APPENDIX D. FREQUENCY ANALYSIS WITH ZEROS

The highly heterogeneous nature of real-world groundwater contamination combined with the need to delimit the plumes often results in datasets that have multiple zero measures. These zero measures complicate the traditional implementation of frequency analysis used to attain the empirical cumulative distribution function (cdf) for contaminant concentrations. In this thesis, the empirical cdf is vital for quantile kriging discussed in Chapters 5 and 6. Guidance for dealing with multiple zero measures can be taken directly from the field of hydrology where zero measures are also prevalent. *Haan* (1977) shows how total probability theory can be used to address this issue. First, let us define the following quantities:

C = random variable representing contaminant concentrations;
S_n = set of nonzero contaminant data;
S_o = set of zero data;
c_i = the i^{th} contaminant measure;
$Pr(c_i \in S_n)$ = probability of observing a nonzero measure;
$Pr(c_i \in S_o)$ = probability of observing a zero measure;
$F(C \leq c_i)$ = empirical cdf for contaminant concentrations;
$Pr(C \leq c_i \mid c_i \in S_n)$ = probability of the i^{th} contaminant observation given the sample is a member of the set of nonzero data, $c_i \in S_n$;
$Pr(C \leq c_i \mid c_i \in S_o)$ = probability of the i^{th} contaminant observation given the sample is a member of the set of zero data, $c_i \in S_o$;
i = the index for the contaminant data measures;
N = total number of contaminant data;
n = total number of nonzero contaminant data;
r_i = the rank assigned to the i^{th} contaminant data measure.

Given these quantities, equation (D.1) utilizes total probability theory to divide the contaminant data set into 2 populations (1) the nonzero data and (2) the zero data when computing the empirical cdf, $F(C \leq c_i)$.

$$F(C \leq c_i) = Pr(C \leq c_i \mid c_i \in S_n)\, Pr(c_i \in S_n) + Pr(C \leq c_i \mid c_i \in S_o)\, Pr(c_i \in S_o) \quad \text{(D.1)}$$

Given a total N contaminant measures of which n are nonzero measures, equations (D.2) and (D.3) illustrate how to compute the probabilities of attaining nonzero and zero measurements, respectively.

$$Pr(c_i \in S_n) = \frac{n}{N} \quad \text{(D.2)}$$

$$Pr(c_i \in S_o) = \frac{N-n}{N} \quad \text{(D.3)}$$

Equation (D.4) shows, as expected, the probability that the j^{th} observation is less than or equal to zero given it has been taken from the set of zero data is equal to 1.

$$Pr(C \leq c_i \mid c_i \in S_o) = \frac{N-n}{N-n} \to 1 \quad \text{(D.4)}$$

Equation (D.5) is the result of substituting equations (D.2) - (D.4) into equation (D.1).

$$F(C \leq c_i) = Pr(C \leq c_i \mid c_i \in S_n)\, \frac{n}{N} + \frac{N-n}{N} \quad \text{(D.5)}$$

The final quantity that remains to be specified in equation (D.5) is the conditional probability of attaining the i^{th} contaminant measure given that the measure is a member of the nonzero data set, $Pr(C \leq c_i \mid c_i \in S_n)$. This quantity is computed by ranking all of the data in ascending order (i.e., from the least to the greatest). When ranking the data all of the zero measures are assigned a rank of 0 (since they have been segregated into a different statistical population) and the nonzero ranks range from 1 to n. Equation (D.6) specifies how to compute $Pr(C \leq c_i \mid c_i \in S_n)$.

$$Pr(C \leq c_i \mid c_i \in S_n) = \frac{r_i}{n+1} \tag{D.6}$$

Finally, substituting equation (D.6) into equation (D.5) yields equation (D.7), which is an appropriate measure of the empirical cdf in the presence of multiple zero measures.

$$F(C \leq c_i) = \frac{r_i}{n+1} \frac{n}{N} + \frac{N-n}{N} \tag{D.7}$$