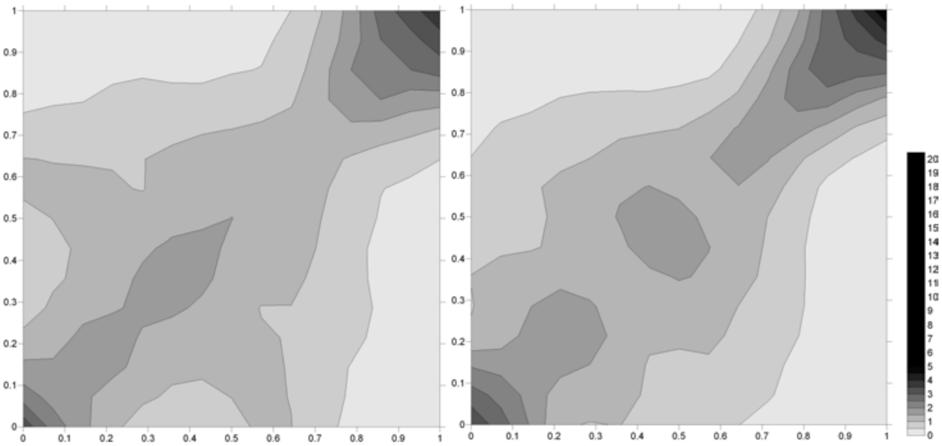
Figure 2: Directional empirical variograms based on original K values from cross-sections AA' (circles) and BB' (triangles). Different shading indicates different lengths used for averaging.

Accounting for Horizontal Layering

As discussed in the introduction, the Borden aquifer is a beach deposit. Such an environment could lead to higher compaction of the sediments, and hence lower K values at the bottom of the deposit than at its top. Both cross-sections were divided into three horizontal layers of equal thickness. The empirical distribution functions (EDFs) for these three layers are shown for both cross-sections on Figure 5.
The top layer exhibits on average higher K-values than the middle layer, and the middle layer higher K-values than the bottom layer. From visual inspection it is evident that the top layer behaves significantly different than the middle and the bottom layer. All of these EDFs are different from each other on a 5% significance level using a Kolmogorov-Smirnov test, except the bottom-layer and middle-layer EDFs of cross-section AA'.

The geologic interpretation and the differing K-distributions for each of the three layers match: it can be assumed that the bottom of the deposit is more compacted and has hence on average a lower K than the top of the deposit. Assuming further that the spatial distribution of K is similar in each of the three layers, an empirical Copula density can be calculated for the entire deposit along each cross-section by using the EDF of each horizontal section as building block (Figure 6). This step assumes that the horizontal spatial dependence structure is similar for all three layers. In vertical direction, the spatial dependence structure is linked via the EDF values, which can be regarded as a valid assumption.

The empirical Copula densities taking the layering into account are shown on Figure 6a in horizontal direction, and on Figure 6b in vertical direction. The rank correlation and symmetry functions that take the layering into account are shown



(a) Horizontal; $|\mathbf{h}| = 2$ m (b) Vertical; $|\mathbf{h}| = 0.05$ m (c)

Figure 3: Empirical spatial Copula density plots for cross-section AA' for the original *K* data-set ("one block").

on Figures 4e and 4f in horizontal direction and on Figures 4g and 4h in vertical direction.

Accommodate horizontal layering has reduced some of the variability, leading to more symmetric empirical Copulas, and also to a slightly decreased rank correlation for short distances in horizontal direction. Taking the layering into account also has explained some of the decrease in effective range, where this decrease has been deduced from the point where the rank correlation– and symmetry functions reach zero. This generally decreased range makes it even possible now to deduce an effective range in horizontal direction, because zero rank correlation is reached now in that direction as well.

Admittedly, introducing layering in a simplistic way as with three layers of equal thickness might not capture all of the layering processes occurring at the site. There might be smaller scale layers both in vertical and horizontal direction ("lenses"). However, empirical Copula densities allow for taking such physical effects into account.

CONCLUSIONS

Using empirical Copulas for geostatistical analysis showed that the spatial dependence structure of the hydraulic conductivity fields based on measurements from cores from two cross-sections of the Borden Aquifer, Ontario is not a normally distributed dependence structure. The novel Copula approach presented here allows for taking additional physical information into account, such as the horizontal layering of the deposit. This has proven to be useful because the systematic part of the variability could be explained by taking the layering into account.

Using this approach for building a geostatistical model influences future interpolation and simulation results positively because it can be considered more physically based. Hydraulic conductivity fields simulated using Copulas will subsequently be used for numerical groundwater flow and contaminant transport simulations and match observed conditions better than currently.

ACKNOWLEDGEMENTS

This research has been funded by the German Research Foundation (DFG) with grant BA 1150/12-1 and has been supported by the International Research Training Group “non-linearities and upscaling in porous media” (NUPUS). Thanks to the great support of co-op students Maria Niedermaier and Philipp Guthke!

REFERENCES

Bárdossy, A (2006). *Copula-based geostatistical models for groundwater quality parameters*. In Water Resources Research, vol. 42, no. W11416. doi:10.1029/2005WR004754.

Cherry, JA, Barker, JF, Feenstra, S and Gillham, RW (1996). *Groundwater and Subsurface Remediation – Research Strategies for In-Situ Remediation*, chap. The Borden Site for Groundwater Contamination Experiments: 1978–1995. Springer Verlag, pp. 101–127.

Frind, E, Muhammad, D and Molson, J (2002). *Delineation of three-dimensional well capture zones for complex multi-aquifer systems*. In Ground Water, vol. 40, no. 6, pp. 585–598.

Joe, H (1997). *Multivariate Models and Dependence Concepts*. No. 73 in Monographs on Statistics and Applied Probability. Chapmann Hall / CRC.

Journel, A and Alabert, F (1989). *Non-gaussian data expansion in the earth sciences*. In TERRA Nova, vol. 1, pp. 123–134.

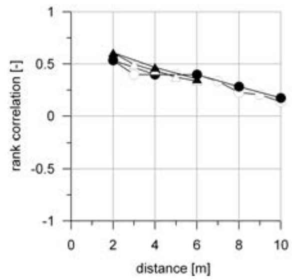
Nelsen, RB (1999). *An Introduction to Copulas*, vol. 139 of *Lecture Notes in Statistics*. Springer Verlag.

Sklar, M (1959). *Fonctions de répartition a n dimensions et leur marges*. In Publ. Inst. Stat. Paris, vol. 8, pp. 229 – 131.

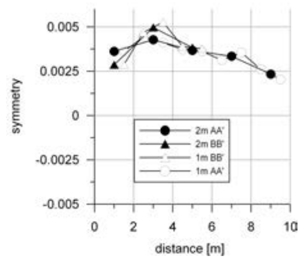
Sudicky, EA (1983). *Migration of contaminants in groundwater at a landfill: A case study, 4. a natural gradient dispersion test*. In Journal of Hydrology, vol. 63, pp. 81–108.

Sudicky, EA (1986). *A natural gradient experiment on solute transport in a sand aquifer: Spatial variability of hydraulic conductivity and its role in the dispersion process*. In Water Resources Research, vol. 22, no. 13, pp. 2069–2082.

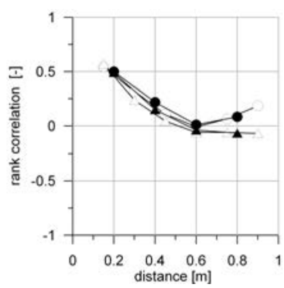
Woodbury, AD and Sudicky, EA (1991). *The geostatistical characteristics of the borden aquifer*. In Water Resources Research, vol. 27, no. 4, pp. 533–546.



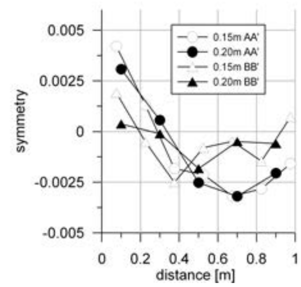
(a) Horizontal RC one block



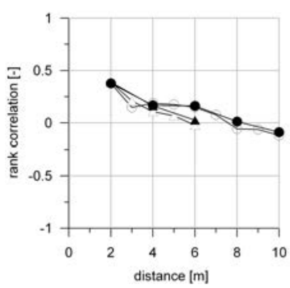
(b) Horizontal Sym one block



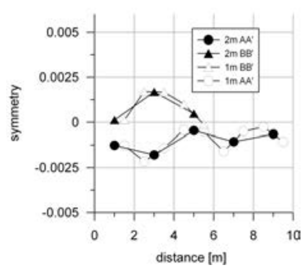
(c) Vertical RC one block



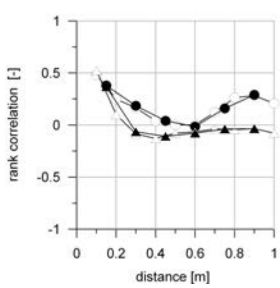
(d) Vertical Sym one block



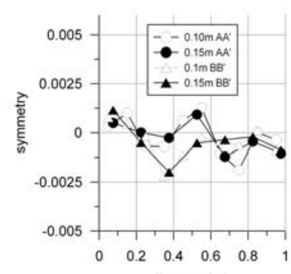
(e) Horizontal RC 3 Layers



(f) Horizontal Sym 3 Layers



(g) Vertical RC 3 Layers



(h) Vertical Sym 3 Layers

Figure 4: Horizontal and vertical rank correlation ("RC") and symmetry ("Sym") functions for the original K data-set ("one block") and with three Layers introduced ("3 Layers"), for cross-sections AA' and BB'.

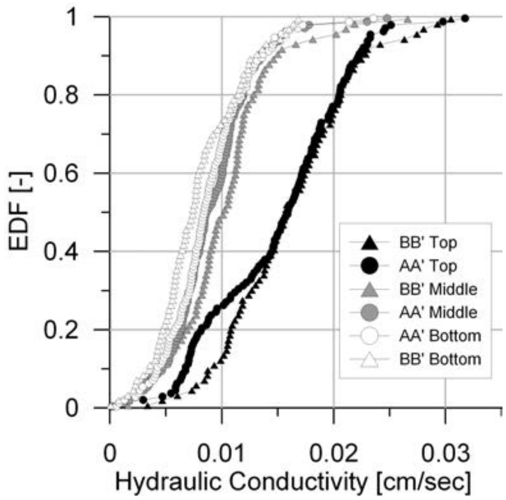
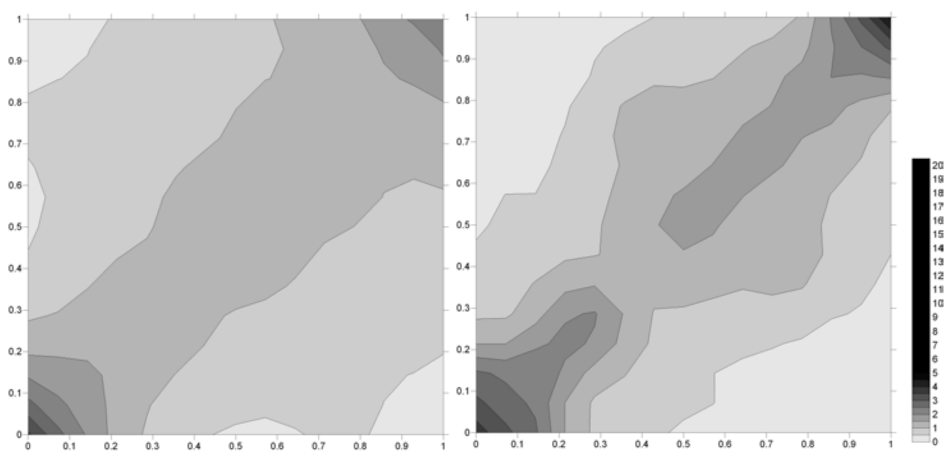


Figure 5: Empirical distribution functions for each of three horizontal layers of equal thickness introduced onto cross-sections AA' and BB'.



(a) Horizontal; $|\mathbf{h}| = 2$ m (b) Vertical; $|\mathbf{h}| = 0.05$ m (c)

Figure 6: Empirical spatial Copula density plots for cross-section AA' with three Layers introduced ("3 Layers").