January 1992

Nonparametric Stratigraphic Interpretation from Drill Log Data

Upmanu Lall

A. I. Ali

Recommended Citation

Lall, Upmanu and Ali, A. I., "Nonparametric Stratigraphic Interpretation from Drill Log Data" (1992). *Reports.* Paper 245.
https://digitalcommons.usu.edu/water_rep/245

Utah Water Research Laboratory

Nonparametric Stratigraphic Interpretation from Drill Log Data

Upmanu Lall and A. I. Ali

Research Report RR-92-HWR-UL/002

October 1992

Utah Water Research Laboratory, Utah State University, Logan UT 84322-8200, USA
(801)-750-3155, FAX : (801)-750-3663

66802

Nonparametric Stratigraphic Interpretation from Drill Log Data

U. Lall, A. Ali
Utah Water Research Laboratory, Utah State University
Logan UT 84322-8200

ABSTRACT

Near surface groundwater contamination is a widespread problem. The potential for contamination of deep aquifers in such areas depends on the vertical hydraulic gradient as well as the extent and location of interconnection between the upper and lower aquifers. In alluvial, sedimentary environments, the geologic units are typically weakly connected lenses or layers of high or low conductivity with variable size, geometry and orientation. Drill logs provide qualitative, local information on such aquifer heterogeneities. A binary (high or low conductivity) indicator function is used to describe the aquifer system. A nonparametric statistical methodology for assessing the probability that a particular location in the aquifer has high or low conductivity using drill log information is developed. The stochastic behavior of the sedimentary process in the vertical is of particular interest. Example applications using data from Lake Bonneville deposits in Salt Lake County, Utah are presented.

INTRODUCTION

Drinking water supplies in many areas are developed from deep aquifers. Shallow aquifers are often contaminated. As water demand increases, at least locally, hydraulic gradients that were historically upward are reversed. Concerns about the likely contamination of the deep aquifers consequently arise. In alluvial, sedimentary environments, shallow and deep aquifer systems are typically separated by discontinuous lenses or layers of markedly different hydraulic conductivity, geometry and size. Preferential pathways of varying size and orientation thus exist for the potential movement of contaminants through the heterogeneous aquifer system. Evaluation of such heterogeneities is a serious challenge.

The primary sources of information on subsurface hydraulic properties are pump tests and drill logs. Existing pump test methodologies are inappropriate for the identification of subsurface heterogeneities since they focus on the estimation of average hydraulic parameters, the aquifer response to pumping is damped and is smoothed over the discontinuities in hydraulic conductivity. Hence pumping tests can provide only limited information as to the spatial variability of hydraulic conductivity. Further, the "inverse problem" of recovering a spatially distributed parameter set from a pump test is inherently ill-posed. Drill logs provide only local, qualitative stratigraphic information. The classical, geologist's stratigraphic sections based on drill logs provide only a smoothed, subjective interpretation of the subsurface. Extracting useful information from such data is clearly a challenge.

We were interested in interpreting drill log information to identify the likely locations and extent of areas of high and low hydraulic conductivity. Such an interpretation could be used for selecting monitoring locations, designing pump tests, and improving the solution of the aquifer parameter identification problem. A probabilistic framework was considered relevant. A binary, indicator function was used to categorize drill log sections as areas of high (1) and low (0) conductivity. A nonparametric, statistical methodology was developed to estimate the probability that a particular location in the aquifer has high or low conductivity. The focus is on the description of a stochastic process for lens occurrence in the vertical, and an interpolation of this process in the horizontal. Nonhomogeneous Markov and Poisson Process models are used, with spatially varying, locally smoothed intensities that are estimated using Kernel methods. Example applications using data from Lake Bonneville deposits in Salt Lake County, Utah are presented.

BACKGROUND AND RELATED RESEARCH

The situation of interest is illustrated through Figure 1. The near surface unconfined aquifer, and the deep confined aquifer have hydraulic conductivities K_u and K_c respectively that are an order of magnitude higher than the hydraulic conductivity K' of the intervening aquitard. A binary classification of hydraulic conductivity (1 for K_u or K_c and 0 for K') can thus be adopted to reflect the potential for water flow and contaminant transport. In alluvial, sedimentary environments, the aquifer system is highly heterogeneous reflecting the sedimentation and erosion history of the environment, which is in turn influenced very locally by prior features (e.g. topography, stream locations) and globally by historical climatic epochs. The alluvial aquifer system is consequently characterized by lenses and layers of media of rather disparate hydraulic conductivity, and of variable size and geometry. The effective degree of hydraulic interconnection between the upper unconfined aquifer and the deeper confined aquifer depends directly on the lateral extent, location and thickness of the high and low conductivity lenses in the strata loosely referred to as the aquitard. Our particular interest was in using drill log information to develop a preliminary assessment of the potential for such interconnection. Specifically, we were interested in a quantitative assessment of the following issues:
(1) the likelihood that a high conductivity zone may exist at a certain depth in a general area.
(2) the likely depth of the interface between high and low conductivity strata in a general area.
(3) the likelihood that a transition from a low to a high or a high to a low conductivity strata may take place in a certain area.

In this paper, these assessments are made directly from drill log information. Integration of different sources of information (e.g. pump tests and drill logs) may be pursued in the future. Some general comments regarding the attributes of drill log information, the binary classification of data, scale and identifiability issues are offered in this section. A brief review of past related work is also offered.

It is apparent from the list of interests above that our focus is primarily on the process in the vertical. We recognize that groundwater contaminant transport is a three dimensional process. However, the concern for deep aquifer contamination, as well as the nature of the available information and depositional processes motivate us to look for properties of the aquifer system in the vertical across the study area, rather than seeking a general three dimensional reconstruction of strata.

Drill log data are usually rather imprecise. The driller may or may not adhere to Unified Soil Classification System (USCS), and may not be particularly diligent in recording the information. Even if the soil types are properly documented, direct inference of hydraulic conductivities from the drill log is infeasible. Johnson and Dreiss (1989) discuss two possible schemes based on interpretations of the USCS for the binary classification of drill log data into high and low conductivity regions. Their table 1 is reproduced as our table 1. Interpretation (a), highlights layers of low conductivity and emphasizes the separation between the high conductivity layers, while interpretation (b) highlights the spatial structure of zones of high conductivity. Thus the choice of the binary classification scheme of material with hydraulic conductivity ranging on a continuum may emphasize different attributes of the data. The sedimentary process rarely leads to a distinct separation between two layers. The depositional process can be viewed as a time series where changes are for the most part gradual. Relatively rapid depositional epochs may however punctuate such a time series. Such a time series is often nonstationary, with clustering of events, and variation in rates of deposition. These features are consequently imparted to the deposits and hence to the aquifer.

High permeability Indicator=1		Low permeability Indicator=0
	Interpretation (a)	
GW, SW, GM, SM		ML, CL, OL
GP, SP, GC, SC		MH, CH, OH
	Interpretation (b)	
GW, SW		GM, SM, ML, CL, OL
GP, SP		GC, SC, MH, CH, OH

G, gravel; S, sand; M, silt; C, clay; O, organics; W, well graded, (i.e. poorly sorted); P, poorly graded; L/H, Low/High plasticity.

TABLE 1. Indicator values for Alternative Interpretations of the united Soil Classification system.(Johnson and Dreiss (1989))

The lesson from these observations is that even though a binary classification system is adopted here, one has to be cognizant of the fact that the classification is somewhat subjective, transitions between the two states occur smoothly rather than sharply, and that statistically we should expect the lens occurrence process to be nonstationarity in the vertical and perhaps also in the horizontal. The above discussion also suggests that there are likely to be multiple scales of variation in the lens occurrence process, i.e. one could have fine scale (e.g. cm or m) fluctuations between high and low conductivity media superimposed on larger scale variations(e.g. 10 or 100 m)

The drill hole typically samples the aquifer only to a few inches (cm.) in lateral extent. The horizontal information provided is consequently very local and subject to a high degree of variability. Given that the spacing between drill logs may range from a few meters to hundreds of meters, one has to be very cautious in essentially extrapolating aquifer attributes between drill logs. While, in a sense the drill log provides virtually continuous (in the sense that breakpoint information is recorded) sampling in the vertical, drill holes may be arbitrarily scattered over the study area. Note that the data developed by the binary classification scheme reflects a discretization at a thickness commensurate with computational resources available and some assessment of the scale of variation in the vertical that is of interest to the investigator. With regard to spatial interpolation we note that in general one cannot expect to resolve variation at scales finer than the observational scale (in the horizontal, and in the vertical).

Our discussion thus far suggests that an attempt at a recovery of the stochastic structure of variation in the vertical, followed by an interpolation of the parameters of such a model in the horizontal may be superior to a direct attempt to interpolate the binary information in all three dimensions. Lattice models (Cressie, 1991, Chapters 6 and 7) are sometimes used to study spatial variability. We found the majority of such models unattractive for our problem upon comparing the assumptions on stationarity and underlying Gaussian or other pre-specified structure with our conceptual observations above.

A problem rather similar to the one considered in this paper was studied by Johnson and Dreiss (1989). We refer the reader to their paper for other work in the same general area, that is less closely related to the objectives of our paper. They used indicator Kriging to interpret drill log data sets to reconstruct the three dimensional spatial variability of sedimentary environments. Indicator Kriging (Journel, 1983, 1989, Isaaks and Srivastava, 1989) is a stochastic interpolation procedure, that interpolates the binary indicator data to effectively assess the probability of exceedance of the threshold used to originally classify the data into the binary scheme. Kriging entails two primary activities. The first is the inference of an appropriate variogram from the data, and the second is the use of this variogram together with the data to develop a Best Linear Unbiased Estimate (and its mean square error) of the parameter of interest at a given location. Second order stationarity of the increments of the spatial random field is typically assumed. While Kriging is de facto synonymous with geostatistics, critical reviews of its strengths and weaknesses are offered by [Yakowitz, 1985] and by Owosina et al (1992). In the three dimensional spatial estimation situation considered by Johnson and Dreiss (1989) we question (1) the ability to effectively identify the correct variogram from the data, (2) the ability to address the nonstationarity and stochastic anisotropy of the process, (3) the ability to estimate the mean square error (MSE) of estimate given that the variogram is incorrect, and (4) whether the striking differences in sampling density in the vertical and the horizontal are properly accounted for. The variogram is critical for effective Kriging. Johnson and Dreiss report that (1) they used a three dimensional, nested, spherical variogram, (2) the range of the variograms estimated by them for their data was very sensitive to the orientation of the search plane reflecting a marked anisotropy in the data, and (3) the nugget was a relatively large fraction of the sill, suggesting weak observational scale dependence. Issues related to nonstationarity of the field were not addressed by them. From the discussion in Owosina et al (1992) we note that (1) the identifiability of the correct variogram from a set of parametric choices is poor even in the two dimensional, isotropic case, (2) the correct form of the variogram can give the right estimate, but with the wrong mean square error, (3) an incorrect variogram can lead to a total breakdown, and (4) the choices (IRF and Universal Kriging) in the nonstationary case are hard to successfully implement. [Yakowitz, 1985] and Owosina et al (1992) propose Kernel regression methods as a nonparametric alternative to Kriging. We note that Journel (1989) advocates a nonparametric approach (in his view Indicator Kriging instead of Ordinary Kriging) for spatial estimation problems. In this paper a nonparametric estimation framework is developed that addresses (in the context of our objectives) some of the objections raised in this paragraph.

METHODOLOGY

Two separate, nonparametric models are developed and presented. The first model (NMP) is an extension of a nonhomogeneous Markov process based on counting processes pioneered by [Aalen, 1978]. It is used to estimate the intensity of transition between two states (e.g. low to high conductivity) at a given depth and related measures such as the transition probability (integrated intensity over an interval). The second model (NPP) is a two stage model, that treats the occurrence of high conductivity material as a nonhomogeneous Poisson process in the vertical, and then uses a local regression scheme to estimate the rate of the Poisson process at any given depth at a location of interest using the rates estimated at the first stage at each of the drill sites. The formulations (NPP-1) and (NPP-2) are presented for this model. This model may be used to estimate the probability of high conductivity material over a user specified interval in the vertical at any location within the study area. Both models are nonparametric and employ kernel function estimators. They differ in their structure and implementation.

Model NMP

The general approach here is to model the process of transitions between the two states (high and low conductivity) in the vertical as evident from the drill log information. An estimator that smooths the transitions recorded in the drill log counts in the vertical and in the horizontal is developed. Our development of this model has been directly influenced by [Keiding, 1989] who based their work on Danish hospital admission and discharge data on the developments in [Aalen, 1978]. We follow their notation for the most part. Let us presume that we sample each drill log over an interval 0 to Z, at a spacing dz (say 1), thereby creating a data set of length (n*nz), where n is the number of drill logs, and nz (=Z/dz) is the number of observations generated in the vertical. To allow for surface topographic variations, and variability in the depth drilled at each hole, this index set and number of observations nz would need to be properly adjusted.

The two states of the Markov Process are 1 corresponding to high conductivity and 0 corresponding to low conductivity. We are interested in the transitions in the vertical between these two states. To ease the presentation, we shall first consider statistical homogeneity and complete spatial randomness in the horizontal. In this case, the statistical attributes of the aquifer are the same at any location in the study area, and vary only by depth. Depending on the relative observational and study scales, this may not be unreasonable for some situations. Also for non point source contamination of the upper aquifer, one may simply be interested in an areal average probability of high conductivity sections of the aquitard at some depth.

Define $Y_1(z)$ as the total number of drill logs recording state 1 at depth z, and $Y_0(z) = n(z) - Y_1(z)$ as the number of drill logs in state 0, where n(z) is the total number (out of n) of valid drill logs at the depth z. Define counting processes $N_{01}(z)$ and $N_{10}(z)$, that count the number of subscripted transitions in [0,z]. Assuming that $Y_k(z) > 0$ for all z, (i.e. a transition is possible out of state k), one can define the following:

$$A_{kl}(z) \quad = \int_0^z \alpha_{kl}(u)du \tag{1}$$

$$\hat{A}^{kl}_{j}(z) = \int_{0}^{z} dN_{kl}(u) / Y_{k}(u)$$

$$= \sum_{Z_j^{kl} \le z} \frac{m_j^{kl}}{Y_k(Z_j^{kl})} \qquad (2)$$

where $A_{kl}(z)$ is the integrated transition intensity from state k to state l, from 0 to z, $\alpha_{kl}(z)$ is the k-l transition intensity at depth z, Z_j^{kl} is the j^{th} depth at which a k-l transition takes place, and m_j^{kl} is the number of such transitions (generally 1).

If Y_k (z) is zero for some z, no drill logs are in state k at depth z, and obviously a transition out of state k at depth is not possible. Also, the transition probability into state k at depth z is zero. Only transitions l-k and l-l need to be considered at depth z in this situation.

The transition intensities $\alpha_{kl}(z)$ are now estimated by kernel smoothing as follows:

$$\hat{\alpha}_{kl}(z) = \frac{1}{h} \int_{0}^{Z} K\left(\frac{z-u}{h}\right) d\hat{A}_{kl}(u) = \sum_{Z_j^{kl} \le z} \frac{1}{h} K\left(\frac{z-Z_j^{kl}}{h}\right) \frac{m_j^{kl}}{Y_k(Z_j^{kl})} \qquad (3)$$

where h is a bandwidth, K(u) is a kernel function, and $u=(z-Z_j^{kl})/h$.

The kernel function is a symmetric probability density function, that has the role of a weight function in this context. The transition intensity at z is estimated through a weighted average of the relative number of transitions in a neighborhood of z, that is determined by the bandwidth h. As the bandwidth increases, one averages over a larger interval and the variance of the estimate decreases, while its bias increases. Silverman (1986) presents a very accessible monograph that focuses on kernel density estimation and on problems similar to the one posed here. Silverman shows that the choice of a kernel function is not critical in terms of the mean square error of estimate. The choice of a kernel function can thus be based on other criteria such as computational and smoothness considerations. It is desirable to use kernels with compact support (i.e. defined on a closed interval, e.g. [-1,1]) to minimize boundary effects in our situation. Boundary effects arise because at the boundaries ($0 \le z \le h$, and $Z\text{-}h \le z \le h$), we do not have a full complement of observations on both sides of the point of estimate. This leads to a biased estimate unless the symmetric kernel used in the interior is appropriately modified. The kernel used in the interior by [Keiding, 1989] is the mean square error (MSE) optimal kernel given by Epanechnikov:

$$K(u) = 0.75\ (1 - u^2) \quad \text{if } |u| \leq 1, 0 \text{ else} \tag{4}$$

This kernel has discontinuous derivatives at the ends of its support, leading to discontinuous derivatives of the estimated function. One can alternately use the bisquare kernel to address this issue, without much increase in the MSE. This is given as :

$$K(u) = \frac{15}{16}(1-u^2)^2 \tag{5}$$

[Keiding, 1989] used a method to adjust for boundary bias that is due to [Gasser, 1979]. However, one can use boundary kernels given by [Müller, 1991], or use reflection to account for boundary effects. Reflection augments the data set by reflecting points across each boundary (e.g. z_i is augmented by $-z_i$ - reflecting across the boundary at 0). The real and the reflected data are then used to form the estimate, with modifications to recognize the increased sample and domain size. Reflection is appropriate if the derivative of the target function is zero at the boundary across which reflection takes place. Müller's (1991) boundary kernels are derived from asymptotic MSE minimization considering the asymmetric neighborhood of estimation. Müller's boundary kernels corresponding to the two kernels given above are :

Left Boundary Epanechnikove kernel

$$K_{left}(q,u)=6(1+u)(q-u)\frac{1}{(1+q)^3}\left\{1+5\left(\frac{1-q}{1+q}\right)^2+10\frac{1-q}{(1+q)^2}u\right\} \tag{6}$$

Left Boundary Bisquare kernel

$$K_{left}(q,u)=30(1+u)^2(q-u)^2\frac{1}{(1+q)^5}\left\{1+7\left(\frac{1-q}{1+q}\right)^2+14\frac{1-q}{(1+q)^2}u\right\} \tag{7}$$

Where:

$$u = \frac{z-z_i}{b}$$

$$q = \frac{z}{b}$$

Right boundary kernels are defined as:

$$K_{right}(q,u)=K_{left}(q,-u) \tag{8}$$

In general the reduction in boundary bias by modifications of the estimator is typically offset by an increase in variance, and one should be cautious in interpreting results in the boundary region.

The bandwidth h may be selected by minimizing the Mean Integrated Square Error (MISE) through cross validation (see Silverman 1986, Ramlau-Hansen, 1983 and [Patil, 1992]) :

$$\mathrm{MISE} = \mathrm{E} \int_{z_0}^{z_1} [\hat{\alpha}(z) - \alpha(z)]^2 \, dz \tag{9}$$

over an interval $[z_0, z_1]$ that excludes the boundary region. The MISE is estimated through cross validation as:

$$\mathrm{CV}(h) = \int \hat{\alpha}^2(z) \, du - 2n \sum_i \hat{\alpha}_{-i}(z_i) \tag{10}$$

where:

$$\int \hat{\alpha}^2(u) \, du = h^{-1} \sum_i \sum_j K^{(2)}\left(\frac{Z_i^{kl} - Z_j^{kl}}{h}\right) \bullet \frac{m_j^{kl}}{Y_k(Z_j^{kl})} \bullet \frac{m_i^{kl}}{Y_k(Z_i^{kl})} \tag{11}$$

where:

$$K^{(2)}\left(\frac{Z_i^{kl} - Z_j^{kl}}{h}\right) = K\left(\frac{Z_i^{kl} - z}{h}\right) \bullet K\left(\frac{z - Z_j^{kl}}{h}\right) \tag{12}$$

$$\sum_i \hat{\alpha}_{-i}(z_i) = \frac{1}{h(n-1)} \sum_{i=1}^{n} \sum_{j=1 \neq i}^{n} K\left(\frac{Z_i^{kl} - Z_j^{kl}}{h}\right) \bullet \frac{m^2(Z_i^{kl})}{Y_k^2(Z_i^{kl})} \tag{13}$$

This yields a global or fixed bandwidth for the estimator. One can select local or variable bandwidths that depend on the intensity and location as well (see for example, [Müller and Wang, 1990], and [Abramson, 1982]).

The transition probabilities can be evaluated by using the basic idea that follows from the state transition probability for a Markov chain is that:

Let $Y_k(z_i)$ is the total number of drill logs recording state k at depth z_1

Let $M_{kl}(z_i)$ is the number of drill logs recording state k at depth z_1 that changes to state l at depth z_2, then a one step transition can be described as:

$P_{kl}(z_i, z_{i+1}) = M_{kl}(z_i) / Y_k(z_i)$

A two step-transition is described as:

$P_{kl}(z_1, z_3)$ = probability (in state l at depth z_{i+3} / in state k at depth z_i)

$= P_{cc}(z_i, z_{i+1}) * P_{cs}(z_{i+1}, z_{i+2}) + P_{cs}(z_i, z_{i+1}) * P_{ss}(z_{i+1}, z_{i+2})$

Generally , two state n-step transition probability matrix can be described by the following:

$$\hat{P}(Z_1, Z_n) = \prod_{i=1}^{n} (I+\Delta\hat{A}(Z_i)) \tag{14}$$

where

$$I+\Delta\hat{A}(z_i) = \left| \begin{matrix} 1-\dfrac{M_i^{kl}}{Y_k(z_i)} & \dfrac{M_i^{kl}}{Y_k(z_i)} \\ \dfrac{M_i^{lk}}{Y_l(z_i)} & 1-\dfrac{M_i^{lk}}{Y_l(z_i)} \end{matrix} \right| \tag{15}$$

Example applications and implementation details of the above approach are presented in the section on Model Applications.

The one dimensional model of [Keiding, 1989] that has been presented so far can be extended to the spatial case in the following manner. Conceptually, the model thus far does a simple average of all transitions in the study area and a weighted average or smoothing of transition counts in the neighborhood of z in the vertical. The estimator for transition intensity is thus local in z but just a global average in x,y. The generalization we consider still focuses on transitions in z, but allows the transition intensity to vary over the location in the x,y plane. We employ the kernel function again to develop a local average at (x,y) in a manner similar to the smoothing in z. Conceptually, this boils down to defining an area in the neighborhood of the point (x,y) at which the estimate is needed and using that data to develop the local estimate. We can estimate the integrated transition intensity as :

$$\begin{aligned} \hat{A}^{kl}_{j,x}(z) &= \sum_{i=1}^{n} \int_{\Omega_i} \int_{0}^{z} dN_{kl}(u) / Y_k(u)\, K_x\left(\left\| \frac{x - x_i}{h_x} \right\| \right) dx \\ &= \sum_{i=1}^{n} \sum_{Z_j^{kl} \le z} \int_{\Omega_i} \frac{m^{kl}_{j,i}}{\beta_i Y_k(Z_j^{kl})} K_x\left(\left\| \frac{x - x_i}{h_x} \right\| \right) dx \\ &= \sum_{Z_j^{kl} \le z} \sum_{i=1}^{n} \frac{m^{kl}_{j,i}}{\beta_i Y_k(Z_j^{kl})} w(x, x_i, h_x) \end{aligned} \tag{16}$$

where $K_x(u)$ is a bivariate kernel function defined over (x,y), Ω_i is an area (e.g., obtained by

Dirichelet tessellation) around the drill site $\mathbf{x_i}$ that contains only that location, $\mathbf{x}$ represents a vector with coordinates as (x,y), $\mathbf{h_x}$ is a bivariate bandwidth vector, β_i is a factor that ensures that the count $Y_k(z)$ is based on the drill sites that are effectively considered in the estimation, $\mathbf{u}$ is given by a modified (scaled by each coordinate bandwidth) Euclidean distance between $\mathbf{x}$ and a drill site $\mathbf{x_i}$, and $w(\mathbf{x}, \mathbf{x_i}, \mathbf{h_x})$ is a weight function obtained by integrating the Kernel over the domain Ω_i .

The bivariate kernel $K_x(u)$ may be taken to be a bivariate Gaussian kernel with covariance matrix $\mathbf{S_x}$ chosen to be the same as the sample covariance matrix for the location indices (x,y). This choice recognizes the orientation and relative spacing of the sampling locations. In this case the bandwidths are specified through a matrix $\mathbf{h_x}$ that is proportional to the covariance matrix $\mathbf{S_x}$ (i.e. $\mathbf{h_x} = h_0\mathbf{S_x}$). Alternately , a bivariate kernel can be developed as the product of two univariate Epanechnikov or Bisquare kernels with different bandwidths in each direction. Procedures for the selection of the bandwidth in this context may be found in [Müller, 1987; Müller, 1988].

The use of the kernel function in the above expression serves to localize the estimate to the neighborhood of the point at which the estimate is desired, since it vanishes beyond a bandwidth of the point for a kernel of compact support. The approach here is kernel regression using the Gasser-Müller estimator. For a discussion of this and other kernel regressors see Owosina et al (1992). Alternate representations for the above equation that do not entail the triangulation needed for defining the area Ω_i can also be developed. Essentially these lead to a modification of the manner in which the weight function $w(\mathbf{x}, \mathbf{x_i}, \mathbf{h_x})$ is specified. Following the development above it is easy to see that the transition intensity and transition probabilities at a point x,y may be given as in terms of $w(\mathbf{x}, \mathbf{x_i}, \mathbf{h_x})$:

$$\hat{\alpha}^{x}_{kl}(z) = \sum_{Z^{kl}_{j} \le z} \sum_{i=1}^{n} \frac{1}{h_z} K_z\left(\frac{z - Z^{kl}_{j}}{h_z}\right) \frac{m^{kl}_{j}}{Y_k(Z^{kl}_{j})} w(\mathbf{x}, \mathbf{x}_i, \mathbf{h}_x) \tag{17}$$

give eqns

The bandwidth selected would be different depending on whether, A_{kl} or α_{kl} is to be estimated. However, the criteria for bandwidth selection is the same in each case. Where the bandwidths **h** in (x,y,z) are estimated using cross validation, the computational burden can be significant. We have not yet pursued such an implementation. We note that the model considers transition intensities through a smoothed counting process in the vertical, and that a Markovian dependence structure is considered in the vertical. The process in the horizontal considers just a local (in the horizontal) estimation of this Markovian dependence structure. Thus the model presented is still a one dimensional model, with areally heterogeneous parameters. The scale of the horizontal average is dictated by the bandwidth $\mathbf{h_x}$. However, note that since the kernel function $K_\mathbf{x}(.)$ has mass concentrated at the origin, the effective scale of such an average is considerable smaller, and as $n \to \infty$, the areal average approaches the point value. We mention this to suggest that the estimator proposed is asymptotically consistent. An approach to the above process that is computationally less demanding is to define the weight function $w(\mathbf{x}, \mathbf{x_i}, \mathbf{h_x})$ as 1 if $\mathbf{x_i}$ belongs to the p^{th} nearest neighborhood of x and 0 else. The p^{th} nearest neighborhood of x is defined as the set of p closest $\mathbf{x_i}$ to $\mathbf{x}$, where closeness is measured in terms of a distance metric (e.g. $d(\mathbf{x},\mathbf{x_i}) = \Sigma_m |x_m - x_{m,i}|$, where m is a co-ordinate direction). The number of nearest neighbors may be chosen by minimizing a MISE criterion, or more simply using a heuristic such as $p = \sqrt{n}$. The number of neighbors p plays the role of the bandwidth in the kernel smoother (indeed this procedure is equivalent to using a rectangular kernel with p=2nh). A small p leads to higher variance and a large p to larger bias. However, in our experience there is a range of p values for which the estimator is relatively insensitive to the actual choice of p. The above heuristic for picking p works reasonably well for the most part. Notice that this scheme leads to a direct local average using moving and different sized (in **x** space) neighborhoods.

In the presentation above we have considered estimation of transition intensities for $0 \le z \le Z$, with transitions considered only in the direction 0→Z. This follows from the perceptual bias that we are interested in what happens as we go deeper into the aquifer. Clearly transitions in the direction Z→0 may be considered just as easily. This would make more sense if for example one is interested in the time series of sedimentation as evidenced by the depositional sequence.

Model NPP

An alternate formulation based on treating the occurrence of high conductivity layers in the vertical as a nonhomogeneous Poisson Process is now presented with reference to Figure 2. The general structure of Model NPP is to first estimate the rate $\lambda(z)$ of an inhomogeneous Poisson process as a function of depth z_{ji} at each drill log site $\mathbf{x_i}$. The next step is to use the data set

($\hat{\lambda}(z_{ji})$, z_{ji} ,x_i, j=1...nc$_i$, i=1..n) to estimate $\hat{\lambda}(z)$ at any arbitrary (x,z) using nonparametric regression. The inherent assumption is that the drill log sequence at all the sites is the consequence of a large scale (larger than study area) depositional time sequence that is nonstationary in time (hence the variable Poisson rate in the vertical), with local (areal) variability in the sequence captured by each drill log. The nonparametric regression of the rate parameter over the domain then recognizes any structure (e.g.that imparted by existing topography at the time of deposition) in the spatial variation of the depositional process. Note that all $\hat{\lambda}(z_{ji})$ values are used to estimate $\hat{\lambda}(z)$ for any z at a point x at which an estimate is desired. However, information from neighboring sites is not used to estimate $\hat{\lambda}(z_j)$ at a drill site in the first stage of the process. The two phase approach, and the consideration of only a vertical rate is necessitated by the fact that we have continuous sampling only in the vertical at a finite set of locations, and hence the there is inadequate information for characterizing a spatial point process in the horizontal direction.

We begin our discussion with a brief review of point processes that is relevant for our presentation. The point process formulation relative to our problem is then presented. A nonparametric technique for estimation of the rate parameter at locations other than the drill log is then briefly discussed.

The Poisson process is the cornerstone on which most of the theory of spatial point processes is built. A spatial point process is a stochastic mechanism which generates a countable set of events or points in an interval, plane or volume. While a three dimensional situation is of interest to us, the sampling of the process is really continuous only in the vertical at each drill log. Since we can only observe the point process at these discrete locations, and because our interest is in the migration of contaminants in the vertical, the spatial point process model of interest is constructed only in the vertical and then its parameters are interpolated in the horizontal.

The homogeneous Poisson Process represents a situation in which occurrences of the event of interest are completely random, independent and uniformly distributed over the space of interest. Given the discussion earlier on the nature of sedimentary processes, it is unlikely that a homogeneous Poisson process is an appropriate descriptor for our problem, since it does not admit clustering of events, and assumes a constant rate of event occurrence over the domain. It is however, interesting to note from Cressie (1991), that finite samples from a homogeneous Poisson process can exhibit apparent clustering, since the probability density of the first nearest neighbor distances for the Poisson process is the Chi-Square, which has a substantial mass near zero. We shall focus on the development of an inhomogeneous or doubly stochastic Poisson process that admits a variable rate of event occurrence. Cox and Isham (1980) define such processes and discuss parametric specifications of the rate of such a process. They point out that parameter estimation and description of such processes is very difficult in a finite domain. However, Diggle (1985), Diggle and Marron (1988), and Solow (1990) present theoretical developments and applications of effective nonparametric methods for describing the rate of the inhomogeneous Poisson process using a finite sample. This is the approach we pursue in this paper. Rather than presenting an extensive review of stochastic processes we refer the reader to Karr (1986), Cox and Isham (1980) and Diggle (1983) for background on point processes and Poisson processes.

At first glance, the rationale for a point process model for alternating layers of high and low conductivity from a continuous drill log may not be obvious, since the events of interest are intervals and not points. However, we see that actually two constructions are possible.

Model NPP-1

The *first* construction termed *NPP-1*, is formed by focusing on the interface between each layer or the transition between layers. The description of the process is accomplished by looking at the count of such transition over the vertical, and by noting the type of the layer at the surface (i.e. the first layer). If the first layer is of type 1, it is obvious that the layer after an odd number of transitions is of type 0, and after an even number of transitions is of type 1. If a sequence of transitions occurs at distances z_1, z_2, ...z_i... from the surface, and one knows the type of each such layer, clearly $(z_i - z_{i-1})$ gives the thickness and type of the layer preceding the i^{th} transition. We notice the parallel with the renewal model construction in the previous section. In that case we focused on the probability or intensity of a transition at a particular depth. In that case we estimated these quantities by averaging across drill logs and using kernel smoothing in the vertical for a drill log. In this case we shall estimate the transition intensity by counting the transitions in the vertical for each drill log, and using kernel smoothing as before. The averaging (or in this case, variation) across drill logs is done in a separate step. Thus, the two models are philosophically very similar, but are cast in somewhat different estimation frameworks. A discretization of the vertical is employed in the renewal model, but is not needed in model NPP-1. A comparative discussion of the attributes and performance of the two approaches is presented in the final section of this paper.

At a drill log, the intensity function of the nonhomogeneous Poisson process is defined as (Diggle , 1983):

$$\lambda(z) = \lim_{dz \to 0} \frac{E[N(dz)]}{dz}$$

where N(dz) is the number of events in an interval dz.

For a homogeneous Poisson Process, $\lambda(z)$ is a constant rate λ defined as the average number of transitions per unit distance. For the inhomogeneous Poisson process following Diggle (1985) we can estimate $\lambda(z)$ through kernel smoothing as follows:

$$\hat{\lambda}_i(z) = \sum_{j=1}^{nt_i} \frac{1}{h_i} K\left(\frac{z-z_{ji}}{h_i}\right) \tag{18}$$

with z_{ij} defined as a transition depth, h_i as a bandwidth, and K(u) as a kernel function, and nt_i is the total number of transitions at drill site i.

Equation 18 above can be readily interpreted as a weighted average of the transition rate over a distance 2h centered about the point at which the estimate is needed. For a constant underlying rate the bandwidth h should tend to infinity, while for a high degree of fluctuation in the rate, h should tend to zero. Practically, for a finite domain, our ability to recover the underlying rate depends on the length, (Z_i), of the domain and the number of transitions (nt_i) encountered (nt_i large relative to Z_i is the most identifiable situation). Clearly if we have only one transition from high to low conductivity layers, or none at a drill log, no estimation is possible with this model.

Diggle and Marron (1988) show the equivalence between kernel intensity and kernel probability density estimation. They show that the same bandwidth is optimal for both under a Mean Square Error criterion. This makes the vast literature of bandwidth estimation for kernel density estimates accessible in this case. We considered Maximum Likelihood Cross Validation (MLCV), Least Squares Cross Validation (LSCV) and the Sheather and Jones (SJ) methods for bandwidth selection. The two cross validation methods are the traditional favorites. However, they suffer from a tendency to undersmooth and from a high degree of variability leading to an MISE convergence rate of $O(n^{-1/10})$ which is substantially worse than the $O(n^{-1/2})$ rate promised by the SJ algorithm. A comparative review of these algorithms may be found in Owosina et al (1992).

A number of bandwidth selection methods have historically been used to guide kernel density estimates. Cross validation methods (Maximum Likelihood and Least Squares Cross Validation, see Silverman (1986), section 3.4) were popular in the 1980's. However, these methods are prone to undersmoothing, and a suffer from high degree of sampling variability, leading to rather poor MSE convergence rates ($O(n^{-1/10})$ for the kernel estimator (see, Hall and Marron (1987)). A resurgence of interest in "plug in" or recursive estimators for h, that use data driven kernel estimates of the probability density f(x) and f"(x) for determining the optimal bandwidth has followed. Such methods were originally proposed by Woodroofe (1970), and pursued by Scott et al (1977), Scott and Factor (1981) and Sheather (1983, 1986). Improvements by Park and Marron (1990), and Sheather and Jones (1991) among others have lent stability to these methods and have led to a MISE convergence rate of h_{opt} of the order of $n^{-5/14}$, as well as a reduction in the size of the constants associated with this rate. In our tests, the Sheather and Jones algorithm generally outperformed the other methods, with a tendency to slightly oversmooth, rather than undersmooth. The latter is a positive aspect, since it reduces the possibility of observing spurious structure in the estimated p.d.f. and leads to reduced variance of estimate. Note from 3.3 that increasing h, increases the bias, but reduces the variance of $f_n(x)$.

The MISE of the fixed kernel estimator and the corresponding optimal bandwidth are given by Silverman (1986), section 3.3 as:

$$\mathrm{MISE}(f_n(z)) = (nh)^{-1} R(K) + 0.25\, h^4 \sigma_K^2 R(f") \tag{19}$$

$$h_{opt} = \{R(K)/(\sigma_K^2 R(f"))\}^{1/5} n^{-1/5} \tag{20}$$

where $R(g) = \int g^2(z)dz$ and $\sigma_g^2 = \int z^2 g(z)dz$

The terms R(K) and σ_K^2 depend only on the known kernel K(z). Consequently, the unknown term in (19) and (20) is R(f"). Sheather and Jones develop a recursive kernel estimate $S(\alpha)$ for R(f") as :

$$S(\alpha) = \{n(n-1)\}^{-1}\alpha^{-5}\sum_{i=1}^{n}\sum_{j=1}^{n} K^{iv}((z-Z_i)/\alpha) \qquad (21)$$

where α is a bandwidth (not equal to h), and $K^{iv}(.)$ is a fourth derivative kernel.

The bandwidth α needs to be estimated for this estimate. A similar procedure can be reapplied for the purpose. Sheather and Jones relate the bandwidths α and h as

$$\alpha(h) = 1.357\ \{S(a)/T(b)\}^{1/7} h^{5/7} \qquad (22)$$

where

$$T(b) = -\ \{n(n-1)\}^{-1} b^{-7}\sum_{i=1}^{n}\sum_{j=1}^{n} K^{vi}((Z_i-Z_j)/b) \qquad (23)$$

$$a=0.92\lambda n^{-1/7} \text{ and } b=0.912\lambda n^{-1/9} \qquad (24)$$

K^{vi} is the sixth derivative kernel, and λ is the sample interquartile range $(Z_{0.75} - Z_{0.25})$, T(b) is an estimate of R(f''') and a, b are bandwidths that are evaluated with reference to a Normal distribution for the derivative kernels considered.

Note that relatively crude estimates (with reference to a known distribution) of the bandwidths used in estimating R(f'') and R(f''') suffice given that the dependence of the MISE expression (20) on these expressions is successively weaker (note the exponents). The optimal bandwidth h_{opt} is now evaluated by computing a and b from the data, evaluating S(a) and T(b), and substituting the expression 23 into 22, and 22 into 21. This leads to a nonlinear expression in terms of h, which is solved using the Newton Raphson method. Sheather and Jones specify the Normal kernel for K(.) and evaluate the derivative kernels as the appropriate derivatives of this kernel.

From 21 observe that knowing the optimal bandwidth h_N for the Normal kernel, the optimal bandwidth h_K for a kernel different from the Normal kernel can be readily evaluated as:

$$h_K = \{(R(K)\sigma_N^2)/(\sigma_K^2 R(N))\}^{1/5}\ h_N \qquad (25)$$

where "N" identifies the Normal kernel, and "K" the kernel of interest.

As in the case of the renewal model the kernel estimator is plagued by boundary bias problems. Solutions similar to those adopted in the previous section are necessitated. Diggle (1985) suggests essentially normalizing the estimator to reflect the area of the kernels "lost" outside the boundaries. Solow (1990) and Diggle and Marron (1988) advocate reflection. We considered reflection as well as the Müller (1991) boundary kernels.

Once the intensity function has been estimated as described above, one can interpolate or regress it to other lateral locations. This scheme is described at the end of this section for both formulations considered. At this stage we point out, that additional information is also needed with this formulation to complete the description. Let us say that we have used the the n drill logs to estimate the intensity function of transitions at an arbitrary point **x** in the spatial domain. One still needs to specify the type of the first layer in the vertical sequence at **x**. This can only be done probabilistically. One possibility for assessing the probability that the top layer is of a certain type is to consider a nonparametric regression of the data set with dependent variable defined through the binary process b_i (1= high, 0=low conductivity layer), and the independent variable as the drill site locations $\mathbf{x_i}$. The estimate b(**x**) then gives the probability that at location **x**, the top layer is of type 1. The transition sequence defined at site **x** is then valid with probability b(**x**). The sequence is reversed with probability 1-b(**x**).

Model NPP-2

Cox and Isham (1980), point out that a special form of the doubly stochastic or inhomogeneous Poisson process is obtained by considering that the point process of interest alternates stochastically between a rate of 1 and a rate of 0, i.e a section where the rate is 1, and one with no points. Their example is of a component of some system that alternates between periods of use and idleness. The parallel of this example with our situation is obvious. To advance an estimation framework, consider as in the renewal model a discretization of the drill log over an interval 0 to Z, at a spacing dz (say 1), thereby creating a data set of length (n*nz), where n is the number of drill logs, and nz (=Z/dz) is the number of observations generated in the vertical. Now we have nz binary (1 or 0) values at each drill log representing whether that dz thickness is of type 1 or 0. Convert this into a data set recorded as z_1, z_2, ...z_i..., where z_i records the depth of the i^{th} location that was coded 1. Essentially we are defining each occurrence of a type 1 element of thickness dz as an event. Note that in reality dz is an interval and not a point. This scheme shall thus have the unfortunate attribute of being dependent on the discretization adopted in the vertical, with equivalence to a formal point process only in the case where dz approaches 0. However, note that the special construction of this point process, considers the rate to simply alternate between 0 and 1. Thus, if we classify dz as type 1, what we are saying is that all points in dz are of type 1, i.e. the rate is 1, which is precisely what we get if we use the interval dz. The assumption of concern in that case is whether each such interval dz is homogeneous. Subject to the accuracy of the drill log, we consequently need to ensure that the dz selected is somewhat smaller than the thinnest uniquely identifiable layer. Noting the subjectivity involved in interpreting drill logs this may not be a serious issue. We also note that in contrast to Cox and Isham's example our construction shall allow rates between 0 and 1 in areas of transition from one type of layer ot another. The transitions considered are smooth rather than abrupt. We feel that this is appropriate given that there is often a gradation of materials rather than abrupt changes, materials of mixed grades are being classified into only two distinct classes, and a discretization is adopted for the classification.

The NPP-2 intensity function is estimated in exactly the same manner as that for NPP-1, using equation 18. The difference is in the interpretation of the two intensity functions. The NPP-1 intensity function gives the rate of transitions in the vertical between the two types of media considered. The NPP-2 intensity function gives the rate at which type 1 material is observed. Its complement gives the rate at which type 0 material is observed. NPP-1 suffers from low sample size if a drill log has very few transitions. NPP-2 uses the same information always through nz values, and does not suffer the same malady. Also, for NPP-2 one does not need to estimate the probability of the vertical sequence at an unsampled point **x**, as was the case for NPP-1. Note that the differences between the models presented are really in terms of the estimation framework and not so much for the conceptual representation, which considers a stochastic structure in the vertical and a deterministic structure in the horizontal.

Spatial Extrapolation of Drill Log Intensity Function:

The situation of interest for spatial interpolation of the rate parameter estimated at each drill log is illustrated in Figure 3. The data available at this stage is the Poisson rate evaluated at a number of elevations at each drill log from the previous step. A finer discretization may be employed at drill logs which exhibit a high degree of variability than in those that are essentially all sand or clay. If contours of $\lambda(z)$ are desired, estimates may be considered at the nodes of a 3 dimensional grid as shown in Figure 3. Such estimates are developed by effectively considering weighted spatial moving averages of the drill log estimates through nonparametric regression.

Cleveland et al (1988) and Cleveland and Devlin (1988) developed weighted locally linear and quadratic regression. The estimator proposed by them considers local linear or quadratic fits using k nearest neighbors of the point at which the estimate is desired. The idea is related to the developments in Stone (1975) and Cleveland (1979).

Cleveland et al (1988) consider the standard multivariate, nonparametric regression situation defined through the general model

$$y_i = g(x_i) + \varepsilon_i \qquad (26)$$

where g(**x**) is the regression function or conditional expectation, and ε_i is assumed to be a normally distributed residual.

An estimate $\hat{g}$(x) is developed nonparametrically (in the global sense) by a weighted local linear or quadratic least squares regression in the neighborhood of **x** defined by its k nearest neighbors. The locally quadratic model fitted locally is :

$$\hat{g}(\mathbf{x}) = \mathbf{T}\,\hat{\beta} \qquad (27)$$

where $\hat{\beta}$ is a $(1+p+p^2)$ x 1 coefficient vector, and **T** is an n x $(1+p+p^2)$ matrix with entries given by:

$$T = \begin{bmatrix} 1 & z(1)_1 & \dots\, z(p)_1 & z(1)_1^2 & \dots & z(j)_1 z(k)_1 & \dots\dots & z(p)_1^2 \\ \dots\,.. & & ..\;.. & .. & \dots & \dots & \dots\dots\;\dots & \\ 1 & z(1)_n & \dots\, z(p)_n & z(1)_n^2 & \dots & z(j)_n z(k)_n & \dots\dots & z(p)_n^2 \end{bmatrix} \qquad (28)$$

For the locally linear model, the last p^2 coefficients in $\hat{\beta}$, corresponding to the quadratic terms, as well as the corresponding elements of **T** are dropped. Further note that only the subset of **T**, that corresponds to the $\mathbf{x_i}$ that are the k-nearest neighbors of **x** is used for the estimation of $\hat{\beta}$. The coefficients $\hat{\beta}$ are determined as the solution to a weighted least squares problem defined as:

$$\min_{\hat{\beta}} \sum_{i \subset K(z)} w_i(z) \left(y_i - \hat{g}(Z_i) \right)^2 \qquad (29)$$

where K(**x**) is the index set of the $\mathbf{x_i}$ that are the k nearest neighbors of **x**, and w_i(**x**) is a weight function defined as :

$$w_i(z) = W\left(\frac{\rho(z, Z_i)}{d_k(z)}\right) \tag{30}$$

where ρ(.) is the Euclidean distance function, $d_k(x)$ is the Euclidean distance from **x** to its k^{th} nearest neighbor in $\mathbf{x}_i$, and W(u) is a tricubic weight function given as:

$$W(u) = \begin{matrix} (1-u^3)^3 & 0 \le u \le 1 \\ 0 & \text{else} \end{matrix} \tag{31}$$

In summary, Cleveland et al (1988) propose a weighted local linear or quadratic regression, on the k nearest neighbors of the point of interest, with a tricubic weight function applied to relative distance of the nearest neighbors from the point. The number of nearest neighbors, k, acts as a smoothing parameter. As k increases, bias increases, but variance decreases. Also, the locally linear model (r=1) leads to higher bias, and lower variance relative to the locally quadratic model (r=2), since there is a lower order approximation, but the number of degrees of freedom of the fit increases. Thus the fit determined by Loess depends on the choice of the number (k) of nearest neighbors and the order (r) of the fit. Cleveland et al (1988) propose a graphical method, based on analyzing an M-plot for the selection of both these parameters.

Loess is a linear estimator, i.e. it depends on $\mathbf{x}_i$, W, ρ, r and k, but not on y_i. The dependence is expressed as:

$$\hat{g}(z) = \sum_{i=1}^{n} l_i(z) y_i \tag{32}$$

The statistic of interest is a scaled version of the mean square error given as:

$$M_{rk} = \frac{\left[E\sum_{i=1}^{n} \left(\hat{g}_{rk}(Z_i) - g(Z_i) \right)^2 \right]}{\hat{\sigma}^2} \tag{33}$$

where E denotes an expectation and $\hat{\sigma}^2$ is the variance estimated as

$$\hat{\sigma}^2 = \hat{\varepsilon}^T \hat{\varepsilon}\, \partial_1 \tag{34}$$

with

$$\partial_1 = \text{trace}\left[(\mathbf{I} - \mathbf{L})(\mathbf{I} - \mathbf{L})^T\right]$$

$$\hat{\mathbf{y}} = \mathbf{L}\mathbf{y}$$

$$\partial_2 = \text{trace}\left[(\mathbf{I} - \mathbf{L})(\mathbf{I} - \mathbf{L})^T\right]^2 \tag{35}$$

Cleveland et al (1988) note that (33) can be written in terms of a bias and a variance contribution following standard arguments, as:

$$M_{rk} = B_{rk} + V_{rk}$$

$$= \frac{\sum_{i=1}^{n}(E\,\hat{g}_{rk}(x_i) - g(x_i))^2}{\hat{\sigma}^2} + \text{trace}\,[L_{rk}L_{rk}^T] \tag{36}$$

Under the ***assumptions*** that Loess is unbiased and the residuals are normally distributed, the approximate distribution of Sqrt(M_{rk}) is obtained as a t distribution with $\partial_1{}^2/\partial_2$ degrees of freedom. Cleveland et al (1988) propose a plot of M_{rk} versus V_{rk}, to choose r and k. Confidence limits on M_{rk} are also plotted under the assumption that the estimator is unbiased (i.e. $B_{rk} = 0$). The plot thus allows us to examine the relative role of bias and variance in the fit, where V_{rk} may also be interpreted as the equivalent degrees of freedom of the fit. Cleveland et al (1988) suggest a visual assessment and choice of values of r,k from which the bias or the variance component increase rapidly. They are opposed to choosing r,k on the basis of simply minimizing M_{rk} or other formal risk criteria.

The most clever aspect of Loess is perhaps the Cleveland et al implementation that aims to reduce the computational burden drastically. They partition the space of independent variables using an algorithm based on p-d trees due to Friedman et al (1977), and compute the nonparametric regression surface only at the vertices of the tree. Blending function, cubic interpolation (due to Barnhill, 1977) is then used to interpolate function values at other points from these vertices. We find this to be an aggressive, but potentially useful strategy that could be used for other estimators (e.g. the multivariate kernel) as well. The p-d tree is a data structure for hierarchical partitioning. It subdivides the sample space by successive cuts orthogonal to the co-ordinate axes in p dimensions. The resulting vertices are organized into a binary tree.

The Loess algorithm implemented by Cleveland et al (1988) may be summarized as:

(1) Scale each original independent variable by dividing by an appropriate scale parameter. Rotate the coordinate axes to align with the principal components of variation (This is similar to what is achieved with the kernel estimator based on (30)).

(2) Identify direction j (of **x**) as the one with the maximum spread.

(3) Identify median$_j$ as μ_j, and divide the cell in half at μ_j.

(4) Recursively apply steps 2 and 3 to each subcell. If a cell is too small, or has too few points (as specified on user input), leave it alone.

(5) Identify and store resulting vertices in a tree.

(6) Identify k nearest neighbors of each vertex, and get the Loess estimate (r=1 or 2). The nearest neighbor information is passed to adjacent vertices to speed identification of nearest neighbors.

(7) Use product tensor interpolation using the vertices of a cell to estimate a value at a point **x** inside a cell. They use cubic interpolation with cross terms set to zero.

APPLICATIONS

The models formulated were applied to data from the Salt Lake Valley, provided by Michele Lemieux, State of Utah, Department of Natural Resources, Division of water rights. Of the three formulations discussed in the previous section, the one of greatest interest because of its ready interpretability and breadth of insight is NPP-2. Results from this application are presented and discussed in this section.

Preliminaries:

Each of the 40 drill logs was first processed according to the classification scheme "b" presented in Table 1. This scheme was chosen because it highlights the occurrence of media of high conductivity. This may be the primary factor of concern in determining the potential for cross migration of contaminants from the upper to the lower aquifer. A discretization of 1 ft (.3048 m) in the vertical was used to develop the classification at each drill log. It is desirable to choose a discretization level for classification that is of the same order (a little smaller) than the thickness of the thinnest layers reported in the drill logs across the site. This enables one to work at an interval resolution for the Poisson Process that is consistent with the effective resolution of drill log interpretation. A 1 or 0 is recorded for each interval, corresponding to high or low conductivity material for every one foot over that interval. Surface elevation information was not available to us. However, the general area at the site has low topographic relief, and we assumed a constant surface elevation over the site.

Phase I

The intensity of sand occurrence in the vertical was estimated using the phase I procedures described for model NPP-2. Estimates of the intensity λ(z) were made over a discretization of 1 ft (.3048 m) in the vertical.The Maximum Likelihood Cross Validation method was used for the band width optimization. Four sample profiles, drill logs 1, 2, 9, and 35 are presented here through Figures 4 to 7. Note that the bandwidth is larger for drill logs that have fewer layers and are more homogeneous, as expected. Each intensity plot also shows that the intensity reflects a moving average of the sand occurrence rate in the vertical. Recalling that we believe the lens occurrence process to be quite heterogeneous in alluvial deposits of interest, and the relatively small sampling domain associated with a drill log, such an averaging is necessary for providing a stable, meaningful, consistent and useful interpretation of the drill log data. Note that as the bandwidth goes to zero, the averaging interval shrinks and we are closer to the original information. In drill logs where there are only a few transitions, the optimal bandwidth is seen to be larger, reflecting the homogeneity. One can see from the figures that drill logs that show sustained and frequent transitions have smaller bandwidths or averaging intervals. The optimal kernel bandwidth is thus intimately related to the scale of fluctuations exhibited by the media and can be viewed as a measure of statistical homogeneity in the vertical. Note also from the figures 4 to 7, that an intensity λ(z) of 0.5 corresponds roughly with the location of the interface between sand and clay materials, particularly for thick sand and clay layers. This is an expected and natural consequence of the weighted moving average nature of the kernel intensity smoother.

From the examples presented above, we see that the kernel intensity smoother provides an interpretation of the drill log data that can be useful for assessing an appropriate averaging scale for the fluctuations in media in the vertical at a drill log, as well as an estimate of the average "sand content" at that averaging scale through the magnitude of the intensity.

Phase 2

The loess algorithm as implemented in S-Plus was used to estimate λ(z) at a grid of size 6000 feet, northing, 6200 feet, easting, and 250 feet in the vertical using the phase 1 results at each drill log. A pictorial representation of the site and the gridded area is shown in Figures 8 and 9. Since loess does not allow extrapolation, results are evaluated at interior grid nodes only.

The data set used for loess is l(zji) vs zji, xi, yi, where (xi, yi) is the drill log location, and zji are the locations of the points at which the phase 1 evaluation of l(zji) was made in the vertical at drill log i. The span (fraction of sample used to compute local regression) and degree (order of local approximation 1= linear, 2 -quadratic) were chosen by using the F statistic to compare alternate values of span and degree. A loess surface is fit for a span and degree and the residual sum of squares is computed. The F-statistic using the proper number of degrees of freedom in each case is used to compare the residual sums of squares (RSS) at a significance level of 0.01. In case the residual sum of squares are significantly different (at the 0.01 level) the span/degree with the lower RSS is selected. Otherwise the span/degree with the lower number of equivalent parameters (higher degrees of freedom) is selected. For the data set explored, a degree of 1, and a span of 0.09 was best. The associated RSS and coefficient of determination R2 were 211.88 and .36 respectively. The R2 suggests that 36% of the variability in the spatial field is explained by loess. We expect linear function (degree=1) to do better in this case because the data is binary and composed of linear segments. The low span of 0.09 is indicative of the heterogeneity of the data set in the three dimensions that are fitted. As in the phase I discussion, the resulting values of λ(z) at the grid represent an averaging over the horizontal and vertical of the sand/clay occurrence process. The effective scale of averaging is determined by the k nearest neighbors used in the estimate (k = span * sample size), which translates into the relative distance between k drill logs in a neighborhood of the point at which an estimate of λ(z) is desired and the sampling frequency in the vertical at those drill logs. In our case a constant sampling frequency in the vertical was used for all drill logs

Cross sectional results for λ(z) are exhibited through contour plots in Figures 10 to 12. Note that the contouring mechanisms reflects a further averaging of the results at the grid scale. So the primary utility of these contours is to get an idea of the relative variation in the homogenized field over the site. Of interest are transitions over the area between sand and clay at different depths. From Figure 10 (Northing-Easting at 3 depths) we see that the aquifer appears to have a higher degree of homogeneity as one goes deeper, with transition from low conductivity material at the north west corner to high conductivity material at the south east corner. While the pattern at the 190 ft (57.9 m) depth is relatively smooth, a consistent dip in the contours at an easting of approximately 1387500 ft. is suggestive of a large scale structure in the aquifer. This structure is not apparent at the 125 or 60 foot depths. The contours λ(z) at the 60 foot depth suggest a very different picture of the aquifer than those at either the 125 or the 190 foot levels. The trends in the horizontal variation of λ(z) as well as the degree of variation or heterogeneity are also markedly different. Potential areas of concern for contaminant migration are in the upper middle and middle of the site rather than at the south east corner as was the case at a depth of 190 feet , or the entire middle as is the case for the 125 ft (38.1 m) depth.

A vertical view at three eastings is shown in Figure 11. For an easting of 1386200, we spot a high conductivity area at a depth of approximately 150 ft in the middle of the site. and a transition to very high conductivity material with depth towards the south end of the site. The high conductivity area increases in size (depth and lateral extent) as we move to the easting of 1387800 ft. At both of these eastings an interesting structure is apparent at a northing of approximately 1.4744*107, where the contours dip with depth suggesting a shift downwards in the clay layers at this northing. This structure has essentially disappeared or actually somewhat reversed by the time we reach an easting of 1389400 ft.

A vertical view at three northings is presented in Figure 12. A marked dip in the vertical in the contours at a easting of approximately 138500 is suggestive of a structure in that area, that is probably related to the ones identified earlier. The middle north of the site at a depth of about 150 feet appears to be the area of greatest concern for contaminant migration in the vertical.

It is interesting to compare the results for the drill logs presented as part of the phase I discussion with the contours presented above. The drill log locations are identified in the contour plot figures. We observe from these comparisons that the contours do show the same general features illustrated in each of the drill log at these locations. As expected a more smoothed and interpolated representation is provided.

In conclusion, we have demonstrated through an example using data from Salt Lake Valley that the modeling approach proposed can be useful in identifying patterns of heterogeneity in aquifer properties across the site, in highlighting underlying structure in the pattern of occurrence of layers of high and low conductivity, and for identifying potential locations and depths of concern for migration of contaminants in the vertical. It is also worth noting that this was achieved with a probabilistic (rate of occurrence of a soil type) representation, rather than a deterministic one (i.e. interpolate layer boundaries from drill log to drill log). The intensity of 0.5 identifies the transition from clay to sand. Thus at the averaging scale of interest, the 0.5 contour (vertical or horizontal) is the horizon at which monitoring or control activities should be focused. There are additional benefits from the probabilistic formulation of the problem.

In case a groundwater flow and contaminant transport model is to be calibrated, the intensity of vertical transition estimated for each cell of a finite difference or finite element model can be used to guide the specification of average vertical hydraulic conductivity or vertical layer definition for the cell. The intensity measure can be interpreted as a percent sand or percent high conductivity material in the cell, averaged over some moving locale in vertical. If an average value of vertical conductivity is assumed for the site, or for each type of media considered, the percent sand interpretation of the intensity, allows a ready perturbation of this gross site quantity into reasonable values for each cell. Further, since the intensity has an interpretation similar to that of the probability density, one can actually generate random profiles of sand/clay in the vertical for each cell that are consistent with the characterization provided by the data through the model. This is achieved simply by sampling from the nonhomogeneous rate Poisson process in the vertical for each cell at a discretization in the vertical. Given $\lambda(z_{jk})$ for cell k, this can be done directly using the algorithm given by Devroye (1986) , p. 250. The distribution function is :-

$$F(\Delta z) = 1\text{-}e^{-\left(A(z+\Delta z)-A(z)\right)} \quad (\Delta z >= 0)$$

Where:

$$A(z) = \int_0^z \lambda(z)\, dz$$

SUMMARY

We have presented a general probabilistic framework for analyzing drill log data. The framework is based on a binary classification scheme for the variation of media in the vertical, the analysis of resulting counting processes, and nonparametric evaluation of variation at different scales. A demonstration of the procedures developed was provided for one of three formulations proposed using data from Salt Lake Valley.

The NPP-2 formulation is perhaps the most directly useful of the ones presented here. However, we stress that conceptually all three formulations are roughly equivalent and will share similar theoretical properties. The differences lie in how the estimates are developed. One could of course directly use loess to regress the binary information. However, such an approach would be devoid of the probabilistic interpretation, the ability to analyze conditional situations (possible for the formulations presented but not demonstrated), the ability to simulate probabilistic realizations, and the lack of exploitation of the time series structure of the depositional process in the vertical. With regard to the NPP formulations it is possible to forego the two step process and directly adopt a kernel estimator to recover l(z, x,y) given the raw binary data along z. The setup for such an estimator is similar to the 3-d extension of the Keiding and Anderson algorithm presented for the Markov process formulation. However, the ready availability of loess led to a focusing of our efforts on the two phase approach. The kernel approach would have the potential advantage that the bandwidths could be locally adapted by direction, location and heterogeneity exhibited, and would be directly interpretable as measures of heterogeneity. However, such an implementation would have called for developmental effort beyond the resources available in this project.

The nonhomogeneous Markov process formulation is potentially useful once it is localized, (e.g. using the kernel estimator) since it provides a direct estimation of the probability of transition to media type k given media type l at some depth. We have a global averaging scheme working and implemented with the Salt Lake data. However, we prefer to release results from this implementation after the local averaging formulation has been tested.

The nonparametric, kernel estimation framework leads to a number of interesting theoretical properties that are not detailed here. Briefly, the use of localized, weighted, moving averages, leads to strongly asymptotically consistent estimates of the properties of the counting process, asymptotic normality of the estimates based on a central limit theorem type behavior, and the possibility of direct interpretation of the data at different scales.

The example application of NPP-2 with the Great Salt Lake data effectively demonstrates how a practitioner may use such models. Structures revealed by the model were not directly obvious from an examination of individual drill logs. In this sense the model serves a useful role for condensing and interpreting information. Its utility for the groundwater modeling process was also discussed.

REFERENCES

[1] Aalen, O. O. a. S. J. (1978). An empirical transition matrix for non-homogeneous Markov Chains based on censored observations. Scand J Statist, 5, 141-150.

[2] Abramson, I. S. (1982). Arbitrariness of the Pilot Estimator in Adaptive Kernel Methods. Journal of Multivariate Analysis, 12, 562-567.

[3] Barnhil, R. E., (1977). Representation and approximation of surfaces. Mathematical software, Vol 3, J. R. Rice, pp 68-119, academic press, London.

[4] Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatter plots. Journal of the American Statistical Association, 74(368), 829-836.

[5] Cleveland, W. S., S. J. Devlin (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. JASA, 83(403), 596-610.

[6] Cleveland, W. S., S. J. Devlin, E. Grosse (1988). Regression by Local Fitting. J. of Econometrics, 37, 87-114.

[7] Cox, D. R. and Isham, V. (1980). Point Processes. Chapman and Hall.

[8] Cressie, N. (1991). Statistics for Spatial Data. New York: John Wiley & Sons Inc.

[9] Cressie, N. A. C. (1991). Statistics for spatial data. New York : J. Wiley.

[10] Devroye, L., (1986). Non-uniform random variate generation. New York: Springer-Verlag.

[11] Diggle, P. (1985). A kernel method for smoothing point process data. Applied Statistics, 34(2), 138-147.

[12] Diggle, P. G. (1983) Statistical analysis of spatial point pattern. Academic Press, London

[13] Diggle, P. a. J. S. M. (1988). Equivalence of Smoothing Parameter Selectors in Density and Intensity Estimation. JASA, 83(403), 793-800.

[14] Friedman, J. H., Bentley, J. L., & Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. ACM Trans Math Soft, 3, 209-226.

[15] Gasser, T., C. Jennen-Steinmetz and J. Engel (1991). Comment on Choosing a Kernel Regression Estimator by Chu and Marron. Stat Sci, 6(4), 419-421.

[16] Hall, P., & Marron, J. S. (1987). Extent to Which Least-Squares Cross-Validation Minimizes Integrated Square Error in Nonparametric Density Estimation. Probab. Th. Rel. Fields, 74, 567-

581.

[17] Isaaks, E., Sirvastava, R. (1989). An Introduction to Applied Geostatistics. Oxford University Press, New York, NY.

[18] Johnson, N. M. and Dreiss, S.J. (1989). Hydrostratigraphic Interpretation Using Indicator Geostatistics, Water Resources Research, 25(12), 2501-2510.

[19] Journel, A. (1983). Non-Parametric estimation of spatial distributions. Math Geology, 15(3):445-468.

[20] Journel, A. (1989). Fundamentals of Geostatistics in Five Lessons. Vol 8, Short Course in Geology, American Geophysical Union, Washington, D.C.

[21] Karr, A. F. (1986). Inference for stationary random fields given poisson samples. Adv. Appl. Prob., 18, 406-422.

[22] Keiding, N., & Andersen, P. K. (1989). Nonparametric estimation of transition intensities and transition probabilities:. Applied Statistics, 38(2), 319-329.

[23] Müller, H. G. (1988). Kernel and probit estimates in quintal bioassay. Journal of the american statistical association, 83(403), 750-759.

[24] Müller, H.-G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. , 82(397), 231-238.

[25] Müller, H.-G. (1991). Smooth optimum kernel estimators near endpoints. Biometrika, 78(3), 521-530.

[26] Müller, H.-G., & Wang, J.-L. (1990). Locally adaptive hazard smoothing. Probability Theory and Related Fields, 85, 523-538.

[27] Owosina, A., Lall, U., Sangoyomi, T. and Bosworth, K., (1992) Methods for assessing the space and time variability of Groundwater Data. Research report RR-92-HWR-UL/001

[28] Park, B. U., & Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. Journal of the American Statistical Association, 85(409), 66-72.

[29] Patil, P. N., M.T. Wells and J.S. Marron (1992). Kernel Based Estimators of Ratio Functions. .

[30] Ramlau and Hansen (1983) Smoothing counting process intensities by means of Kernel function. Annals of statistics, Vol 11, p 453-466.

[31] Scott, D. W., & Factor, L. E. (1981). Monte Carlo Study of Three Data-Based Nonparametric Probability Density Estimators. , 76(373), 9-15.

[32] Scott, D. W., Tapia, R. A., and Thompson, J. R. (1977). Kernel Density Estimation Revisited. J. Nonlinear Analysis Theory Meth. Applic. 1 339-372.

[33] Sheather, S. (1983). A data-based algorithm for choosing the window width when estimating the density at a point. Computational Statistics and Data Analysis, 1, 229-238.

[34] Sheather, S. J. (1986). An improved data-based algorithm for choosing the window width when estimating the density at a point. Computational Statistics and Data Analysis, 4, 61-65.

[35] Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society B, 53(3), 683-690.

[36] Silverman, B. W. (1986). Density estimation for statistics and data analysis. New York:

Chapman & Hall.

[37] Solow, A. R. (1990). The Nonparametric Analysis of Point Process Data: The Freezing History of Lake Konstanz. Journal of Climate, vol 4.

[38] Stone C. J. (1975). Nearest neighbour estimator of a nonlinear regression function. Proceedings of Computer Science and Statistics, 8th symposium, p413-418.

[39] Woodroofe, M. (1970). On Choosing Delta-Sequence. Ann. Math. Statist. 41 1665-1671.

[40] Yakowitz, S. J. (1985). Nonparametric density estimation, prediction, and regression for markov sequences. Journal of the American Statistical Association, 80(389), 215-221.

[41] Yakowitz, S. J., & Szidarovsky, F. (1985). A comparison of Kriging with nonparametric regression method. Journal of Multivariate Analysis, 16(1), 21-53.

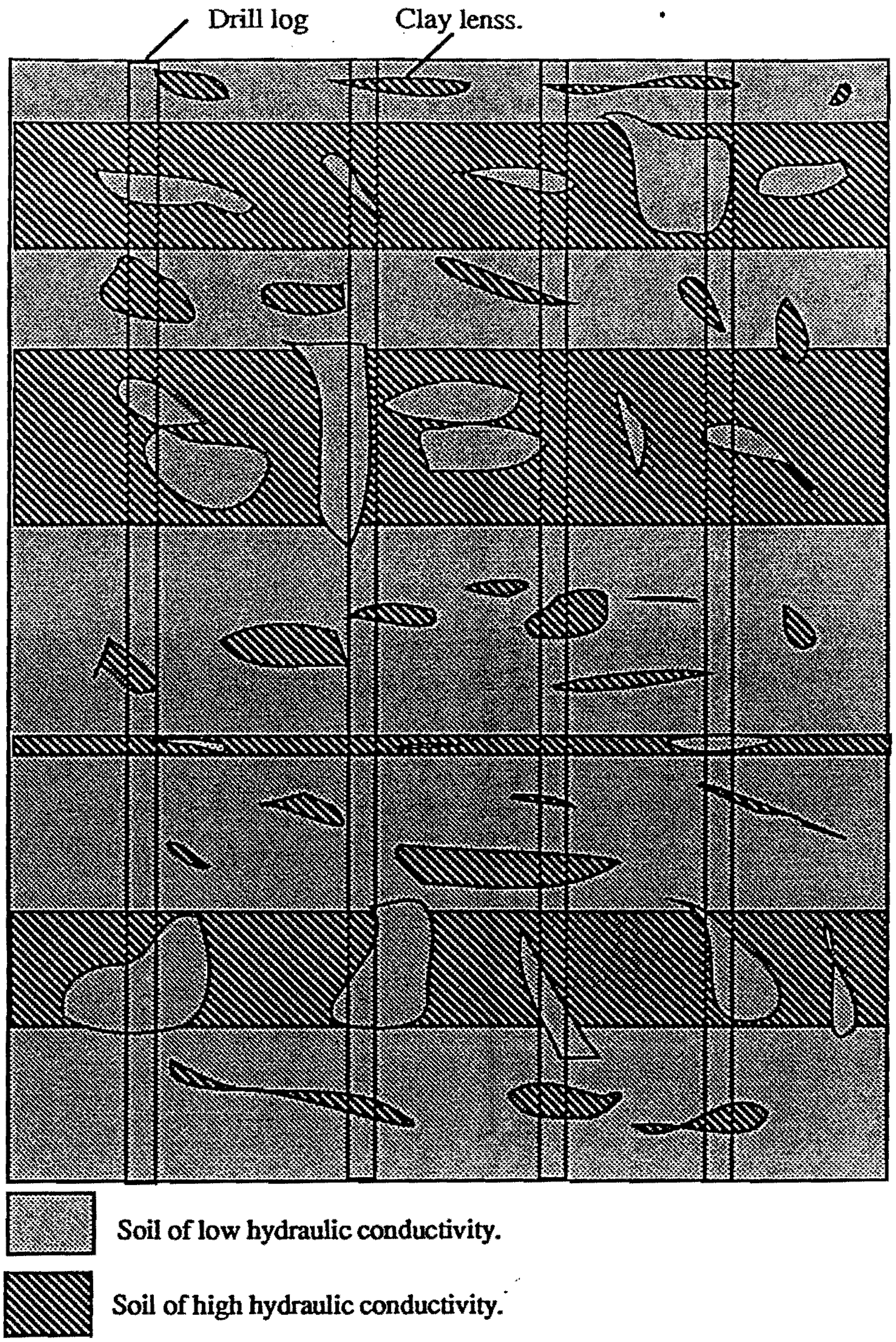

Drill log
Clay lenss.
▒ Soil of low hydraulic conductivity.
▨ Soil of high hydraulic conductivity.

Figure 1. Soil heterogeneitty in alluvial, sedimentary environments

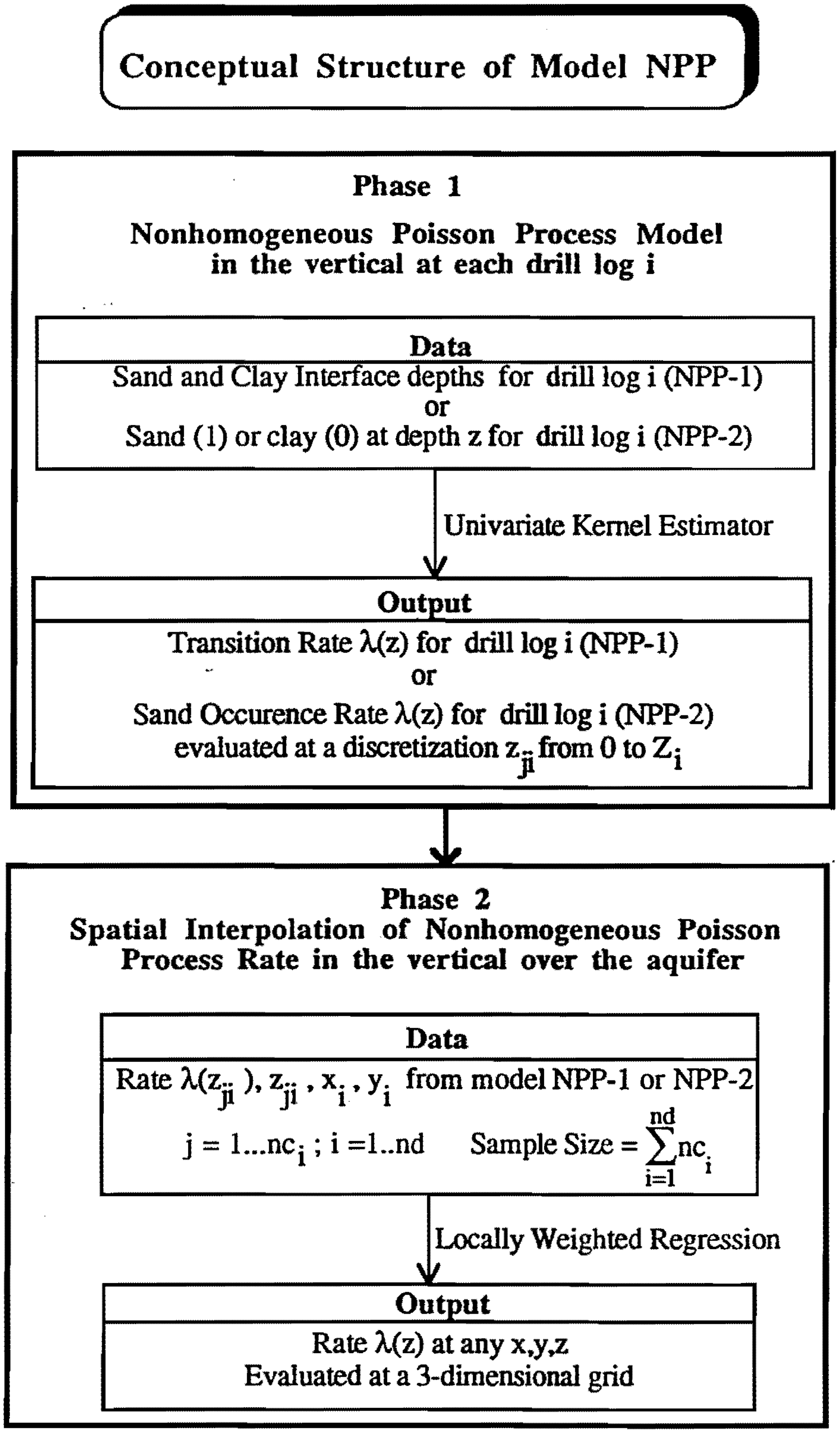

Conceptual Structure of Model NPP
Phase 1
Nonhomogeneous Poisson Process Model in the vertical at each drill log i
Data
Sand and Clay Interface depths for drill log i (NPP-1)
or
Sand (1) or clay (0) at depth z for drill log i (NPP-2)
Univariate Kernel Estimator
Output
Transition Rate $\lambda(z)$ for drill log i (NPP-1)
or
Sand Occurence Rate $\lambda(z)$ for drill log i (NPP-2) evaluated at a discretization z_{ji} from 0 to Z_i
Phase 2
Spatial Interpolation of Nonhomogeneous Poisson Process Rate in the vertical over the aquifer
Data
Rate $\lambda(z_{ji})$, z_{ji}, x_i, y_i from model NPP-1 or NPP-2
$j = 1...nc_i$; $i = 1..nd$
Sample Size = $\sum_{i=1}^{nd} nc_i$
Locally Weighted Regression
Output
Rate $\lambda(z)$ at any x,y,z
Evaluated at a 3-dimensional grid

Figure 2 Conceptual Structure of Model NPP

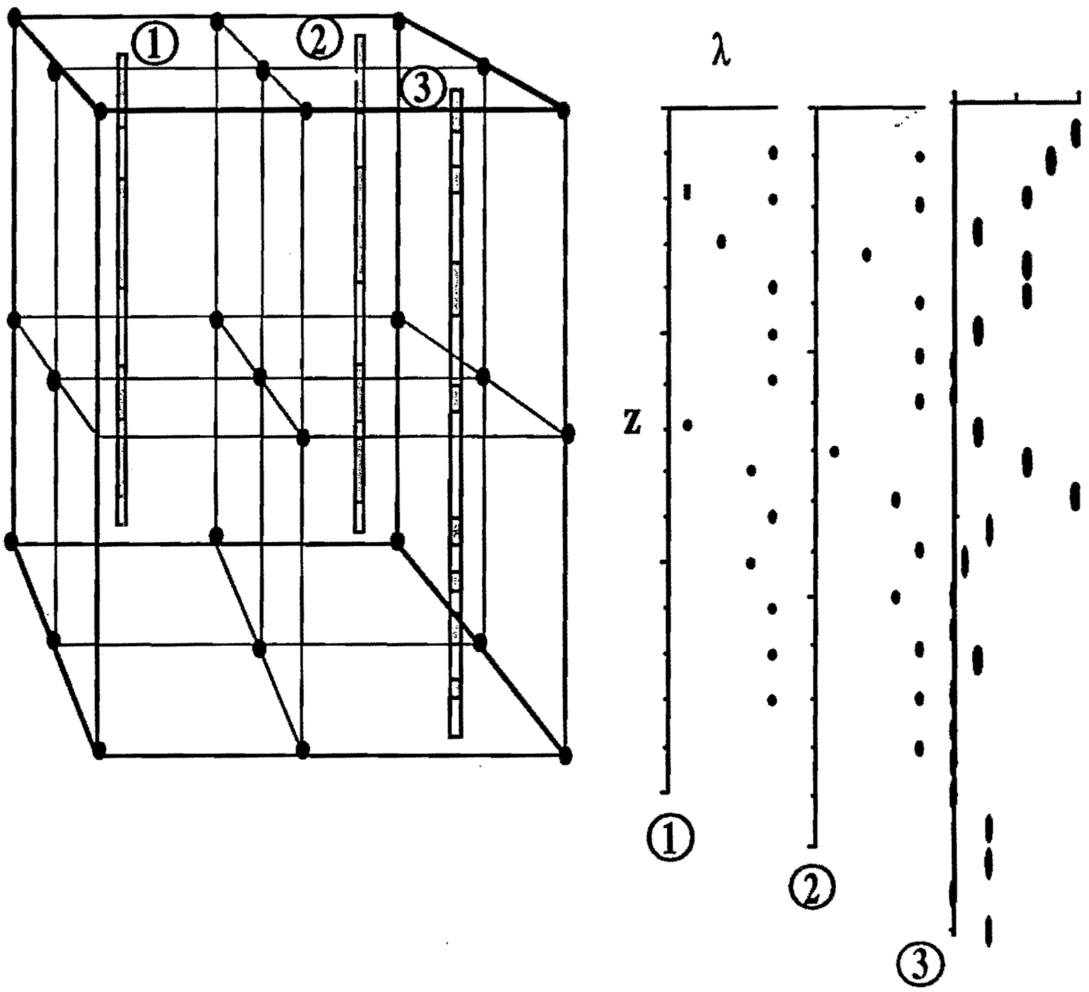
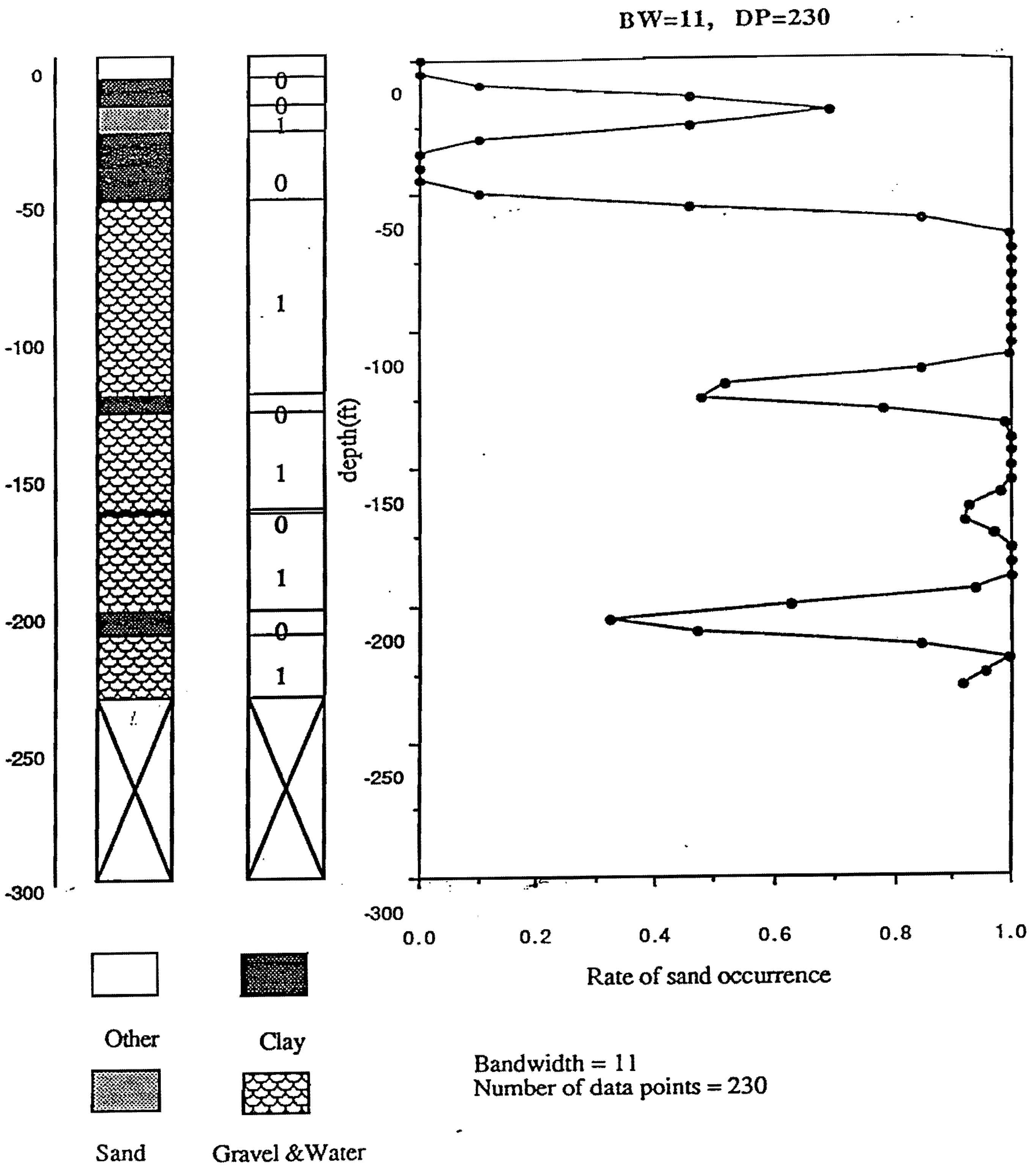


Figure 3 Spatial Interpolation of drill log rates λ(z) onto a grid for contouring



Bandwidth = 11
Number of data points = 230

Figure 4 Drill Log 1, Showing (a) Drill log Profile, (b) 0,1 profile (interpretation b, table1) (c) Kernel smoothed profile for intensity with depth.

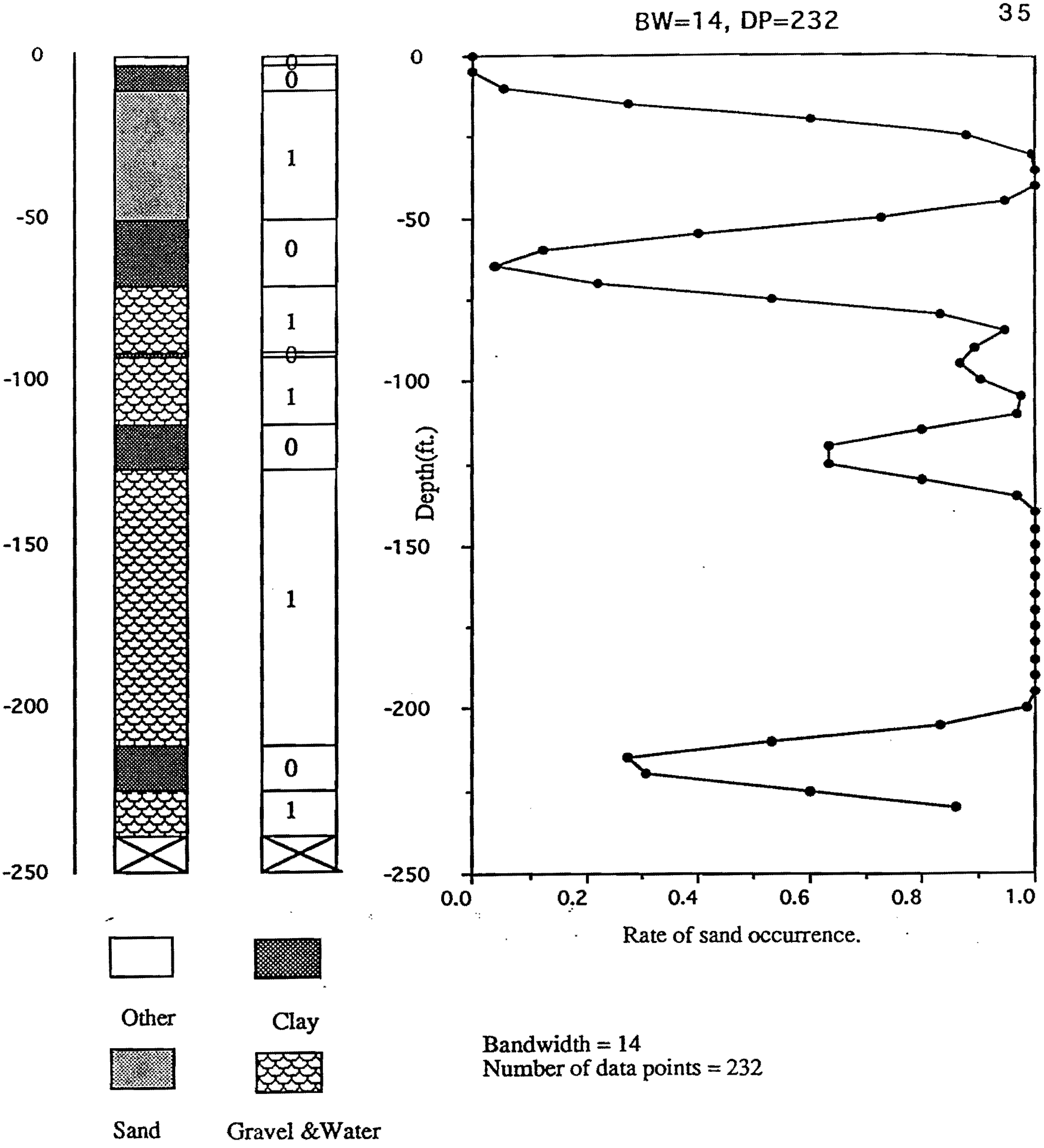


Figure 5 Drill Log 2, Showing (a) Drill log Profile, (b) 0,1 profile (interpretation b, table1)

(c) Kernel smoothed profile for intensity with depth.

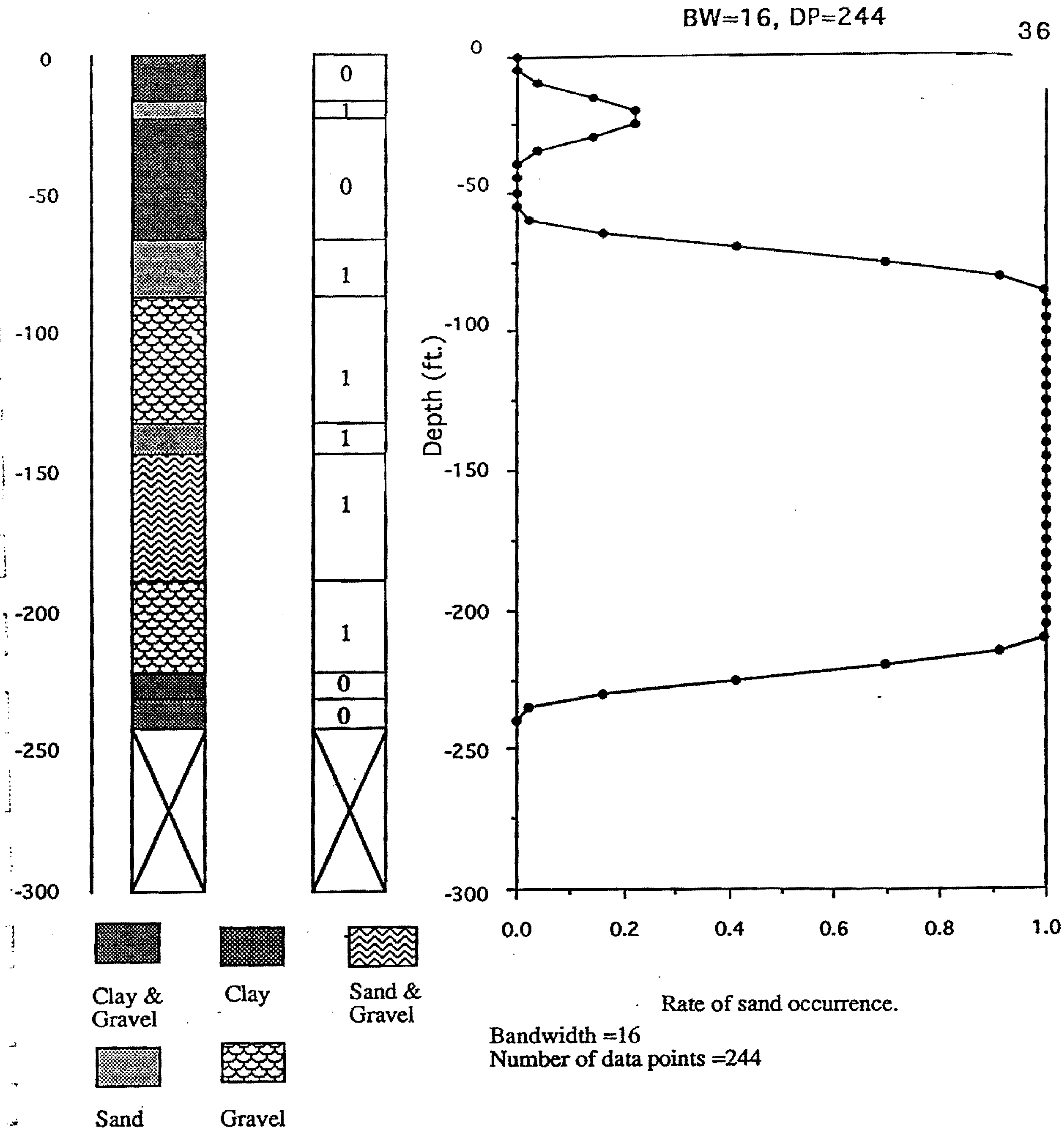


Figure 6 Drill Log 9, Showing (a) Drill log Profile, (b) 0,1 profile (interpretation b, table1)

(c) Kernel smoothed profile for intensity with depth.

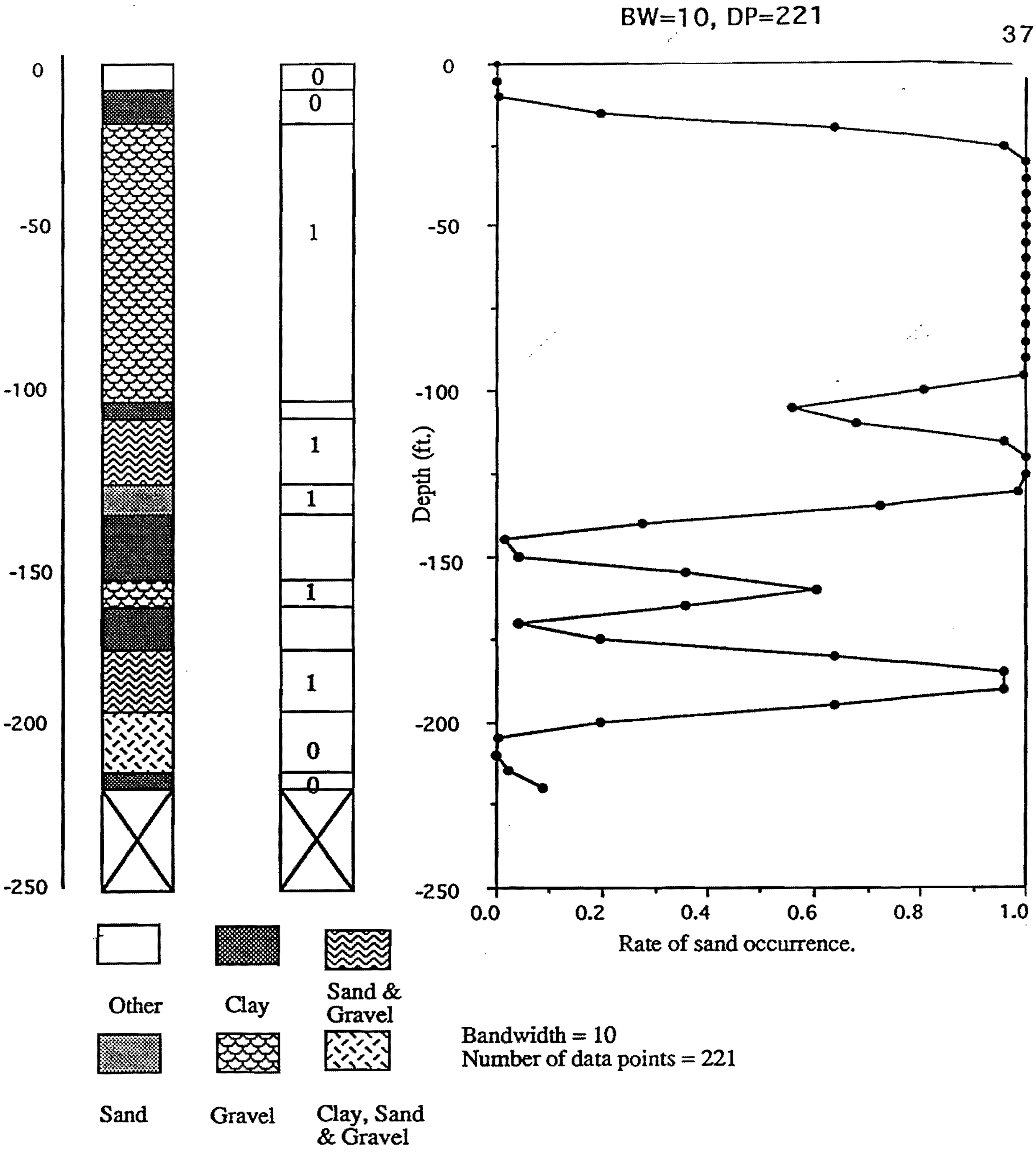


Figure 7 Drill Log 35, Showing (a) Drill log Profile, (b) 0,1 profile (interpretation b, table1)

(c) Kernel smoothed profile for intensity with depth.

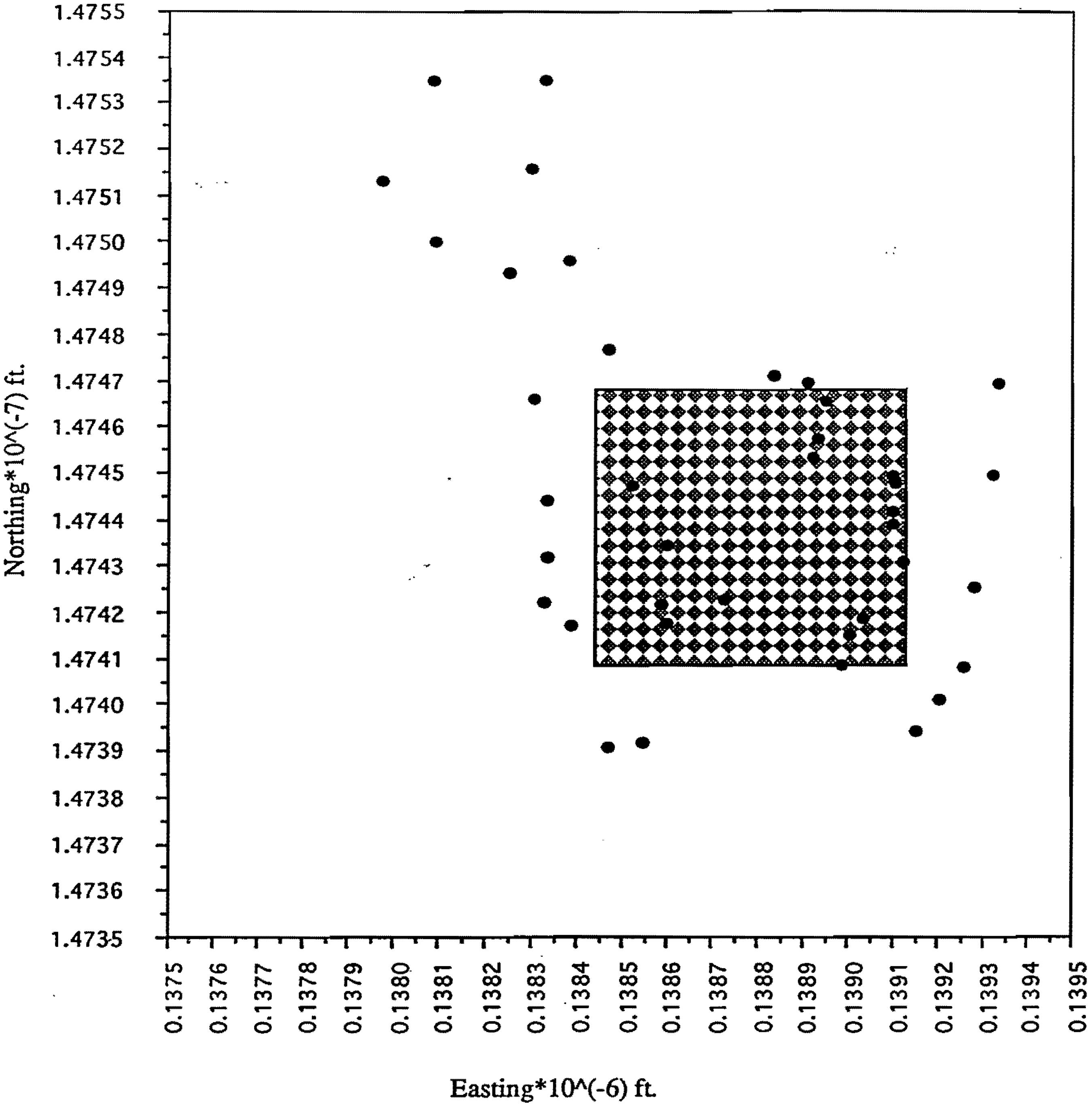


Figure 8. Wells and the estimating grid locations

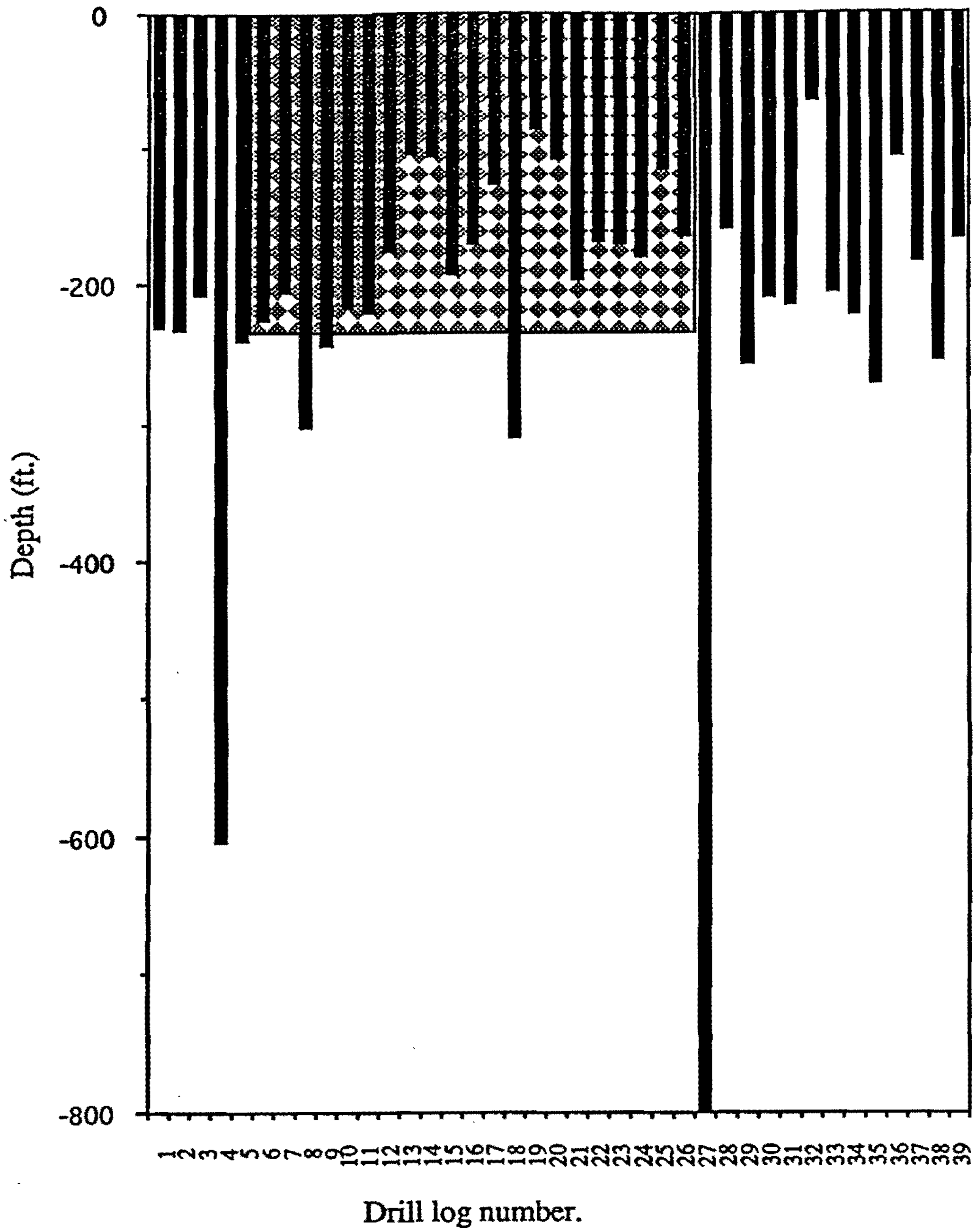


Figure 9. Vertical plan for the wells locations and the estimating grid in the northing direction.

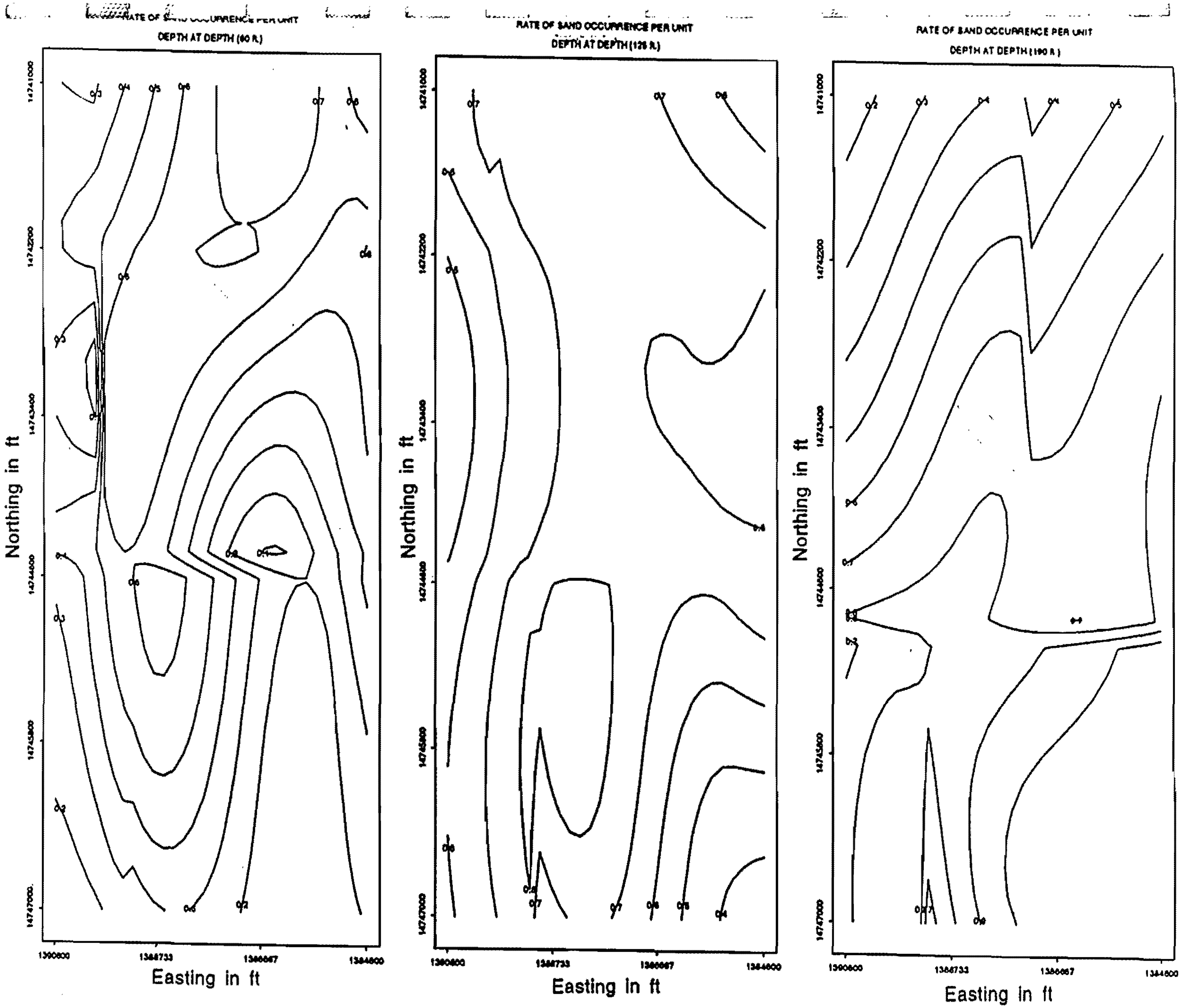


Figure 10. Intensities contour lines in Northing-Easting plan at different values of depth.

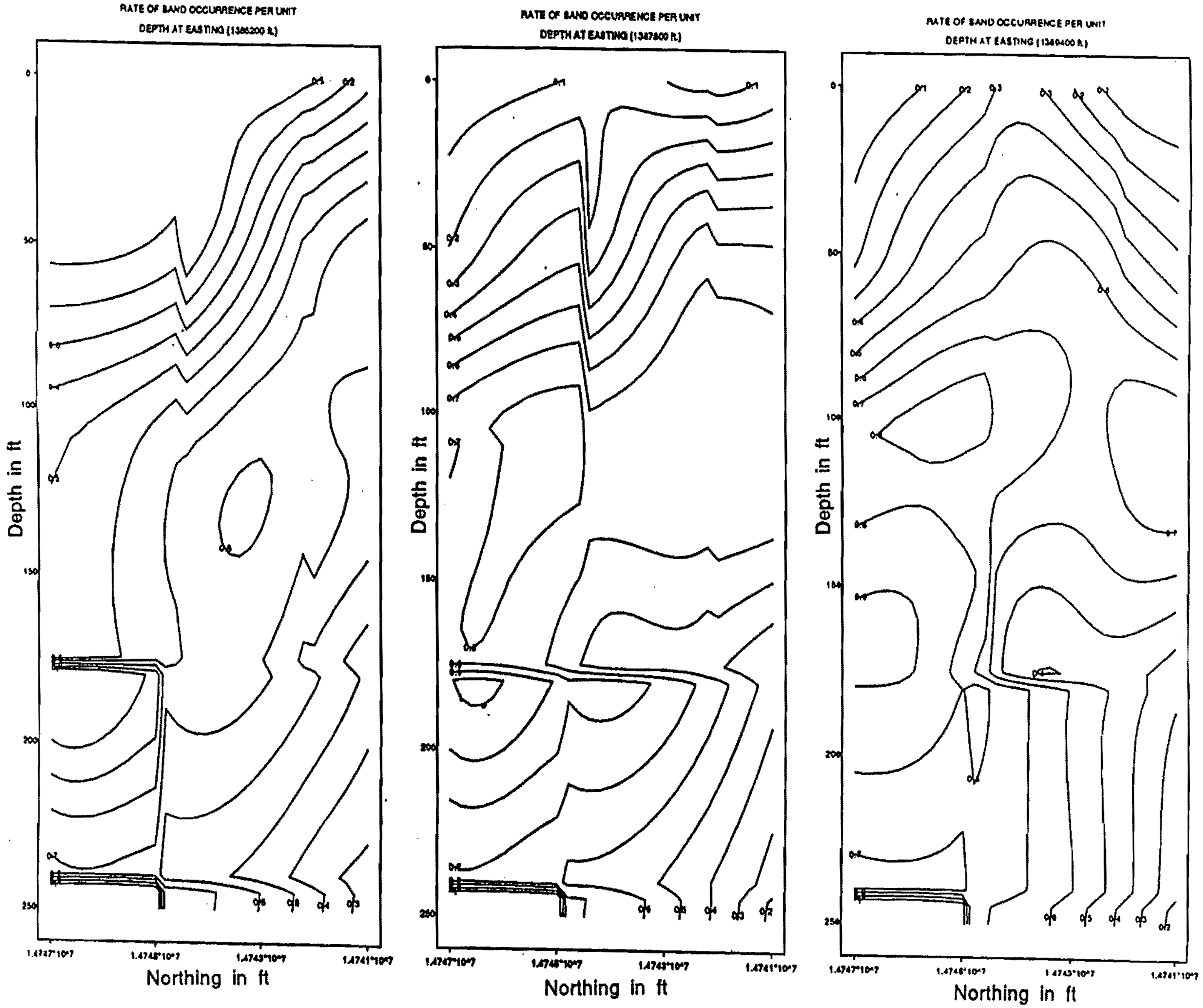


Figure 11. Intensities contour lines in Vertical-Northing plan at different values of Easting.

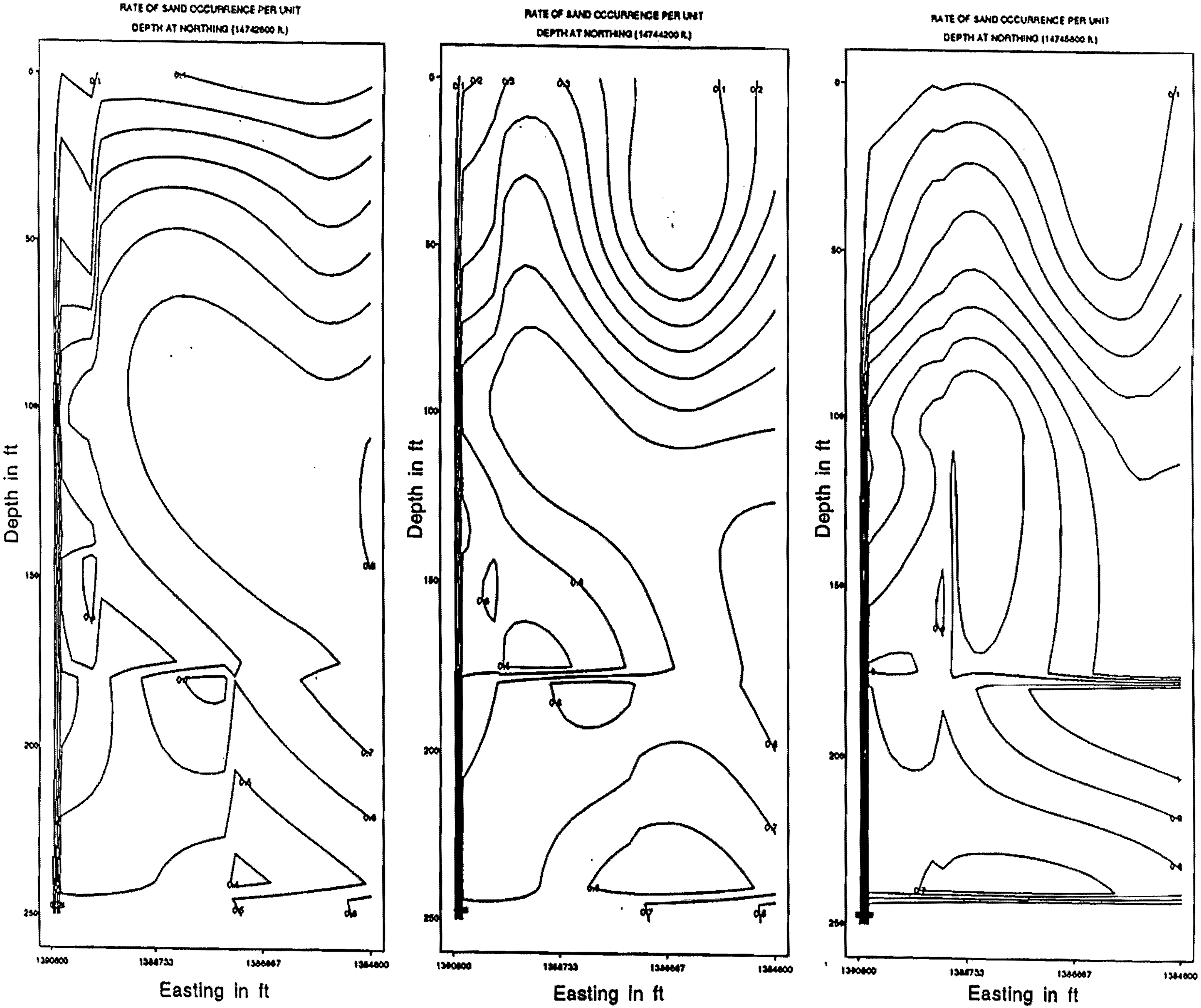


Figure 12. Intensities contour lines in Vertical-Easting plan at different values of Northing