

Sergey S. Rabotyagov, Todd Campbell, Manoj Jha, Philip W. Gassman, Jeffrey Arnold, Lyubov Kurkalova, Silvia Secchi, Hongli Feng, and Catherine L. Kling. 2010. Least cost control of agricultural nutrient contributions to the Gulf of Mexico hypoxic zone. *Ecological Applications* 20:1542-1555.

Appendix A. Algorithm description and additional results.

The language and logic of evolutionary algorithms

Beginning in 1950's and 1960's computer scientists came to a realization that the theory of biological evolution can be used as an optimization tool for engineering problems. Since the field of evolutionary computation owes its origins to observations of biological evolution, the terminology used has its analogs in biology, although, typically, the entities used to describe an optimization problem are much simpler than the real biological entities bearing the same name. A *genome* (or a *chromosome*) refers to a complete collection of *genes* and fully describes an *individual* (a candidate solution in an optimization problem). A set of possible values that any gene can take is referred to as an *allele set*, or *alphabet*. Often, a genome representing a candidate solution is a one-dimensional array, or vector. A gene then is an element of this array and encodes a particular element of a candidate solution. A value of a gene comes from its allele set, also a vector. Analogous to haploid organisms in real biology, *offspring* is created from two parent individuals. During sexual reproduction, *recombination* (*crossover*) occurs: the offspring's genome consists of portions of each of the two parents' genomes. As in biological evolution, offspring are subject to *mutation*: a random substitution of a gene's value with a value from its allele set.

In this study, the following correspondence between the terminology of evolutionary algorithms and entities related to nonpoint source pollution is made. Table A1 provides the necessary terms:

TABLE A1. Terminology of evolutionary algorithms in relation to watershed optimization

Evolutionary computation term	Its interpretation in a nonpoint source pollution setting
Allele set	A set of mutually exclusive land use options and conservation practices
Individual (genome)	A distinct allocation of conservation practices and land use options in the watershed
Gene	Spatial unit of analysis (HRU)

In this application of evolutionary algorithms to spatial optimization, a genome is a vector of length F , where F is the number of spatial decision-making units. Each element of the vector (gene) is encoded with a value from the allele set A , and denotes a particular land use option.

As in biological evolution, individuals at every *generation* form *populations*, and are characterized by their *fitness*—a score which measures how well each individual is solving the optimization problem at hand (for example, a value of an objective function). Individuals possessing higher fitness scores are more likely to be selected for reproduction and therefore are more likely to pass along the characteristics associated with the candidate solutions they represent.

While there are many variations of evolutionary algorithms, most that can be called “genetic algorithms” have the following elements in common: populations of individual

solutions, selection for reproduction according to fitness levels, crossover to produce new solutions (offspring), and random mutation of new offspring.

Given that in order to characterize the tradeoffs outlined above, a multiobjective optimization problem needs to be addressed, we turn to a class of evolutionary algorithms designed to solve multiobjective problems. Recent years have seen emergence of several multiobjective evolutionary algorithms. We use an algorithm called Strength Pareto Evolutionary Algorithm 2 (SPEA2), developed by Zitzler and Thiele (Zitzler et al. 2002).

The search process starts with a population of candidate solutions from which a new population is created by the process of selection, crossover, and mutation. The fitness score of each individual in the population is a function of how many other individuals in the population it dominates (in the sense of Pareto) and by how many individuals it is dominated by. The algorithm provides an approximate solution to the main multiobjective minimization problem by preserving Pareto-nondominated individuals, by eliminating Pareto-dominated solutions, and by iteratively creating new candidate solutions and assessing how well they perform on the multiple objectives. Furthermore, the algorithm takes into account the degree of “crowding” around an individual in order to preserve the diversity in the population and to explore a greater region of the objective space.

Fitness assignment

An individual \mathbf{i} is assigned a strength value $S(\mathbf{i})$ which equals to the number of solutions it dominates:

$$S(\mathbf{i}) = \left| \{ \mathbf{j} \mid \mathbf{j} \in \mathbf{P}_t \cup \bar{\mathbf{P}}_t \wedge \mathbf{i} \succ \mathbf{j} \} \right|, \quad (\text{A.1})$$

where $\bar{\mathbf{P}}_t$ is the original population at generation t , \mathbf{P}_t is the temporary population created, $|\cdot|$ denotes the cardinality of a set, and \succ corresponds to the Pareto dominance relation. On the basis of this definition of strength values, the raw fitness for individual \mathbf{i} is calculated:

$$R(\mathbf{i}) = \sum_{\mathbf{j} \in \bar{\mathbf{P}}_t \cup \mathbf{P}_t, \mathbf{j} \succ \mathbf{i}} S(\mathbf{j}). \quad (\text{A.2})$$

Thus, the raw fitness of an individual is determined by the strength of the dominators (individuals that dominate \mathbf{i}). Then, the raw fitness value of $R(\mathbf{i}) = 0$ corresponds to a nondominated individual, while a high raw fitness value corresponds to an individual that is dominated by many other individuals (which in turn dominate other individuals). In light of this interpretation, fitness minimization used in the formulation of the algorithm makes intuitive sense. Figure A1 demonstrates the fitness assignment process and highlights the fact that individuals that are located in the “crowded” areas of the objective space get a higher raw fitness value, and therefore are less likely to be selected into a future generation. For instance, point F dominates points B, C, and A, and therefore gets a strength value of 3. Since point F is nondominated, its raw fitness is zero. Point D, on the other hand, dominates only A, and thus gets the strength value of one, but is dominated by point G, which itself dominates 3 points. Thus, point D gets the raw fitness value of 3. Point A is the ‘worst’ point in the objective space, as it is associated with the highest cost and pollution levels. It itself does not dominate any other points, but is dominated by points F, G (with a strength value of 3), H (with a strength value of 2), D (with a strength value of 1), and E (with a strength value of 1). Therefore, the raw fitness value for point A is $3+3+2+1+1=10$. Recalling that in this algorithm, individuals with the lower fitness scores are considered ‘more fit’, it is clear that individual A is far less likely to survive into the next generation than, for example, point F.

Such assignment of raw fitness scores also takes into account the relative ‘isolatedness’ of candidate solutions in the objective space. Conceptually, one would like the resulting Pareto-optimal frontier to span a large portion of the objective space. Therefore, candidate solutions on the interior of the frontier are somewhat less preferred than those close to the edges. In the figure, for example, while both points B and C are dominated, point C is dominated by both points F and G by virtue of its ‘interior’ location in the objective space; whereas point B is dominated only by point F and not by point G: its pollution level is lower than that of G. As a result, point B has a raw fitness score of 3 as opposed to the score of 6 for C, and its ‘genetic makeup’ is therefore less likely to be eliminated in the subsequent generations.

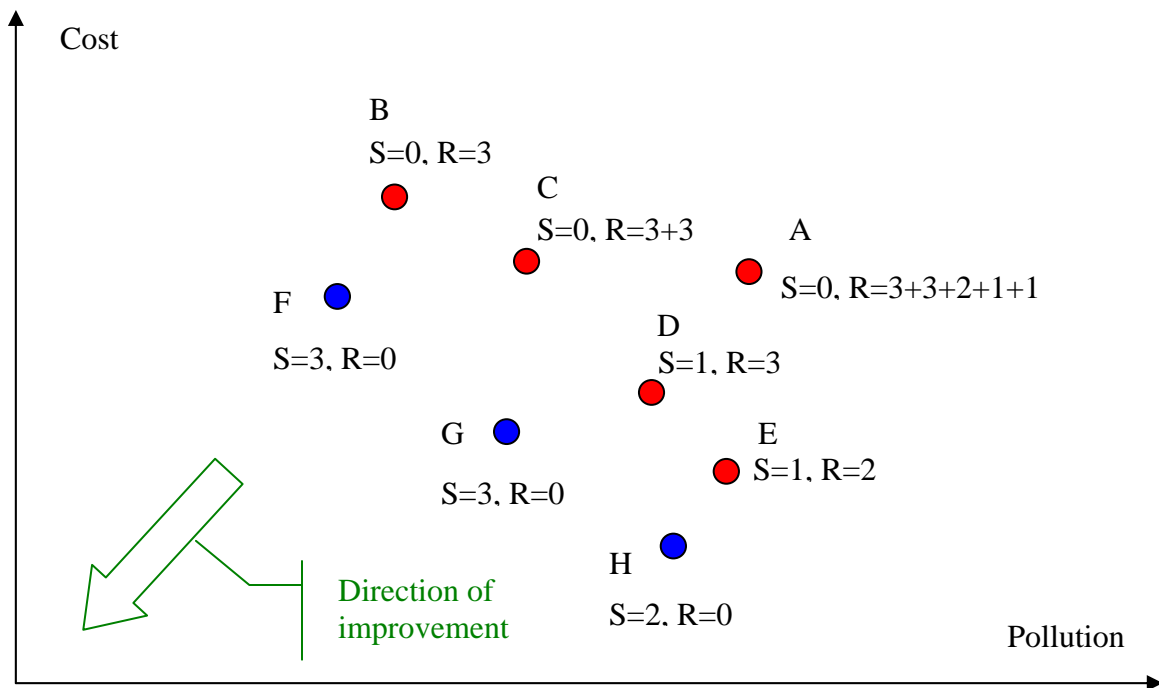


FIG. A1. Raw fitness assignment in SPEA2.

Finally, while the raw fitness score assignment outlined above incorporates some information on the location of the solutions in the solution space, additional density information is also incorporated into the calculation of a fitness score. Density estimation technique is used to further differentiate between individuals that are located in the “crowded” areas of the objective space (less preferred) from those located in the relatively sparse areas of the objective space (more preferred). The density estimation technique used in SPEA2 is an adaptation of the k -th nearest neighbor method, where the density at any point is a decreasing function of the distance to the k -th nearest data point. For each individual \mathbf{i} , we calculate the distances (in objective space) to all the individuals in the population and the temporary population, and store them in a list. After sorting the list in an increasing order, the k -th element yields the distance, denoted as σ_i^k . k is chosen to equal to the square root of the sum of the initial population size and the size of the temporary population ($\sqrt{40+12} \approx 7$). An additional measure of distance was incorporated into the algorithm in order to preserve diversity in the objective space. In each generation, the distance from a given individual to the center of the cube defined by the endpoints of the frontier was established. The purpose of this calculation is to further reward individuals who are located closer to the edges of the frontier, and thus prevent loss of diversity.

This distance is denoted as σ_i^c . The density is computed as:

$$D(\mathbf{i}) = \frac{1}{\sigma_i^k + 0.25\sigma_i^c + 2}, \quad (\text{A.3})$$

where 2 is added to the denominator to ensure that the value of the density is greater than zero and less than one. Given the raw fitness score and the estimated density, the fitness of an individual \mathbf{i} is calculated as:

$$F(\mathbf{i}) = R(\mathbf{i}) + D(\mathbf{i}). \quad (\text{A.4})$$

For the HRUs which were observed to have the relevant conservation practice in the baseline, the allele set was constrained. Reduced fertilizer (RF) in the table above refers to a 20 percent reduction in nitrogen fertilizer application. The allele set is constructed to take into account the fact that many of the practices considered are not mutually exclusive and can be implemented jointly on any given field.

TABLE A2. Field-level Conservation Options (Unconstrained Allele Set).

Option number	Allele description
1	Conventional Till (CT)
2	Ridge Till (RT)
3	Mulch Till (MT)
4	No Till (NT)
5	CT+Contour
6	RT+Contour
7	MT+Contour
8	NT+Contour
9	CT+Grassed Waterway
10	RT+Grassed Waterway
11	MT+Grassed Waterway
12	NT+Grassed Waterway
13	CT+Terraced
14	RT+ Terraced
15	MT+Terraced
16	NT+Terraced
17	CT+RF
18	RT+RF
19	MT+RF
20	NT+RF
21	CT+Contour+RF
22	RT+Contour+RF
23	MT+Contour+RF
24	NT+Contour+RF
25	CT+Grassed Waterway+RF
26	RT+Grassed Waterway+RF
27	MT+Grassed Waterway+RF
28	NT+Grassed Waterway+RF
29	CT+Terraced+RF

30	RT+Terraced+RF
31	MT+Terraced+RF
32	NT+Terraced+RF
33	Land retirement

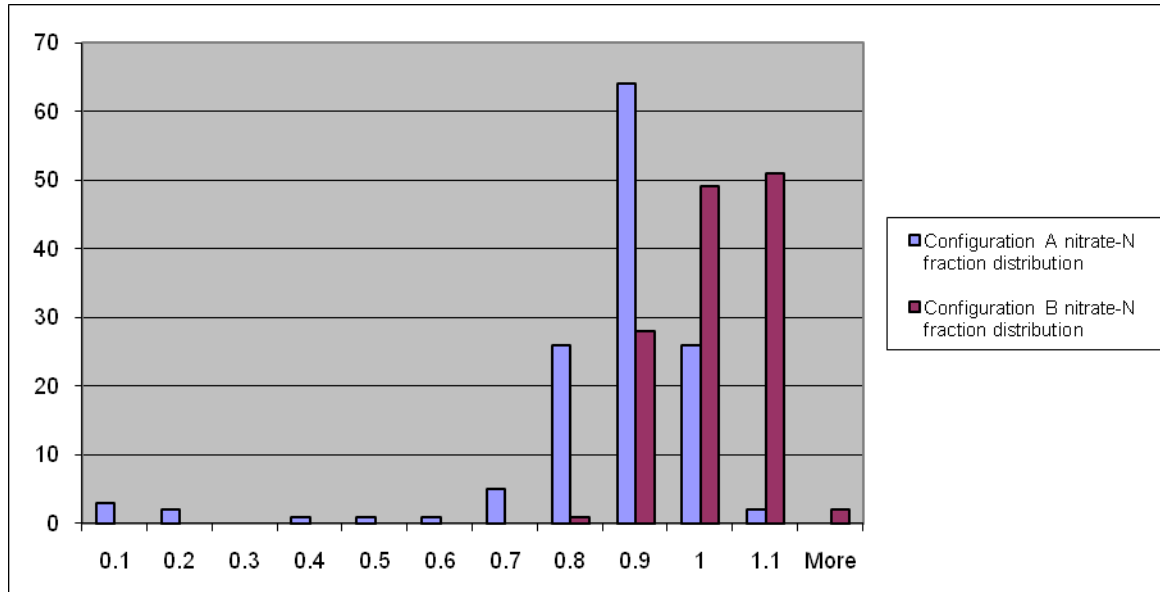


FIG. A2. Histogram of subbasin nitrate loadings, by share of baseline loadings.

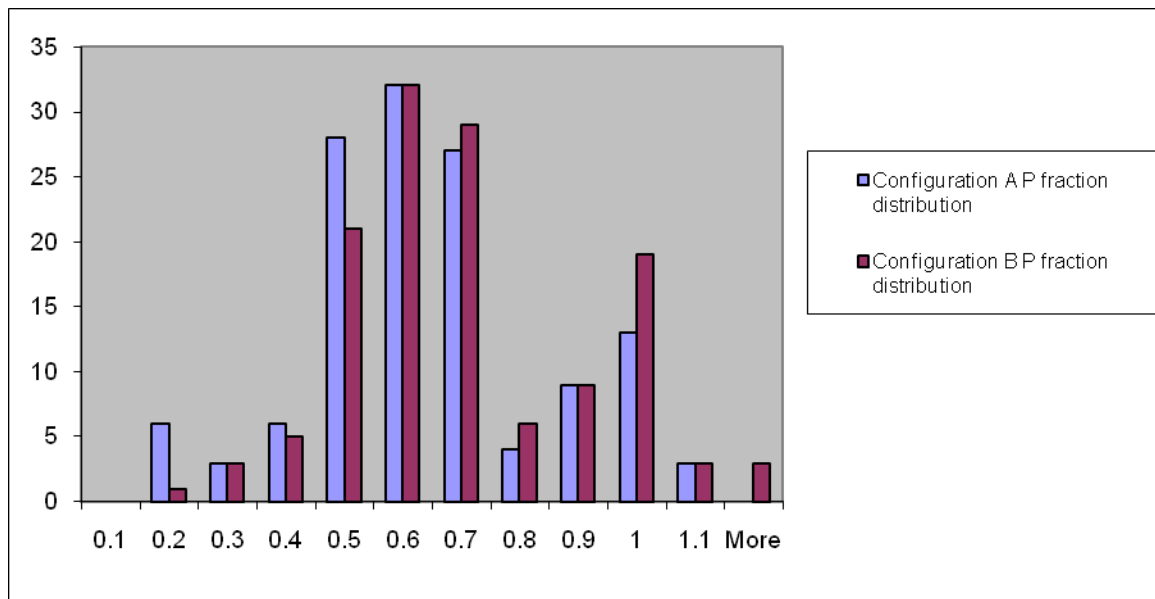


FIG. A3. Histogram of subbasin phosphorus loadings, by share of baseline loadings.

LITERATURE CITED

Zitzler E., M. Laumanns, and L. Thiele. 2002. SPEA2: Improving the Strength Pareto Evolutionary Algorithm for Multiobjective Optimization, in *Evolutionary Methods for Design, Optimisation, and Control*, CIMNE, Barcelona, Spain, pp. 95–100.