

Artificial Intelligence Methods in Atmospheric and Ocean Sciences

Genetic Algorithms II: More Advanced Techniques and Applications

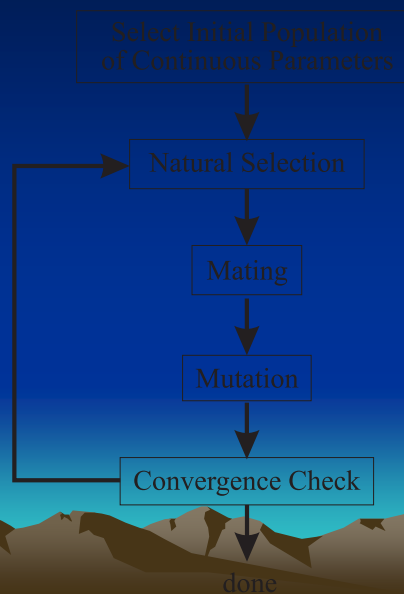
Sue Ellen Haupt
Applied Research Laboratory
The Pennsylvania State University

Nearly Anything can be posed as
an optimization problem

Outline

- I. Coding Considerations
 - A. Variable Representation
 - B. Selection and Mating Techniques
 - C. Writing a Fitness Function
 - D. GA parameter selection
 - E. Parallel Genetic Algorithms
- II. More Applications
 - A. Solving PDEs
 - B. Contingency Tables
 - C. Air Pollution Receptor Model
- III. Other Environmental Science Applications
- IV. A Look to the Future

Review



I. Important factors in implementation

- Choosing binary or continuous
- Selection Method
- Population size
- Mutation rate
- Coding cost function
- Parallel implementation issues

A. Variable Representation

- Binary
 - Good for discrete parameters
 - Difficult to decide truncation when representing real numbers
 - Analyzed the most
- Continuous
 - Ideal for representing real numbers
 - Can represent to machine precisions
 - Crossover a bit more complex

B. Selection and Mating

- How many to replace? $N_{keep} = X_{rate} \times N_{pop}$
- Example: $X_{rate} = 50\%$
- Selection Methods vary – best if based on rank or cost



Crossover

- Binary GA
 - Single point
 - Two point
 - Three parents – two points
 - Uniform (bits exchanged via random mask)

More Crossover

- Continuous GA
 - Simple Swap

$$parent_1 = [p_{m1}, p_{m2}, p_{m3}, p_{m4}, p_{m5}, p_{m6}, \dots, p_{mN_{var}}]$$

$$parent_2 = [p_{d1}, p_{d2}, p_{d3}, p_{d4}, p_{d5}, p_{d6}, \dots, p_{dN_{var}}]$$

$$offspring_1 = [p_{m1}, p_{m2}, \uparrow p_{d3}, p_{d4}, \uparrow p_{m5}, p_{m6}, \dots, p_{mN_{var}}]$$

$$offspring_2 = [p_{d1}, p_{d2}, \uparrow p_{m3}, p_{m4}, \uparrow p_{d5}, p_{d6}, \dots, p_{dN_{var}}]$$
 - Along Axes

$$offspring_1 = parent_1 - \beta (parent_1 - parent_2)$$

$$offspring_2 = parent_2 + \beta (parent_1 - parent_2)$$
 - Uniform random numbers

$$offspring_1 = parent_1 - [\beta_1 (p_{m1} - p_{d1}), \beta_2 (p_{m2} - p_{d2}), \dots, \beta_{N_{var}} (p_{mN_{var}} - p_{dN_{var}})]$$

$$offspring_2 = parent_2 + [\beta_1 (p_{m1} - p_{d1}), \beta_2 (p_{m2} - p_{d2}), \dots, \beta_{N_{var}} (p_{mN_{var}} - p_{dN_{var}})]$$

C. Writing the Cost Function

- Fitness function determines how long algorithm takes to run.
- Try to minimize time in cost function (or write to make parallel)
- Multi-objectives – often choose to weight:

$$cost = \sum_{n=1}^N w_n f_n$$

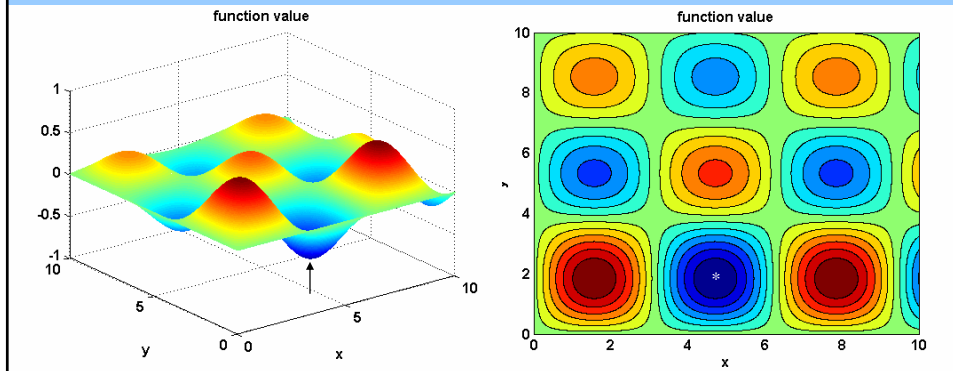
D. GA Parameters

- Often want to minimize the number of calls to an expensive cost function
- Crossover rate and selection method don't make a big difference
- Choice of population size and mutation rate do make a big difference

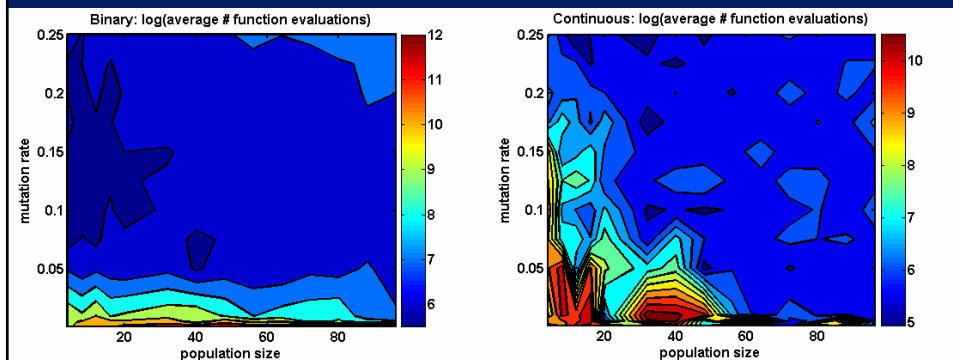
Example: Bes-sin

Find the minimum of: $f(x, y) = \sin(x)J_1(y)$

Subject to: $0 \leq x \leq 10$ and $0 \leq y \leq 10$



Sensitivity to Mutation and Population Size



- Problem Dependent

- In general, smallest number of cost function evaluations for:

- small population size (order 8-16)
- moderately large mutation rate (.15-.25)

E. Parallel GAs

Motivation

- Speedup important for expensive cost functions
- Co-evolving subpopulations
- Mimics nature

Ways to Parallelize

Master-Slave

Master controls cost function evaluations done by slave processors

- Easy to implement
- No subpopulation evolution

Island GA

Subpopulations evolve in parallel with some migration between

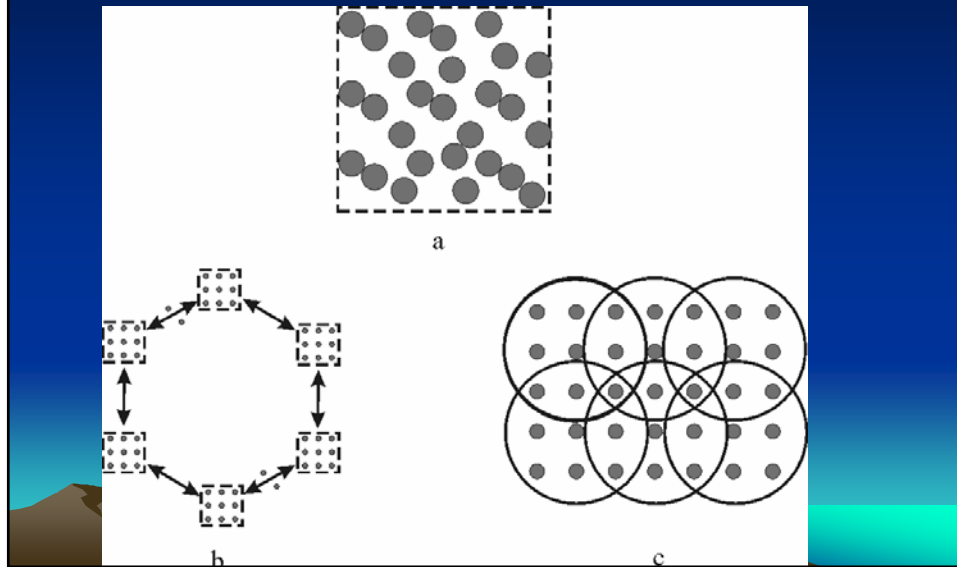
- More difficult
- Increase local diversity

Cellular GA

Chromosomes on Nodes. Communicate Only with nearest Neighbors

- More difficult
- Slow dispersion of information

Pictorially



Speedup

$$T_p = \frac{N_{pop} T_f}{P} + \rho(P-1)T_c$$

where: T_f = the time to evaluate the fitness of one chromosome

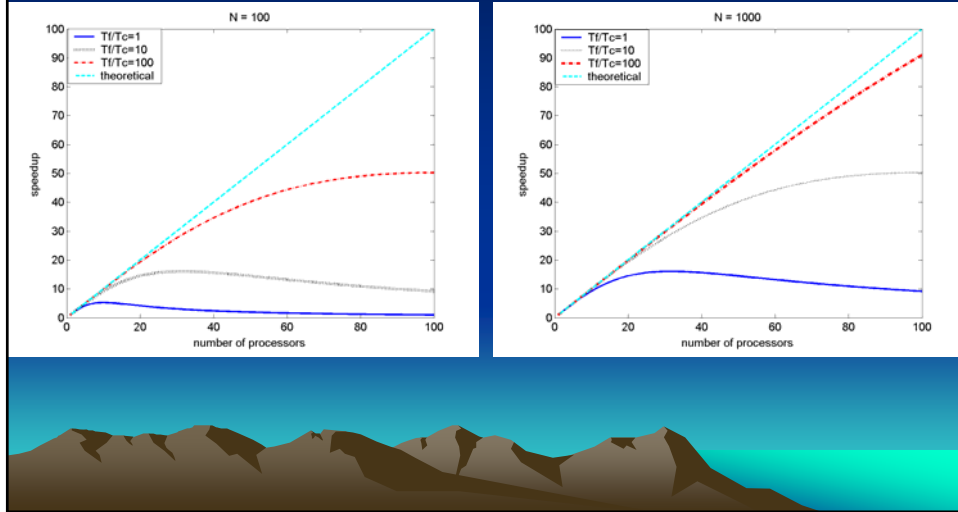
T_c = the average time to communicate with one processor

N_{pop} = population size

P = number of processors

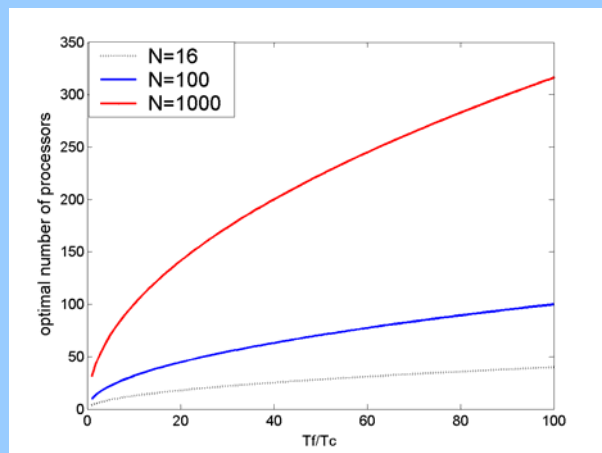
ρ = parameter dependent on selection and parallelization method

Speedup



Optimum Processors

$$P_{opt} = \sqrt{\frac{N_{pop} T_f}{\rho T_c}}$$



II. Applications

A. Solving Partial Differential Equations with a GA

Super Korteweg de Vries Equation:

$$u_t + \alpha u u_x + \mu u_{xxx} - \nu u_{xxxxx} = 0$$

Models Shallow Water Waves near critical value of surface tension.

Use steadily translating form, $X = x - ct$:

$$(\alpha u - c)u_x + \mu u_{xxx} - \nu u_{xxxxx} = 0$$

Expand in Fourier Series:

$$u(X) \approx u_K(X) = \sum_{k=1}^K \cos(kX)$$

Applying Galerkin discretization, equation becomes:

$$\sum_{k=1}^K [-k(\alpha u - c) + \mu k^3 + \nu k^5] a_k \sin(kX) = 0$$

Genetic Algorithm Application:

Code parameters, a_k , into chromosome

Minimize Cost:

$$\text{cost}(u_K) = \text{abs} \left\{ \sum_{k=1}^K [-k(\alpha u - c) + \mu k^3 + \nu k^5] a_k \sin(kX) \right\}$$

Use: Real Genetic Algorithm

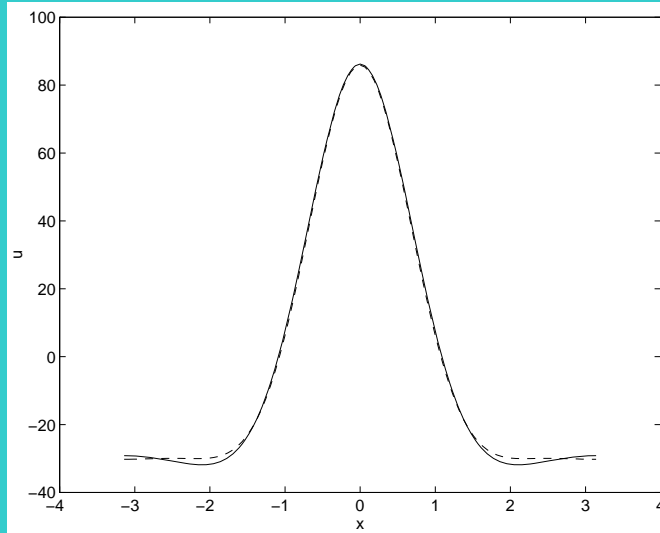
mutation rate of 0.2

k=6 (evaluating function at 6 points)

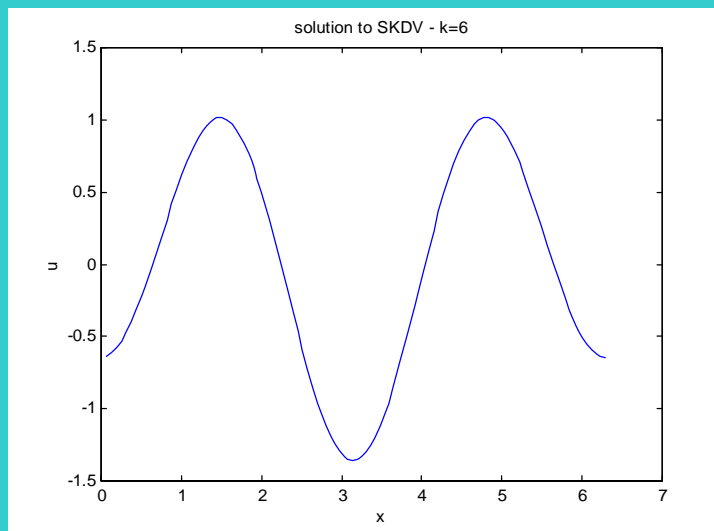
initial population size=500

population size=100

70 iterations



A cnoidal wave of the SKDV with phase speed of $c=14.683$. The solid line is an “exact” solution due to Boyd (1986) and the dashed line is the GA solution which approximates it.



Double cnoidal wave of the SKDV as found by the genetic algorithm.

Results:

The Genetic Algorithm was able to find several solutions to the Super Korteweg de Vries Equation.

Prospects:

This is just a simple one dimensional equation. We were able to find several solutions. Unlike most techniques for such problems, we did not require an excellent first guess. Genetic algorithms may prove useful for solving other higher dimensional PDEs.

B. Contingency Tables

Predict occurrence or non-occurrence of Hail

| Forecast /Actual | Hail | No Hail |
|-------------------------|-------------|----------------|
| Hail | a | b |
| No Hail | c | d |

Marzban, C. and S.E. Haupt, 2005: On Genetic Algorithms and Discrete Performance Measures, AMS 4th Conference on Artificial Intelligence, San Diego, CA, paper 1.1.

- In the past, Neural Nets have shown some success at training predictive networks
- Such networks optimized via Least Squares
- What if could optimize directly on Skill Scores used to evaluate success?
- Genetic Algorithm capable of optimizing on those skill scores

$$C = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{matrix} N_0 \text{ (nonevents)} \\ N_1 \text{ (events)} \end{matrix}$$

1. Fraction correct (FRC): $FRC = \frac{a + d}{N_0 + N_1}$

2. Critical Success index (CSI): $CSI = \frac{d}{b + N_1}$

3. Heidke's Skill Statistic (HSS):

$$HSS = \frac{2(ad + bc)}{N_0(b + d) + N_1(a + c)}$$

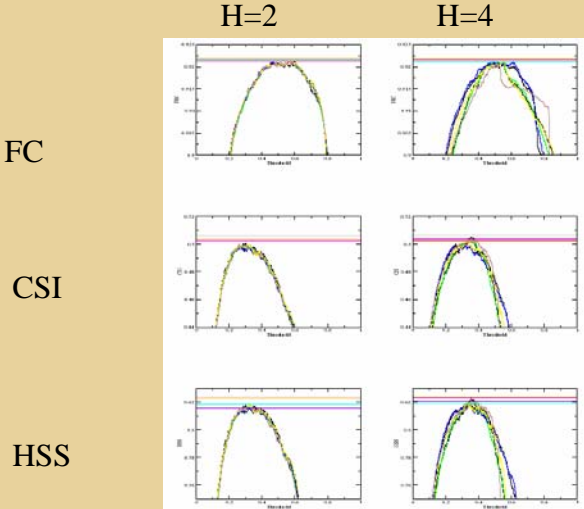
Comparison of GA with Conjugate Gradient

Table 1. The average performance values and confidence intervals, for the different measures, and for H=2 and H=4.

| H = 2 | | |
|---------|-----------------|-----------------|
| Measure | GA | CG |
| FRC | 0.92163±0.00019 | 0.92130±0.00006 |
| CSI | 0.50370±0.00105 | 0.49970±0.00037 |
| HSS | 0.62076±0.00266 | 0.61768±0.00038 |

| H = 4 | | |
|---------|-----------------|-----------------|
| Measure | GA | CG |
| FRC | 0.92157±0.00022 | 0.92109±0.00018 |
| CSI | 0.50360±0.00147 | 0.50208±0.00146 |
| HSS | 0.62192±0.00158 | 0.61958±0.00106 |

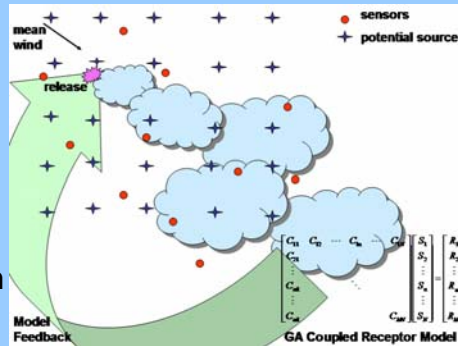
Results



The fraction correct, CSI, and HSS as obtained from five different initializations of conjugate gradient with H=2 (left) and H=4 (right). The horizontal lines are the corresponding scores from GA.

C. Receptor Model of Air Pollution

- Given data on
 - Pollutant monitored at receptors
 - Potential source characteristics and emission rates
 - Meteorological conditions
- Determine
 - Percentage of pollutant from each source
 - Actual wind direction for plume to match

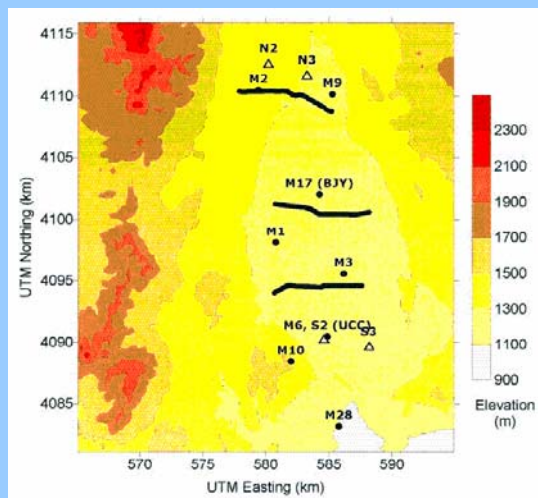


Source Characterization

- Need: Characterize source of airborne contaminant
 - amount
 - location
 - type
 - transport characteristics – wind field
 - Tools:
 - Receptor Models
 - Dispersion Models
 - Approach: Combine the tools
 - Linking Mechanism: Genetic Algorithm (GA)
1. Coupling of a Receptor and Dispersion Model with a Genetic Algorithm – Haupt 2005 (Atm.Env.)
 2. Approach verified through Monte Carlo runs using a Gaussian Plume Dispersion Model and synthetic data – Haupt, Young, Allen 2006 (JAMC)
 3. Incorporated SCIPUFF and Tested with Field Data – Allen, Haupt, Young 2006 (JAMC, in press)
 4. Use GA to also evolve correct wind direction and source locations – Allen, Young, Haupt 2006 (Atm. Env., in press)

Dipole Pride 26 (DP26) Data Set

- Yucca Flats, Nevada Test Site
- Instantaneous releases of sulfur hexafluoride (SF6)
- 90 sampler bags along three lines
- Four sources
- 17 field tests



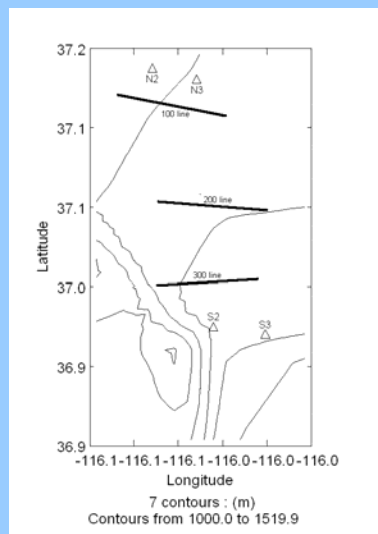
From Chang, et al. (2003)

Dipole Pride Data

- optimize agreement over multiple receptors

$$RMS = \frac{\sum_{r=1}^{90} \sqrt{\sum_{m=1}^M \log_{10} ((C_{mnr} \cdot S_n - R_{mr})^2)}}{\sum_{r=1}^{90} \sqrt{\sum_{m=1}^M \log_{10} ((R_{mr})^2)}}$$

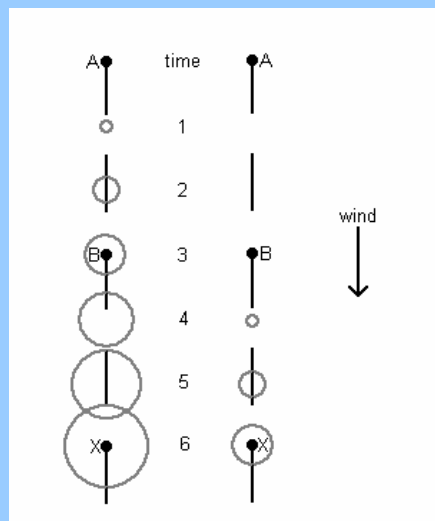
- Correct source and time pinpointed 64% of time
- Within one time period and one source rest of time



Issue of time vs. Source Strength

Specific Source Characterization

- 4x4 grid of candidate sources with emission time uncertainty
- The coupled model cannot easily distinguish upwind location from emission time
- The model can typically identify the crosswind source location within $.02^\circ$ and emission time within 40 minutes
- Strength characterizations are less accurate



Multi-Stage Process

Derived from performance optimization results

Four steps

1. Determine possible emission times with coarse source grid
2. Identify source location at each time from step 1 with finer location grid
3. Pinpoint emission time more accurately than step 1 using locations from step 2
4. Source strength analysis using results from steps 2 and 3; Filter out sources with weak strengths as they are likely not actual emitters

Multi-Stage Process

Results

- Model run with data from 14 field tests
- Number of sources: 6/14 correct, 5/14 slightly overestimated
 - Experimental error likely on remaining three tests
- Typical result: Location within one grid point ($.02^\circ$), emission time within 40 minutes (1 time period)
- Additional fine-tuning of the process could improve performance

Current Model

- Directly Back-calculates 7 parameters
 - 2D location (x,y)
 - Effective Release Height
 - Source Strength
 - Release Time
 - Wind Speed
 - Wind Direction
- Using information theory to determine how many receptors needed and effects of Noise

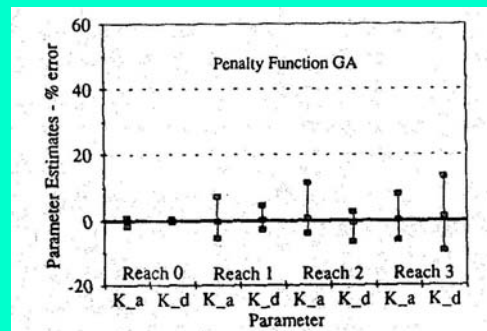
Long, K.J., S.E. Haupt, G.S. Young, and C.T. Allen, 2007: Characterizing Contaminant Source and Meteorological Forcing using Data Assimilation with a Genetic Algorithm, Fifth Conference on Artificial Intelligence Applications to Environmental Science at AMS Annual Meeting, San Antonio, TX, Jan. 16, Paper number 4.3.

III. What has been done in Environmental Sciences

- Need for optimization
- Often need to fit model to observed data
- If know functional form of model, can use GA to fit parameters
- Once modeled → Predictions

Calibrating Water Quality Models

- Mulligan and Brown (1998)
 - GA used to estimate parameters of model
 - GA better than traditional techniques
 - GA also gives information about search space useful for confidence regions
- Molian and Loucks (1995)
- McKinney and Lin (1994)
- Rogers and Dowla (1994)
- Simpson, et al. (1994)



Range of parameter estimates of 100 GA runs

Managing Groundwater Resources

Peralta with Aly, Shieh, and Fayad -
 Combined GAs with neural networks and simulated annealing to fit parameters to optimize pumping locations and schedules for groundwater treatment

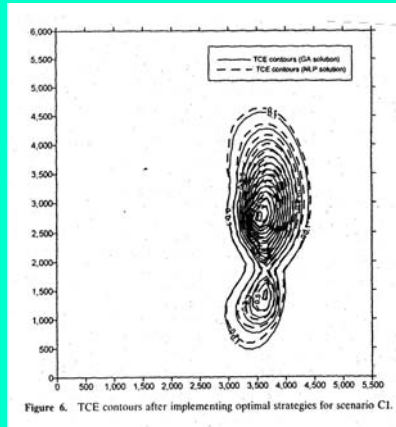


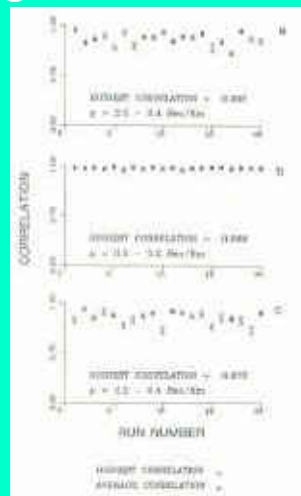
Figure 6. TCE contours after implementing optimal strategies for scenario C1.

GA minimizes aquifer contamination of trichloroethylene

Identifying underground rock layers

Jervis & Stoffa (1993); Jervis, et al. (1996); Sen & Stoffa (1992a,b); Chunduru, et al. (1995); Boschetti, et al. (1995, 96, 97); Porsani, et al. (2000)

- Seismic information of application of current to obtain potential difference.
- Use data to fit model – highly nonlinear

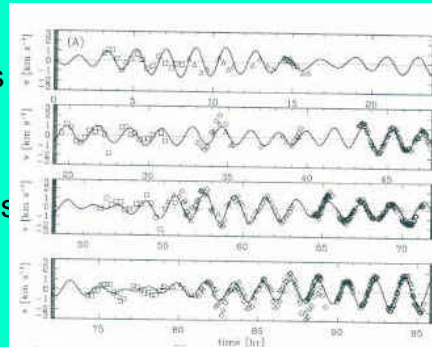


High correlations result from 20 parallel GA runs

Astrophysics Applications

Charbonneau (1995)

- Modeling rotation curves of galaxies
- Extracting pulsation periods of Doppler velocities in spectral lines
- Optimizing a model of hydrodynamic wind



Genetic fit to velocity variation observations

Other Applications include:

- Optimizing design of an oceanographic experiment (Barth 1982)
- Location of array of sensors in ocean after drifting (Porto, et al. 1995)
- Finding source of monitored air pollutants given data on source regions and wind patterns (Cartwright and Harris 1993)
- Locating hypocenter of earthquake (Minister, et al. 1995)

IV. So - Where are GAs Going?

Everywhere

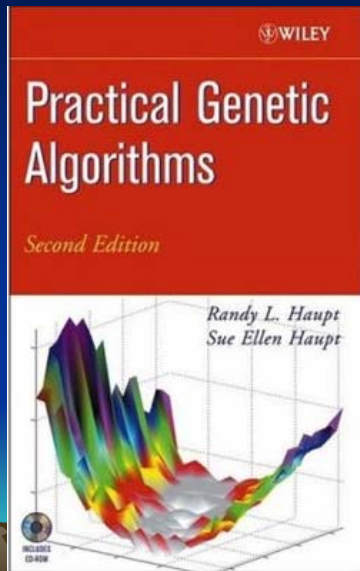
Current Research:

- Optimizing Parameters
- Non-Obvious Applications
- Combining with Simulations
- Environmental Applications
- Parallel Applications

Single Word Summary:

Versatile

For More Information:



Second Edition
includes MATLAB code
(some HPF Fortran)
haupts2@asme.org