# Optimized Sampling Frequencies for Weld Reliability Assessments of Long Pipeline Segments

Cesar J. Carrasco[1]
and
Vladik Kreinovich[2]

[1]Department of Civil Engineering
[2]Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
contact email vladik@utep.edu

# 1. Introduction

- *Inspecting pipelines:* important problem (economy, environmental dangers).
- *How it is done:*
  - *Engineering part:* excavations, ultrasonic (and X-ray) measurements.
  - *Statistical part:* from measurements, we estimate $a_i$ and $\sigma_i$ for relevant quantities.
  - *Estimating $p$:* we use known known models to translate these $a_i$ and $\sigma_i$ into the probability of the pipeline failure $p$.
  - *Decision making:* if $p \geq p_0$, the pipeline must be repaired.
- *Fact:* due to limited sampling, we only have a (confidence) interval $[\underline{p}, \overline{p}]$ of possible values of $p$.
- *Regulations:* repaired when $p \geq p_0$ is possible, i.e., when $\overline{p} \geq p_0$.
- *Fact:* pipeline repairs are extremely expensive.
- *Conclusion:* our estimates for $p$ must be as accurate as possible.
- *Problem:* allocate given resources between different possible measurements so as to provide the most accurate estimation of the probability failure.
- *How it is solved now:* exhaustive search of all possible combinations of different measurements $n_1, \ldots, n_k$.
- *Drawback:* time consuming.

# 2. Statistical Approach to Pipeline Reliability Assessment: Motivations

- *Quantities that describe the reliability of a pipeline* $x_1, \ldots, x_k$:
  - the pipe's thickness,
  - parameters describing pipe deformation,
  - different degrees of corrosion, etc.
- *Fact:* the value of each $x_i$ randomly varies from point to point.
- *Why normal distributions:*
  - the value of each $x_i$ is caused by a large number of different independent factors;
  - to extend the pipeline's service, pipelines are designed in such a way that the effect of these major factors is minimized;
  - thus, the state of the pipe is affected by the large number of relatively small difficult-to-exclude processes;
  - due to the Central Limit Theorem, we have a normal distribution.
- *Known:* a normally distributed $x_i$ is uniquely determined by its mean $a_i$ and standard deviation $\sigma_i$.
- *Fact:* for different types of pipelines, there exist models $f$ that estimate the probability of failure $p$:

$$p = f(a_1, \ldots, a_k, \sigma_1, \ldots, \sigma_k).$$

# 3. To Make a Meaningful Decision About the Pipeline, We Must Also Know How Accurate Is the Reliability Estimate

- *How estimates are done:*
  - *Measurements:* for each $x_i$, we perform $n_i$ measurements at different places along the pipeline;
  - *Statistical analysis:* based on the results $x_k^{(1)}, \ldots, x_k^{(n_k)}$ of these measurements, we compute:

$$\widetilde{a}_k = \frac{x_k^{(1)} + \ldots + x_k^{(n_k)}}{n_k};$$

$$\widetilde{\sigma}_k = \sqrt{\frac{(x_k^{(1)} - \widetilde{a}_k)^2 + \ldots + (x_k^{(n_k)} - \widetilde{a}_k)^2}{n_k - 1}}.$$

  - *Estimating probability of failure:*

$$\widetilde{p} = f(\widetilde{a}_1, \ldots, \widetilde{a}_k, \widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_k).$$

- *Problem:* since $\widetilde{a}_i \approx a_i$ and $\widetilde{\sigma}_i \approx \sigma_i$, we have $\widetilde{p} \approx p$.
- *Corollary:* even when $\widetilde{p} < p_0$, it is possible that $p \ge p_0$.
- *Resulting problem:* to make a correct decision on the pipeline's state, we must have information about the estimation error $\Delta p \stackrel{\text{def}}{=} \widetilde{p} - p$.

# 4. Formulas for the Accuracy of the Reliability Estimate

- *Fact:* estimation errors $\Delta a_i \stackrel{\text{def}}{=} \widetilde{a}_i - a_i$ and $\Delta\sigma_i \stackrel{\text{def}}{=} \widetilde{\sigma}_i - \sigma_i$ are small.
- *Conclusion:* we can safely ignore terms which are quadratic (and of higher order) in $\Delta a_i$ and $\Delta\sigma_i$: $\Delta p = \sum\limits_{i=1}^k \dfrac{\partial f}{\partial a_i} \cdot \Delta a_i + \sum\limits_{i=1}^k \dfrac{\partial f}{\partial \sigma_i} \cdot \Delta\sigma_i$.
- *Known:*
  - for a reasonably large number of measurements, the estimation errors $\Delta a_i$ and $\Delta\sigma_i$ are independent and (almost) normally distributed;
  - standard statistical estimates are un-biased – so the mean values of the estimation errors is 0;
  - $\sigma[\Delta a_i] = \dfrac{\sigma_i}{\sqrt{n_i}}$; $\sigma[\Delta\sigma_i] = \dfrac{\sigma_i}{\sqrt{2n_i}}$.
- *Conclusion:* $\Delta p$ is a linear combination $\sum\limits_{j=1}^{2k} \alpha_j \cdot \xi_j$ of independent normally distributed random variables with 0 means and known standard deviations $\sigma[\xi_j]$.
- *Resulting formula:* $\sigma^2 = \sum\limits_{j=1}^{2k} \alpha_j^2 \cdot \sigma[\xi_j]^2$, i.e.:

$$\sigma^2 = \sum_{i=1}^k \left(\frac{\partial f}{\partial a_i}\right)^2 \cdot \frac{\sigma_i^2}{n_i} + \sum_{i=1}^k \left(\frac{\partial f}{\partial \sigma_i}\right)^2 \cdot \frac{\sigma_i^2}{2n_i}.$$

# 5. Measurement Planning

- *Idea:* once we performed a few measurements, and came up with the estimates $\widetilde{a}_i$, $\widetilde{\sigma}_i$, $\widetilde{p}$, and $\sigma$, then:
  - with probability 95%, we have $p \le \widetilde{p} + 2\sigma$;
  - with probability 99.9%, we have $p \le \widetilde{p} + 3\sigma$;
  - in general, $p \le \widetilde{p} + k_0\sigma$ for an appropriate $k_0$.
- *Conclusion:* $\sigma \le \sigma_0 \stackrel{\text{def}}{=} \dfrac{p_0 - \widetilde{p}}{k_0}$, or, equivalently, $\sigma^2 \le \varepsilon_0 \stackrel{\text{def}}{=} \sigma_0^2$.
- *We have shown:* $\sigma^2 = \sum\limits_{i=1}^{k} \left(\dfrac{\partial f}{\partial a_i}\right)^2 \cdot \dfrac{\sigma_i^2}{n_i} + \sum\limits_{i=1}^{k} \left(\dfrac{\partial f}{\partial \sigma_i}\right)^2 \cdot \dfrac{\sigma_i^2}{2n_i}$.
- *Constraint:* $\sum\limits_{i=1}^{k} \dfrac{b_i}{n_i} \le \varepsilon_0$, where $b_i \stackrel{\text{def}}{=} \left(\dfrac{\partial f}{\partial a_i}\right)^2 \cdot \sigma_i^2 + \left(\dfrac{\partial f}{\partial \sigma_i}\right)^2 \cdot \dfrac{\sigma_i^2}{2}$.
- *Objective function:* $\sum\limits_{i=1}^{k} c_i \cdot n_i + c_0 \cdot \max\limits_i n_i$, where:
  - $c_i$ is the cost of $i$-th measurement, and
  - $c_0$ is the cost of a single excavation.
- *Resulting optimization problem:* minimize $\sum\limits_{i=1}^{k} c_i \cdot n_i + c_0 \cdot \max\limits_i n_i$ under the constraint $\sum\limits_{i=1}^{k} \dfrac{b_i}{n_i} \le \varepsilon_0$.

# 6. First Result: Finding the Exact Optimum is Computationally Intractable (NP-Hard)

- *Reminder:* minimize $\sum_{i=1}^{k} c_i \cdot n_i + c_0 \cdot \max_i n_i$ under the constraint $\sum_{i=1}^{k} \frac{b_i}{n_i} \le \varepsilon_0$.
- *First result:* even a simplified version of this problem is NP-hard, when:
  - we fix $n < n'$, and
  - for each $i$, we either perform $n$ measurements, or $n'$ measurements.
- *Proof:* by reduction to a known NP-hard subset sum problem:
  - given integers $s_1, \ldots, s_k > 0$ and an integer $s > 0$,
  - check whether it is possible to find a subset of this set of integers whose sum is equal to exactly $s$, i.e., whether $\exists y_i \in \{0, 1\}$ for which $\sum s_i \cdot y_i = s$.
- *Reduction:* $y_i = 1$ if we choose $n'$, and $y_i = 0$ if we choose $n$;
- $c_0 = 0$, $c_i \stackrel{\text{def}}{=} \frac{s_i}{n' - n}$; $b_i \stackrel{\text{def}}{=} \frac{s_i \cdot n \cdot n'}{n' - n}$, and $\varepsilon_0 \stackrel{\text{def}}{=} s_0 + \sum_{i=1}^{k} \frac{b_i}{n'}$
- *Meaning of NP-hardness:* we cannot find the optimal $n_i$ faster than by time-consuming trying of all possible combinations of $n_i$.
- *Uncertainty in the input makes faster algorithms possible:*
  - in practice, the parameters $b_i$, $c_i$, and $c_0$ are only approximately known;
  - it is therefore reasonable, instead of looking for an optimal solution, to only look for an asymptotically optimal one.

# 7. New Algorithm: Motivations

- *Known:*
  - discrete optimization problems are more difficult than
  - the continuous ones – in which the values $n_i$ can take arbitrary real values.
- *Resulting approximate algorithm:*
  - solve the corresponding continuous optimization problem;
  - round off $n_i$ to the nearest integers.
- *Accuracy:*
  - rounding's absolute error is $\le 0.5$, so
  - relative error is $\le \dfrac{0.5}{n_i} \sim \dfrac{1}{n_i}$.
- *How to solve the continuous problem:* for optimal $n_i$,
  - adding small $\Delta n_i$ for which $n'_i = n_i + \Delta n_i$ satisfy the constraints
  - should not decrease the value of the objective function.

# 8. Resulting Algorithm

- Sort $i$ in the increasing order of $b_i/c_i$; computation time $O(k\cdot\log(k))$.
- For every $t=1,\ldots,k$, compute

$$Z_t \stackrel{\rm def}{=} \sum_{j-1}^{t-1} \sqrt{b_j\cdot c_j}+\sqrt{\sum_{l=t}^k b_l}\cdot\sqrt{\sum_{l=t}^k c_l+c_0};$$

  from $Z_t$ to $Z_{t+1}$, each sum changes by only one term, so we only need a constant number of terms to find each of $k$ values $Z_t$ – to the total of $O(k)$.

- Find $t$ for which $Z_t$ is the smallest; for this $t$, compute $\sqrt{\lambda}=Z_t/\varepsilon_0$.
- Compute $n_i=\sqrt{\lambda}\cdot\sqrt{\dfrac{b_i}{c_i}}$ for $i<t$ and for $i\ge t$,

$$n_i=\sqrt{\lambda}\cdot\sqrt{\frac{\sum\limits_{l=t}^k c_l}{\sum\limits_{l=t}^k c_l+c_0}}.$$

- This algorithm requires computation time

$$O(k\cdot\log(k))+O(k)=O(k\cdot\log(k)).$$

# 9. Acknowledgments

This work was supported in part:

- by NSF grant EAR-0225670,
- by NASA under cooperative agreement NCC5-209,
- by NIH grant 3T34GM008048-20S1,
- by Army Research Lab grant DATM-05-02-C-0046,
- by Star Award from the University of Texas System,
- and by Texas Department of Transportation grant No. 0-5453.