Propagation and Provenance of Probabilistic and Interval Uncertainty in Cyberinfrastructure-Related Data Processing and Data Fusion

Paulo Pinheiro da Silva, Aaron Velasco, Martine Ceberio, Christian Servin, Matthew G. Averill, Nicholas Del Rio, Luc Longpré, and Vladik Kreinovich

University of Texas, El Paso, TX 79968, USA
contact email vladik@cs.utep.edu

1. Cyberinfrastructure: A Brief Overview

- Practical problem: need to combine geographically separate computational resources.
- Centralization of computational resources – traditional approach to combining computational resources.
- Limitations of centralization:
  - need to reformat all the data;
  - need to rewrite data processing programs: make compatible w/selected formats and w/each other
- Cyberinfrastructure – a more efficient approach to combining computational resources:
  - keep resources at their current locations, and
  - in their current formats.
- Technical advantages of cyberinfrastructure: a brief summary.

2. Data Processing vs. Data Fusion

- *Practically important situation:* difficult to measure the desired quantity y with a given accuracy.
- *Data processing*:
  - measure related easier-to-measure quantities $x_1, \ldots, x_n$;
  - estimate y from the results $\widetilde{x}_i$ of measuring x_i as $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.
- *Example:* seismic inverse problem.
- *Data fusion:*
  - measure the quantity y several times;
  - combine the results $\widetilde{y}_1, \ldots \widetilde{y}_n$ of these measurements.
- *Specifics of cyberinfrastructure:* first looks for stored results $\widetilde{x}_i$ (corr., $\widetilde{y}_i$), measure only if necessary.
- *Combination of data processing and data fusion.*

3. Need for Uncertainty Propagation, and for Provenance of Uncertainty

- *Need for uncertainty propagation.*
  - main reasons for data processing and data fusion: accuracy is not high enough;
  - we must make sure that after the data processing (data fusion), we get the desired accuracy.
- *In cyberinfrastructure* this is especially important:
  - accuracy varies greatly, and
  - we do not have much control over these accuracies.
- *Need for the provenance of uncertainty:*
  - sometimes, the resulting accuracy is still too low;
  - it is desirable to find out which data points contributed most to the inaccuracy.

4. Uncertainty of the Results of Direct Measurements: Probabilistic and Interval Approaches

- Manufacturer of the measuring instrument (MI) supplies Δ_i s.t. $|\Delta x_i| \le \Delta_i$, where $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$.
- The actual (unknown) value x_i of the measured quantity is in the interval $\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.
- *Probabilistic uncertainty:* often, we know the probabilities of different values $\Delta x_i \in [-\Delta_i, \Delta_i]$.
- *How probabilities are determined:* by comparing our MI with a much more accurate (standard) MI.
- *Interval uncertainty:* in two cases, we do not determine the probabilities:
  - cutting-edge measurements;
  - measurements on the shop floor.
- In both cases, we only know that $x_i \in [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.

5. Typical Situation: Measurement Errors are Reasonably Small

- *Typical situation:*
  - direct measurements are accurate enough;
  - the resulting approximation errors Δx_i are small;
  - terms which are quadratic (or of higher order) in Δx_i can be safely neglected.
- *Example:* for an error of 1%, its square is a negligible 0.01%.
- *Linearization:*
  - expand f in Taylor series around the point $(\widetilde{x}_1, \ldots, \widetilde{x}_n)$;
  - restrict ourselves only to linear terms:

$$\Delta y = c_1 \cdot \Delta x_1 + \ldots + c_n \cdot \Delta x_n,$$

where $c_i \stackrel{\text{def}}{=} \dfrac{\partial f}{\partial x_i}$.

6. Case of Data Processing

- *Propagation (probabilistic case):* if Δx_i are independent with st. dev. σ_i (and 0 mean), then Δy has st. dev.

$$\sigma^2 = c_1^2 \cdot \sigma_1^2 + \ldots + c_n^2 \cdot \sigma_n^2.$$

- *Provenance:*
  - we know which component σ^2 comes from the i-th measurement;
  - we can predict how replacing the i-th measurement with a more accurate one ($\sigma_i^{\text{new}} \ll \sigma_i$) will affect σ^2.

- *Propagation of interval uncertainty:*

$$\Delta = |c_1| \cdot \Delta_1 + \ldots + |c_n| \cdot \Delta_n.$$

- We can predict how replacing the i-th measurement with a more accurate one ($\Delta_i^{\text{new}} \ll \Delta_i$) will affect Δ.

7. Propagation of Probabilistic Uncertainty Through Data Fusion

- *Situation:* we know several results $\widetilde{y}_1, \ldots, \widetilde{y}_n$ of measuring the same quantity y with st. dev. σ_i:

$$\rho_i(y) = \frac{1}{\sqrt{2\pi} \cdot \sigma_i} \cdot \exp\left(-\frac{(y - \widetilde{y}_i)^2}{2\sigma_i^2}\right).$$

- *Resulting probability density:*

$$\rho(y) = \rho_1(y) \cdot \ldots \cdot \rho_n(y) = \text{const} \cdot \exp\left(-\sum_{i=1}^{n} \frac{(y - \widetilde{y}_i)^2}{2\sigma_i^2}\right).$$

- *Maximum Likelihood Estimate:* $\rho(y) \to \max$, hence

$$\widetilde{y} = \frac{1}{\sum\limits_{i=1}^{n} \frac{1}{\sigma_i^2}} \cdot \sum_{i=1}^{n} \frac{\widetilde{y}_i}{\sigma_i^2}.$$

8. Propagation of Probabilistic Uncertainty Through Data Fusion (cont-d)

- *Reminder:*

$$\widetilde{y} = \frac{1}{\sum\limits_{i=1}^n \frac{1}{\sigma_i^2}} \cdot \sum_{i=1}^n \frac{\widetilde{y}_i}{\sigma_i^2}.$$

- *Resulting st. dev. σ for $\widetilde{y}$:* $\widetilde{y}$ is a linear combination of independent normal $\widetilde{y}_i$, hence its st. dev. is:

$$\sigma^2 = \frac{1}{\left(\sum\limits_{i=1}^n \frac{1}{\sigma_i^2}\right)^2} \cdot \sum_{i=1}^n \frac{\sigma_i^2}{\sigma_i^4} = \frac{1}{\left(\sum\limits_{i=1}^n \frac{1}{\sigma_i^2}\right)^2} \cdot \sum_{i=1}^n \frac{1}{\sigma_i^2} = \frac{1}{\sum\limits_{i=1}^n \frac{1}{\sigma_i^2}}.$$

- *Simplified expression:*

$$\frac{1}{\sigma^2} = \sum_{i=1}^n \frac{1}{\sigma_i^2}.$$

- *Provenance:* we can predict how replacing σ_i with a "more accurate" value $\sigma_i^{\text{new}} \ll \sigma_i$ affects σ.

9. Propagation of Interval Uncertainty Through Data Fusion

- *Situation:* we know several results $\widetilde{y}_1, \ldots, \widetilde{y}_n$ of measuring the same quantity y with bounds Δ_i.
- *Analysis:* the unknown (actual) value y belongs to n intervals $\mathbf{y}_i \stackrel{\text{def}}{=} [\widetilde{y}_i - \Delta_i, \widetilde{y}_i + \Delta_i]$.
- *Conclusion:* the range $\mathbf{y}$ of possible values of y is the *intersection* $\mathbf{y} = [\underline{y}, \overline{y}] = \mathbf{y}_1 \cap \ldots \cap \mathbf{y}_n$ of intervals $\mathbf{y}_i$:
$$[\max(\widetilde{y}_1 - \Delta_1, \ldots, \widetilde{y}_n - \Delta_n), \min(\widetilde{y}_1 + \Delta_1, \ldots, \widetilde{y}_n + \Delta_n)].$$
- *Provenance – a problem:* if we replace Δ_i with the same new value $\Delta_i^{\text{new}} \ll \Delta_i$, we may get different accuracies.
- *Example:* $\mathbf{y}_1 = [-1, 1]$, $\mathbf{y}_2 = [-2, 2]$, and $\mathbf{y} = [-1, 1]$. If we use $\Delta_2^{\text{new}} = 1 \ll \Delta_2 = 2$, we may get:
  - $\mathbf{y}_2 = [-1, 1]$; then $\mathbf{y} = [-1, 1]$ is unchanged.
  - $\mathbf{y}_2 = [0, 2]$; then $\mathbf{y} = [0, 1]$ is much narrower.

10. Pre-Estimating the Accuracy of Data Fusion Under Interval Uncertainty: A Problem

- *We know:* the i-th measurement error $\Delta y_i \in [-\Delta_i, \Delta_i]$.
- *Fact:* different values Δy_i lead to different intersections

$$\mathbf{y} = [\underline{y}, \overline{y}] = \bigcap_{i=1}^{n} [(y + \Delta y_i) - \Delta_i, (y + \Delta y_i) + \Delta_i].$$

- *Reasonable assumptions:*
  - Δy_i is uniformly distributed on $[-\Delta_i, \Delta_i]$;
  - Δy_i and Δy_j $(i \neq j)$ are independent;
  - we allow a small probability p_0 of mis-estimation.
- *Formulation of the problem:* find the smallest Δ s.t.:
  - the probability to have $\overline{y} \leq y + \Delta$ is at least $1 - p_0$, and
  - the probability to have $\underline{y} \geq y - \Delta$ is also $\geq 1 - p_0$.

11. Pre-Estimating the Accuracy of Data Fusion Under Interval Uncertainty: Solution

- *Resulting formula:* when fusion is efficient ($\Delta \ll \Delta_i$), we get $\dfrac{1}{\Delta} = \text{const} \cdot \sum\limits_{i=1}^{n} \dfrac{1}{\Delta_i}$, with $\text{const} = 2|\ln(p_0)|$.
- *Example:* for $\Delta_1 = \ldots = \Delta_n$, we get $\Delta = \dfrac{\text{const}}{n} \cdot \Delta_1$.
- *Prob. case:* $\dfrac{1}{\sigma^2} = \text{const} \cdot \sum\limits_{i=1}^{n} \dfrac{1}{\sigma_i^2}$, w/$\Delta_i$ instead of σ_i^2.
- *Observation:* for prob. uncertainty, $\sigma \sim \dfrac{\text{const}}{\sqrt{n}} \cdot \sigma_1$.
- *Data processing:* $\Delta = \sum\limits_{i=1}^{n} |c_i| \cdot \Delta_i$ vs. $\sigma^2 = \sum\limits_{i=1}^{n} |c_i|^2 \cdot \sigma_i^2$.
- $\sim$: $\|$ and sequential resistors $\dfrac{1}{R} = \sum\limits_{i=1}^{n} \dfrac{1}{R_i}$, $R = \sum\limits_{i=1}^{n} R_i$.

12. Optimal Data Processing and Data Fusion

- *Problem:* find the least expensive way to guarantee the given accuracy σ or Δ.
- *Costs:* $c_i^{\rm prob}(\sigma_i)=\dfrac{C_i}{\sigma_i^{\alpha_i}}$ and $c_i^{\rm int}(\Delta_i)=\dfrac{C_i}{\Delta^{\alpha_i}}$.
- *Case of data fusion:* we measure the same quantity, so $C_1=\ldots=C_n$ and $\alpha_1=\ldots=\alpha_n$.
- *Optimal data fusion:* minimizing cost, we get $\sigma_1=\ldots=\sigma_n=\sqrt{n}\cdot\sigma$ and $\Delta_1=\ldots=\Delta_n=n\cdot\Delta$.
- *Optimal data processing: probabilistic case.*

$$\sigma_i=\left(\frac{\alpha_i\cdot C_i}{2\lambda\cdot c_i^2}\right)^{1/(2+\alpha_i)},\ \text{with}\ \sum_{i=1}^n c_i^2\cdot\left(\frac{\alpha_i\cdot C_i}{2\lambda\cdot c_i^2}\right)^{2/(2+\alpha_i)}=\sigma^2.$$

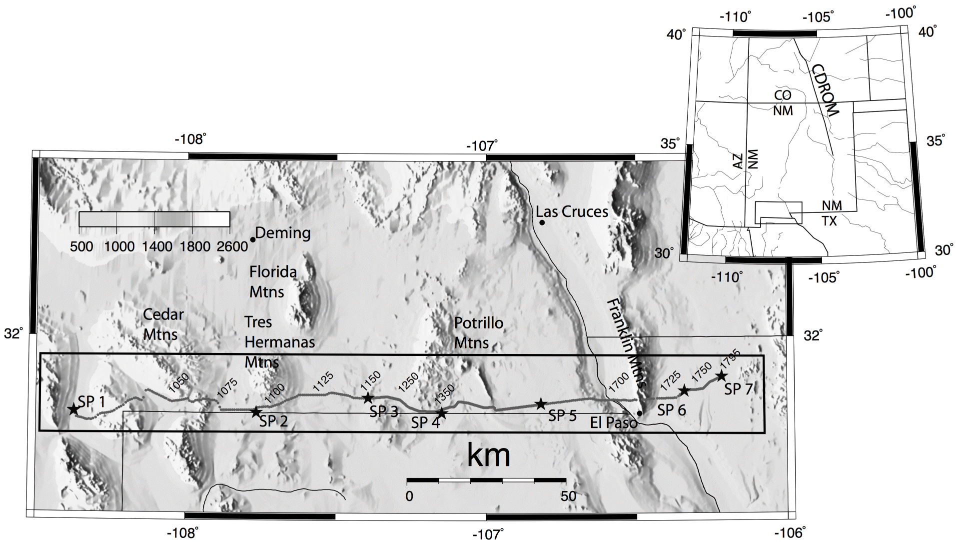
- *Optimal data processing: interval case.*

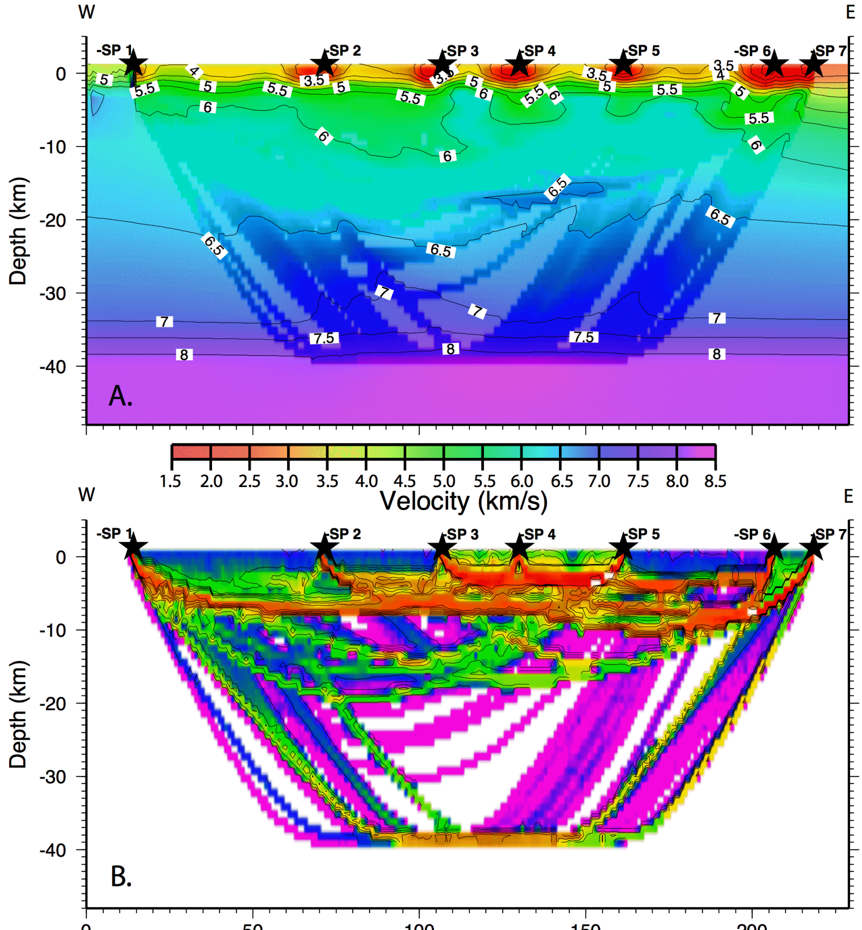
$$\Delta_i=\left(\frac{\alpha_i\cdot C_i}{\lambda\cdot|c_i|}\right)^{1/(1+\alpha_i)},\ \text{with}\ \sum_{i=1}^n |c_i|\cdot\left(\frac{\alpha_i\cdot C_i}{\lambda\cdot|c_i|}\right)^{2/(2+\alpha_i)}=\Delta.$$

13. Beyond Probabilistic and Interval Uncertainty

- *Up to now:* we considered two extreme situations:
  - *probabilistic* uncertainty, when we know all the probabilities;
  - *interval* uncertainty, when we have no information about the probabilities.
- *Fact:* probabilistic situation is a particular case of the interval situation.
- *Conclusion:* interval bounds are wider.
- *In practice:* often, we have partial information about probabilities.
- *As a result:*
  - probabilistic bounds are too narrow,
  - interval bounds are too wide.
- *We need:* intermediate bounds.

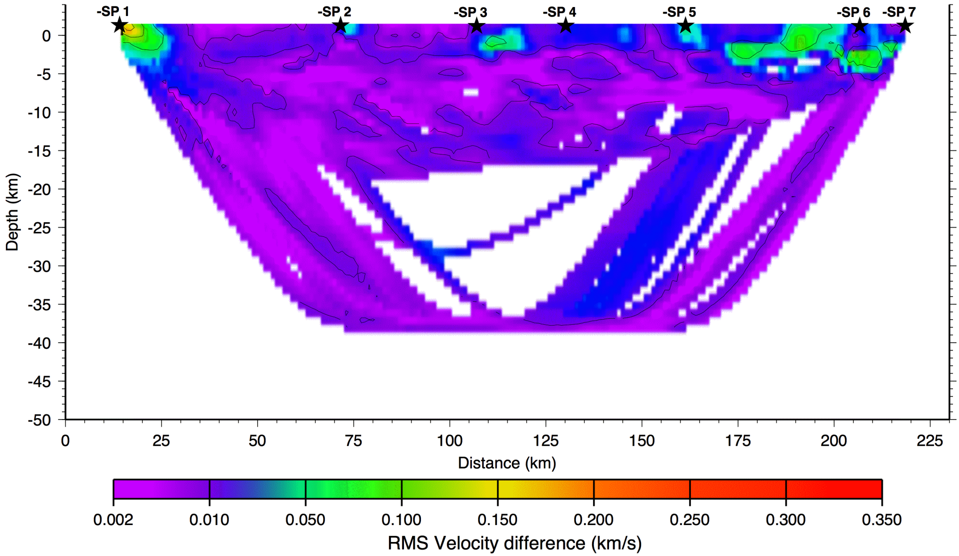
14. Case Study: Seismic Inverse Problem in the Geosciences



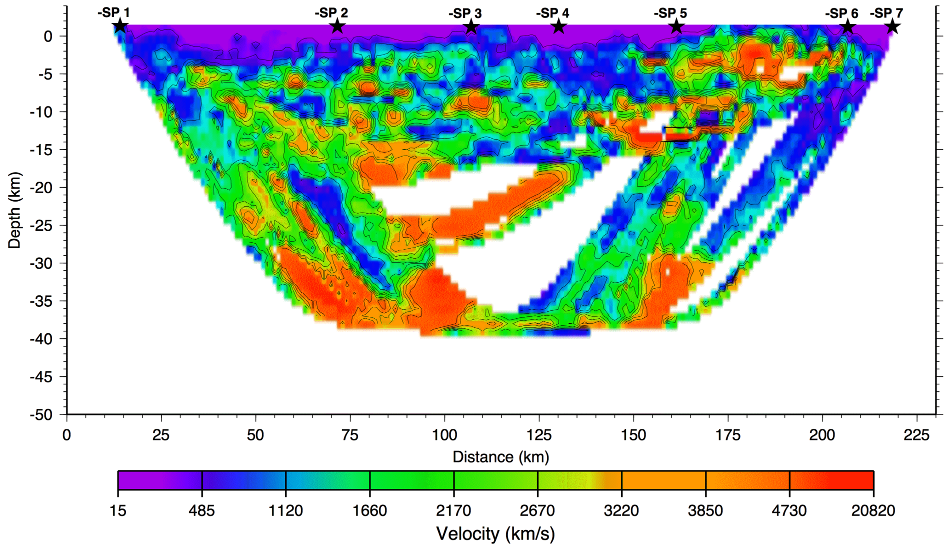


Title Page
Page 16 of 25
Go Back
Full Screen
Close
Quit

15. Estimating Uncertainty, First Try: Probabilistic Approach



16. Estimating Uncertainty, Second Try: Interval Approach



17. Towards a Better Estimate: Revisiting Estimation Algorithms Under Probabilistic and Interval Uncertainty

- *Linearization:* $\Delta y = \sum\limits_{i=1}^{n} c_i \cdot \Delta x_i$, where $c_i \stackrel{\text{def}}{=} \dfrac{\partial f}{\partial x_i}$.
- *Formulas:* $\sigma^2 = \sum\limits_{i=1}^{n} c_i^2 \cdot \sigma_i^2$, $\Delta = \sum\limits_{i=1}^{n} |c_i| \cdot \Delta_i$.
- *Numerical differentiation:* n iterations, too long.
- *Monte-Carlo approach:* if Δx_i are Gaussian w/σ_i, then $\Delta y = \sum\limits_{i=1}^{n} c_i \cdot \Delta x_i$ is also Gaussian, w/desired σ.
- *Advantage:* # of iterations does not grow with n.
- *Interval estimates:* if Δx_i are Cauchy, w/$\rho_i(x) = \dfrac{\Delta_i}{\Delta_i^2 + x^2}$, then $\Delta y = \sum\limits_{i=1}^{n} c_i \cdot \Delta x_i$ is also Cauchy, w/desired Δ.

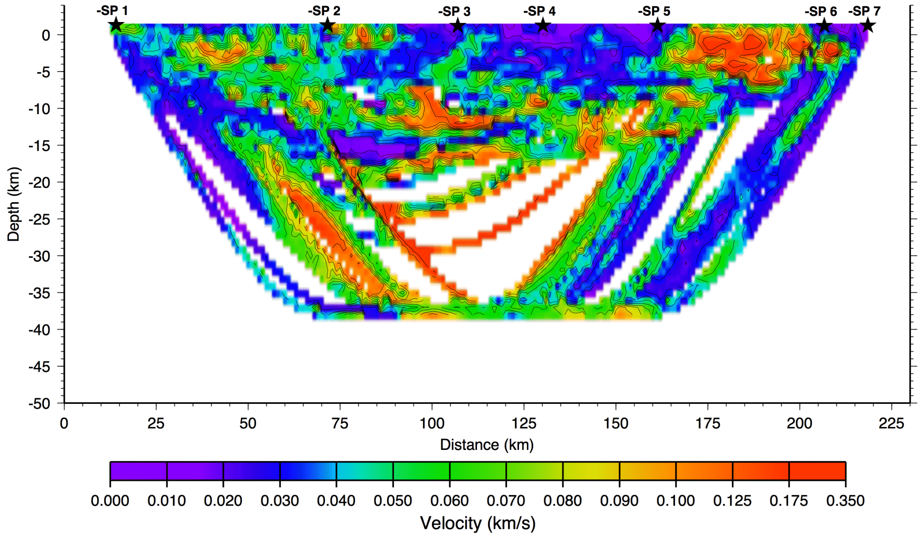
18. Resulting Fast (Linearized) Algorithm for Estimating Interval Uncertainty

- Apply f to $\widetilde{x}_i$: $\widetilde{y} := f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$;
- For $k = 1, 2, \ldots, N$, repeat the following:
  - use RNG to get $r_i^{(k)}$, $i = 1, \ldots, n$ from $U[0, 1]$;
  - get st. Cauchy values $c_i^{(k)} := \tan(\pi \cdot (r_i^{(k)} - 0.5))$;
  - compute $K := \max_i |c_i^{(k)}|$ (to stay in linearized area);
  - simulate "actual values" $x_i^{(k)} := \widetilde{x}_i - \delta_i^{(k)}$, where $\delta_i^{(k)} := \Delta_i \cdot c_i^{(k)}/K$;
  - simulate error of the indirect measurement:
  $$\delta^{(k)} := K \cdot \left(\widetilde{y} - f\left(x_1^{(k)}, \ldots, x_n^{(k)}\right)\right);$$
- Solve the ML equation $\sum\limits_{k=1}^{N} \dfrac{1}{1 + \left(\dfrac{\delta^{(k)}}{\Delta}\right)^2} = \dfrac{N}{2}$ by bisection, and get the desired Δ.

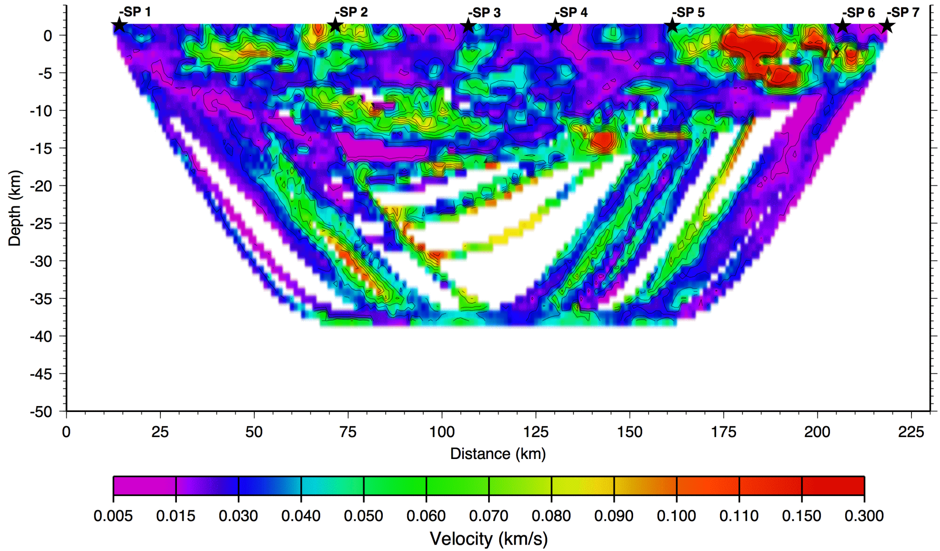
19. A New (Heuristic) Approach

- *Problem:* guaranteed (interval) bounds are too high.
- *Gaussian case:* we only have bounds guaranteed with confidence, say, 90%.
- *How:* cut top 5% and low 5% off a normal distribution.
- *New idea:* to get similarly estimates for intervals, we "cut off" top 5% and low 5% of Cauchy distribution.
- *How:*
  - find the threshold value x_0 for which the probability of exceeding this value is, say, 5%;
  - replace values x for which $x > x_0$ with x_0;
  - replace values x for which $x < -x_0$ with $-x_0$;
  - use this "cut-off" Cauchy in error estimation.
- *Example:* for 95% confidence level, we need $x_0 = 12.706$.

20. Heuristic Approach: Results with 95% Confidence Level



21. Heuristic Approach: Results with 90% Confidence Level



22. Conclusions

- *In the past:* communications were much slower.
- *Conclusion:* use centralization.
- *At present:* communications are much faster.
- *Conclusion:* use cyberinfrastructure.
- *Related problems:*
  - gauge the the uncertainty of the results obtained by using cyberinfrastructure;
  - which data points contributed most to uncertainty;
  - how an improved accuracy of these data points will improve the accuracy of the result.
- *We described:* algorithms for solving these problems.
- *Additional problem:* what if interval estimates are too wide and probabilistic estimates are too narrow.

23. Acknowledgments

This work was supported in part by:

- by National Science Foundation grants HRD-0734825, EAR-0225670, and EIA-0080940,
- by Texas Department of Transportation contract No. 0-5453,
- by the Japan Advanced Institute of Science and Technology (JAIST) International Joint Research Grant 2006-08, and
- and by the Max Planck Institut für Mathematik.