Handling Provenance, Including Mathematical Proofs, in Cyberinfrastructure-Oriented Data Processing

Ann Q. Gates[1], Olga Kosheleva[1],
Vladik Kreinovich[1], Sa-aat Niwitpong[2],
Paulo Pinheiro da Silva[1], and Leonardo Salayandia[1]
[1]University of Texas, El Paso, Texas 79968, USA
contact email vladik@utep.edu
[2]King Mongkut's University of Technology North Bangkok

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:*

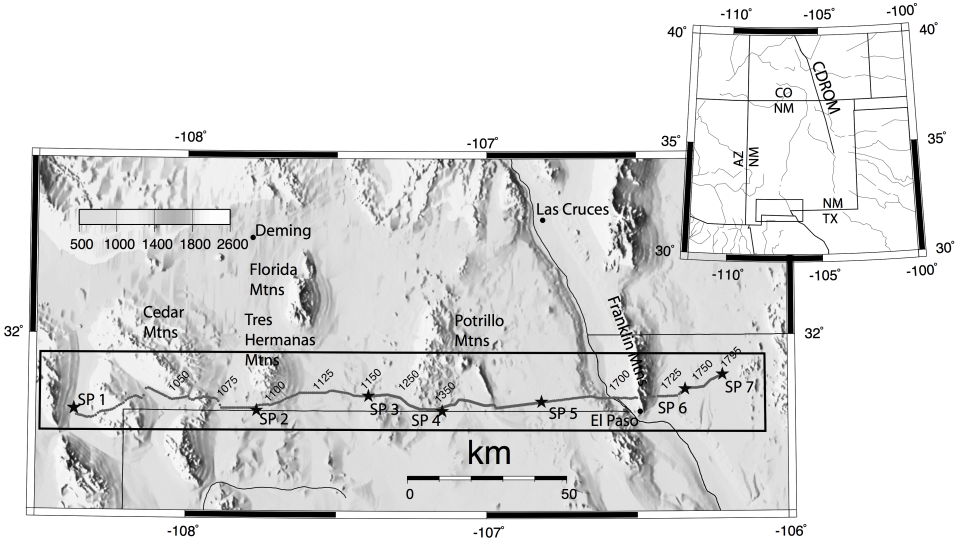
$$\Delta = |c_1| \cdot \Delta_1 + \ldots + |c_n| \cdot \Delta_n.$$

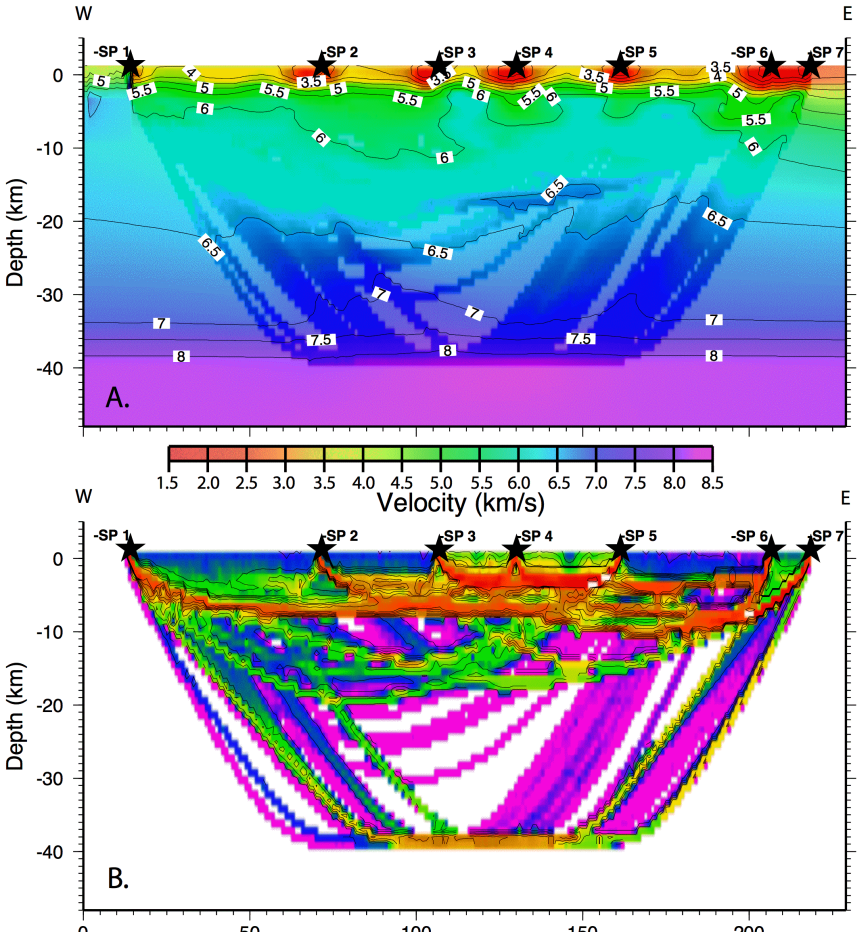
- We can predict how replacing the i-th measurement with a more accurate one ($\Delta_i^{\text{new}} \ll \Delta_i$) will affect Δ.

7. Beyond Probabilistic and Interval Uncertainty

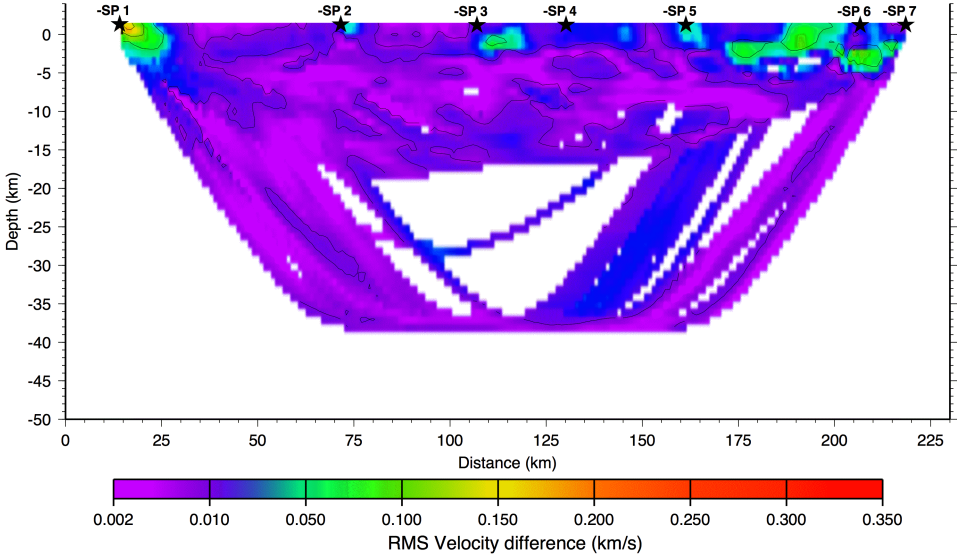
- *Up to now:* we considered two extreme situations:
  - *probabilistic* uncertainty, when we know all the probabilities;
  - *interval* uncertainty, when we have no information about the probabilities.
- *Fact:* probabilistic situation is a particular case of the interval situation.
- *Conclusion:* 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.

8. Case Study: Seismic Inverse Problem in the Geosciences

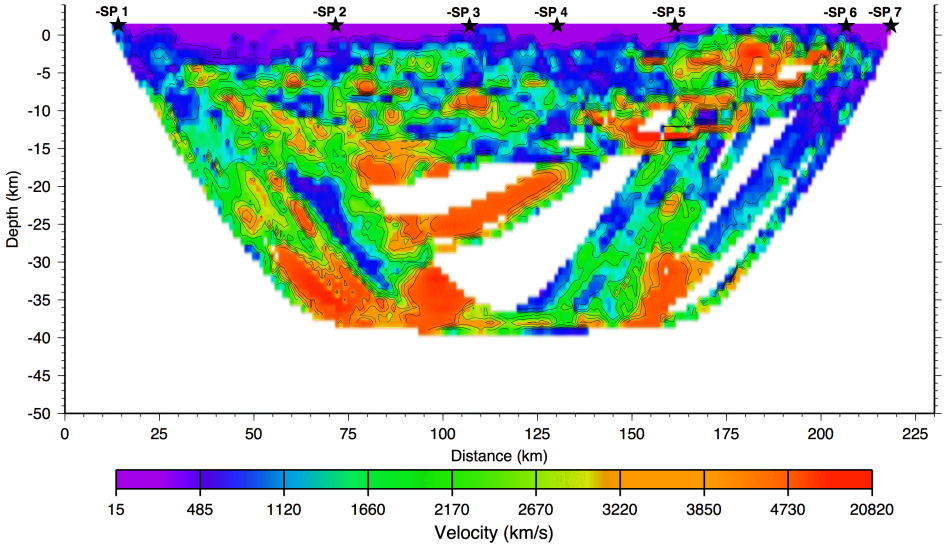




9. Estimating Uncertainty, First Try: Probabilistic Approach



10. Estimating Uncertainty, Second Try: Interval Approach



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

- *Linearization:* $\Delta y = \sum_{i=1}^{n} c_i \cdot \Delta x_i$, where $c_i \stackrel{\text{def}}{=} \dfrac{\partial f}{\partial x_i}$.
- *Formulas:* $\sigma^2 = \sum_{i=1}^{n} c_i^2 \cdot \sigma_i^2$, $\Delta = \sum_{i=1}^{n} |c_i| \cdot \Delta_i$.
- *Numerical differentiation:* n iterations, too long.
- *Monte-Carlo approach:* if Δx_i are Gaussian w/σ_i, then $\Delta y = \sum_{i=1}^{n} c_i \cdot \Delta x_i$ 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 $\Delta y = \sum_{i=1}^{n} c_i \cdot \Delta x_i$ is also Cauchy, w/desired Δ.

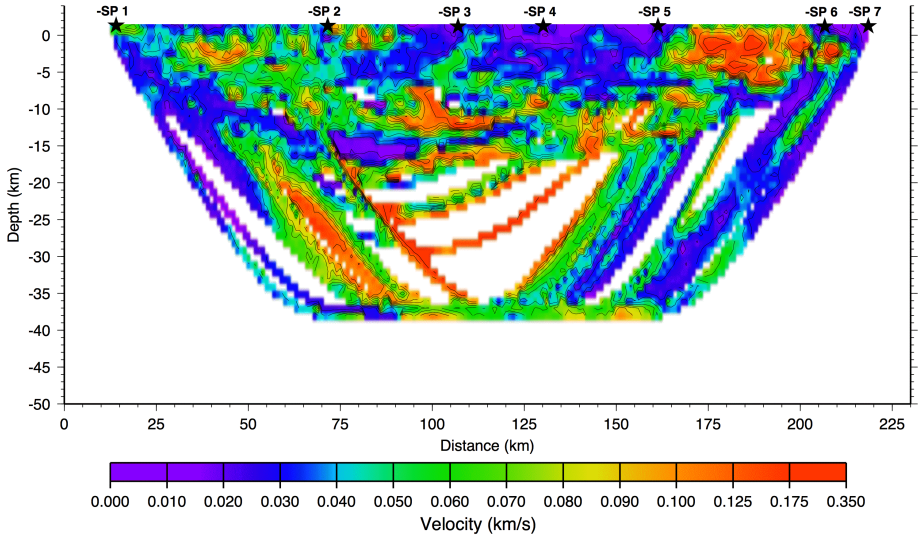
12. Resulting Fast (Linearized) Algorithm for Estimating Interval Uncertainty

- 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 $\sum_{k=1}^{N} \frac{1}{1 + \left(\frac{\delta^{(k)}}{\Delta} \right)^2} = \frac{N}{2}$ by bisection, and get the desired Δ.

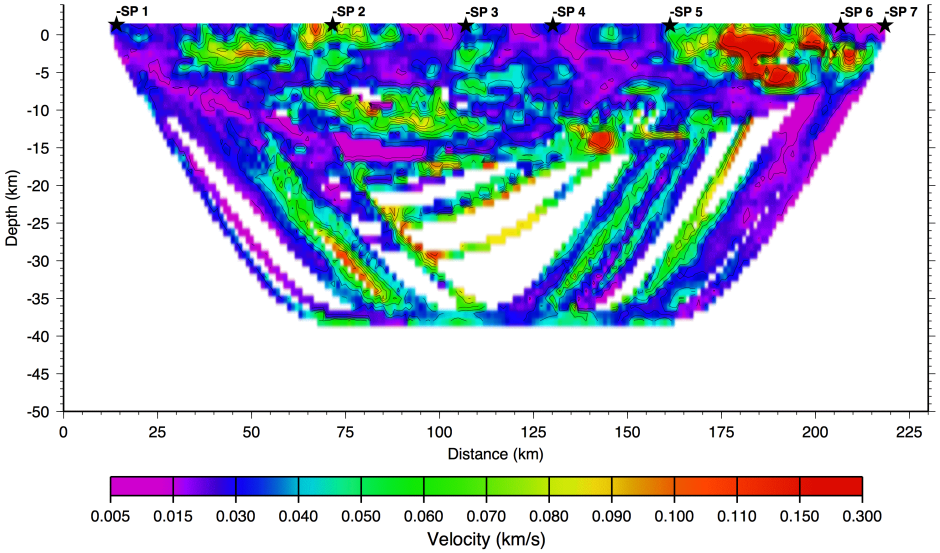
13. A New (Heuristic) Approach

- *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 $x_0 = 12.706$.

14. Heuristic Approach: Results with 95% Confidence Level



15. Heuristic Approach: Results with 90% Confidence Level



16. Proofs as a Particular Case of Provenance

- For handling algorithms provenance, we need to handle different types of such provenance ranging:
  - from expert opinion on heuristic techniques
  - to experimental confirmation of semi-heuristic numerical methods
  - to formal proofs of algorithm correctness.
- In line with the main ideas behind cyberinfrastructure, it is desirable:
  - to combine and process these provenances
  - without moving them to a central location.
- This necessitates, e.g., a need to keep the proof of correctness of the combined algorithm de-centralized.

17. 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.

18. Acknowledgments

This work was supported in part:

- by NSF grant HRD-0734825 and
- by Grant 1 T36 GM078000-01 from the National Institutes of Health.