Lance R. Curtis

Engineering enthusiast | Metallurgist | Mechanical failure analyst | Reliability engineer | Author

www.linkedin.com/in/lancercurtis/

2009 Missing Data Project

Documentation for LinkedIn Profile
29 October 2013

Disclaimer

This presentation summarizes a project performed by the author while employed at General Electric (GE).

The presentation which documents the work performed with actual data sets is proprietary to GE and therefore not in possession of the author.

This presentation contains only a high-level summary of that work. No GE proprietary information appears in this presentation, and no GE equipment was used or accessed in the preparation of this presentation.

www.linkedin.com/in/lancercurtis/

Contents

www.linkedin.com/in/lancercurtis/

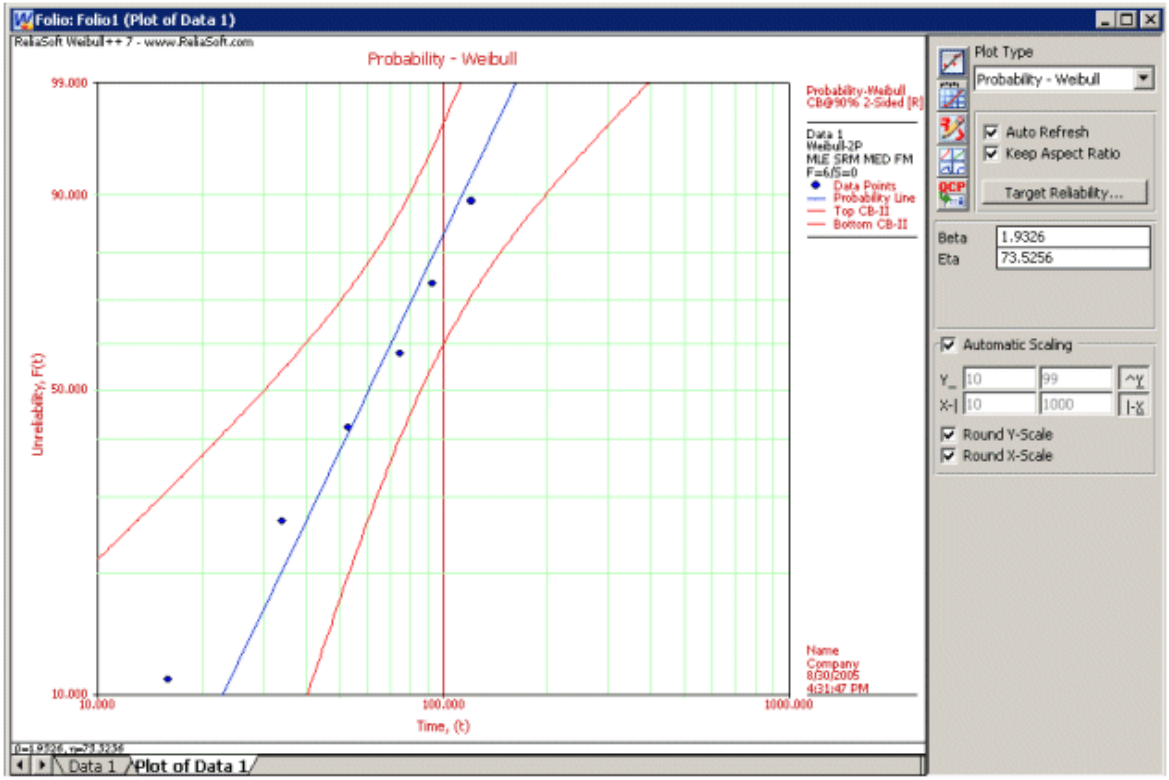
Background

- Gas turbines supply electricity for the power grid and certain manufacturing plants.
- In general terms, an unplanned outage occurs when a turbine goes offline unexpectedly.

- **Unplanned outages can be very costly and may require an offline unit (which might not be available) to come online to maintain the previously supplied load.**

www.linkedin.com/in/lancercurtis/

Predicting unplanned outage risk



- Data on turbine operation is collected and analyzed using statistical distributions to make models.
- Engineers like the Weibull distribution; it can be adjusted to mimic several other distributions.

Weibull models can predict unplanned outage risk, but the models are only as good as the data supplied to create them.

General problem

- Data collection practices can vary greatly.

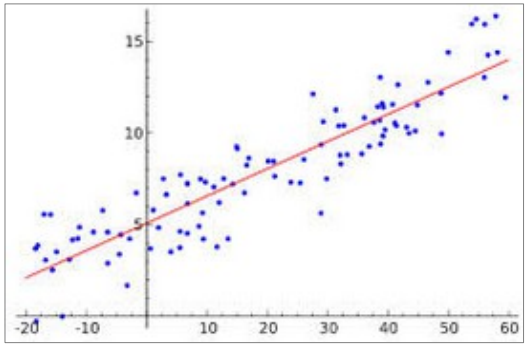
  - Some sites are really good at collecting data.
  - Other sites either don't care about collecting data or only report data that makes them look good because of cultural considerations.

- "Garbage in, garbage out" is the rule of thumb when making predictive models.

  - Making good models is easy when data sets are complete with good data.
  - But what happens when data sets are incomplete or questionable? Good models are still needed!

Specific problem

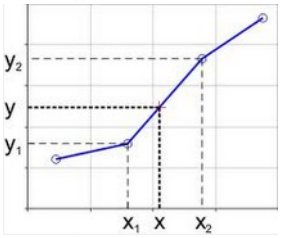
- Engineers need to estimate missing data to provide a more complete data set for modeling.
  - Quality documents required use of linear regression.



  - Linear regression is fine if only a handful of data points are missing, but when thousands of estimated data points are needed, linear regression is *very* time consuming.
- A more efficient way to estimate missing data that doesn't sacrifice efficacy was needed.

Proposed solution

- Linear interpolation and extrapolation are specific forms of linear regression and should provide adequate results.



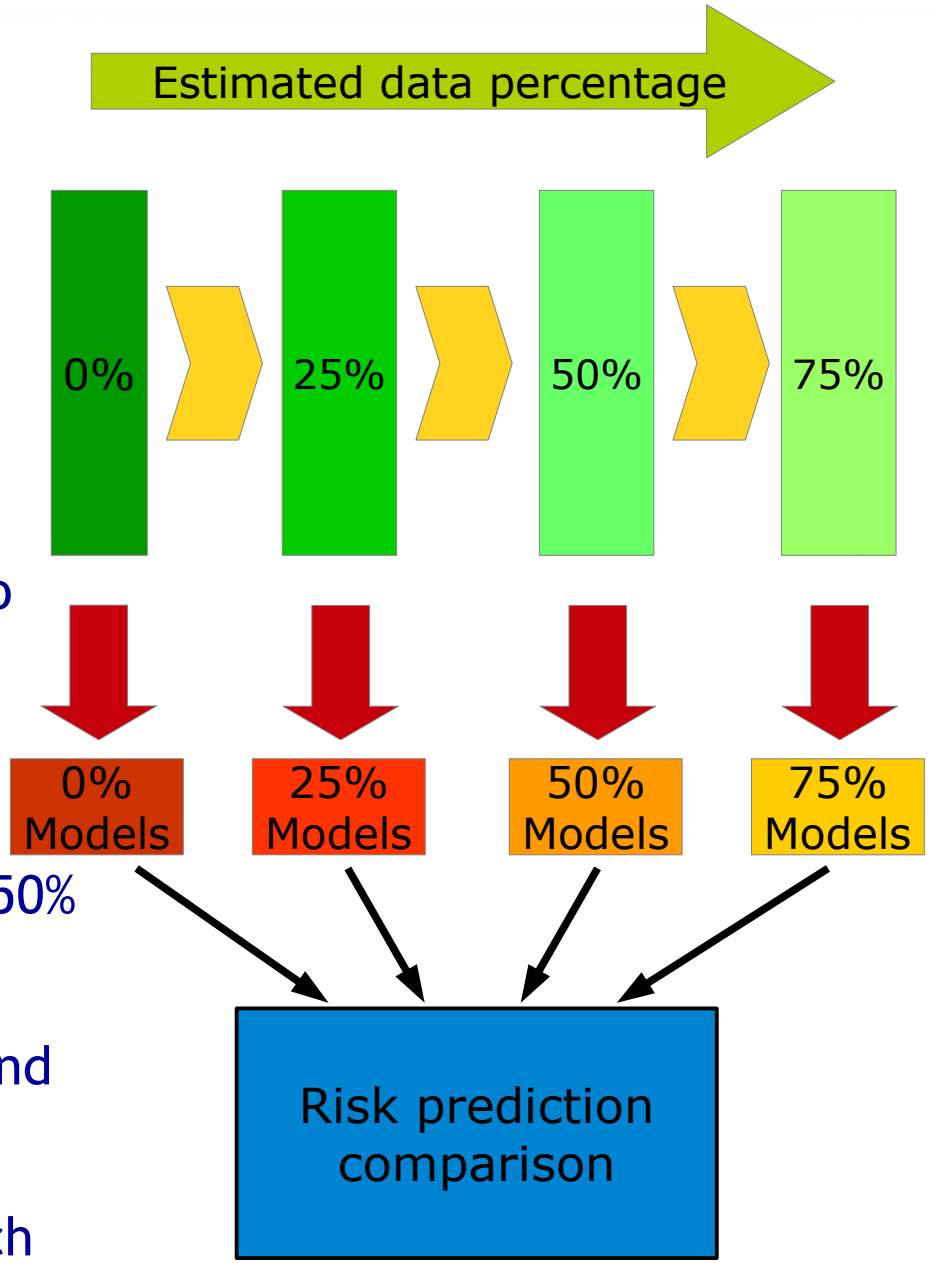
  - Interpolation & extrapolation can be incorporated quickly into Excel spreadsheets used for managing data sets.

- What results when interpolated and extrapolated data are used in modeling?

  - You don't need to be Einstein to realize that the models using estimated data need to predict a similar risk profile as models made from actual good data sets.

Study design

- **Use a fleet with a well-established data set to make multiple models for comparison.**
- The data set for the well established fleet has no estimated data (0%) and so serves as the baseline.
- Randomly replace 25% of the baseline data set with data estimated via interpolation. Repeat the process for 50% and 75% of the baseline data set.
- Make models using the 0%, 25%, 50%, and 75% data sets.
- Compare the predictive abilities of each model.



Study results

- Models made from interpolated data sets showed no significant difference in predictive ability.

  - Interpolation of missing data points can be used to any extent desired without sacrificing effectiveness.

- Models made from increasingly extrapolated data sets showed increasing differences from the baseline.

  - Extrapolation of missing data points should be used with extreme caution.

Process improvement

- **Process completion time was improved by up to 99%!**
  - The worst case data set with extensive missing data required 80 manhours to complete using linear regression.
  - The same data set required only 2 manhours to complete using interpolation.
- Quality documents were updated to reflect the new, proven process.