

Additive-Tree-Based Learning and Prediction for Large Recurrent Event Data

LUNCH AND LEARN

Thursday, Oct 1, 2026 | 12pm ET

Thank you for attending!



Xiao Liu

David M. McKenney Family Associate Professor
H. Milton Stewart School of Industrial and Systems
Engineering (ISyE)

Associate Director, Brook Byers Institute of
Sustainable Systems (BBISS)

Georgia Institute of Technology

<https://sites.google.com/site/liuxiaosite1/>

About our Related Course Series

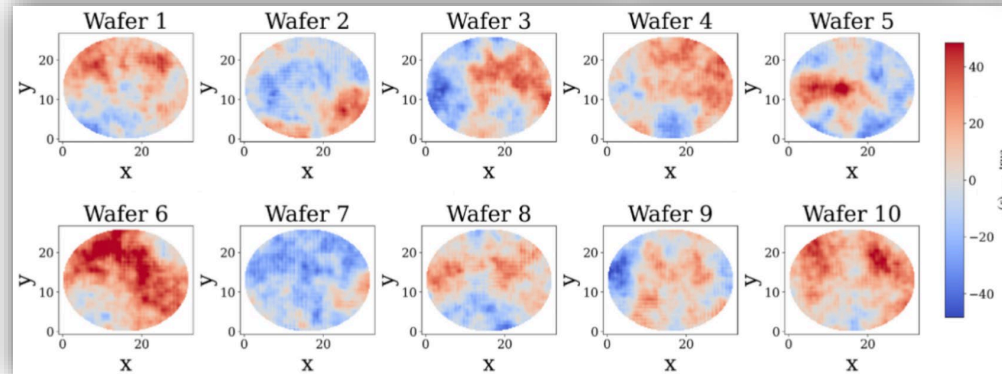
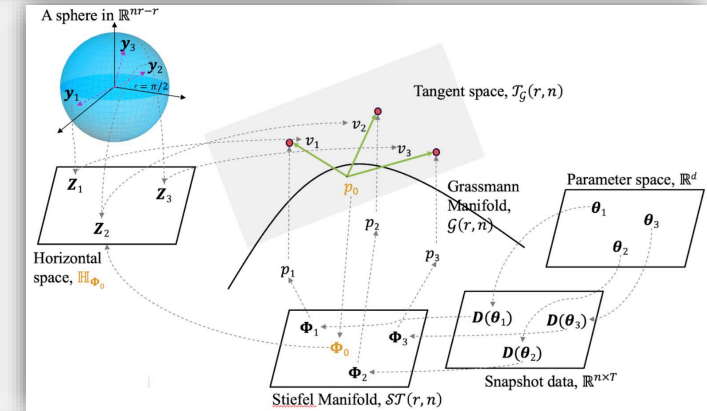
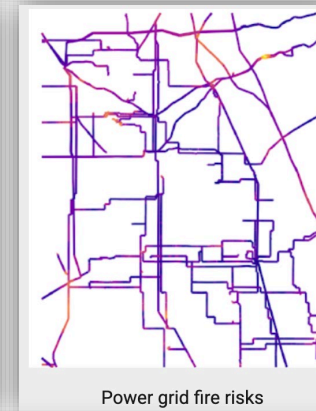
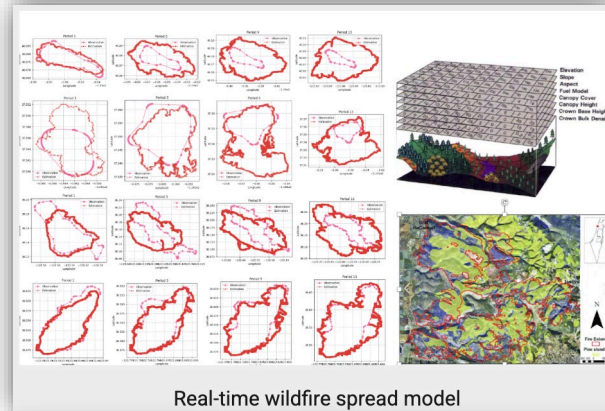
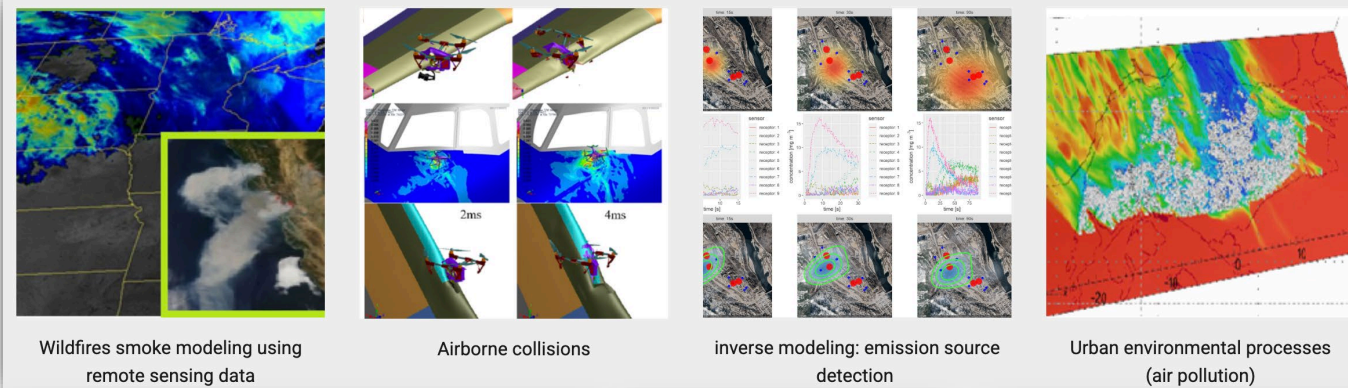
scl.gatech.edu/gaiascp



Supply Chain and
Logistics Institute

About My Research

- Statistical Learning for Dynamical Systems
- Statistical Learning for Reduced-Order Modeling
- Wildfires-Energy-Environment-Resilience
- Wafer Chip Warpage Modeling and Control



About this Webinar

- A useful and flexible statistical learning method for modeling and prediction of **recurrent event data**
- Main references:
 - Liu, X., and Pan, R. (2020), "Analysis of Large Heterogeneous Repairable System Reliability Data with Static System Attributes and Dynamic Sensor Measurement in Big Data Environment", Technometrics, 62, 206-222.
 - Liu, X., and Pan, R. (2021), "Boost-R: Gradient Boosting for Recurrent Event Data", Journal of Quality Technology, Special Issue-Artificial Intelligence & Statistics for Quality Technology, 53, 545-565.
 - Code is available

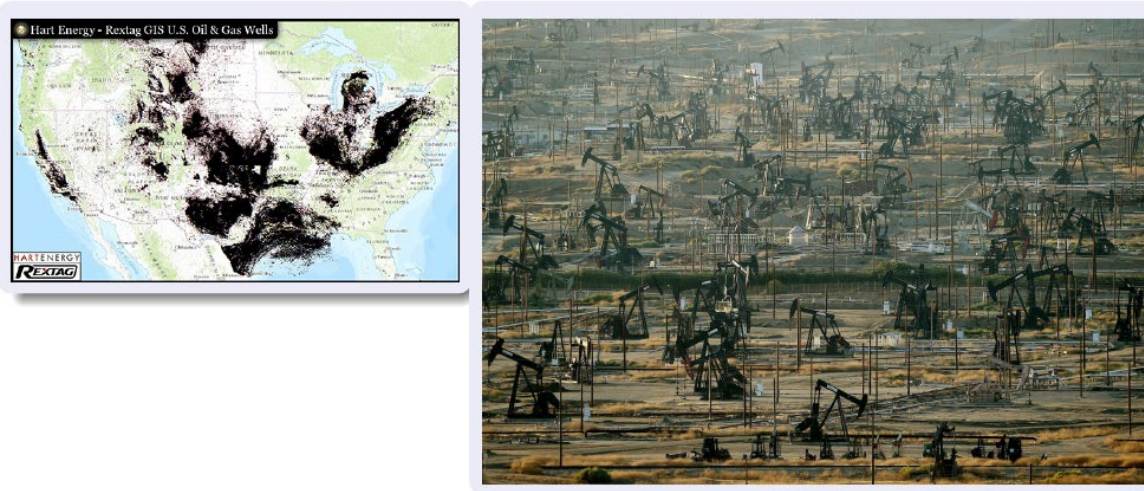
Xiao Liu, ISyE, GT (xiao.liu@isye.gatech.edu)



What is recurrent event data?

Motivating Example: Oil and Gas Well Failure Data

- A modified data set from a multinational oil and natural gas service company which provides predictive well maintenance service.
- Failure data, system attributes, and sensor measurement for 8232 oil and gas wells in California.



xiao.liu@isye.gatech.edu

2026

3 / 52

Static Well Attributes

- A number of 8 well attributes, $x_1 \sim x_8$, are available for reliability analysis.
 - ▶ well size, nominal operating settings/conditions, geo-locations, etc.
 - ▶ heterogeneity and subpopulations

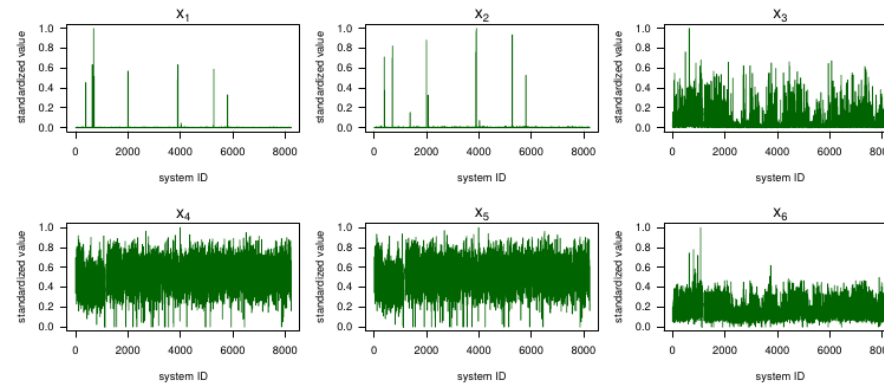


Figure: Standardized values of $x_1 \sim x_6$ for all 8232 systems.

Static Well Attributes (Geo-locations)

- Covariates, x_7 and x_8 , respectively specify the latitude and longitude of each well.
- Neighboring wells share common (unobserved) environmental conditions such as temperature and humidity variation, soil type, contamination, etc.
- Spatial trend

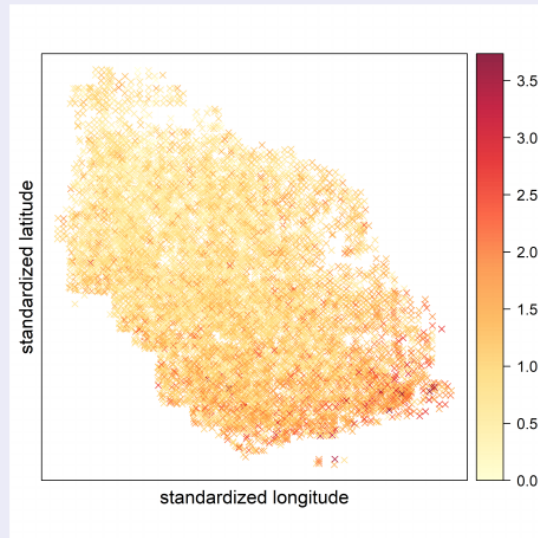


Figure: Spatial locations of a total number of 8232 systems (the legend shows the average annual failures for each system)

Static Well Attributes (Geo-locations)

- Covariates, x_7 and x_8 , respectively specify the latitude and longitude of each well.
- Neighboring wells share common (unobserved) environmental conditions such as temperature and humidity variation, soil type, contamination, etc.
- Spatial trend

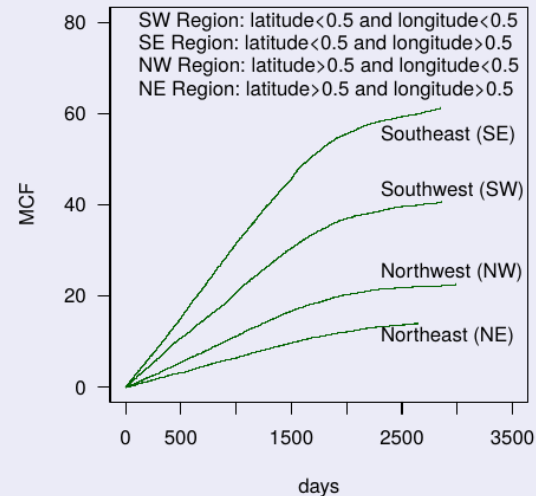


Figure: Estimated Mean Cumulative Function (MCF) for systems over four geo-regions [Nelson(1995)]

Dynamic Sensor Measurement

- Seven dynamic sensor measurement of system working conditions:
 - ▶ $z_1(t) \sim z_5(t)$: dynamic torque and load measurement including gear torque, load range, etc.
 - ▶ $z_6(t) \sim z_7(t)$: monitored gearbox and structural stress.

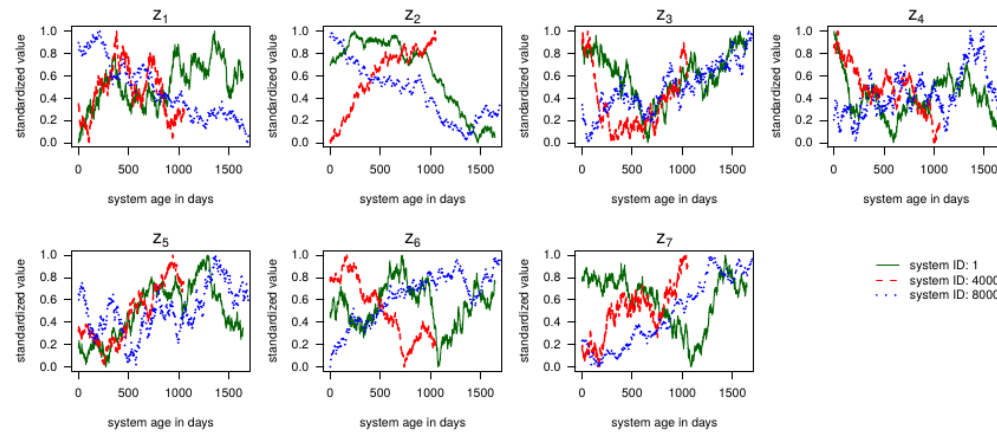
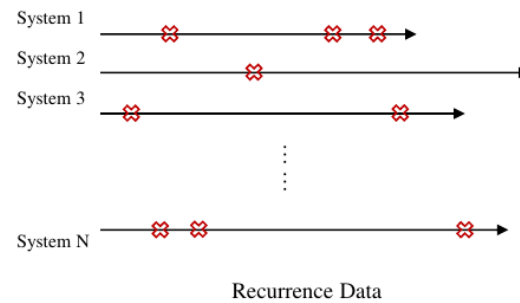


Figure: Sensor measurement in terms of torque, load and stress, $z_1 \sim z_7$, for three selected systems: system #1, system #4000, and system #8000.

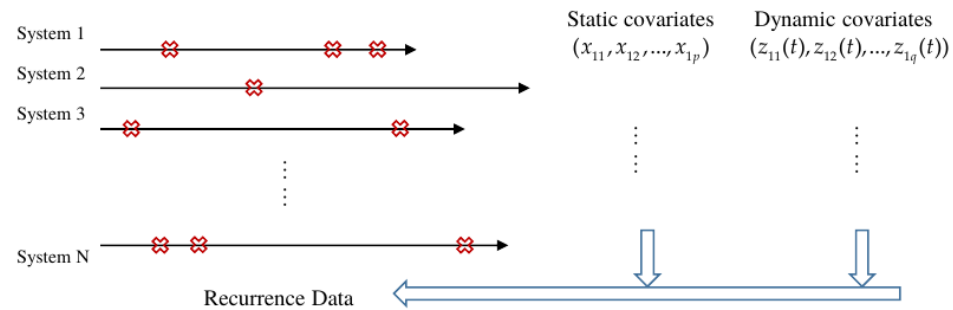
Introduction

- Advances of sensing and IoT technologies are re-shaping the landscape of statistical reliability modeling and analysis.
- For a large fleet of heterogeneous repairable systems:
 - failure/repair times



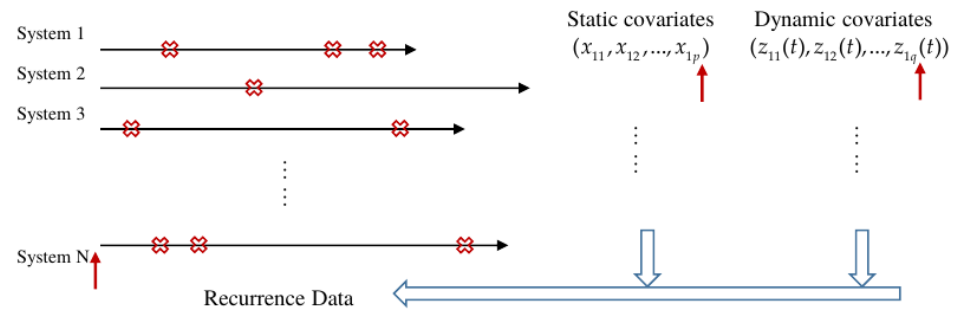
Introduction

- Advances of sensing and IoT technologies are re-shaping the landscape of statistical reliability modeling and analysis.
- For a large fleet of heterogeneous repairable systems:
 - failure/repair times
 - a set of static and dynamic covariates



Introduction

- Advances of sensing and IoT technologies are re-shaping the landscape of statistical reliability modeling and analysis.
- For a large fleet of heterogeneous repairable systems:
 - failure/repair times
 - a set of static and dynamic covariates



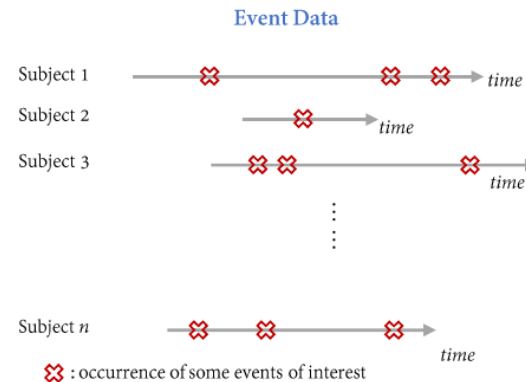
Introduction

- Advances of sensing and IoT technologies are re-shaping the landscape of statistical reliability modeling and analysis.
- For a large fleet of heterogeneous repairable systems:
 - failure/repair times
 - a set of **static** and **dynamic** covariates
- **Static** covariates: system attributes
 - nominal working conditions, geo-locations, etc.

Introduction

- Advances of sensing and IoT technologies are re-shaping the landscape of statistical reliability modeling and analysis.
- For a large fleet of heterogeneous repairable systems:
 - ▶ failure/repair times
 - ▶ a set of **static** and **dynamic** covariates
- **Static** covariates: system attributes
 - ▶ nominal working conditions, geo-locations, etc.
- **Dynamic** covariates: multi-channel sensor measurement of system operating and environmental conditions.
 - ▶ sometimes referred to as the SOE data [Meeker and Hong(2014)], representing a key trend of modern reliability analysis in the age of Big Data.

Recurrence Data from Interdisciplinary Domains



Examples of events of interest from interdisciplinary domains:

*Cyber-attack and fraudulent activities, Natural disasters,
Extreme weather events, Engineering system failures,
Disease occurrence, Traffic accidents, etc.*

Heterogeneous Feature Information

- **Static** subject attributes
 - (disaster and weather): geo-locations, topography, land use, ...
 - (healthcare): Age, gender, family history, diet, ...
 - (cyber security): geo-locations, nationalities, hackers' attributes, ...
 - (transportation): highway id, road width, number of lanes, ...
 - (engineering): dimension, nominal operating parameters, ...
 - (online retail): Season, age, gender, location, holidays, ...
- **Dynamic** signals
 - (disaster and weather): station measurements, radar data, ...
 - (healthcare): ECG signals, monitored pressure, cholesterol, ...
 - (cyber security): internet traffic, ...
 - (transportation): traffic volume, precipitation, weather, ...
 - (engineering): load, stress, rotational speed, ...
 - (online retail): economic trends, price fluctuations, ...
- **Unstructured** information
 - (disaster and weather): images/text on social media, ...
 - (healthcare): patient record (unstructured) text, ...
 - (cyber security): geopolitics events, voice/text on social media, ...
 - (transportation): images/text on social media, ...
 - (engineering): expert opinion, engineering knowledge, ...
 - (online retail): social media data, government regulations, ...

Analysis of Recurrence Data

- Nonparametric graphical methods for estimating the Mean Cumulative Function (MCF)
[Nelson(1995), Meeker and Escobar(1998)].
 - ▶ flexible and widely used in practice.
 - ▶ how can one model the complex relationship between failure processes and covariates?

Analysis of Recurrence Data

- **Nonparametric** graphical methods for estimating the Mean Cumulative Function (MCF)
[Nelson(1995), Meeker and Escobar(1998)].
 - ▶ flexible and widely used in practice.
 - ▶ how can one model the complex relationship between failure processes and covariates?
- **Parametric** models based on point and counting processes
[Xie(1991), Fleming and Harrington(1991), Anderson et al.(1993), Meeker and Escobar(1998), Rigdon and Basu(2000)].
 - ▶ Nonhomogeneous Poisson Process (NHPP), renewal process, piecewise exponential model, modulated power law process, bounded intensity model, as well as their variations
[Pulcini(2001), Pan and Rigdon(2009), Yang, et al.(2012), Ye, et al.(2013b), Mittman, et al.(2018)].

Challenges

- Some fundamental challenges arise in the modern Big Data analytics environment.

Challenges

- Some fundamental challenges arise in the modern Big Data analytics environment.
- **Heterogeneity**. Impossible to ignore individual heterogeneity in analyzing large recurrence data [Ye, et al.(2013), Mittman, et al.(2018)].
 - ▶ Subpopulations exist due to individual attributes.
 - ▶ Different operating and environmental conditions.
 - ▶ Different dependence structure between event processes and covariates.

Challenges

- Some fundamental challenges arise in the modern Big Data analytics environment.
- **Heterogeneity**. Impossible to ignore individual heterogeneity in analyzing large recurrence data [Ye, et al.(2013), Mittman, et al.(2018)].
 - ▶ Subpopulations exist due to individual attributes.
 - ▶ Different operating and environmental conditions.
 - ▶ Different dependence structure between event processes and covariates.
- **Model Specification**. Difficult to force and validate any parametric or semi-parametric assumptions.
 - ▶ Interactions and nonlinear effects of covariates.
 - ▶ Both the effects of covariates and the parametric link between failure processes and covariates may vary across subpopulations.

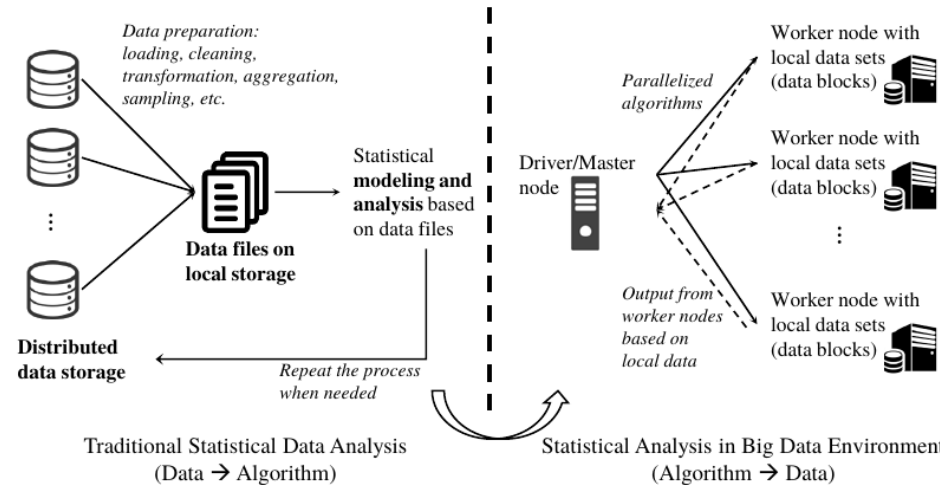
Challenges (cont.)

- **Model Complexity and Interpretability.** Many of these covariates are redundant from either the statistical modeling or domain knowledge perspectives, although an increasingly larger number of covariates has been made available.

Challenges (cont.)

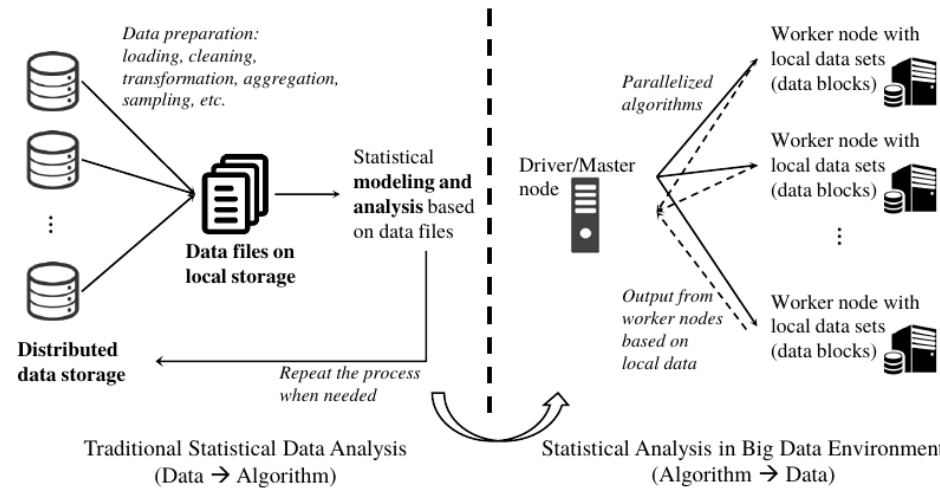
- **Model Complexity and Interpretability.** Many of these covariates are redundant from either the statistical modeling or domain knowledge perspectives, although an increasingly larger number of covariates has been made available.
- **The Issue of “Data Locality”.** The ways data are stored on modern Big Data infrastructure affects how statistical data analysis can be performed efficiently.

“Data Locality” in the Age of Big Data



- The large data are stored on distributed storage systems, such as the Hadoop Distributed File System, as a common practice in industry today.
- Reliability, Scalability, Efficiency, Security.

“Data Locality” in the Age of Big Data



- **Algorithm → Data**

- ▶ MapReduce programming model, Apache Spark, GPU computing.

Tree-Based Methods

- The work investigates the integration of the modern statistical learning (i.e., the **Random Forests (RF)** or **Boosting**) and classical repairable system reliability analysis (i.e., the nonparametric MCF estimator, and parametric NHPP models).



source: <https://www.stat.berkeley.edu/~breiman/RandomForests/>

Tree-Based Methods

- The work investigates the integration of the modern statistical learning (i.e., the Random Forests (RF) or Boosting) and classical repairable system reliability analysis (i.e., the nonparametric MCF estimator, and parametric NHPP models).

Tree-Based Methods

- The work investigates the integration of the **modern statistical learning** (i.e., the **Random Forests (RF)** or **Boosting**) and **classical repairable system reliability analysis** (i.e., the nonparametric MCF estimator, and parametric NHPP models).
- **Random Forests (RF)**: ensembles (forests) from de-correlated base learners (trees)—one of the most popular ensemble tree methods which has gained tremendous success in practice [Breiman(2001)].
 - ▶ Bagging + Random node split

Tree-Based Methods

- The work investigates the integration of the **modern statistical learning** (i.e., the **Random Forests (RF)** or **Boosting**) and **classical repairable system reliability analysis** (i.e., the nonparametric MCF estimator, and parametric NHPP models).
- **Random Forests (RF)**: ensembles (forests) from de-correlated base learners (trees)—one of the most popular ensemble tree methods which has gained tremendous success in practice [Breiman(2001)].
 - ▶ Bagging + Random node split
- **Boosting Trees**: sequentially fitting a sequence of “weaker learners” in which each tree explains the data variation not explained by previous trees [Chipman et al.(2010)].

Tree-Based Methods

- The work investigates the integration of the modern statistical learning (i.e., the Random Forests (RF) or Boosting) and classical repairable system reliability analysis (i.e., the nonparametric MCF estimator, and parametric NHPP models).
- Random Forests (RF): ensembles (forests) from de-correlated base learners (trees)—one of the most popular ensemble tree methods which has gained tremendous success in practice [Breiman(2001)].
 - ▶ Bagging + Random node split
- Boosting Trees: sequentially fitting a sequence of “weaker learners” in which each tree explains the data variation not explained by previous trees [Chipman et al.(2010)].
- Nonparametric MCF and Parametric NHPP: one of the most successful statistical methods for analyzing repairable system reliability data (i.e., recurrence data) in industry [Nelson(1995), Meeker and Escobar(1998), Doganaksoy et al.(2006), Zuo, et al.(2012)].

Tree-Based Methods

- The work investigates the integration of the modern statistical learning (i.e., the Random Forests (RF) or Boosting) and classical repairable system reliability analysis (i.e., the nonparametric MCF estimator, and parametric NHPP models).
- Random Forests (RF): ensembles (forests) from de-correlated base learners (trees)—one of the most popular ensemble tree methods which has gained tremendous success in practice [Breiman(2001)].
 - ▶ Bagging + Random node split
- Boosting Trees: sequentially fitting a sequence of “weaker learners” in which each tree explains the data variation not explained by previous trees [Chipman et al.(2010)].
- Nonparametric MCF and Parametric NHPP: one of the most successful statistical methods for analyzing repairable system reliability data (i.e., recurrence data) in industry [Nelson(1995), Meeker and Escobar(1998), Doganaksoy et al.(2006), Zuo, et al.(2012)].
- Other tree-based methods for survival data: [Hothorn et al.(2004), Hothorn et al.(2006), Fan et al.(2006), Fan et al.(2009), Ishwaran et al.(2008), Chipman et al.(2010), Bou-Hamad(2011)].

The Main Idea

- Traditional binary regression tree: $T(\mathbf{x}; \boldsymbol{\theta}) = \sum_{j=1}^J \mu_j I(\mathbf{x} \in R_j)$, where $\boldsymbol{\theta} = (R_j, \mu_j)_{j=1}^J$.

The Main Idea

- Traditional binary regression tree: $T(\mathbf{x}; \boldsymbol{\theta}) = \sum_{j=1}^J \mu_j I(\mathbf{x} \in R_j)$, where $\boldsymbol{\theta} = (R_j, \mu_j)_{j=1}^J$.
- For recurrent event data, $\mathfrak{D} = (\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i(t))$ where $|\mathfrak{D}| = n$, $\mathbf{x}_i \in \mathbb{R}^p$, $\mathbf{y}_i \in \mathbb{R}^{r_i+1}$ and $\mathbf{z}_i(t) : [0, \infty) \rightarrow \mathbb{R}^q$, a point process $\mu_j(t)$ is needed to capture the event recurrence pattern of the individuals on a tree node j :

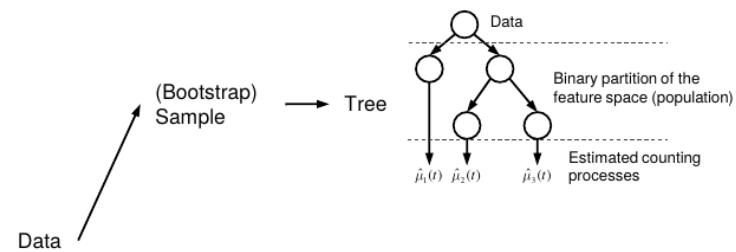
The Main Idea

- Traditional binary regression tree: $T(\mathbf{x}; \boldsymbol{\theta}) = \sum_{j=1}^J \mu_j I(\mathbf{x} \in R_j)$, where $\boldsymbol{\theta} = (R_j, \mu_j)_{j=1}^J$.
- For recurrent event data, $\mathfrak{D} = (\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i(t))$ where $|\mathfrak{D}| = n$, $\mathbf{x}_i \in \mathbb{R}^p$, $\mathbf{y}_i \in \mathbb{R}^{r_i+1}$ and $\mathbf{z}_i(t) : [0, \infty) \rightarrow \mathbb{R}^q$, a point process $\mu_j(t)$ is needed to capture the event recurrence pattern of the individuals on a tree node j :

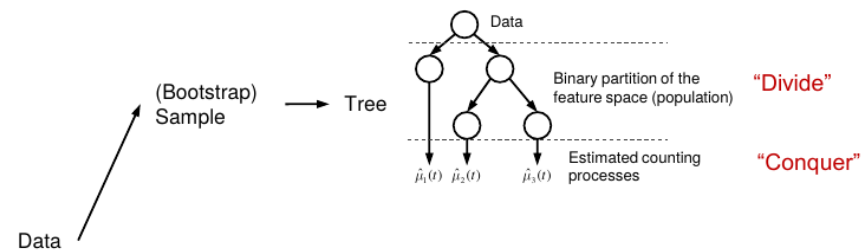
$$f_K(\mathbf{x}, \mathbf{z}(t)) = \sum_{k=1}^K T(\mathbf{x}, \mathbf{z}(t)), \quad T \in \mathfrak{T} \quad (1)$$

where $\mathfrak{T} = \{T(\mathbf{x}, \mathbf{z}(t)) = \mu_{g(\mathbf{x}, \mathbf{z}(t))}(t)\}$ is the space of the trees, $\mu(t) : [0, \infty) \rightarrow \mathbb{N}^+$ is a counting process, and g is some mapping that maps $(\mathbf{x}, \mathbf{z}(t))$ to a tree node.

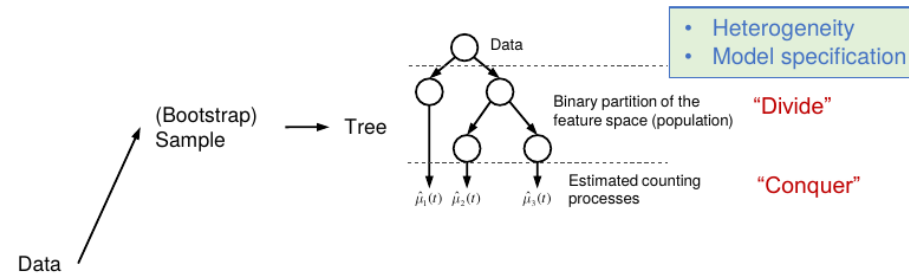
The Main Idea of RF-R



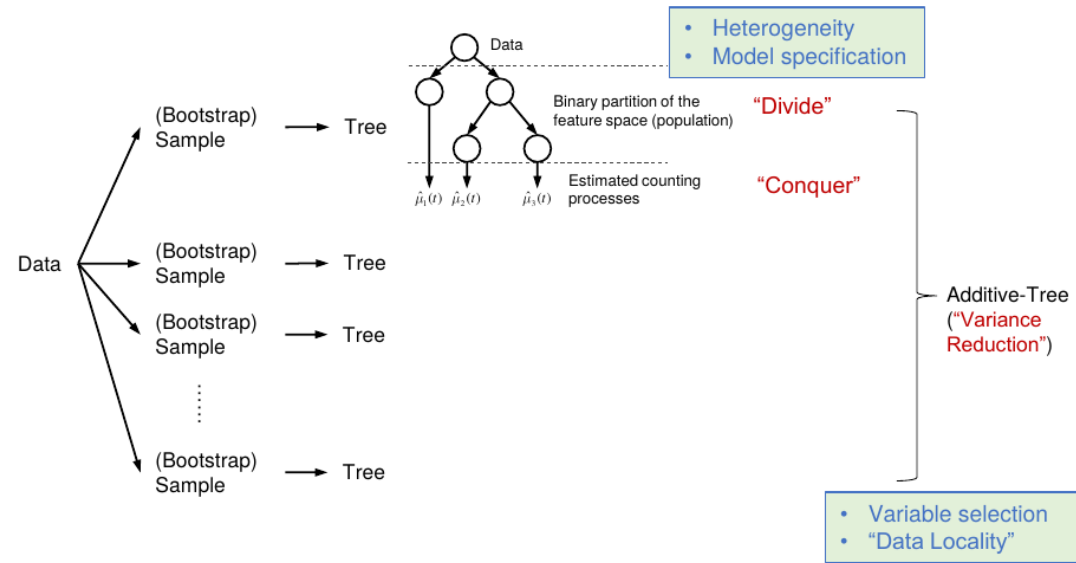
The Main Idea of RF-R



The Main Idea of RF-R



The Main Idea of RF-R



The Main Idea of Boost-R

The objective function

$$\text{minimize} \quad \sum_{i=1}^n l(\mu_i(t), \hat{\mu}_i(t)) + \sum_{k=1}^K \Omega(T^{(k)}) \quad (2)$$

where $\Omega(T^{(k)}) = \gamma D + \frac{1}{2} \lambda \sum_{d=1}^D \|\hat{\omega}_d(t)\|_2^2$ and l is a differentiable convex function.

The Main Idea of Boost-R (Gradient Boosting)

- Discretize $\mu_i(t)$ at equally-spaced times $\{t_j\}_{j=1}^q$, and let $\{\mu_{i,j}\}_{j=1}^q$ be the values of $\mu_i(t)$ at the discrete time points.
- Let $\{\hat{\mu}_{i,j}^{(k-1)}\}_{j=1}^q$ be the estimates from the first $(k-1)$ ensemble trees.
- Let $\{\omega_{r(\mathbf{x}_i),j}^{(k)}\}_{j=1}^q$ be the prediction generated from the k th tree for individual i .

Gradient Boosting

$$\text{minimize } \sum_{i=1}^n \left\{ \sum_{j=1}^q \tilde{l}(\mu_{i,j}, \hat{\mu}_{i,j}^{(k-1)}) + \omega_{r(\mathbf{x}_i),j} \Delta + \gamma D + \frac{1}{2} \lambda \sum_{d=1}^D \|\omega_{d,\cdot}\|_2^2 \right\} \quad (3)$$

The Main Idea of Boost-R (Gradient Boosting)

Gradient Boosting

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^n \left\{ \sum_{j=1}^q \tilde{l}(\mu_{i,j}, \hat{\mu}_{i,j}^{(k-1)}) + \omega_{r(\mathbf{x}_i),j} \right\} \Delta + \gamma D + \frac{1}{2} \lambda \sum_{d=1}^D \|\omega_{d,\cdot}\|_2^2 \\ & \approx \sum_{i=1}^n \left\{ \sum_{j=1}^q \tilde{l}(\mu_{i,j}, \hat{\mu}_{i,j}^{(k-1)}) + g_{i,j} \omega_{r(\mathbf{x}_i),j} + \frac{1}{2} h_{i,j} \omega_{r(\mathbf{x}_i),j}^2 \right\} \Delta \\ & \quad + \gamma D + \frac{1}{2} \lambda \sum_{d=1}^D \|\omega_{d,\cdot}\|_2^2 \end{aligned} \quad (4)$$

where $g_{i,j} = \partial_{\hat{\mu}_{i,j}^{(k-1)}} \tilde{l}(\mu_{i,j}, \hat{\mu}_{i,j}^{(k-1)})$ and $h_{i,j} = \partial_{\hat{\mu}_{i,j}^{(k-1)}}^2 \tilde{l}(\mu_{i,j}, \hat{\mu}_{i,j}^{(k-1)})$.

The Main Idea of Boost-R (Gradient Boosting)

Let $I_d = \{i | r(\mathbf{x}_i) = d\}$

$$\text{minimize } \sum_{d=1}^D \sum_{j=1}^q \left(g_{\cdot,j} \omega_{d,j} + \frac{1}{2} (h_{\cdot,j} + \lambda) \omega_{d,j}^2 \right) \Delta + \gamma D \quad (5)$$

where $g_{\cdot,j} = \sum_{i \in I_d} g_{i,j}$ and $h_{\cdot,j} = \sum_{i \in I_d} h_{i,j}$.

Let $\Delta \rightarrow \infty$, the gradient boosting at any tree node is solved by

$$\text{minimize } \sum_{d=1}^D \int \left(g_{\cdot,j} \omega_{d,j} + \frac{1}{2} (h_{\cdot,j} + \lambda) \omega_{d,j}^2 \right) dt + \gamma D \quad (6)$$

Numerical Example 1

- A simple example which involves 200 systems.
- For each system, a number of 10 system attributes are respectively sampled from the unit interval $[0, 1]$.
- The failure data of system i are simulated from a Homogeneous Poisson Process (HPP) with the following intensity:

$$\lambda_i = \begin{cases} 0.01 & \text{if } 0 \leq x_{i,1}, x_{i,2} \leq 0.5 \\ 0.10 & \text{if } 0.5 < x_{i,1}, x_{i,2} \leq 1 \\ 0.05 & \text{otherwise} \end{cases} \quad (7)$$

- Covariates $x_3, x_4, \dots, x_{10}$ are purposely made redundant.

Numerical Example 1: Covariate Selection

- Covariate importance: the average decrease of the OOB prediction C-index after the values of a particular covariate have been randomly permuted.

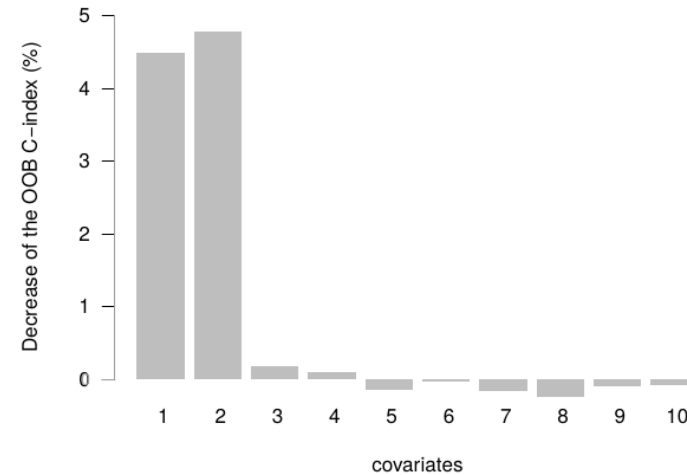
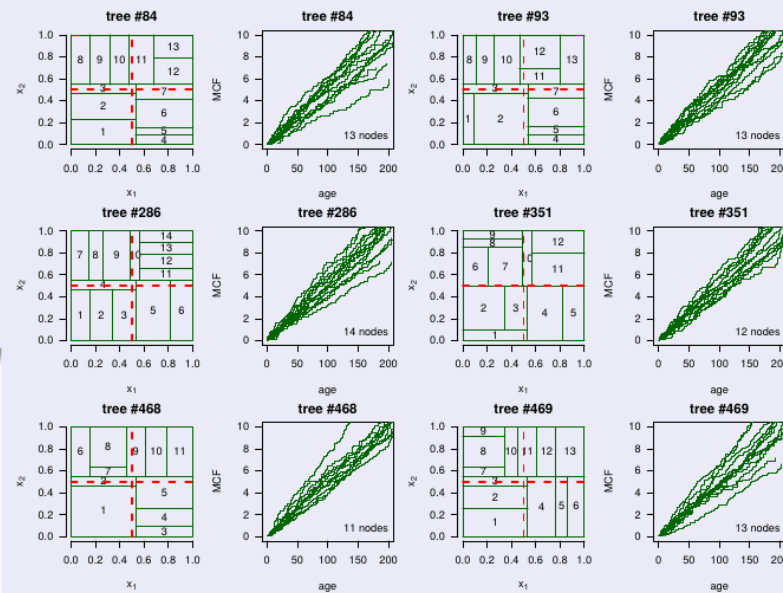


Figure: Covariate importance for numerical example 1

Numerical Example 1: A Closer Look

- We re-run the RF-R algorithm retaining only covariates x_1 and x_2 , and visualize the binary partition of the covariate space by each tree.



Tree Pruning

- Using full-grown trees seldom costs much in terms of model performance [Hastie et al.(2009)].
 - ▶ When trees are grown deep, each individual tree usually has low bias with high variance.
 - ▶ Bagging effectively reduces the variance of the ensemble estimator
- Control the depth of a tree:
 - ▶ If the node splitting is based on the log-rank test: stop further splitting a node if the minimum p -value is larger than a certain threshold
 - ▶ Use a larger value of d_0 in the RF-R algorithm, which is the minimum number of systems with failures in a node.

Numerical Example 1: Comparison

- A cross-validation-based comparison study considering the following methods:
 - ▶ *RF-R*: the proposed RF-R approach.
 - ▶ *MCF*: the nonparametric estimation for MCF without utilizing any system attributes.
 - ▶ *MCF-K*: For a system in the testing data set, its MCF is estimated using only the data from its K nearest “neighboring” systems in the training data sets.
 - ▶ *HPP*: MLE assuming HPP with a log-linear intensity function, $\log(\lambda_i) = \beta_0 + \sum_{j=1}^p x_{i,j} \beta_j$ and $p = 10$.

Numerical Example 1: Comparison

- A cross-validation-based comparison study considering the following methods:
 - ▶ *RF-R*: the proposed RF-R approach.
 - ▶ *MCF*: the nonparametric estimation for MCF without utilizing any system attributes.
 - ▶ *MCF-K*: For a system in the testing data set, its MCF is estimated using only the data from its K nearest “neighboring” systems in the training data sets.
 - ▶ *HPP*: MLE assuming HPP with a log-linear intensity function, $\log(\lambda_i) = \beta_0 + \sum_{j=1}^p x_{i,j} \beta_j$ and $p = 10$.
- We also perform the comparison based on the failure data exactly simulated from an HPP with the log-linear intensity function:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} \quad (8)$$

where $\beta_0 = 0.01$, $\beta_1 = 2$ and $\beta_2 = 0.5$.

Navigation icons: back, forward, search, etc.

Numerical Example 1: Comparison

- A cross-validation-based comparison:

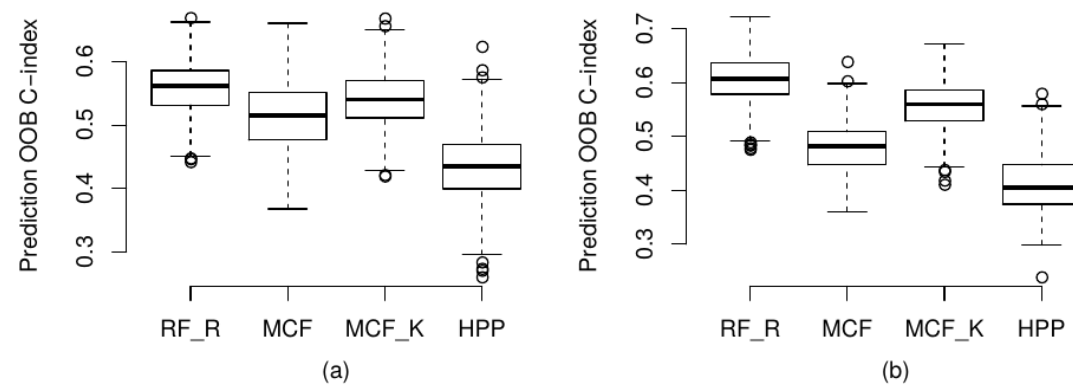


Figure: Cross-validation-based comparison of the prediction C-index for four different approaches. Panel (a) on the left shows the comparison results based on the data simulated by (7), and panel (b) on the right shows the results based on the data simulated by (8)

Numerical Example 2

- We consider both static and dynamic covariates.
- For each system: 10 static covariates randomly sampled from $[0, 1]$, and a dynamic covariate $z(t)$ simulated from a Brownian motion.
- The failure times of system i are simulated from a NHPP using the thinning method [Lewis and Shedler(1979)] with the intensity function:

$$\lambda_i(t) = \exp\{\beta_{i,0} + \beta_{i,1}z_i(t)\} \quad (9)$$

and

$$\exp\{\beta_{i,0}\} = \begin{cases} 0.01 & \text{if } 0 \leq x_{i,1}, x_{i,2} \leq 0.5 \\ 0.10 & \text{if } 0.5 < x_{i,1}, x_{i,2} \leq 1 \\ 0.05 & \text{otherwise} \end{cases} \quad \beta_{i,1} = \begin{cases} 0.5 & \text{if } 0 \leq x_{i,1}, x_{i,2} \leq 0.5 \\ 0.1 & \text{if } 0.5 < x_{i,1}, x_{i,2} \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

- Covariates $x_3, x_4, \dots, x_{10}$ are purposely made redundant.

Numerical Example 2: Covariate Selection

- Covariate importance: the average decrease of the OOB prediction C-index after permuting the values of a particular covariate.

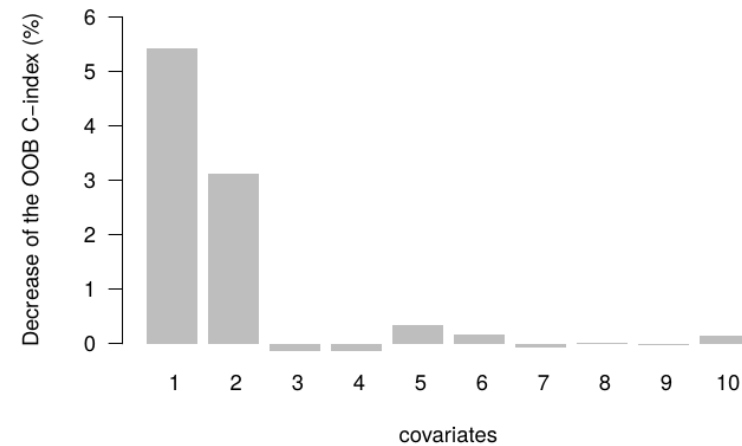


Figure: Importance of the ten system attributes, measured as the decrease of the OOB C-index (%), based on the data simulated from (9) and (10)

Numerical Example 2: A Closer Look

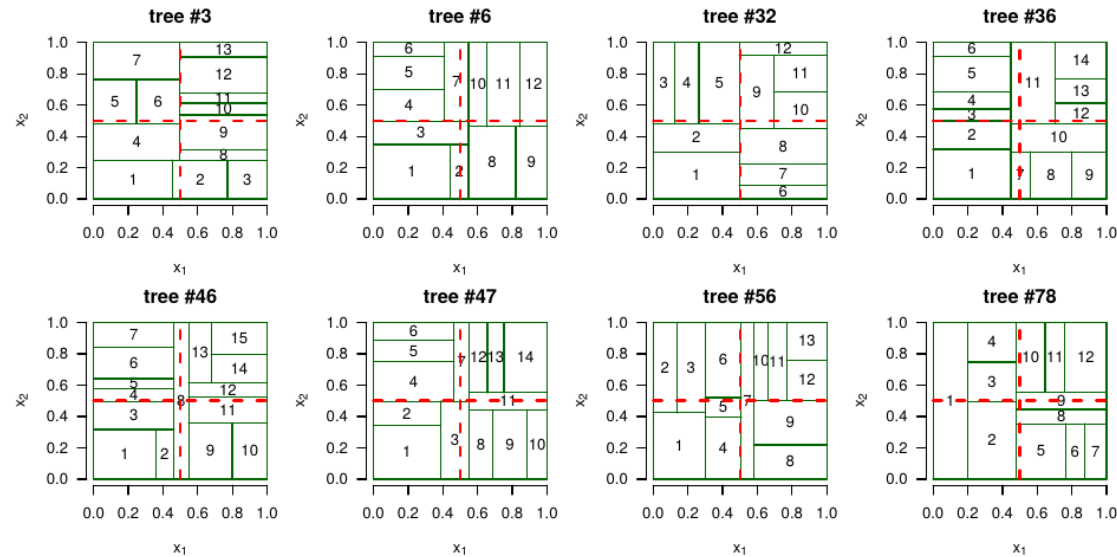


Figure: Binary partitions of the covariate space by eight randomly chosen trees from the ensemble forests with 100 trees. The read dash lines indicate the true classification of failure intensity defined in (9) and (10)

Numerical Example 2: Comparison

- A cross-validation-based comparison study considering the following methods:
 - ▶ *RF-R*: the proposed RF-R approach.
 - ▶ *MCF*: the nonparametric estimation for MCF without utilizing any system attributes.
 - ▶ *MCF-K*: For a system in the testing data set, its MCF is estimated using only the data from its K nearest “neighboring” systems in the training data sets.
 - ▶ *NHPP*: MLE assuming a NHPP model with the log-linear intensity function, $\lambda_i(t) = \exp\{\beta_{i,0} + \sum_{j=1}^{10} \beta_{i,j}x_{i,j} + \beta_{i,11}z_i(t)\}$.

Numerical Example 2: Comparison

- A cross-validation-based comparison study considering the following methods:
 - ▶ *RF-R*: the proposed RF-R approach.
 - ▶ *MCF*: the nonparametric estimation for MCF without utilizing any system attributes.
 - ▶ *MCF-K*: For a system in the testing data set, its MCF is estimated using only the data from its K nearest “neighboring” systems in the training data sets.
 - ▶ *NHPP*: MLE assuming a NHPP model with the log-linear intensity function, $\lambda_i(t) = \exp\{\beta_{i,0} + \sum_{j=1}^{10} \beta_{i,j}x_{i,j} + \beta_{i,11}z_i(t)\}$.
- We also perform the comparison based on the failure data exactly simulated from a NHPP with the log-linear intensity function:

$$\lambda_i(t) = \exp\{\beta_0 + \beta_1x_{i,1} + \beta_2x_{i,2} + \beta_3z_i(t)\} \quad (11)$$

where $\beta_0 = \log(0.01)$, $\beta_1 = 2$, and $\beta_2 = \beta_3 = 0.5$.

Navigation icons: back, forward, search, etc.

Numerical Example 2: Comparison

- A cross-validation-based comparison:

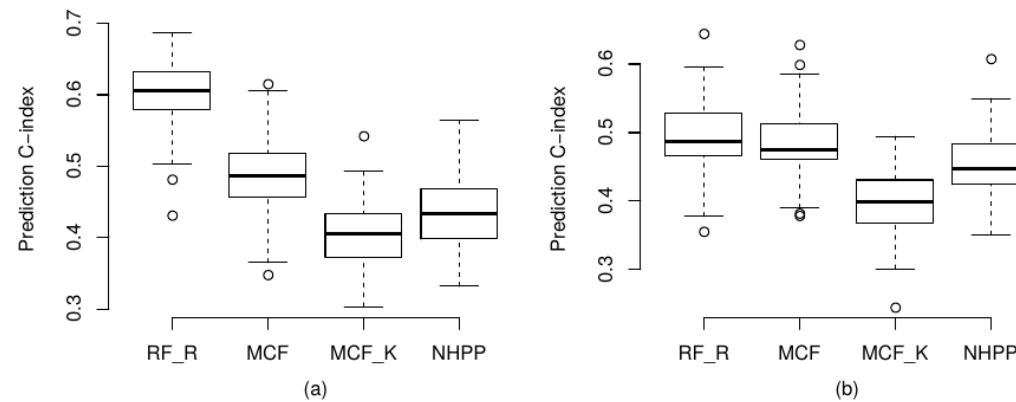
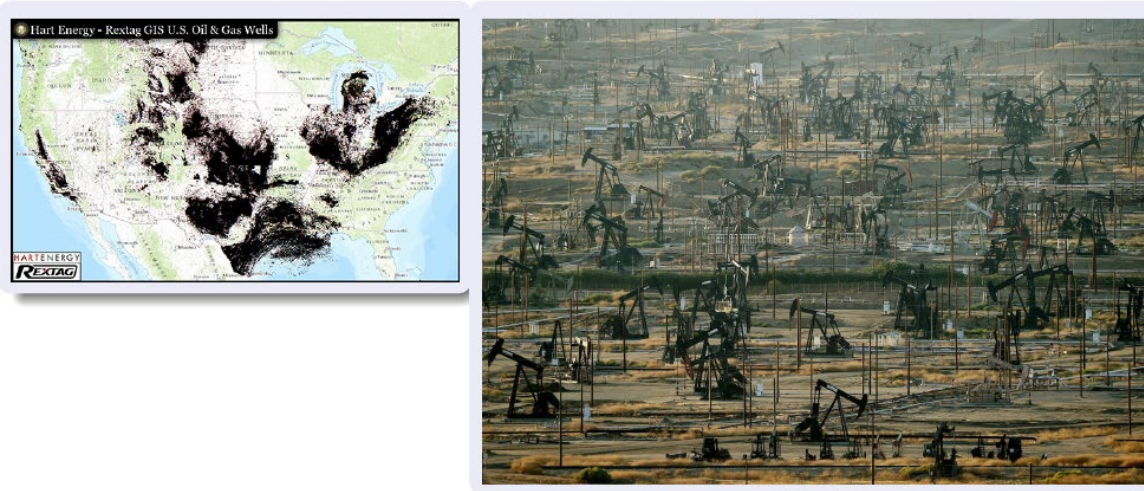


Figure: Cross-validation-based comparison of the prediction C-index for four different approaches. Panel (a) on the left shows the comparison results based on the data simulated by (9), and panel (b) on the right shows the results based on the data simulated by (11)

Case Study: Oil and Gas Well Failure Data

- A modified data set from a multinational oil and natural gas service company which provides predictive well maintenance service.
- Failure data, system attributes, and sensor measurement for 8232 oil and gas wells installed in the California area.



xiao.liu@isye.gatech.edu

2026

40 / 52

Case Study: Static Well Attributes

- A number of 8 covariates, $x_1 \sim x_8$, are available for reliability analysis.
 - ▶ well size, nominal operating settings/conditions, geo-locations, etc.
 - ▶ heterogeneity

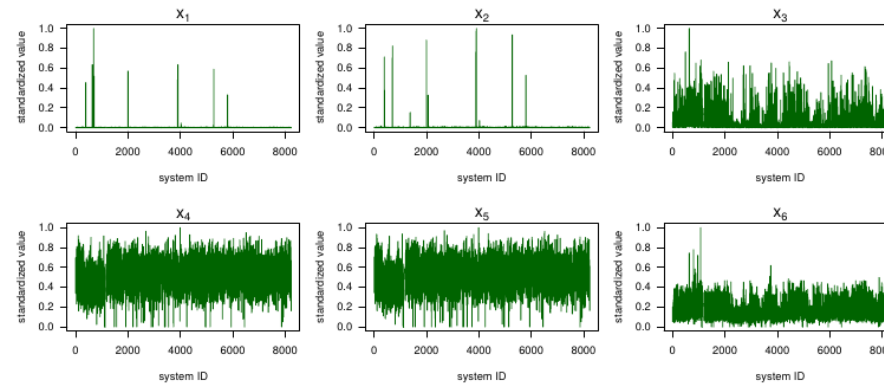


Figure: Standardized values of $x_1 \sim x_6$ for all 8232 systems.

Case Study: Geo-locations

- Covariates, x_7 and x_8 , respectively specify the latitude and longitude of each well.
- Neighboring wells share common (unobserved) environmental conditions such as temperature and humidity variation, soil type, contamination, etc.
- Spatial trend

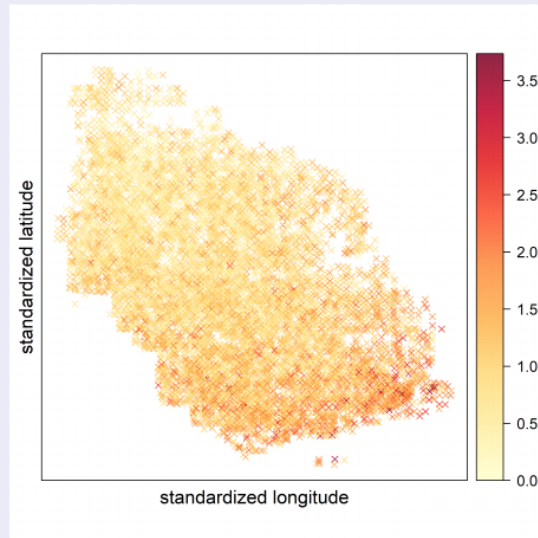


Figure: Spatial locations of a total number of 8232 systems (the legend shows the average annual failures for each system)

Case Study: Geo-locations

- Covariates, x_7 and x_8 , respectively specify the latitude and longitude of each well.
- Neighboring wells share common (unobserved) environmental conditions such as temperature and humidity variation, soil type, contamination, etc.
- Spatial trend

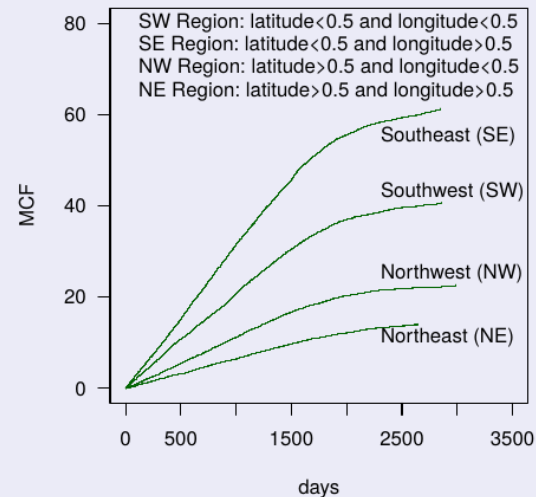


Figure: Estimated Mean Cumulative Function (MCF) for systems over four geo-regions [Nelson(1995)]

Case Study: Dynamic Sensor Measurement

- Seven dynamic sensor measurement of system working conditions:
 - ▶ $z_1(t) \sim z_5(t)$: dynamic torque and load measurement including gear torque, load range, etc.
 - ▶ $z_6(t) \sim z_7(t)$: monitored gearbox and structural stress.

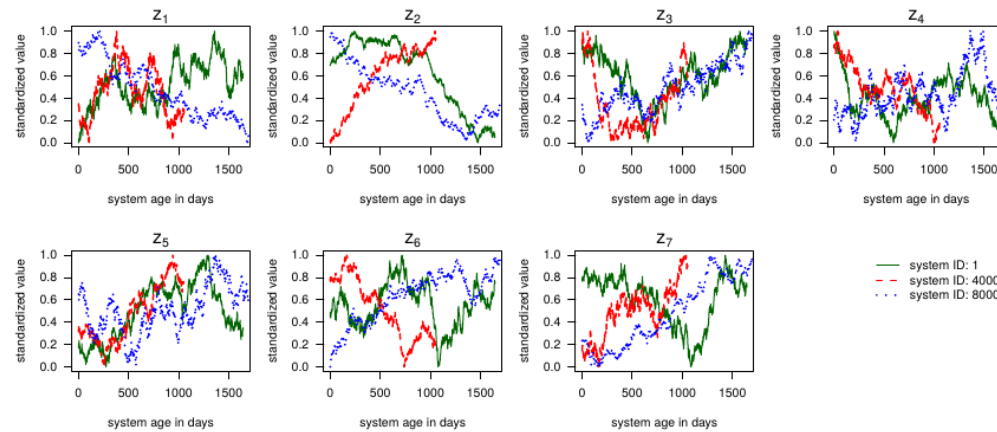


Figure: Sensor measurement in terms of torque, load and stress, $z_1 \sim z_7$, for three selected systems: system #1, system #4000, and system #8000.

Case Study: Covariate Selection

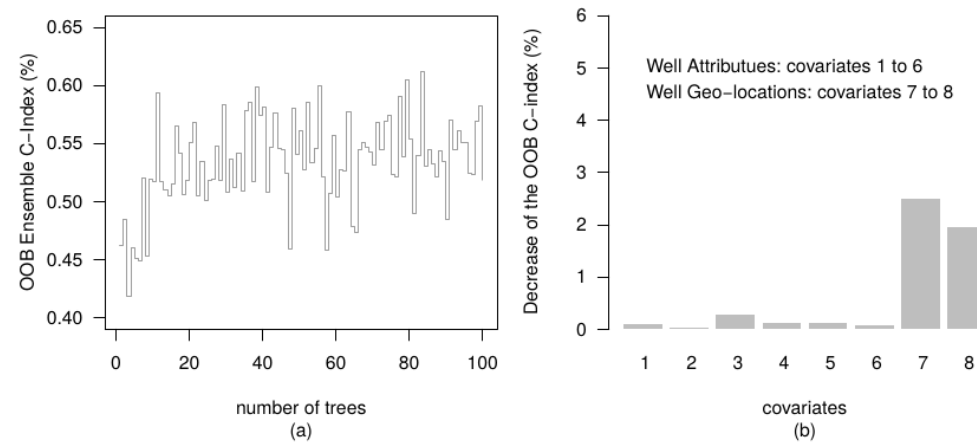
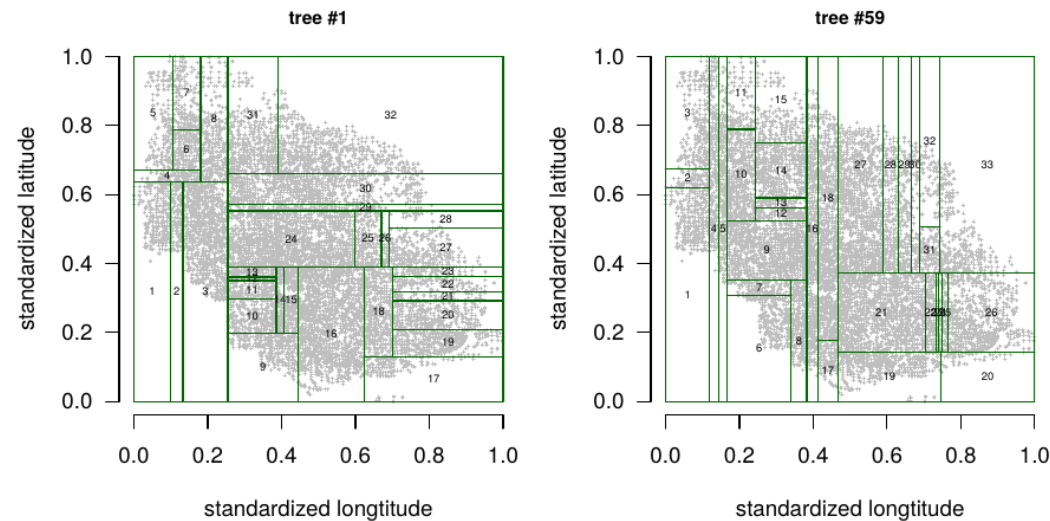


Figure: The left panel (a) shows the OOB ensemble prediction C-index against the number of trees in a forest for the first 100 trees, and the right panel (b) shows the importance of the eight static system attributes measured by the average decrease of the OOB prediction C-index

Case Study: A Closer Look

- We re-run the RF-R algorithm by only retaining the geo-location information



Case Study: Effects of Dynamic Covariates

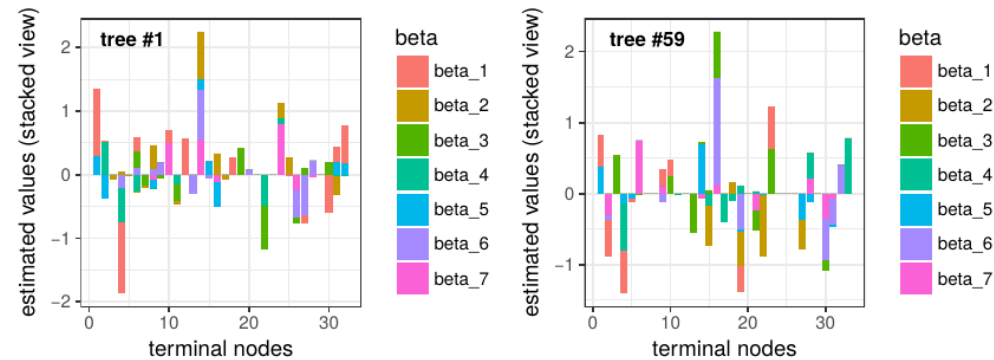
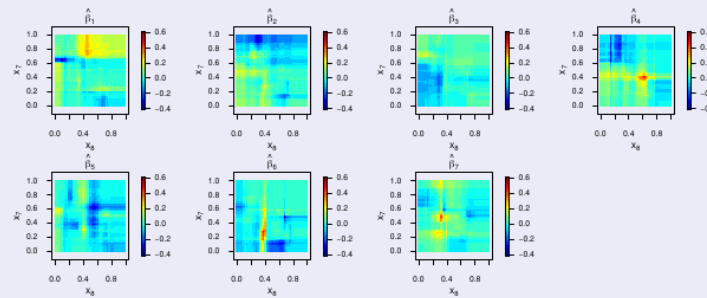


Figure: $\hat{\beta}_1 \sim \hat{\beta}_7$, on each terminal node of tree #1 and tree #59.

- The Lasso penalty successfully imposes sparsity on the terminal nodes of a tree.
- The intensity functions of systems from different geo-regions may have different dependence structures on system operational and environmental conditions.

Case Study: Effects of Dynamic Covariates

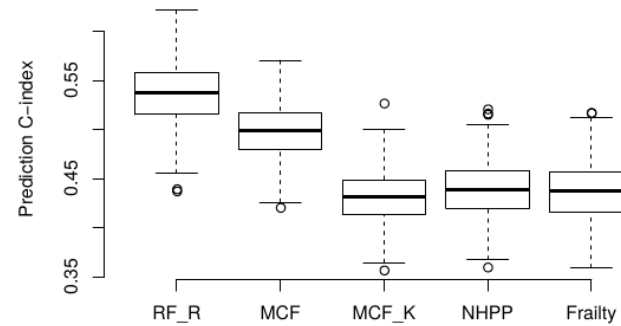
- A spatially aggregated view of the averaged $\hat{\beta}_1 \sim \hat{\beta}_7$ from the ensemble trees over the spatial domain.
- The failure intensities at different geo-locations are dominated by different operational and environmental conditions.



Case Study: Comparison

- A cross-validation-based comparison study considering the following methods:
 - ▶ *RF-R*: the proposed RF-R approach.
 - ▶ *MCF*: the nonparametric estimation for MCF without utilizing any system attributes.
 - ▶ *MCF-K*: For a system in the testing data set, its MCF is estimated using only the data from its K nearest “neighboring” systems in the training data sets.
 - ▶ *NHPP*: MLE assuming a NHPP model with the log-linear intensity function.
 - ▶ *Frailty*: Gamma Frailty

Case Study: Comparison



- The proposed RF-R outperforms.
- The parametric approach assuming NHPP suffers from the presence of a large number of redundant covariates.
- We would rather not to use any covariate information than incorporating a large number of redundant covariates.

References

xiao.liu@isye.gatech.edu

2026

51 / 52

-  ANDERSON, P. K. AND BORGAN, O. AND GILL, R. D., AND KEIDING, N. (1993). *Statistical Models based on Counting Processes*, Springer-Verlag, New York.
-  BACCHETTI, P. AND SEGAL, M.(1995). Survival Trees with Time-Dependent Covariates: Application to Estimating Changes in the Incubation Period of AIDS. *Lifetime Data Analysis*, **1**, 35-47.
-  BOU-HAMAD, I. AND LAROCQUE, D. AND BEN-AMEUR, H. AND MASSE, L. AND VITARO, F. AND TREMBLAY, R.(2009). Discrete-Time Survival Trees. *Canadian Journal of Statistics*, **37**, 17-32.
-  BOU-HAMAD, I.(2011). A Review of Survival Trees. *Statistics Surveys*, **5**, 44-71.
-  BREIMAN, L.(2001). Random Forests. *Machine Learning*, **45**, 5-32.
-  CHIPMAN, H. A. AND GEORGE, E. I. AND MCCULLOCH, R. E.(2006). Bart: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, **4**, 266–298.
-  DOGANAKSOY, N. AND HAHN, G. J. AND MEEKER, W. Q.(2006). How to Analyze Reliability Data for Repairable Products. *Quality Progress*, **39**, 93–95.
-  FAN, J. AND SU, X. G., AND LEVINE, R. AND NUNN, M. AND LEBLANC, M. (2006). Trees for Censored Survival Data by Goodness of Split, with Application to Tooth Prognosis. *Journal of the American Statistical Association*, **101**, 959–967.
-  FAN, J. AND NUNN, M. AND SU, X. G.(2009). Multivariate Exponential Survival Trees and Their Application to Tooth Prognosis. *Computational Statistics and Data Analysis*, **53**,1110–1121.
-  FLEMING, T. R. and HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*, John Wiley & Sons, New York.
-  HARRELL, F. AND CALIFF, R. AND PRYOR, D. AND LEE, K. AND ROSATI, R.(2018). Evaluating the Yield of Medical Tests. *Journal of American Medicine Association*, **247**, 2543–2546.
-  HASTIE, T. AND TIBSHIRANI, R. AND FREDMAN, J.(2009). *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*, Springer.

Navigation icons: back, forward, search, etc.

-  HONG, Y. AND ZHANG, M. AND MEEKER, W. Q.(2018). Big Data and Reliability Applications: The Complexity Dimension. *Journal of Quality Technology*, **50**, 00–00.
-  HOTHORN, T. AND LAUSEN, B. AND BENER, A. AND RADESPIEL-TROGER, M.(2004). Bagging survival trees. *Statistics in Medicin*, **23**, 77–91.
-  HOTHORN, T. AND BÜHLMANN, P. AND DUDOIT, S. AND MOLINARO, A. AND VAN DER LAAN, M.(2006). Survival Ensembles. *Biostatistics*, **7**, 355–373.
-  ISHWARAN, H. AND KOGALUR, U. B. AND BLACKSTONE, E. H. AND LAUER, M. S.(2008). Random Survival Forests. *The Annals of Applied Statistics*, **2**, 841–860.
-  ISHWARAN, H. AND KOGALUR, U. B.(2010). Consistency of Random Survival Forests. *Statistics Probability Letter*, **80**, 1056–1064.
-  ISHWARAN, H. AND KOGALUR, U. B. AND GORODESKI, E. Z. AND MINN, A. J. AND LAUER, M. S.(2010). High-Dimensional Variable Selection for Survival Data. *Journal of the American Statistical Association*, **105**, 205–217.
-  KLEIN, J. P. AND MOESCHBERGER, M. L.(2005). *Survival Analysis – Techniques for Censored and Truncated Data*, Springer.
-  LEWIS P. AND SHEDLER, G.(1979). Simulation of Nonhomogenous Poisson Processes by Thinning. *Naval Research Logistics Quarterly*, **26**, 403–413.
-  MEEKER, W. Q. and ESCOBAR, L. A. (1998). *Statistical Methods for Reliability Data*, John Wiley & Sons, New York.
-  MEEKER, W. Q. AND HONG, Y.(2014). Reliability Meets Big Data: Opportunities and Challenges. *Quality Engineering*, **26**, 102–116.
-  MITTMAN, E. T. AND LEWIS-BECK, C. AND MEEKER, W.(2018). A Hierarchical Model for Heterogenous Reliability Field Data. *The Annals of Applied Statistics*, to appear.

-  NELSON, W.(1995). Confidence Limits for Recurrence Data: Applied to Cost or Number of Product Repairs. *Technometrics*, **37**, 147–157.
-  PAN, R. AND RIGDON, S. E.(2009). Bayes Inference for General Repairable Systems. *Journal of Quality Technology*, **41**, 82–94.
-  PULCINI, G.(2001). A Bounded Intensity Process for the Reliability of Repairable Equipment. *Journal of Quality Technology*, **33**, 480–492.
-  RIGDON, S. E. AND BASU, A. P.(2000). *Statistical Methods for the Reliability of Repairable Systems*, John Wiley & Sons, New York.
-  TERENCE, A. AND DONG, S. F. AND DRAPER, D.(2017). GPU-Accelerated Gibbs Sampling: A Case Study of the Horseshoe Probit Model. *arXiv*, **1608.04329v4**.
-  ZUO, J. AND WU, H. AND MEEKER, W. Q.(2012). Asymptotic Properties of Some Estimators of the Mean Cumulative Function. *Journal of Statistical Planning and Inference*, **142**, 2943–2952.
-  YE, Z. AND HONG, Y. AND XIE, Y.(2013). How Do Heterogeneities in Operational Environments Affect Field Failures. *The Annals of Applied Statistics*, **7**, 2249–2271.
-  YE, Z. AND MURTHY, D. N. P. AND XIE, M. AND TANG, L. C.(2013). Optimal Burn-in for Repairable Products Sold with a Two-Dimensional Warranty. *IEEE Transactions on Reliability*, **45**, 164–176.
-  YANG, Q. AND HONG, Y. AND CHEN, Y. AND SHI, J.(2012). Failure Profile Analysis of Complex Repairable Systems With Multiple Failure Modes. *IEEE Transactions on Reliability*, **66**, 180–191.
-  XIE, M.(1991). *Software Reliability Modelling*, World Scientific Publisher, Singapore.

Upcoming Courses

Generative AI Application for Supply Chain Professionals

Oct 19, 2026 – Oct 21, 2026 | Georgia Tech Savannah Campus

World Class Sales and Operations Planning

Oct 12, 2026 – Oct 14, 2026 | Virtual (Instructor-led)

Supply Chain Optimization and Prescriptive Analytics

Nov 2, 2026 – Nov 5, 2026 | Virtual (Instructor-led)



**Supply Chain and
Logistics Institute**

www.scl.gatech.edu

 [Follow us](#)

[Subscribe to our Newsletter](#)