Advances in Systems Science and Applications (1995) **Special Issue I**, 7-12

Variance and Uncertainty Measures of Population Diversity Dynamics

MARK A. BEDAU*†, MARTIN ZWICK† AND ALAN BAHM‡

*Department of Philosophy, Reed College, U.S.A.
†Systems Science Ph.D. Program, Portland State University, U.S.A.
‡Photon Kinetics, Portland OR, U.S.A.

We define variance and uncertainty measures of population diversity. Both measures have precise decompositions that we can exploit in analysis of evolutionary dynamics. We discuss how these measures are related and how they how they can be observed in artificial and natural evolving systems.

1. Introduction

Evolving systems have a two-tier structure: a micro level consisting of individuals whose behavior is influenced by their genomes and whose interaction with their environment is governed by some explicit dynamics; and a macro level consisting of the population as a whole whose higher-level dynamics emerges statistically from the underlying micro processes. Following traditional population genetics, our approach to studying evolutionary dynamics is to define and observe macro-scale measures that reflect important aspects of a system's evolution (Packard 1989; Bedau and Packard 1992; Bedau, Ronneburg and Zwick 1992; Bedau and Bahm 1994; Bedau, Giger and Zwick 1995). We prefer measures that can be observed in a wide variety of artificial and natural settings, for this facilitates the search for universal features of evolutionary dynamics. We think that the important *stable* features of an evolving population are dynamical; evolution is a dynamic equilibrium. By studying these measures in numerical simulations of simple evolving systems, we can respect the context-dependence of fitness and other macro-level measures of evolution, we can explicitly treat many loci and many alleles per loci at the micro-level, and we can observe the *dynamical* properties of the measures.

Various kinds of *diversity* are interesting macro-scale measures (Bedau, Ronneburg and Zwick 1992); one is the diversity of the genetic information in the population. Population diversity is interesting partly because it reflects exactly that aspect of the system that evolution directly changes—the genetic structure of the population. Thus, the dynamics of population diversity is a central aspect of the intrinsic dynamics of evolution. While diversity itself is definable directly in terms of micro properties, its dynamics are in general unpredictable from the underlying micro dynamics and thus reflect an emergent property of the system.

In addition, population diversity is both cause and effect of a population's evolving interaction with its environment. For example, it is reasonable to suppose that a pop-

ISSN 1078-6236

ulation's level of diversity is both a response to the complexity of its environment and a predictor of the population's aggregate future performance, so one might hypothesize that relatively high levels of population diversity are associated with greater population fitness if and only if the population's environment is sufficiently complex to make survival difficult but not so complex that survival becomes a matter merely of chance.

In this paper we define two kinds of measures of population diversity. One set of measures is based on variance, the other is based on information-theoretic uncertainty. Our ultimate goal is to develop methods that facilitate quantitative comparison of diversity dynamics across a variety of artificial and natural systems.

2. Measures of Diversity

The variance and uncertainty measures of population diversity can be applied to evolutionary systems in which (i) the population (or subpopulation) of interest consists of individuals which share some number of genetic loci, and (ii) at each locus each individual possesses one of some number of possible alleles shared across the population. The uncertainty measures make no assumptions about the relationships among the the loci or the alleles. The variance measures, however, can be defined only if the alleles at each locus share some common metric and this common metric applies univocally across all loci. To make the discussion of the variance measures more concrete, we will make two further assumptions: (iii) each locus genetically encodes a type of behavior that is triggered whenever a given type of local environmental condition is sensed (i.e., the genome is a sensorimotor mapping), and (iv) each behavior type is a specific magnitude of (x, y) displacement in the two-dimensional local environment. Under these assumptions, all the alleles at all the loci share a spatial metric, since there is a spatial distance between the behaviors encoded by any two alleles. For examples of systems in which these assumptions hold, see Bedau and Packard 1992; Bedau, Ronneburg and Zwick 1992; Bedau and Bahm 1994; and Bedau, Giger and Zwick 1995.

2.1. Variance as a Measure of Diversity

To reflect metric information about the quantitative similarity of a population's alleles, we define total diversity as the mean squared deviation between the average movement of the whole population, averaged over all individuals and over all loci, and the individual movements of particular individuals encoded at particular loci, i.e.,

$$D = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} [(x_{ij} - \bar{x}^{IJ})^2 + (y_{ij} - \bar{y}^{IJ})^2] \tag{2.1}$$

where I is the number of individuals i, J is the number of loci j, (x_{ij}, y_{ij}) is the movement vector of individual i encoded at locus j, and $(\bar{x}^{IJ}, \bar{y}^{IJ})$ is the displacement of the population averaged over all individuals i and loci j.

Diversity D is naturally decomposable into different components. We collect alleles into "groups" and measure diversity both within and between groups, as is done in the analysis of variance (Iversen and Norpath 1976). D is then the sum of within-group and between-group diversity components. We define a "group" as the set of alleles in the population for a particular locus (Bedau, Ronneburg and Zwick 1992; Bedau and Bahm

1994).† Then total diversity D can be decomposed as follows:

$$D = W + B \tag{2.2}$$

where W is the within-locus diversity and B is the between-locus diversity. Formally, the components of the total diversity are defined as follows:

$$W = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}[(x_{ij} - \bar{x_j}^I)^2 + (y_{ij} - \bar{y_j}^I)^2] \tag{2.3}$$

$$B = \frac{1}{J}\sum_{j=1}^{J}[(\bar{x_j}^I - \bar{x}^{IJ})^2 + (\bar{y_j}^I - \bar{y}^{IJ})^2] \tag{2.4}$$

where $(\bar{x_j}^I, \bar{y_j}^I)$ is the displacement of the population in locus j averaged over all individuals i. These diversity measures are summarized in Table 1.

Imagine a plot collecting the spatial displacements (genetically encoded behaviors) of different individuals under different environmental conditions. This plot will show a collection of clusters of displacements, where each cluster is the set of populational responses to a particular environmental condition (i.e., the behavioral responses encoded a a particular locus). The within-locus diversity is the average spread within each cluster, which measures how similar are individual responses to the same environmental condition. The between-locus diversity is the spread between the centers of the different clusters, which measures the variability in the average populational response to different conditions.

2.2. Uncertainty as a Measure of Diversity

In some settings, alleles lack a common metric so it is impossible to define measures such as D, W, and B of eqs. (2.1), (2.3) and (2.4). To deal with this more general case, we can treat alleles merely as nominal variables. Where K is the number of all possible alleles, index each possible allele a by some variable, k. Uncertainty measures of genetic diversity are defined from the distribution of probability values $p(a_k)$ for the different k. We here adopt the uncertainty measure of information theory, which has a direct link to (likelihood-ratio) chi-square analysis as well as having some useful decompositional properties analogous to those of variance diversity above.

Analogous to the total variance diversity D, we define the total uncertainty diversity $\mathcal{D}$ as the information-theoretic uncertainty of the alleles, in terms of the usual Shannon entropy expression:

$$\mathcal{D} = H(a) = -\sum_{k=1}^{K} p(a_k)\ \log_2\ p(a_k) \tag{2.5}$$

(We shall refer to this as "uncertainty" to avoid contributing to the commonly made error of assuming that Shannon "entropy" has any necessary connection to the Second Law of Thermodynamics.) This expression is more commonly encountered as a measure of ecosystem diversity (Margalef 1968, Wilson and Bossert 1971), where $p(a_k)$ is the

† One can also define a group as the set of behavioral responses of an individual, and which yields the decomposition $D = W_b + B_b$, where W_b and B_b are the within- and between-individual diversities. We will not discuss these diversities here; see Bedau, Ronneburg and Zwick 1992.

probability that any biological individual is a member of species k. In our populational use of this expression, a distribution over alleles is used instead.

$\mathcal{D}$ is the total uncertainty of what a particular agent will do (where it will move to) under the action of a particular allele at a particular locus. That is, if an allele is randomly selected at any locus of any individual, $\mathcal{D}$ is the uncertainty about what that allele would be. The uncertainty arises from two sources: variation in the responses of different agents encoded at a given locus, and variations in the loci.

We decompose $\mathcal{D}$ into within- and between-locus uncertainties, as follows:

$$\mathcal{D} = \mathcal{W} + \mathcal{B} \tag{2.6}$$

where $\mathcal{W} = H(a|l)$ and $\mathcal{B} = I(a{:}l)$. $H(a|l)$ is the uncertainty of the allele a, given the particular locus l, i.e., the within-locus (allelic) uncertainty. $I(a{:}l)$ is mutual information (i.e., also called the information-theoretic "transmission") between locus specification and allele specification, i.e., the amount of allele uncertainty associated with locus uncertainty and thus removed by specifying the locus (i.e., environmental condition) of interest, in short, the between-locus (allelic) uncertainty.

Operationally, given $\mathcal{D}$, its components are evaluated by directly measuring two other quantities: the uncertainty $H(l)$ of the alleles, i.e., the uncertainty about the locus to be selected,

$$H(l) = -\sum_{j=1}^{J} p(l_j) \ \log_2 \ p(l_j) \tag{2.7}$$

and the joint uncertainty $H(a, l)$ of the alleles and loci,

$$H(a, l) = -\sum_{j=1}^{J}\sum_{k=1}^{K} p(l_j, a_k) \ \log_2 \ p(l_j, a_k) \tag{2.8}$$

where, as in the previous section, J is the number of loci. ¿From these quantities, we can calculate the within- and between-locus uncertainties as follows:

$$\mathcal{W} = H(a|l) = H(a, l) - H(l) \tag{2.9}$$

$$\mathcal{B} = I(a{:}l) = H(a) - H(a|l) = H(a) - H(a, l) + H(l) \tag{2.10}$$

It is also possible to define these quantities directly in terms of probabilities, as follows:

$$\mathcal{W} = -\sum_{j=1}^{J} p(l_j) \sum_{k=1}^{K} p(a_k|l_j) \ \log_2 \ p(a_k|l_j) \tag{2.11}$$

$$\mathcal{B} = \sum_{j=1}^{J}\sum_{k=1}^{K} p(l_j, a_k) \ \log_2 \ \frac{p(l_j, a_k)}{p(l_j)p(a_k)} \tag{2.12}$$

These uncertainty measures are summarized in Table 1.

Since, according to the assumption (iii) made at the outset of Section 2, the loci l_j correspond to the environments that *could exist* in principle, the distribution of $p(l_j)$ is flat, with $p(l_j) = \frac{1}{J}$. Under these conditions, $H(l)$ is constant and maximal; specifically, $H(l) = \log_2 J$. Hence, our uncertainty calculations treat the population independently of the environments that actually exist in the world, or, yet more specifically, the en-

Table 1. Summary of variance and uncertainty measures of diversity.

	Name	Visualization
D	Total variance	Variance of all alleles at all loci.
W	Within-locus variance	Variance of the alleles at a given locus, averaged over all loci.
B	Between-locus variance	Variance of the *average* allele value for different loci.
$\mathcal{D}$	Total uncertainty	Uncertainty of all alleles at all loci.
$\mathcal{W}$	Within-locus uncertainty	Uncertainty of alleles at a given locus, averaged over all loci.
$\mathcal{B}$	Between-locus uncertainty	Mutual information between alleles and loci.

vironments that the agents actually encounter, and thus measure the diversity of the population's *potential* behavior.

One could alternatively interpret the l_j as the loci that are actually used, i.e., the environments that are *encountered*, in which case $H(l)$ would measure the uncertainty of the *encountered* environments and $H(a)$ would measure the uncertainty of the population's *actual* behavior. On this interpretation, $p(l_j)$ would not be flat and $H(l)$ could vary in time, becoming in fact a measure of environmental diversity (variability). A third variant would be to base the definition of $H(l)$ on an interpretation of l_j as the environments that exist in the entire world, irrespective of the frequency with they are actually encountered by the agents. Bedau, Giger and Zwick (1995) study these alternative interpretations of l_j, and Bedau (1994) exploits these alternative interpretations of $H(l)$.

2.3. Comparisons of Variance and Uncertainty Measures

We should note that, in the special case of gaussian distributions, there is an analytical relationship between variance and information-theoretic uncertainty, and hence between our variance and uncertainty measures (Shannon and Weaver 1949, Garner and McGill 1956). However, this relationship has limited applicability since in general we cannot assume that distributions of behavioral responses are gaussian.

Consideration of some specific cases in which both variance and uncertainty diversity can be measured can illuminate their similarities and differences, and thus highlight the distinctive value of each.

To start with, consider two simple possibilities. At one extreme, the population consists of genetically identical individuals (i.e., "clones"). We have then, for variance diversity, $(x_{ij}, y_{ij}) = (\bar{x_j}^I, \bar{y_j}^I)$, for all i, or, for uncertainty diversity, $p(a_k|l_j) = 1$ for some $k = k'$ and 0 for all $k \neq k'$, for all j. In this situation, all of the total diversity shows up in the range of responses of this single genotype to different environmental conditions. Thus, we have $D = B$ and $W = 0$, as well as $\mathcal{D} = \mathcal{B}$ and $\mathcal{W} = 0$. At the other extreme, all alleles are present in the population with equal frequency at each locus, creating a flat distribution of alleles. We have, then, $(\bar{x_j}^I, \bar{y_j}^I) = (\bar{x}^{IJ}, \bar{y}^{IJ})$, for all j, or $p(a_k|l_j) = \frac{1}{K}$ for all j. In this case, the overlap between the responses at any two environments is total. All of the total diversity shows up in the range of individual responses to particular environmental conditions. In this case, we have $D = W$ and $B = 0$, as well as $\mathcal{D} = \mathcal{W}$ and $\mathcal{B} = 0$. Thus, in these two extreme conditions, variance and uncertainty diversity yield equivalent decompositions.

In general, though, variance and uncertainty diversity are not equivalent. For example, consider a "corner post" population in which, for each environmental condition, responses are equally distributed among movements to the four corners of output space. In this case, for every environmental condition, the clump of responses at each "corner post" is maximally distant from the populational mean response, which in located at the center of the output space. Thus, D and W are maximal; on the other hand, since each distribution is peaked at a small fraction of the possible values, $\mathcal{D}$ and $\mathcal{W}$ are relatively low. (As it happens, in this case, $B = \mathcal{B} = 0$). By contrast, the maximal value for $\mathcal{D}$ and $\mathcal{W}$ would occur in the flat distribution. There is thus no general equivalence between the variance and uncertainty measures for total and within-locus diversity.

Similarly, B and $\mathcal{B}$ might well differ. Contrasting two situations brings this out. First, consider a case in which the average populational responses for each locus, the $(\bar{x_j}^I, \bar{y_j}^I)$, are close together, but the responses for each locus are so tightly clustered that there is no overlap among the clusters, or, in the uncertainty analysis, a case in which, for all k, $p(l_j|a_k) = 1$ for some $j = j'$ and 0 for all $j \neq j'$. In this case, B is low due to the closeness of each locus's centroid, while $\mathcal{B}$ is high because of the lack of overlap among the clusters. Second, consider a case in which the average populational responses for each locus are fairly disparate, but the clusters for each locus are so broad that they all overlap substantially. In this case, the distance between the centroids would make B high, while the overlap among the clusters would make $\mathcal{B}$ low.

3. Conclusion

We have defined a family of variance and uncertainty measures of population diversity. When these measures of diversity are observed in simple models of evolution, diversity exhibit various interesting dynamics (Bedau, Ronneburg and Zwick 1992; Bedau and Bahm 1994; Bedau, Giger and Zwick 1995), such as gradual trends and sudden shifts, some of which depend on mutation rate. The ultimate interest of these diversity measures depends on the extent to which such observations reveal fundamental features of evolving systems in general.

References

Bedau, M. A. (1994). The Evolution of Sensorimotor Functionality. In P. Gaussier and J. -D. Nicoud, ed., *From Perception to Action*. New York: IEEE Press.

Bedau, M. A., Bahm, A. (1994). Bifurcation Structure in Diversity Dynamics. In R. Brooks and P. Maes, eds., *Artificial Life IV*. Cambridge, Mass.: Bradford/MIT Press.

Bedau, M. A., Giger, M., Zwick. M. (1995). Evolving Diversity of Population and Environment in Static Resource Models. *Advances in Systems Science and Applications* **Special Issue I**.

Bedau, M. A., Packard, N. H. (1992). Measurement of Evolutionary Activity, Teleology, and Life. In C. G. Langton, C. E. Taylor, J. D. Farmer, and S. Rasmussen, eds., *Artificial Life II*. SFI Studies in the Sciences of Complexity, Vol. X. Reading, Mass.: Addison-Wesley.

Bedau, M. A., Ronneburg, F., Zwick, M. (1992). Dynamics of Diversity in an Evolving Population, *Parallel Problem Solving from Nature* **2**, 95–104.

Garner, W. R., McGill, W. J. (1956). The Relation Between Information and Variance Analyses, *Psychometrika* **21**, 219–228.

Iversen, G. R., Norpoth, H. (1976). *Analysis of Variance*. Beverly Hills, Calif.: Sage Publications.

Margalef, R. (1968). *Perspectives in Ecological Theory*. Chicago: Univ. of Chicago Press.

Packard, N. H. (1989). Intrinsic Adaptation in a Simple Model for Evolution. In C. G. Langton, ed., *Artificial Life*. SFI Studies in the Sciences of Complexity, Vol. VI. Reading, Mass.: Addison-Wesley.

Shannon, C. E., Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana, Ill.: Univ. of Illinois Press.

Wilson, E. O., Bossert, W. H. (1971). *A Primer of Population Biology*. Sunderland, Mass.: Sinaur.