Natural Variability of Benthic Species Composition

Dean Billheimer **Tamre Cardoso** **Elizabeth Freeman**

Peter Guttorp **Hiu-Wan Ko** **Mariabeth Silkey**

NRCSE

Technical Report Series

NRCSE-TRS No. 001

Natural Variability of Benthic Species Composition in the Delaware Bay

Dean Billheimer *

University of Alaska, Fairbanks

Tamre Cardoso

Elizabeth Freeman

Peter Guttorp

Hiu-Wan Ko

Mariabeth Silkey

University of Washington, Seattle

March 11, 1998

Abstract

Biological monitoring of aquatic biota is used to assess the impact of changes in the

*We are grateful to Melissa Hughes of EPA Delaware and Tony Olsen of EPA Corvallis for making these data available to us, and to Wendy Meiring for a careful reading of the paper and many valuable suggestions. This research had partial support from the United States Environmental Protection Agency under a cooperative agreement with the University of Washington. This document has not undergone Agency review, and reflects in no way official Agency policy.

environment. Critical to the development of a sound biological monitoring protocol is the choice of organisms and characteristics to be monitored. In order to make accurate interpretations of change, some description of natural variability of the system is necessary. We introduce a state-space model for compositional monitoring data, and illustrate how one can incorporate spatial structure and covariates to assess natural variablity. The methods are illustrated on benthic survey data from Delaware Bay, using species composition at the genus level. The distribution of benthic macroinvertebrates in Delaware Bay is significantly dependent on salinity. There is residual spatial dependence in the data after accounting for the salinity effect.

Key Words: Biological monitoring; benthic invertebrates; spatial model; state-space model

1. Introduction

Direct observation of organisms living within an ecosystem is key to evaluating the health of the system, and to understanding the processes occurring within it. As an assessment method, biological monitoring, i.e., direct measurement of changes in a habitat using the number and distribution of individuals or species, captures both episodic and cumulative effects of changes in the environment. A critical issue in the development of a biological monitoring protocol is the choice of organisms and characteristics to be monitored (Marmorek et al., 1988; Spellerberg 1991). The efficiency, productivity and relative abundance of organisms within a biological community are all potential measures of ecosystem health. In addition, wise selection of organisms with a variety of life history characteristics can reveal the effects of environmental phenomena at multiple temporal and spatial scales.

We propose using relative abundance of different groups of species to monitor the ecological condition of an ecosystem. We present a rationale for grouping species into classes

with similar life history/disturbance-response characteristics. In addition, we present methods of statistical analysis for evaluating the natural variability in the relative abundance of these groups. This approach is illustrated by an analysis of the composition of benthic invertebrates collected from the Delaware Bay.

1.1 Methodologies

A sound biological measure of ecological change must provide a high precision determination of both large- and small-scale disturbances. Ecological systems exhibit large natural annual variation in abundance and biomass. Hence changes in total abundance do not necessarily measure the health of a system. An examination of patterns of change in relative numbers of taxa in the system can, however, be more enlightening. Current debate examines the issue of how best to evaluate the health of aquatic, estuarine or marine ecosystems. Fore et al. (1995) reviews and compares four major approaches to biological ecosystem assessment: similarity and diversity indices; pollution tolerance indices based on indicator species; mulitimetric indices; and, multivariate ordination and classification methods.

Diversity indices combine information about species abundance and species richness into univariate summaries of the biological health of the ecosystem. In a study of benthic macroinvertebrate populations, Warwick (1986) looked at the distribution of biomass and abundance in polluted and unpolluted sites. He found that unpolluted communities tended to have higher diversity in numbers of individuals among species, while polluted sites had increased diversity in biomass. Similarity and diversity indices do not account for differing life history characteristics of the organisms comprising the index. Further, these indices provide no information about the type of distribution, stage of succession, or species composition of a biological community (Spellerberg, 1991, p.125). Consequently, such indices are not well suited to differentiate between natural variability in species abundance and variation

due to environmental impacts. Dennis et al. (1979) concludes that while diversity indices exhibit weak relationships with ecological change at a given site, they provide inadequate information about the nature of the change.

Pollution tolerance indices assign a pollution tolerance value to every species and calculate an index score for a site as a function of the number of individuals of each tolerance class. This "canary in the coal mine" approach can be useful, especially when the species of interest are known opportunists of environmental degradation. However, this tool suffers from sampling problems (it is usually easy to demonstrate the presence of an indicator species, but much more difficult to determine its absence). As with similarity and diversity indices, pollution tolerance indices are limited in geographic scope (Schwinghammer, 1988), and rely on variation in absolute abundance measures of the species of interest (Gray and Pearson, 1982).

Multimetric indices examine a multitude of biological attributes, thereby integrating information from ecosystem, community, population and individuals. Each component metric measures an attribute of the assemblage that is the product of evolutionary and biogeographic processes at a site (Karr, 1995). These individual metrics are all based upon the natural history of the system, and each contributes to a univariate summary of the condition of the sampled area (Deegan et al., 1992).

Multivariate methods can combine physical, chemical and biological information into a single matrix from which patterns can be sought. Aschan (1990) implemented principal component analysis and ordination in a study of softbottom macrofauna. The results of such analyses can be heavily driven by the inevitable preponderance of null values in data sets used for this approach (Fore et al., 1995). When all variables are weighted equally, these methods do not take advantage of known natural history of the ecosystem.

Changes in community composition offer a high potential for success in overall ecosys-

tem assessment. In contrast to the various indices mentioned above, this method does not rely on abundance measures (which exhibit high natural variability). Further, community composition is less reductionist in its approach than are univariate indices. Insight into the ecological structure of a community can be retained by knowledge of the relative abundance of its component species. Such insight is lost when the information is combined into a single number. Community composition is directly related to the *biological* response of the system. This differs from multivariate approaches which tend to favor inclusion of chemical data that may or may not be biologically relevant. Shifts in species composition within a community have been identified as valuable early warning indicators of the effects of pollution (Patrick, 1972; Schindler et al., 1985; Marmorek et al., 1988; Guttorp, 1993). To date, one limitation in the use of community compositions for ecological assessment has been the lack of statistical methods available for compositional data with spatial and/or temporal dependence.

In this paper we introduce methods for quantifying the natural variability of compositional data observed in estuarine benthic communities, allowing for the effect of abiotic covariates on this variability. In section 2 we describe the data set used to illustrate the methods. Section 3 describes criteria used to group species. Preliminary data analysis is presented in section 4, and the statistical model outlined in section 5. Results of the modeling effort are found in section 6.

2. Benthos data description

The data used in this paper consist of information from benthic surveys conducted by the US Environmental Protetion Agency as part of the Ecological Monitoring and Assessment Program (EMAP) in 1990 across the Delaware Bay. Contents of benthic grab samples were identified to genus, and where possible, to species. Three grab samples were made at each

location. Data collection procedures in 1990 were specifically geared to investigate the distribution of benthic populations across the Bay, and included 19 sites supplementary to the six baseline stations sampled as part of the nationwide EMAP protocol. These six baseline stations and six other synoptic stations have subsequently been revisited annually. Corresponding conductivity–temperature–density sensor (CTD) measures of salinity, dissolved oxygen, depth, temperature, pH, fluorescence, light transmission, and conductivity were also made during the benthic sampling.

Figure 1 about here

The location of the 25 sampling stations visited in 1990 are shown in Figure 1. Sites are located on a regular hexagonal grid according to the EMAP sampling protocol (Overton et al., 1990). Average abundance in three samples at each visit to a station is summarized, for species occurring at 15 or more stations, in Table 1.

Table 1 about here

Among the 11 species shown in Table 1, the 1990 average abundances in the three samples range from 0.33 to 623. *Mediomastus ambiseta* (family Capitellidae) dominates many stations. The average abundance of this species is 152.96 with the maximum (623), occurring at station 23. Next most common is *Tellina agilis* (family Tellinidae); its average abundance is 22.43. Some of the benthic conditions, recorded at the time of sampling , are summarized in Table 2. Dissolved oxygen, temperature, and pH do not vary widely among the sampling stations.

Table 2 about here

3. Selection criteria for taxonomic groupings

Efforts to assimilate community structure often result in deliniation of species according to feeding strategy. Word et al. (1977) develop an index based on numbers of infaunal benthic invertebrates in four categories: suspension feeders, surface detritus feeders, surface deposit feeders, and sub-surface deposit feeders. Karr (1981) divides a freshwater fish community into suckers and darters.

Although all the organisms sampled in this study live in or on the benthos, they each specialize in their manner of retrieving nutrients from their environment. Suspension feeders extend their polyps to strain food from the water column. Deposit feeders generally have two siphons, the longer one for acquiring food off the estuarine floor, and the other, shorter one, for depositing its feces. Deposit feeders can make the surface unlivable for suspension feeders, by either fouling their polyps with feces, or otherwise kicking up the detritus on the benthic floor. Suspension feeders are generally found in areas of higher water velocity, where food for deposit feeders does not accumulate and thus fewer deposit feeders are found. Thus, these two distinct feeding strategies form a natural division in the community structure. Neither of these strategies dominates the other for survival in environments with different degrees of chemical pollution, they are optimized for differing sediment grain size conditions.

We sought a third grouping, composed of creatures that are particularly hardy in environmentally stressed ecosystems. These are the bloodworms. Hemoglobin in the blood of these organisms allows them to make more efficient use of oxygen than other polycheates; hence, they gain advantage over other genera where oxygen depletion occurs.

Our choice of diagnostic taxa was restricted to those organisms occurring in sufficient numbers and at sufficiently many sites to reveal spatial structure. From the sixteen most prevalent genera, three taxonomic groupings were formed according to life history charac-

teristics: tolerant, intolerant and suspension feeders. The tolerant group consists of the predatory bloodworms *Glycera, Glycinde,*and the sediment feeders *Mediomastus* and *Heteromastus.*The intolerant group consists of sediment eating amphipods (*Corophium* and *Ampelisca*) and the bivalves *Tellina* and *Mulinia.* These deposit feeders are particularly sensitive to the health of the benthic sediments. The suspension feeders in our study are *Polydora, Paraprionospio, Streblospio,* and *Spiochaetopterus.*

4. Data analysis

We first examine relationships of species group composition and benthic conditions in the sampling sites.

Figure 2 about here

The composition shown in Figure 2 varies with salinity. There is a substantial degree of spatial coherence, in that neighboring sites tend to have similar compositions. *Tellina agilis* dominates the high salinity areas while *Mediomastus ambiseta* dominates in regions of mid-range salinity. At sites where *Mediomastus ambiseta* is found, it is at least an order of magnitude greater in abundance than the next most prevalent species. In our evaluation, compositions including *Mediomastus ambiseta* are less informative due to its overwhelming abundance. In addition, it does not have an obligate feeding strategy, which led us to eliminate it as a component in our grouping.

Figure 3 about here

Figure 3 is a ternary diagram corresponding to Figure 2. From this representation it is clear that we have large contributions of the pollution insensitive bottom feeders when the salinity

is low, while high salinity is associated with a low proportion of suspension feeders. Site 11, in the river mouth (and hence with low salinity) is dominated by suspension feeders and intolerant bottom feeders, while these two groups are almost entirely absent at site 23. The latter site appears quite different from its neighbors, which may be an indication of a local disturbance to the bay near this station. In particular, there is a surprisingly low proportion of intolerant species. In the 1960's, 300 ton of DDT was deposited into the water along the nearby Cape May (Alan Mearns, personal communication), which may in part explain the peculiar observation at this site. Besides salinity, other covariates such as dissolved oxygen, temperature and depth, were examined, but no clear relationship to group composition was found.

Three independent samples were collected at each station. The three samples at each station can be used to check whether the natural variability of the data is larger than the multinomial variability. This is often the case for biological populations, as pointed out by, e.g., Pollard (1975, p. 129). At each station, a chi-square test statistic of the hypothesis of equal proportions for the three samples was computed and compared to a reference distribution. Usually, the test statistics calculated in this way are expected to be χ^2 distributed under the null hypothesis of equal proportions. However, since the expected counts of suspension feeders at many stations were less than 5, this null distribution may not be appropriate. Instead, we employed a small Monte Carlo simulation (using 100 repetitions) to determine the null distribution. Four out of the 25 sampling stations did not have test statistics computed because they either contained no suspension feeders or no tolerant species. In the remaining 21 sampling stations, 12 had test statistics greater than the 95th percentile of the reference distribution. Thus, we conclude that the variability of the counts tends to be super-multinomial.

5. Statistical model

In order to explain the super-multinomial variability and the spatial dependence exhibited in the previous section, we adopt a state-space approach for modeling benthic compositions. For each benthic sample we posit an unobservable "state" composition vector describing the proportion of organisms attributable to each group. Conditional upon the state, counts of organisms are assumed multinomial. The effect of covariates, such as salinity or dissolved oxygen, as well as spatial structure is incorporated in the state distribution. We develop a conditional autoregressive model (CAR; Besag, 1974; Mardia, 1988) to define a spatial prior distribution for the state compositions. Markov chain Monte Carlo (see, e.g., Besag et al., 1995) is used to provide information about the posterior distribution of sample site compositions, logistic normal model parameters, and covariates.

Aitchison (1986) describes statistical analysis methods for compositional data with independent observations. These methods rely on the additive logratio transform to map observations from the $(k-1)$–dimensional simplex (∇^{k-1}, the space of k–category proportion vectors) to $(k-1)$–dimensional Euclidean space ($\Re^{k-1}$). Assuming that the transformed data are $(k-1)$–dimensional multivariate normal induces the logistic normal distribution on ∇^{k-1}.

Central to the choice of the additive logratio transform is a perturbation operator whose effect is to combine two composition vectors to produce a third composition (Aitchison, 1982). This operator can be used to produce a structure for noise on ∇^{k-1} that is more natural than the usual additive noise model used in other areas of statistics. The usual statistical model partitions observations into an average level plus independent noise. Our approach decomposes observations into a level (location in the simplex) perturbed by independent noise. Further, the location parameter may be decomposed into an overall location which is

in turn perturbed by the effect of a covariate. By operating directly on proportions, we gain insight and interpretability in evaluating modeling results.

5.1 Perturbations and the logistic normal distribution

We begin by describing several operations and transformations that are central to our statistical models for compositions. The development follows Aitchison (1986) and is shown here to aid the presentation. Suppose that $\mathbf{z}$ is a k–vector of proportions. That is, $0 < z_i < 1$, for all $i = 1, 2, ...k$, and $\sum_{i=1}^{k} z_i = 1$. We say that $\mathbf{z}$ is an element of the $(k-1)$–dimensional simplex ($\mathbf{z} \in \nabla^{k-1}$).

Definition 1 *Composition Operator ($\mathcal{C}$)*

Suppose $\boldsymbol{\alpha}$ is a k–dimensional vector in positive Euclidean space ($\Re_+^k$). Define $\mathcal{C}(\boldsymbol{\alpha})$ by the following operation:

$$[\, \mathcal{C}(\boldsymbol{\alpha}) \,]_i = \frac{\alpha_i}{\sum_{j=1}^{k} \alpha_j}$$

where $[\, \mathcal{C}(\boldsymbol{\alpha}) \,]_i$ denotes the i^{th} element of the k–vector ($i = 1, 2, ..., k$).

Thus, the composition operator normalizes a positive k–vector to sum to one, and $\mathcal{C}(\boldsymbol{\alpha}) \in \nabla^{k-1}$.

Definition 2 *Perturbation Operator*

Let $\mathbf{z}$ be a k–part composition and $\boldsymbol{\alpha}$ be a k–vector with positive elements. Define the perturbation operator as follows:

$$\mathbf{z} \circ \boldsymbol{\alpha} = \mathcal{C}(\mathbf{z} \cdot \boldsymbol{\alpha}) \quad \textit{where } (\cdot) \textit{ denotes element-wise multiplication.}$$

Thus, the composition $\mathbf{z}$ is mapped to a location in ∇^{k-1} by the perturbing vector $\boldsymbol{\alpha}$.

Aitchison (1986, section 2.8, p 42) shows that the perturbation operation is a one-to-one transformation between ∇^{k-1} and ∇^{k-1}, with an inverse transformation; perturbation by $\boldsymbol{\alpha}^{-1} = (1/\alpha_1, 1/\alpha_2, ..., 1/\alpha_k)$. Further, the effect of any perturbing vector $\boldsymbol{\alpha}$ is the same as that for the composition $\mathcal{C}(\boldsymbol{\alpha})$. So, without loss of generality, we need only consider perturbing vectors in ∇^{k-1}.

In general, one may consider the perturbation operator to define an "addition" operator on the $(k-1)$–dimensional simplex. By adding the inverse of a composition, we also obtain a "subtraction" operation. This analogy with simple mathematical operations on $\Re$ leads to the corresponding multiplication and division analogs.

Definition 3 *Scalar Multiplication*

Define multiplication of a composition $\mathbf{z}$ *by a scalar* u *in the following way*

$$\mathbf{z}^u = \mathcal{C}(z_1^u, z_2^u, ..., z_k^u)$$

This defines a "multiplication" operator that is consistent with the perturbation "addition" analogy. Aitchison (1986, section 6.9, p. 125) shows that the perturbation operation leads to the logistic normal distribution as the limit distribution of a sequence of perturbations by independent noise. This distribution was introduced by Aitchison and Shen (1980). Its use in the analysis of compositional data is chronicled by Aitchison (1986). The density function and the relevant properties of the logistic normal distribution are summarized here following the development of Aitchison (1986, Chapter 6, pp. 112–125). To begin, we first define the additive logistic transformation.

Definition 4 *The additive logistic transformation is the one-to-one transformation of* $\mathbf{y} \in \Re^{k-1}$ *to* $\mathbf{z} \in \nabla^{k-1}$ *defined by*

$$z_i = \frac{\exp(y_i)}{\sum_{j=1}^{k-1} \exp(y_j) + 1} \qquad , \ (i = 1, ..., k-1)$$

$$\textit{and} \quad z_k = \frac{1}{\sum_{j=1}^{k-1} \exp(y_j) + 1}$$

The Jacobian of the additive logistic transformation is $(\prod_{i=1}^{k} z_i)^{-1}$. The inverse of this transformation is the additive logratio transformation (alr).

$$y_i = \log\left(\frac{z_i}{z_k}\right)$$

Denote the inverse of the alr transformation (i.e., the additive logistic transformation of Definition 4) by alr$-1(\cdot)$.

Definition 5 *A k–part composition $\mathbf{z}$ has a logistic normal distribution, denoted $L^{k-1}(\boldsymbol{\mu}, \Sigma)$, when $\mathbf{y} = (y_1, ..., y_{k-1})$ has a $(k-1)$–dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ.*

The density function for $L^{k-1}(\boldsymbol{\mu}, \Sigma)$ is written as follows: For $\mathbf{z} \in \nabla^{k-1}$

$$f(\mathbf{z} \mid \boldsymbol{\mu}, \Sigma) = \left(\frac{1}{2\pi}\right)^{\frac{k-1}{2}} \mid \Sigma \mid^{-\frac{1}{2}} \left(\frac{1}{\prod_{i=1}^{k} z_i}\right) \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})' \Sigma^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu})\right]$$

where

$$\boldsymbol{\theta} = \text{alr}(\mathbf{z}) = \log\left(\frac{\mathbf{z_{-k}}}{z_k}\right)$$

and $\mathbf{z_{-k}} = (z_1, z_2, ..., z_{k-1})'$. The i^{th} element μ_i of $\boldsymbol{\mu}$ can be interpreted as $E\{\log(z_i/z_k)\}$, and the $(i,j)^{th}$ element σ_{ij} of Σ as $\text{cov}\{\log(z_i/z_k), \log(z_j/z_k)\}$. Hence, $\boldsymbol{\mu}$ and Σ are the mean vector and covariance matrix for alr$(\mathbf{z})$ (i.e., the multivariate logit) which follows a multivariate normal distribution.

To aid interpretation, the location parameter $\boldsymbol{\mu}$ can be expressed as a composition via the additive logistic transformation. That is,

$$\text{alr}^{-1}(\boldsymbol{\mu}) = \boldsymbol{\xi} \ , \ \text{where} \quad \boldsymbol{\xi} \in \nabla^{k-1}.$$

As a point on the simplex, this value is directly interpretable as a composition. This is much simpler to interpret than $\boldsymbol{\mu}$, a multivariate vector of expected logits. The inverse additive

logratio transform does not preserve the mean and mode properties of $\boldsymbol{\mu}$ for multivariate normal logits. However, the inverse additive logratio transform is monotone in each of the $k-1$ components of $\boldsymbol{\mu}$. As a consequence, $\boldsymbol{\xi} = \text{alr}^{-1}(\boldsymbol{\mu})$ can be interpreted as a component-wise multivariate median for the logistic normal distribution in ∇^{k-1}. Finally, Aitchison (1986, section 5.5, pp. 93–96) shows that the logistic normal density is invariant to permutations of the components of the composition vector $\mathbf{z}$. Thus, the density, and subsequently any inference based on the density, is not affected by the ordering of groups in $\mathbf{z}$.

5.2 Conditional autoregressive spatial model

The logistic normal model, in conjunction with the conditional multinomial observation model, is used to describe the variability between samples from a given site. We incorporate spatial structure between sites by specifying a Markov random field for the prior distribution of logistic normal model parameters. We use a conditional autoregressive model (CAR; Besag, 1974; Mardia, 1988) to construct the prior distribution. Mardia (1988) describes the theoretical background for a multivariate normal Markov random field specification. We briefly review Mardia's result and outline the method of implementation. For full technical details, we refer the interested reader to Billheimer and Guttorp (1995).

Typically, a CAR model is specified via the conditional distribution of the observation at site j, given all of the other sites. We let $\mathbf{x}_j$ denote a p-variate observation at site j, where j indexes sites on a regular spatial lattice, $j = 1, 2, ..., n$. The mean parameter at site j given all other sites is

$$E\{\mathbf{x}_j \mid \mathbf{x}_{-j}\} = \boldsymbol{\mu}_j + \sum_{r \in \delta j} \Lambda_{jr} \, (\mathbf{x}_r - \boldsymbol{\mu}_r)$$

where δj is the set of neighbors of site j, and $\mathbf{x}_{-j}$ denotes the observationss of all sites except site j. The conditional variance matrix for $\mathbf{x}_j$ given $\mathbf{x}_{-j}$ is

$$\text{Var}\,(\mathbf{x}_j \mid \mathbf{x}_{-j}) = \Gamma_j$$

Note that Γ_j and Λ_{jr} are $(k-1) \times (k-1)$ matrices, and Γ_j is positive definite for all j. Assuming $\mathbf{x}_j \mid \mathbf{x}_{-j}$ is conditionally multivariate normal for all n sites, Mardia's result (1988) shows the joint distribution of $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ is np multivariate normal with mean vector

$$\boldsymbol{\mu}^{'} = (\boldsymbol{\mu}_1^{'}, \boldsymbol{\mu}_2^{'}, ..., \boldsymbol{\mu}_n^{'})$$

and variance matrix

$$\Sigma = \{\text{Block}(-\Gamma_j^{-1}\Lambda_{jr})\}^{-1}$$

provided $\Lambda_{jr}\Gamma_r^{'} = \Gamma_j\Lambda_{rj}^{'}$ (for symmetry of Σ), and $\text{Block}(-\Lambda_{jr})$ is positive definite (define $\Lambda_{jj} = -I_{k-1}$). The term "Block" refers to a large matrix comprised of sub-matrices, each of dimension $(k-1) \times (k-1)$, where the $(j,r)^{th}$ sub-matrix of the large matrix is $-\Gamma_j^{-1}\Lambda_{jr}$. (Note that in the symmetry condition we correct a typographic error in Mardia, 1988.)

Mardia shows that the form of $\mid \Sigma \mid$ can be simplified to

$$\mid \Sigma \mid^{-\frac{1}{2}} = \left(\prod_{j=1}^{n} \mid \Gamma_j \mid \right)^{-\frac{1}{2}} \mid \text{Block}\,(-\Lambda_{jr}) \mid^{\frac{1}{2}} .$$

(Again, this expression corrects a typographic error in Mardia, 1988.) The spatial model for species compositions uses this multivariate normal as the prior distribution for the location parameters for the logistic normal distributions.

5.3 Covariates

To incorporate the effect of covariates into the model, the location parameter, $\boldsymbol{\mu}$, may depend on explanatory variables. For a scalar covariate x_j measured at site j, $\boldsymbol{\mu}_j$ can be replaced in the density expression by $\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1(x_j - \bar{x})$. Here, $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are vectors in $\Re^{k-1}$, and $\bar{x}$ is the

mean of the observed covariate values. This parameterization allows interpretation of $\boldsymbol{\beta}_0$ as the overall location, and $\boldsymbol{\beta}_1$ as the change in location for a unit increase in x. Equivalently, the regression expression $\boldsymbol{\mu}_j = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1(x_j - \bar{x})$ can be written as a perturbation of compositions. This is accomplished by taking the inverse additive logratio transformation of both sides,

$$\text{alr}^{-1}(\boldsymbol{\mu}^j) = \text{alr}^{-1}(\boldsymbol{\beta}_0) \circ \text{alr}^{-1}(\boldsymbol{\beta}_1)^{(x_j - \bar{x})}.$$

This equation is more conveniently written in the following form,

$$\boldsymbol{\xi}_j = \boldsymbol{\xi} \circ \boldsymbol{\gamma}^{u_j}$$

where $\boldsymbol{\xi}_j = \text{alr}^{-1}(\boldsymbol{\mu}^j)$, $\boldsymbol{\xi} = \text{alr}^{-1}(\boldsymbol{\beta}_0)$, $\boldsymbol{\gamma} = \text{alr}^{-1}(\boldsymbol{\beta}_1)$, and $u_j = x_j - \bar{x}$. In this parameterization, $\boldsymbol{\xi}$ is the overall location on the simplex. Now the role of the regression composition parameter, $\boldsymbol{\gamma}$, is clear: the location parameter for site j is the overall location ($\boldsymbol{\xi}$) perturbed by $\boldsymbol{\gamma}$ (for $u_j = 1$). Thus the effect of the covariate, $\boldsymbol{\gamma}$, is directly interpretable as a composition. It is the amount by which a location is shifted by a unit increase in the covariate, via a perturbation. Finally, deviations in $\boldsymbol{\gamma}$ from the identity composition, $\mathcal{I}_{k-1} = (1/k, 1/k, ..., 1/k)$ indicate the direction and magnitude of the change. Through this parameterization and the perturbation operator, regression parameters can be interpreted by their effect on compositions. This is more easily interpretable than the alternative version on the log-odds scale that results from the additive logratio transform.

5.4 Implementation for Delaware Bay benthic composition

To implement the model developed above as the prior distribution for the Delaware Bay data, several simplifying assumptions are made. First, because sites may have different numbers of neighbors (from 1 to 6 "first order" neighbors), assume that the prior conditional variance

at site j depends on the number of neighbors as follows:

$$\Gamma_j = \frac{1}{n_j}\,\Gamma$$

where n_j is the number of neighbors of site j. The site composition (adjusted for the covariate) is predicted with greater precision as the number of neighbors increases. The matrix Γ describes the relative variability and covariance relationships between the different groups (given the neighboring sites). This assumption provides a mechanism for allowing increased variability at "edge" sites.

Combining this assumption with the symmetry condition $\Lambda_{jr}\Gamma_r^{'} = \Gamma_j\Lambda_{rj}^{'}$ implies that Λ_{jr} can be simplified to the following form

$$\Lambda_{jr} = \begin{cases} \Lambda_j & \text{if} \quad r \in \delta j \\ -I_{k-1} & \text{if} \quad r = j \\ 0_{(k-1)\times(k-1)} & \text{otherwise.} \end{cases}$$

As a further simplification, we assume that $\Lambda_j = \lambda/n_j \;\; I_{k-1}$. This means that the spatial dependence between neighboring sites is the same for all k groups of organisms (actually the same for all pairs of logits $\log([\mathbf{z}_j]_i/[\mathbf{z}_j]_k)$ and $\log([\mathbf{z}_r]_i/[\mathbf{z}_r]_k)$). Further, the diagonal structure of Λ_j implies that $\log([\mathbf{z}_j]_i/[\mathbf{z}_j]_k)$ and $\log([\mathbf{z}_r]_m/[\mathbf{z}_r]_k)$ are conditionally independent, given all other logits at all other sites. Note that the final assumption, $\Lambda_{jr} = \lambda/n_j\, I_{k-1}$ (when site r is a neighbor of site j), combined with $\Gamma_j = \Gamma/n_j$ implies that the spatial dependence is the same for all neighbor pairs, regardless of direction. With the limited number of sites available, we did not consider it feasible to attempt more elaborate spatial dependence structures.

We consider the state composition for the t^{th} sample ($t = 1, 2, ..., T_j$) at site j ($j = 1, 2, ..., n$), $\mathbf{z}_{jt}$, to be a (unobservable) realization from a logistic normal distribution. The location parameter for this distribution is comprised of a CAR multivariate normal spatial

process ($\boldsymbol{\theta}_j$) plus the effect of a (centered) covariate at site j ($\boldsymbol{\beta} u_j$). That is,

$$z_{jt} \sim L_{k-1}\left(\boldsymbol{\theta}_j + \boldsymbol{\beta}\, u_j,\ \Psi\right)$$

where $\boldsymbol{\beta}$ is the regression parameter vector describing the effect of the covariate, and Ψ describes the within site variance-covariance structure.

Expressions for the observation density (likelihood) and prior distributions complete the model specification. The observed group counts are assumed conditionally multinomial given the unobservable site composition, $\mathbf{z}_{jt}$.

$$p(\mathbf{y}_{jt} \mid \mathbf{z}_{jt}, \sum_{i=1}^{k}[y_{jt}]_i) \;=\; \frac{\left(\sum_{i=1}^{k}[y_{jt}]_i\right)!}{\prod_{i=1}^{k}[y_{jt}]_i!} \prod_{i=1}^{k} [\mathbf{z}_{jt}]_i^{[y_{jt}]_i}$$

where $[\cdot]_i$ denotes the i^{th} component of the vector.

Prior distributions are required for $\lambda,\ \boldsymbol{\beta},\ Q = \Gamma^{-1}, R = \Psi^{-1}$ and $\boldsymbol{\mu}$, the overall level of the spatial process. We assume the following prior distributions:

$$\pi(\lambda) = \text{Uniform}(-1,\ 1)$$

$$\pi(\boldsymbol{\beta}) = N_{k-1}\left(\,0_{k-1},\, a\mathcal{N}\right)$$

$$\pi(\boldsymbol{\mu}) = N_{k-1}\left(\,0_{k-1},\, b\mathcal{N}\right)$$

$$\pi(Q) = \text{Wishart}(\,[c\mathcal{N}]^{-1},\, \rho_1)$$

$$\pi(R) = \text{Wishart}(\,[d\mathcal{N}]^{-1},\, \rho_2)$$

where $\mathcal{N} = I_{k-1} + \mathbf{j}_{k-1}\mathbf{j}_{k-1}'$. Here I_{k-1} is an identity matrix of dimension $(k-1)$, and $\mathbf{j}_{k-1}$ is a $(k-1)$ vector of ones. Typical choices for a, b, c, and d are $a = b = c = d = 1$. These values specify proper, but diffuse, prior distributions for $\boldsymbol{\beta}$, and $\boldsymbol{\mu}$. Their alr transformed location parameters are centered at $\mathcal{I}_{k-1}$. The prior distributions for Q and R are centered at the "null" precision matrix (i.e., compositions formed from independent bases; see Billheimer and Guttorp, 1995, for details). The hyperparameters ρ_1 and ρ_2 must be at least $(k-1)$ to make $\pi(Q)$ and $\pi(R)$ proper distributions.

5.5 Markov chain Monte Carlo implementation

MCMC is used to obtain a Markov chain realization from the joint posterior distribution. The algorithm updates $\mathbf{z}$'s, $\boldsymbol{\theta}$'s, $\boldsymbol{\mu}$, λ, $\boldsymbol{\beta}$, Q, and R each conditional on all other parameters (and on the data, $\mathbf{y}$). Hastings' algorithm (1970) for compositions, described in Billheimer and Guttorp (1995), is used to update the $\mathbf{z}$'s. The spatial dependence parameter, λ, is updated via a symmetric, uniform proposal density and Metropolis algorithm acceptance probability (Metropolis, et al., 1953). Gibbs updating (Geman and Geman, 1984) is used for all other model parameters. Details of the MCMC implementation are described in Billheimer and Guttorp (1995) .

6. Modeling results

The statistical model described in section 5 was used to analyze the benthic compositions of Delaware Bay. The model uses a spatial structure defining neighbors of station j as those stations (when present) at the vertices of a hexagon centered at j. Any hexagon with a "missing" vertex (i.e., no station) simply has fewer neighbors. For example (see Figure 1), station 20 has six neighbors, namely stations {4, 5, 17, 18, 21, 24}, while station 13 has only two neighbors, stations 11 and 10. In addition to spatial structure, the model includes salinity as a covariate (centered to have mean zero).

Inference about the site compositions, the spatial dependence parameter (λ), and the salinity regression parameter vector ($\boldsymbol{\beta}$) resulted from a MCMC run with a burn-in of 200 cycles, and a collection phase of 20,000 cycles. Graphical inspection of realizations and diagnostics evaluating MCMC performance (Raftery and Lewis, 1992, 1995) indicate that 20,000 cycles are adequate to evaluate the posterior distribution. The MCMC realizations suggest partial confounding of the salinity gradient with the spatial structure of the observations.

Such confounding makes separation of the salinity and spatial effects difficult.

Figure 4 about here

The point estimate and 95% credible regions for the salinity effect are shown in Figure 4. The point estimate for this composition is (0.34, 0.38, 0.28). The "no effect" regression composition, $\mathcal{I}_{k-1}$, falls just at the boundary of the 95% credible region. Because the vast majority of the estimated posterior density is displaced from $\mathcal{I}_{k-1}$, this result suggests an association between salinity and benthic composition. The point estimate can be interpreted in the following way: an increase in salinity of 1 ppt (part per thousand) has the effect of perturbing a benthic composition by (0.34, 0.38, 0.28) (over the observed range of 15–30 ppt salinity). This point estimate indicates that as salinity increases, the proportion of suspension feeders decreases and are replaced by pollution intolerant organisms. This result quantifies the earlier graphical interpretation of the association between salinity and benthic invertebrate composition in section 3.

***Figure 5 about here

The realized values of the spatial dependence parameter (λ) are shown in Figure 5. This figure suggests that there is spatial similarity between neighboring sites (i.e., $\lambda > 0$). The median value for the distribution is 0.60, while the observed mean is 0.63. The observed mode is about 0.80. Nearly 93% of the realized values are positive.

To evaluate further the evidence of spatial dependence, a Bayes factor was computed using the Savage density ratio (see Kass and Raftery, 1995 for a review). This ratio compares the prior density for λ with the posterior density; both evaluated at $\lambda = 0$ (spatial independence). A large value for the ratio indicates that the posterior density is shifted away from zero, and that the data provide evidence against spatial independence. The posterior density was

approximated using a kernel density estimator with the MCMC realizations of λ. Note that these realizations approximate the posterior distribution of λ integrated over all other parameters. The kernel estimator resulted in a value of 0.26 for the posterior density at $\lambda = 0$. The prior distribution for λ, Uniform(-1, 1), gives a prior density of 0.5. Hence, the Bayes factor is $0.5/0.26 = 1.9$. This value indicates moderate evidence of positive spatial dependence.

It is important to note that the spatial dependence and effect of salinity are estimated simultaneously. Salinity is a spatially varying covariate that (generally) increases along the gradient from river to ocean across the estuary. The observed spatial dependence is present while the effect of salinity is included in the statistical model. Thus, λ denotes spatial dependence beyond that explained by the salinity gradient.

We assess model adequacy for describing within site and between site variability. To evaluate within site variability, we omit from the data one randomly selected sample from each of the 25 sites. The remaining samples at each site (in conjunction with the statistical model) are used to construct 95% prediction regions for the omitted compositions. Figure 6 shows the results of this prediction.

***Figure 6 about here

The omitted data are well predicted by the statistical model. All hold-out samples with benthic invertebrates (24 of 25) exhibited compositions inside the prediction regions. The sample from one site (site #12) had no tolerant, intolerant or suspension feeding organisms in the sample. Hence, there is no observed composition to check the prediction.

To assess model adequacy for between site variability, all data from a given site were omitted, and prediction regions constructed for the benthic composition at that site. These regions were constructed via the MCMC algorithm by replacing the benthic counts for all

samples at site j with zeros. Benthic counts at other sites were unchanged. The zero counts maintain the neighborhood structure for site j, and allow its composition to be updated as a regular part of the MCMC algorithm. This is the recommended method for accommodating missing observations for MCMC (Besag et al., 1995). A 95% prediction region was constructed from the MCMC realizations for the hold-out site composition. Once this region was defined, a multinomial random vector with sample size equal to the median number of organisms for the omitted site was generated. A single multinomial vector was constructed for each MCMC realization in the region. Finally, a convex hull circumscribing the composition of the multinomial vectors was used to construct the 95% prediction region for the omitted benthic composition.

Figures 7 and 8 show the 95% prediction regions for sites 17 and 20. These sites were randomly selected from the five sites {5, 17, 20, 21, 22} having six neighbors (all other sites had 5 or fewer neighbors).

Figure 7 about here

Figure 8 about here

These figures indicate that the spatial regression model adequately predicts compositions at sites with omitted data. The observed benthic compositions from all samples fall in their respective prediction region for each of the sites.

We also use the statistical model to predict the composition at site #23. Recall that this site was identified as "ecologically disturbed" in the exploratory analysis. Benthic counts from this site were withheld from the data, and a 95% prediction region for the sample composition was constructed. These results are shown in figure 9.

Figure 9 about here

The figure shows that the 95% prediction region covers a large portion of the ternary diagram. This large region is due in part to the relatively small number of neighbors of site 23 (4 neighbors), and the large differences in the observed compositions at these neighbor sites. In spite of the large area of coverage, the observed sample compositions at site #23 are not contained in the prediction region. The observed compositions exhibit a greater proportion of pollution tolerant organisms, and smaller proportions of intolerant and suspension organisms than would be expected at this site. This result supports our contention of a local disturbance near site #23.

We deduce that the statistical model is a useful description of baseline variability of benthic population composition in the Delaware Bay. In subsequent work we will examine how data from later years can be interpreted relative to this baseline measure.

References

[1] Aitchison, J. (1982). "The statistical analysis of compositional data (with discussion)." *J. R. Statist. Soc. B.,* **44,** 139–177.

[2] Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman & Hall, New York.

[3] Aitchison, J. and Shen, S. M. (1980). "Logistic-normal distributions: some properties and uses." *Biometrika,* **67,** 261–272.

[4] Aschan, M. (1990). "Changes in softbottom macrofauna communities along environmental gradients." *Ann. Zool. Fennici* **27:** 329–336.

[5] Besag, J. E. (1974). "Spatial interaction and the statistical analysis of lattice systems" (with Discussion). *J. R. Statist. Soc. B,* **36,** 192–236.

[6] Besag, J. E., Green, P. J., Higdon, D. M. and Mengersen K. (1995). "Bayesian computation and spatial systems" (with Discussion). *Statist. Sci.,* **10,** 3–66.

[7] Billheimer, D. D. and Guttorp, P. (1995). "Spatial statistical models for discrete compositional data". Technical Report, Dept. of Statistics, University of Washington, Seattle.

[8] Dennis, B., Patil, G. P., and Rossi, O. (1979). "The sensitivity of ecological diversity indices to the presence of pollutants in aquatic communities." In *Environmental Biomonitoring, Assessment, Prediction, and Management,* (ed. Cairns, J. Jr., Patil, G. P., Waters, W. E.), pp. 379–413, International Cooperative Publishing House, Burtonsville, MD.

[9] Deegan,, L. A., Finn, J. T., Ayvasian, S. G., and Ryder, C. (1993): "Feasibility and application of the index of biotic integrity to Massachusetts estuaries (EBI)" Final Project Report to Massachusetts Executive Office of Environmental Affairs, Department of Environmental Protextion, North Grafton, MA.

[10] Fore, L. S., Karr, J. R., and Wisseman, R. W. (1995). "A benthic index of biotic integrity for streams in the Pacific northwest." Submitted for publication.

[11] Guttorp, P. (1993). "Statistical analysis of biological monitoring data." In G.P. Patil, C.R. Rao (editors): *Multivariate Environmental Statistics*, 165–174. Amsterdam: North-Holland.

[12] Geman, S. and Geman, D. (1984). "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images." *IEEE Trans. Pattern Anal. Machine Intell.,* **6,** 721-741.

[13] Gray, J.S. and Pearson, T.H. (1982). "Objective selection of sensitive species indicative of pollution-induced change in benthic communities. I. Comparative methodology", *Mar. Ecol. Prog. Ser.*, **9**, 111–119.

[14] Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika* **57:**97–109.

[15] Karr, J.R. 1981. "Assessment of biotic integrity using fish communities." *Fisheries* **6***:*21–27.

[16] Karr, J.R. 1995. "Ecological integrity and ecological health are not the same." In P. Schulze (ed.): *Engineering within ecological constraints.*National Academy of Engineering. Washington: National Academy Press.

[17] Kass, R. E. and Raftery, A. E. (1995). "Bayes factors." *J. Amer. Statist. Assoc.,* **90,** 773–795.

[18] Mardia, K. V. (1988). "Multidimensional multivariate Gaussian Markov random fields with applications to image processing." *J. Multivariate Anal.* **24,** 265–284.

[19] Marmorek, D. R., Bernard, D. P., and Ford, J. (1988). "Biological monitoring for acidification effects: U.S.–Canadian workshop." *U.S. Environmental Protection Agency Report.* Environmental Research Laboratory, U.S. Environmental Protection Agency. Corvallis, Oregon.

[20] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller E. (1953). "Equations of state calculations by fast computing machines." *J. Chemical Physics* **21,**108–1091.

[21] Overton, W. S., White, D., and Stevens, D. K. (1990). "Design Report for Emap: Environmental Monitoring and Assesment Program." EPA/600/3-91/053, U.S. Environmental Protection Agency, Washington, D.C.

[22] Patrick, R. (1972). "Aquatic communities as indices of pollution." in *Indicators of Environmental Quality,* (ed. Thomas, W. A.), pp. 93–100. Plenum Press, New York.

[23] Pollard, J.H. (1975): *Mathematical Models for the Growth of Human Populations.* Cambridge Univeristy Press, Cambridge.

[24] Raftery, A. E. and Lewis, S. M. (1992). "How many iterations in the Gibbs sampler?" in *Bayesian Statistics 4.* (ed. Bernardo, J., Berger, J., Dawid, A. P. and Smith, A. F. M.). Oxford University Press, pp. 765–776.

[25] Raftery, A. E. and Lewis, S. M. (1995). "The number of iterations, convergence diagnostics and generic Metropolis algorithms." In *Practical Markov Chain Monte Carlo* (ed. Gilks, W. R., Spiegelhalter, D. J. and Richardson, S.). Chapman & Hall, London.

[26] Schindler, D. W., Mills, K. H., Malley, D. F., Findlay, D. L., Shearer, J. A., Davies, I. J., Turner, M. A., Linsey G. A., and Cruikshank, D. R. (1985). "Long-term ecosystem stress: the effects of years of experimental acidification on a small lake." *Science,* **228,** 1395–401

[27] Schwinghamer, P. (1988). "Influence of pollution along a natural gradient and in a mesocosm experiment on biomass–size spectra of benthic communities",*Mar. Ecol. Prog. Ser.*, **46**, 199–206.

[28] Spellerberg, I. P. (1991). *Monitoring Ecological Change.* Cambridge University Press, Cambridge.

[29] Warwick, R.M., (1986). "A new method for detecting pollution effects on marine macrobenthic communities", *Marine Biology*, **92**, 557–562.

[30] Word, Jack Q., Meyers, B. L., and Mearns, A. J. (1977). "Animals that are indicators of marine pollution." In Annual report 1977, Coastal Water Research Project. El Segundo, California.

Table 1: Abundance of organisms at Delaware Bay in 1990

Family	Genus	Species	# Stations	Mean	sd	Min	Max
Capitellidae	*Mediomastus*	*ambiseta*	23	152.96	189.51	0.33	623.00
Tellinidae	*Tellina*	*agilis*	19	22.43	23.34	0.33	82.33
Scaphandridae	*Acteocina*	*canaliculata*	20	16.18	15.42	0.33	64.00
Spionidae	*Streblospio*	*benedicti*	19	14.70	20.35	0.33	80.33
Ampeliscidae	*Ampelisca*	*verrilli*	16	9.90	13.73	0.33	43.33
Goniadidae	*Glycinde*	*solitaria*	18	5.60	5.68	0.33	17.00
Capitellidae	*Heteromastus*	*filiformis*	15	5.33	5.87	0.33	19.33
Idoteidae	*Edotea*	*triloba*	20	4.75	10.57	0.33	39.33
Orbiniidae	*Leitoscoloplos*	*robustus*	17	3.24	3.28	0.33	11.33
Mactridae	*Mulinia*	*lateralis*	22	2.12	3.77	0.33	19.00
Diastylidae	*Oxyurostylis*	*smithi*	16	1.51	1.58	0.33	7.00

Table 2: Benthic conditions for the 25 sampling stations in 1990.

Covariate (unit)	Mean	sd	Min	Max
Dissolved oxygen (mg/l)	6.62	1.16	5.10	9.80
Temperature (^0C)	24.56	1.17	21.77	26.44
Salinity (ppt)	24.05	4.88	15.46	30.82
pH (pH units)	7.92	0.19	7.50	8.20
Light transmission (%)	46.58	17.99	1.00	76.00
Depth (m)	6.94	5.24	1.40	21.70

Figure legends

FIGURE 1. 1990 sampling stations in the Delaware Bay.

FIGURE 2. Star-plot of proportions of tolerant (up), intolerant (down to right) and suspension feeders (down to left). The length of each line corresponds to the proportion of the respective group at that sampling station. The lighter contour lines correspond to the gradient of salinity.

FIGURE 3. The data from Figure 2 shown in the simplex. The proportion of tolerant species, for example, can be read off an axis perpendicular to the bottom side, with 0 at the bottom side and 1 at the top apex, while the proportion of intolerant species is represented on an axis having 0 at the left side of the triangle and 1 at the lower left apex. The observations are coded with respect to salinity value at the station.

FIGURE 4. Point estimate and 95% credible region shown in the simplex for salinity regression parameter. The point estimate is (0.34, 0.38, 0.28). A covariate with no effect would have a point estimate falling in the center (1/3, 1/3, 1/3) of the simplex.

FIGURE 5. Histogram of MCMC realizations for the spatial dependence parameter, λ. The observed median of λ is 0.60, and the mode is about 0.80. The x indicates the mean of the realizations of 0.53.

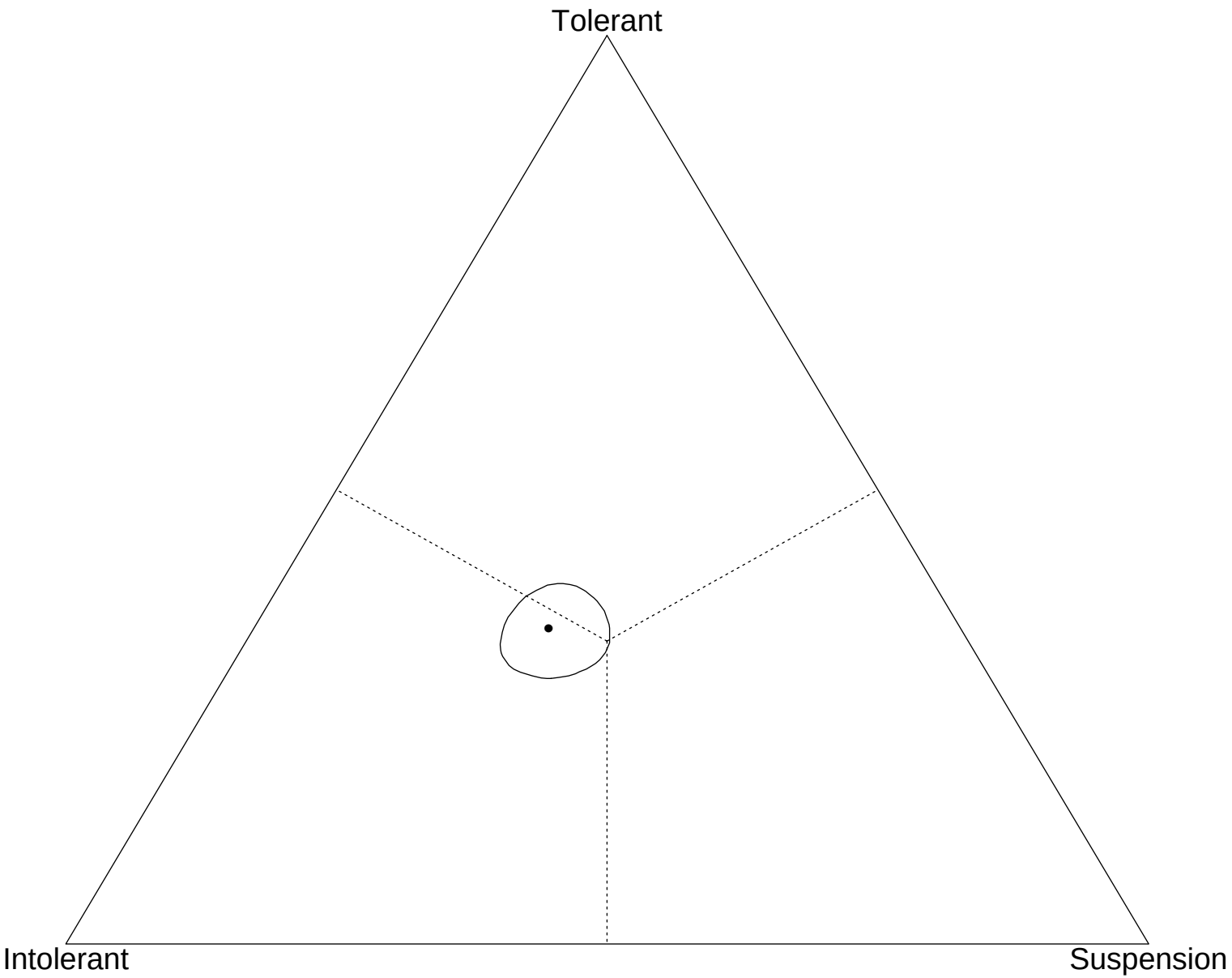
FIGURE 6. 95% prediction regions for compositions of hold-out samples for each of the sites. The dot corresponds to the observed composition of the hold-out sample.

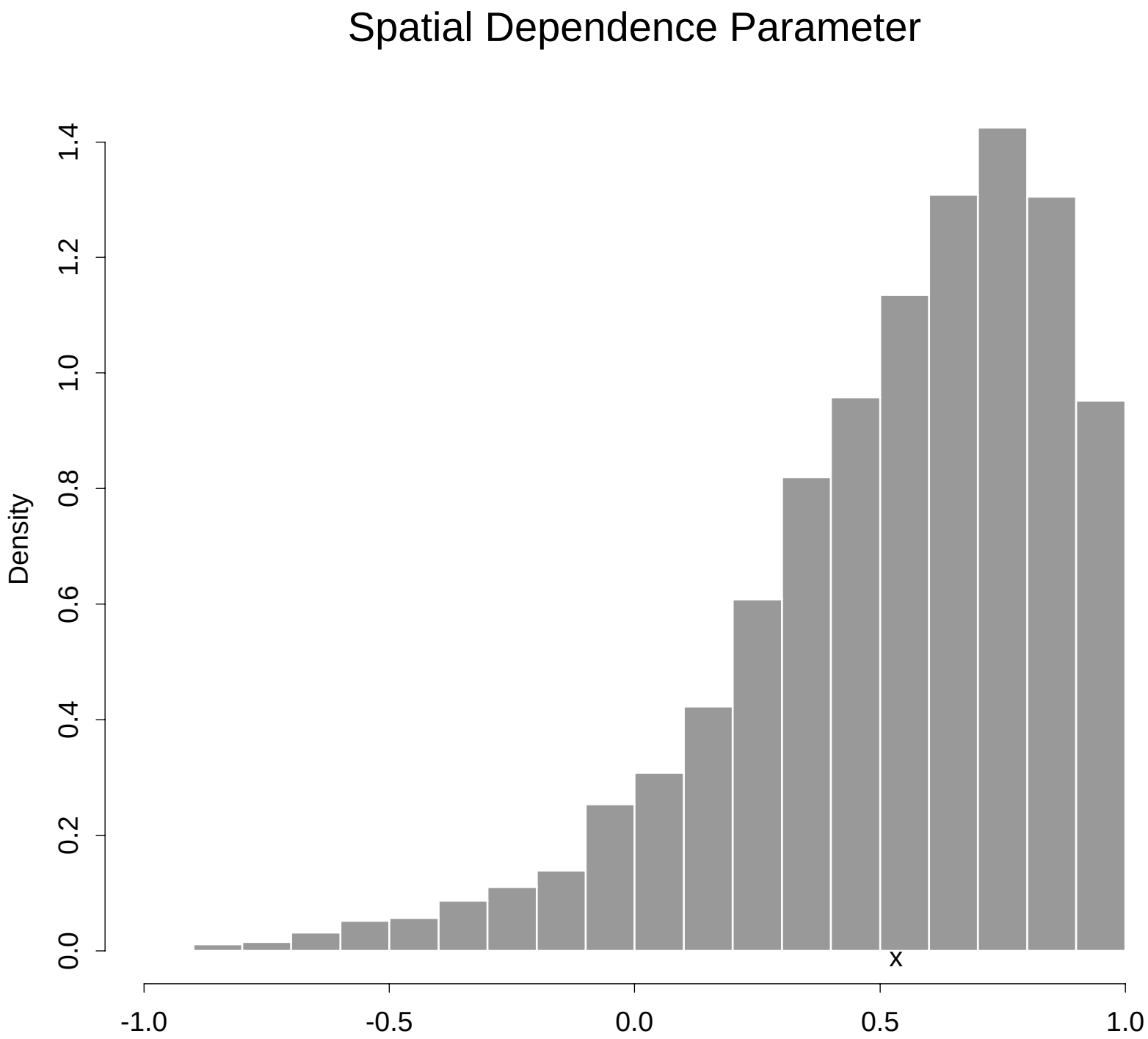
FIGURE 7. 95% prediction region for the composition at site 17 based on the remaining sites, leaving out site 17 data. The dots correspond to the observed sample values.

FIGURE 8. 95% prediction region for the composition at site 20 based on the remaining sites, leaving out site 20 data. The dots correspond to the observed sample values.

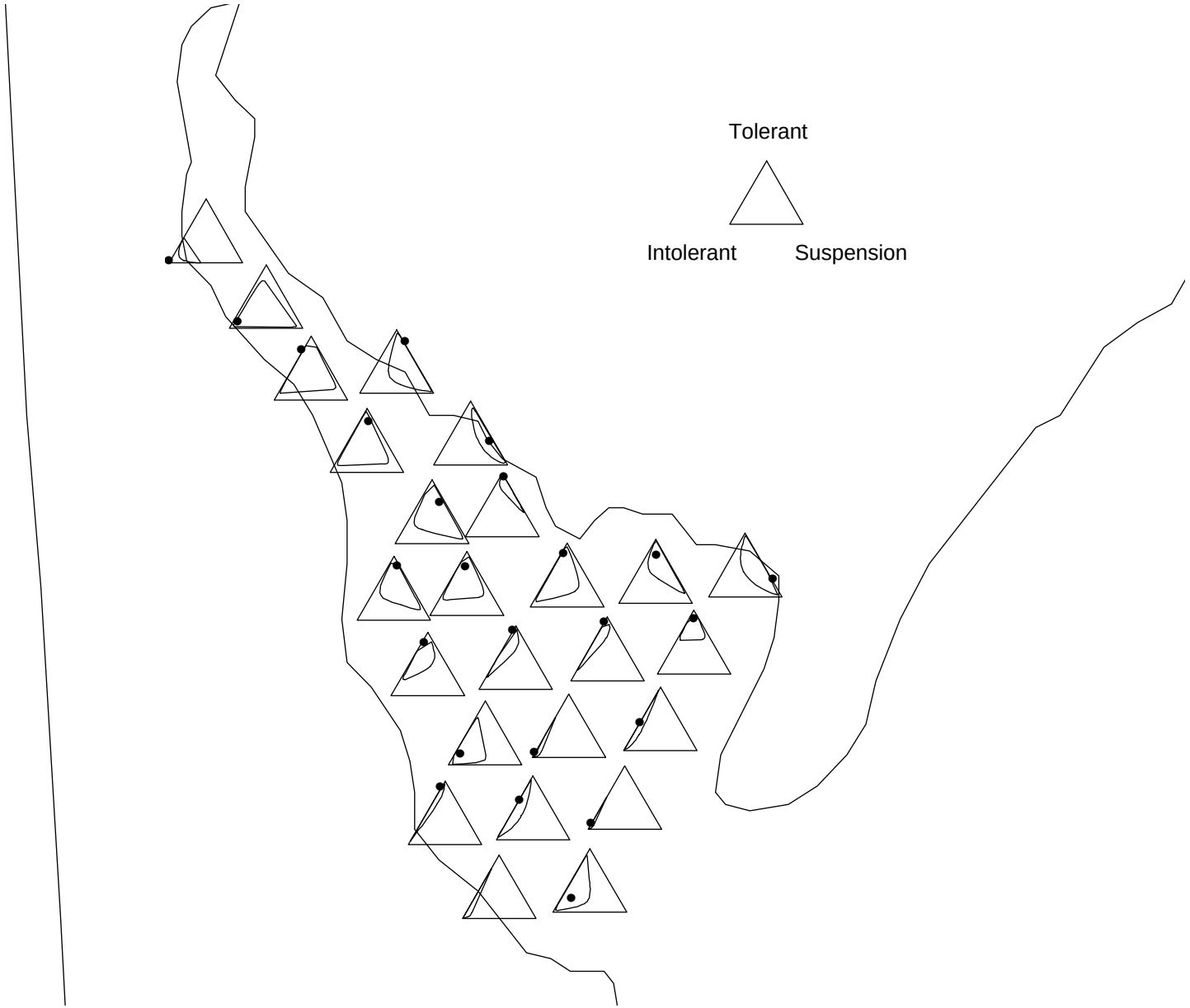
FIGURE 9. 95% prediction region for the composition at site 23 based on data at the remaining sites. Observed sample values outside the region suggest a disturbance near the site.

95% Credible Region for Salinity Regression Composition

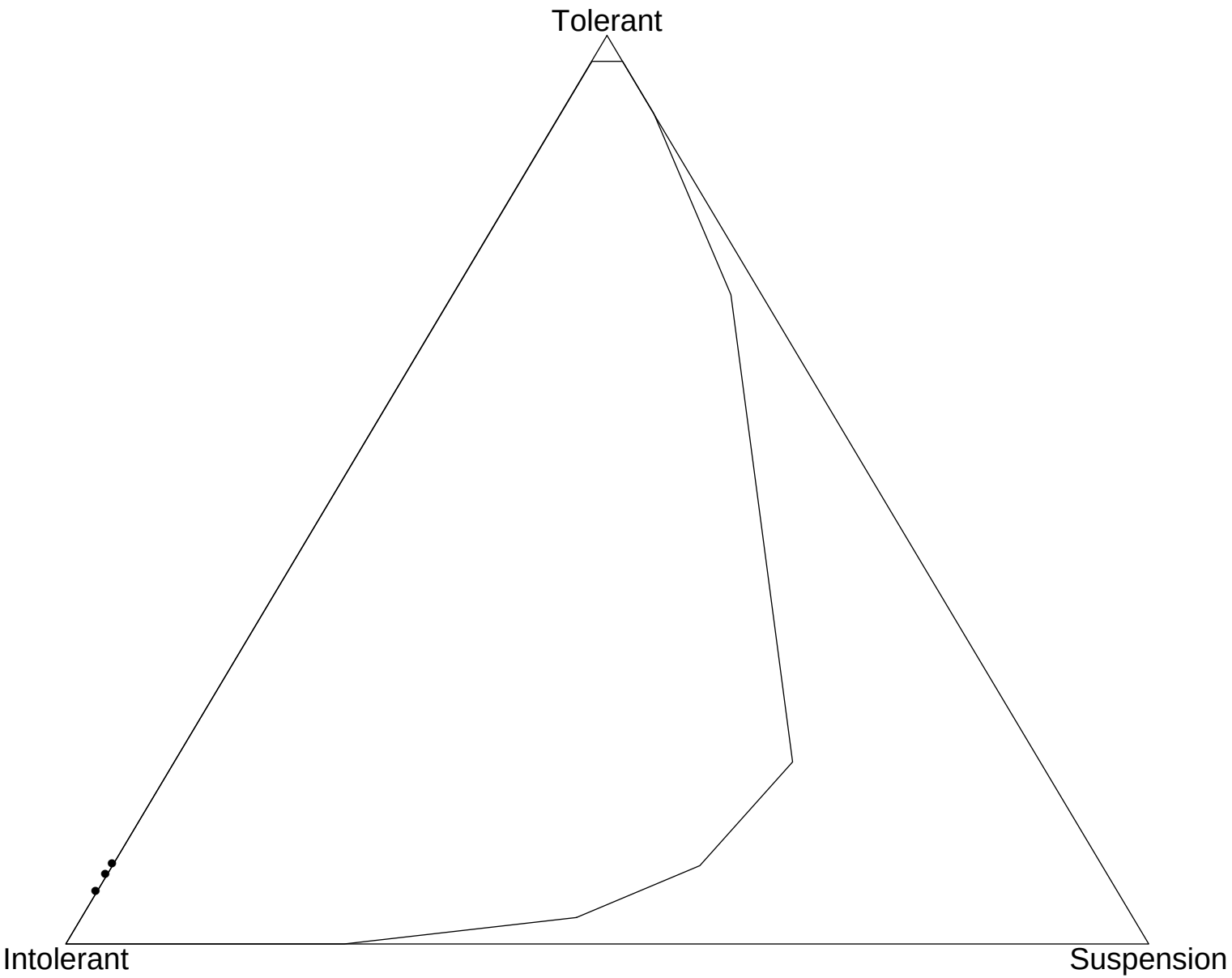



Spatial Dependence Parameter
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
-1.0
-0.5
0.0
0.5
1.0
X

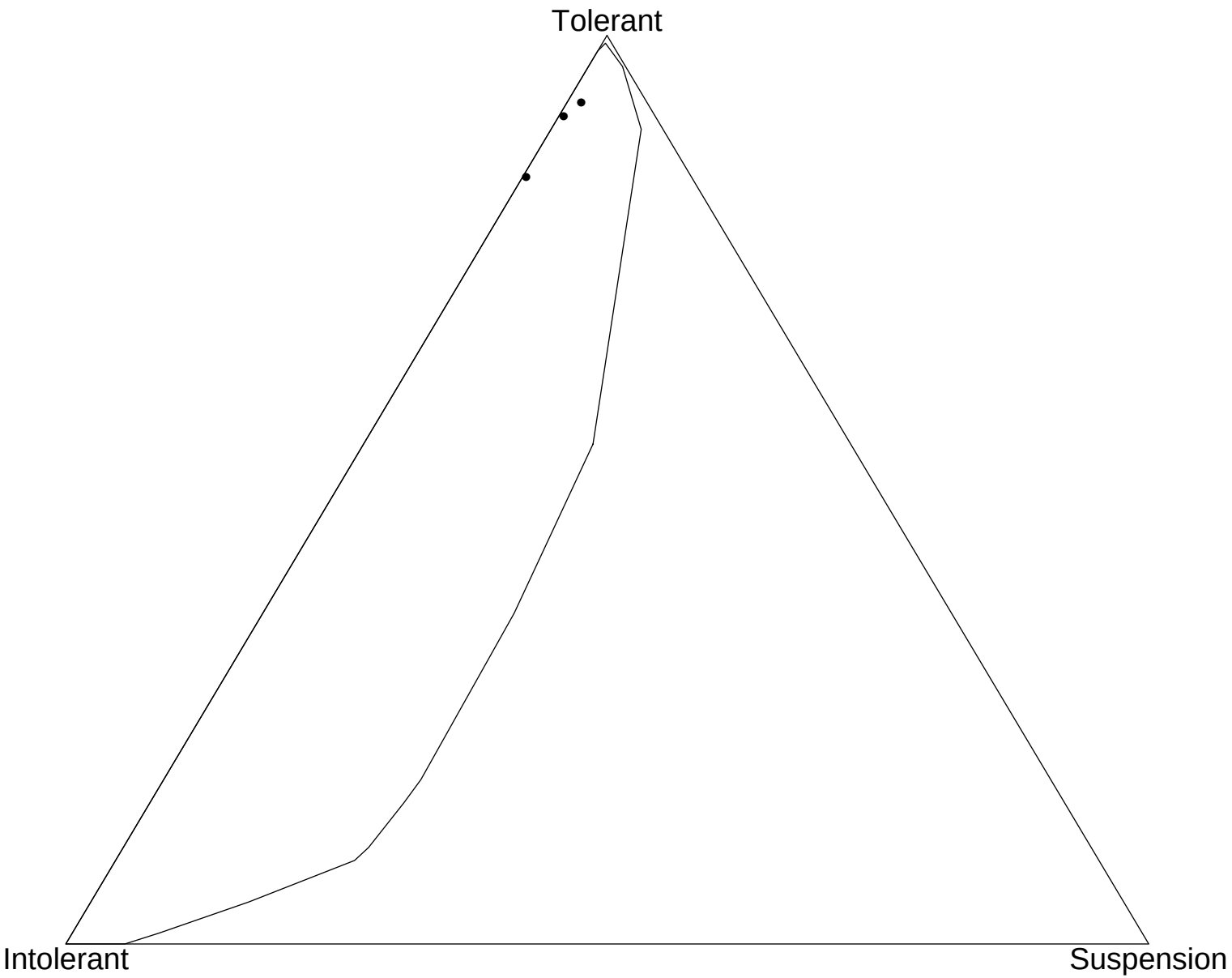
95% Prediction Regions for Hold-out Sub-Sample Compositions



95% Prediction Region Site 17



95% Prediction Region Site 20



95% Prediction Region Site 23