Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira, Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p26462-cd

A Reliability Assessment Methodology for Products in Cold Environments Based on Bayesian Nonparametric Theory

Jiageng Li
School of Reliability and Systems Engineering, Beihang University, China. E-mail: zy2414126@buaa.edu.cn

Jun Yao
School of Reliability and Systems Engineering, Beihang University, China. E-mail: yao2jun1@sina.com

Xuetong Ku
School of Reliability and Systems Engineering, Beihang University, China. E-mail: zy2414117@buaa.edu.cn

Guo Chen
AECC Sichuan Gas Turbine Establishment, China. E-mail: jia187590chen@2925.com

Huan Lu
AECC Sichuan Gas Turbine Establishment, China. E-mail: jia187590huan@2925.com

Chunping Hu
China Aviation Engine Nanjing Aero-Propulsion Co., Ltd., China. E-mail: jia187590hu@2925.com

Xiaohui Wang
School of Reliability and Systems Engineering, Beihang University, China. E-mail: xiaohuiw@buaa.edu.cn

Xiaohong Wang
School of Reliability and Systems Engineering, Beihang University, China. E-mail: wxhong@buaa.edu.cn

Xiaogang Li
School of Reliability and Systems Engineering, Beihang University, China. E-mail: lxg@buaa.edu.cn

Ruyue Li
School of Reliability and Systems Engineering, Beihang University, China. E-mail: lry6688@buaa.edu.cn

Cryogenic reliability assessment for reusable liquid rocket engine components is challenged by small sample size, heterogeneous failure modes, right censoring, and systematic under-coverage of remaining useful life (RUL) prediction intervals under limited observations. This paper presents two complementary, independently deployable tools tailored for this regime: (1) a Bayesian nonparametric stratification method via a Dirichlet Process Gaussian Mixture Model (DP-GMM) for heterogeneity-aware reliability estimation under right censoring; (2) a Gaussian process (GP)-based RUL prediction pipeline with interval-based conformal calibration, which restores near-nominal interval coverage under sparse observations. We further clarify the applicable boundary of mode-aware RUL prediction in small-sample settings. A cryogenic-physics-inspired simulation study and external validation on the NASA C-MAPSS benchmark demonstrate the effectiveness and robustness of the proposed methods.

Keywords: Bayesian nonparametrics, right censoring, Gaussian processes, remaining useful life, conformal prediction.

1. Introduction

Reliability assessment for reusable liquid rocket engines is severely constrained by scarce field/ground-test samples, heterogeneous degradation behaviors, and right-censored incomplete observations. As core flow control components of liquid rocket propulsion systems, cryogenic valves and seals repeatedly undergo extreme thermal cycling between ambient temperature and -253°C/-183°C cryogenic conditions, coupled with high-pressure shock, mechanical wear, and cyclic fatigue. In engineering practice, maintenance and mission-planning decisions must be made with only dozens to hundreds of time-to-event records and a handful of partially observed degradation trajectories, as full-life cycle tests of cryogenic components are extremely costly and time-consuming.

1.1. *Engineering motivation and failure heterogeneity*

Reusable cryogenic valve and seal components can fail via multiple competing mechanisms (e.g., fatigue damage, seal aging/leakage, brittle fracture). In small-sample settings, engineering teams often cannot predefine a reliable taxonomy of failure modes, yet pooling all samples into a single lifetime model blurs heterogeneity and biases reliability estimates: a single model will underestimate the risk of high-failure-rate subpopulations, while overestimating long-life unit reliability, leading to unplanned mission failures or excessive maintenance costs. This motivates an unsupervised stratification method that does not require a prespecified number of modes, and can provide mode-conditional reliability curves under right censoring.

1.2. *Problem setting and core challenges*

We consider two natural data sources in cryogenic reliability studies: (1) right-censored failure-time records with operating covariates, with the core goal of estimating censoring-aware survival curves; (2) partial degradation trajectories of a small number of units, with the core goal of estimating RUL distribution under sparse observation. Three unsolved challenges dominate this small-sample regime: (i) heterogeneity pooling blurs mechanism-specific survival behavior; (ii) right censoring requires specialized survival estimation; (iii) sparse observation leads to systematic under-coverage of RUL prediction intervals.

1.3. *Contributions*

This work makes two primary pragmatic contributions for small-sample cryogenic reliability assessment, plus one exploratory boundary analysis:

(1) **Heterogeneity-aware reliability estimation (Tool A)**. A lightweight variational Bayes DP-GMM for latent stratum discovery without predefining mode numbers, paired with interpretable mode-conditional Kaplan–Meier (KM) reliability curves with bootstrap bands.
(2) **Trustworthy RUL interval prediction (Tool B)**. A unit-level leave-one-unit-out (LOO) conformal calibration protocol for GP-based RUL prediction, which restores near-nominal empirical coverage across observation ratios under sparse data.

Exploratory boundary. We quantify the applicable scope of mode-aware GP routing, highlighting its limitation by small per-mode sample sizes and routing error.

Organization. Related work, methodology, experiments, discussion, and conclusion follow.

2. Related Work

Reusable aerospace PHM and benchmarks. Prognostics and health management (PHM) for safety-critical aerospace assets drives demand for robust RUL prediction under variable operating regimes. The NASA C-MAPSS benchmark is widely used to validate RUL methods under multiple fault modes (Saxena et al., 2008), while existing surveys highlight persistent calibration challenges in small-sample engineering practices (Si et al., 2011; Lei et al., 2018). However, few studies target the extreme cryogenic environment and small-sample regime of reusable liquid rocket engines.

Heterogeneity and censoring-aware reliability. Dirichlet process (DP) mixture models enable flexible clustering without predefining component

counts (Ferguson, 1973; Antoniak, 1974), capturing population heterogeneity obscured by single lifetime models. The Kaplan–Meier estimator is the gold standard for right-censored survival analysis (Kaplan and Meier, 1958); conditioning KM on DP-GMM inferred strata quantifies survival differences across latent subpopulations. However, most existing DP mixture applications in reliability focus solely on clustering accuracy, not the statistical significance of survival separation, which is the core engineering requirement.

Probabilistic prognostics and calibrated uncertainty. Gaussian process regression is widely used for degradation modeling, as it yields full predictive distributions rather than only point forecasts (Rasmussen and Williams, 2006). However, nominal GP intervals often suffer from under-coverage under small samples and extrapolation. Conformal prediction provides a distribution-free calibration layer to restore empirical coverage under exchangeability assumptions (Vovk et al., 2005; Romano et al., 2019). Unit-level calibration for time series avoids invalidation from within-unit temporal dependence, aligning with aerospace maintenance decision-making.

Three critical gaps remain: (1) most studies assume a single lifetime model, obscuring failure heterogeneity; (2) prognostics models are poorly calibrated under sparse observation; (3) end-to-end pipelines are underspecified, limiting reproducibility. This work directly addresses these gaps.

3. Methodology

Figure 1 summarizes the proposed workflow. Failure-time records are clustered into latent modes via DP-GMM for mode-conditional KM reliability estimation. Degradation trajectories are modeled with GP for probabilistic RUL prediction, followed by interval-based conformal calibration to ensure nominal coverage.

3.1. *Mode Discovery via DP-GMM*

We use a variational Bayes truncated DP-GMM for unsupervised latent failure mode discovery. Let x_i denote the standardized feature vector for failure-time sample i, with latent mode label $z_i \in \{1, \ldots, K\}$. The mixture model is defined as:

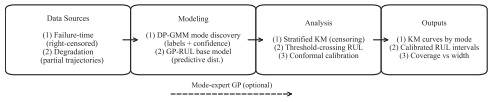

Fig. 1. Overview of the proposed workflow.

$$p(x_i) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k) \tag{1}$$

The Dirichlet Process prior encourages model sparsity, and inference yields posterior responsibilities $r_{ik} = p(z_i = k \mid x_i)$. We use hard assignment $\hat{z}_i = \arg\max_k r_{ik}$ for stratification, with assignment confidence $\max_k r_{ik}$ for stability diagnostics.

Implementation details. DP-GMM is implemented via scikit-learn's BayesianGaussianMixture with Dirichlet-process prior, full covariance, and truncation $K = 10$. Input features include temperature, stress level, reuse count, vibration, pressure drop, energy consumed, and physics-inspired proxies (Arrhenius factor, brittleness indicator, Paris-law crack growth proxy). All features are standardized before clustering, and sensitivity analysis confirms downstream KM stratification is robust to $K \in \{5, 8, 10, 15\}$.

Reproducibility (compact). We fix a deterministic random seed for clustering and GP fitting, and use standard scikit-learn settings (no custom solvers). We report assignment confidence $\max_k r_{ik}$ and restrict KM analysis to samples above a confidence threshold to avoid over-interpreting low-support strata.

3.2. *Mode-Conditional Reliability under Right Censoring*

For each inferred mode k, we estimate the survival function $\hat{S}_k(t)$ via the Kaplan–Meier estimator:

$$\hat{S}_k(t) = \prod_{t_j \le t} \left(1 - \frac{d_{k,j}}{n_{k,j}}\right) \tag{2}$$

where $d_{k,j}$ is the number of failures at time t_j in mode k, and $n_{k,j}$ is the number of at-risk units just prior to t_j. We report per-mode event counts and 90% bootstrap confidence bands for all survival curves.

3.3. *GP-Based RUL Prediction*

For degradation trajectory modeling, we fit a GP regression $y(x) \sim \mathcal{GP}(m(x), K(x, x'))$ per unit, with input $x = [t, \text{temperature}, \text{reuse\_count}]$. The kernel follows a trend+residual+noise structure:

$$K = K_{\text{lin}} + K_{\text{Matern}} + K_{\text{white}} \tag{3}$$

where K_{lin} captures the global degradation trend, K_{Matern} ($\nu = 3/2$) models residual fluctuations, and K_{white} fits measurement noise. All inputs are standardized within each unit fit.

RUL is defined as the remaining time until the health indicator crosses a fixed failure threshold τ. We sample correlated future trajectories from the GP's joint predictive distribution (300 paths in our experiments), compute first threshold-crossing times, and aggregate to form an empirical RUL distribution. Hyperparameters are optimized by maximizing the log-marginal likelihood with 10 random restarts.

Threshold and grid. In our simulator, the health indicator is the simulated performance metric and failure occurs when it crosses $\tau = 20$ (Table 1); forecasting proceeds on the cycle index with a unit-step grid.

3.4. *Interval-Based Conformal Calibration*

Let $[l(x), u(x)]$ be the nominal $1-\alpha$ RUL prediction interval from the GP. On the calibration set, we compute the non-negative conformity score for each unit:

$$s_i = \max\{l(x_i) - y_i,\ y_i - u(x_i),\ 0\} \tag{4}$$

which quantifies how much the true RUL y_i falls outside the base interval.

We perform unit-level LOO calibration to maximize small-sample data utilization: for each target unit i, we compute the empirical quantile $\hat{q}_{-i}$ from the conformity scores of all remaining units, with $\hat{q}$ as the k-th order statistic ($k = \lceil (n_{\text{cal}} + 1)(1-\alpha) \rceil$). Calibration is repeated independently for each observation ratio (0.5/0.7/0.9). The final calibrated interval is:

$$[l'(x), u'(x)] = [l(x) - \hat{q},\ u(x) + \hat{q}] \tag{5}$$

Unit-level calibration avoids invalidation from within-unit temporal dependence, and yields valid marginal coverage under the exchangeability assumption for aerospace components from the same production batch and operating regime.

Protocol clarity. For each observation ratio, we re-run unit-level LOO calibration independently: each target unit is predicted from its observed prefix, while conformity scores and the quantile are computed from the remaining units only.

3.5. *Exploratory Mode-Aware Routing*

We include an exploratory coupling check: infer a unit's mode label from its observed prefix via a DP-GMM on a compact per-unit feature vector, then route prediction to a per-mode GP expert (with global GP fallback for small modes). Full results are in supplementary materials; mode-aware routing offers no consistent performance gain under small per-mode samples, due to routing error and data scarcity.

4. Experiments and Results

4.1. *Simulation Data and Metrics*

We use a cryogenic-physics-inspired simulation generator with heterogeneous latent modes and right censoring, producing 230 failure-time records (200 failures, 30 censored) and 25 degradation trajectories (823 time points). Table 1 summarizes the simulation logic, and Table 2 lists default parameters (code available at https://anonymous.4open.science/r/esrel-2026-B470). Sensitivity analysis on temperature variation, censoring fraction, and failure-time noise confirms all core conclusions are robust (log-rank pass rate 1.0, conformal coverage 0.919–0.976 across perturbations).

We evaluate: mode discovery quality (ARI, NMI), KM curve separation (log-rank p-value), RUL point accuracy (MAE, RMSE), and interval quality (empirical coverage, mean width).

Table 1. Simulator logic framework.

Component	Logic (summary)
Latent failure modes	Four latent modes (mixture weights 0.40/0.30/0.20/0.10) as ground truth.
Operating covariates	Temperature, stress level, reuse count.
Physics-inspired features	Arrhenius factor, brittleness indicator, Paris-law-style crack growth proxy.
Failure-time generation	Mode-dependent baseline cycle count modulated by covariates plus noise; brittle-fracture mode includes early failures.
Censoring mechanism	Right censoring $\approx 15\%$, applied preferentially to lower-stress, longer-life realizations.
Degradation trajectories	Health indicator decreases per cycle at mode-dependent rate modulated by covariates plus measurement noise.
Failure definition	Health indicator threshold crossing (performance ≤ 20).

Table 2. Simulator default parameters.

Parameter	Meaning	Default (mode 0/1/2/3)
C	Paris coefficient	1.0e-4 / 0.5e-4 / 2.0e-4 / 0.8e-4
m	Paris exponent	3.0 / 2.5 / 3.5 / 2.8
E_a (kJ/mol)	Activation energy	30 / 60 / 20 / 40
Cycle-count prior	Mean $\pm$ std cycles	45$\pm$8 / 35$\pm$6 / 25$\pm$5 / 55$\pm$10
Censoring fraction	Right-censoring rate	≈ 0.15

4.2. *Mode Discovery and Reliability Estimation*

DP-GMM with $K = 10$ yields $ARI = 0.345$, $NMI = 0.417$—moderate values expected given overlapping covariates mimicking real mixed failure mechanisms. Critically, 96.1% of samples have assignment confidence ≥ 0.8, and bootstrap refitting confirms KM curve separation is significant in 100% of runs (median $p = 2.2 \times 10^{-18}$).

Interpreting moderate ARI. Our core engineering goal is risk stratification (discovering subpopulations with significantly different survival behavior), not perfect recovery of mechanistic taxonomy. Ground-truth modes are deliberately overlapped in covariate space, making perfect label recovery infeasible. We therefore rely on three validity checks: (i) significant KM curve separation; (ii) high assignment confidence; (iii) bootstrap stability. Mapping strata to physical failure mechanisms requires additional post-test inspection, which is outside the scope of this data-driven stratification.

We restrict KM analysis to samples with confidence ≥ 0.6 (227/230 samples) and merge low-support modes. Global log-rank test confirms significant separation across strata ($\chi^2 = 50.15$, $df = 3$, $p = 7.4 \times 10^{-11}$). Table 3 reports per-stratum counts, and Figure 2 shows mode-conditional KM curves with 90% bootstrap confidence bands.

Table 3. Per-stratum sample sizes and events.

Stratum	Total	Failures	Censored
Mode 0	112	111	1
Mode 2	52	27	25
Mode 9	22	22	0
Merged minor modes	41	38	3

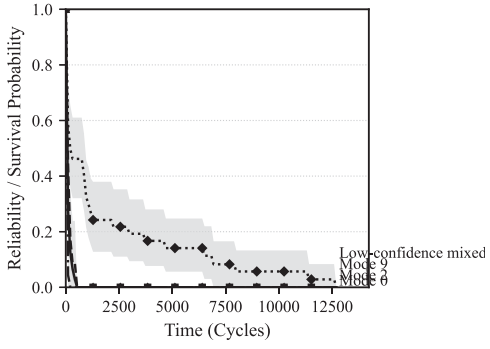

Fig. 2. Mode-conditional KM survival curves with 90% bootstrap confidence bands.

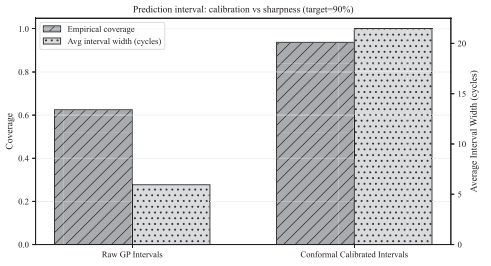

Fig. 3. Raw vs conformalized RUL interval quality at ratio 0.7.

4.3. *Calibrated RUL Prediction Results*

At observation ratio 0.7, raw GP 90% intervals have an average coverage of only 0.783±0.159, while conformal calibration restores near-nominal coverage to 0.917±0.072.

Table 4 summarizes performance across all observation ratios, including the GP warning rate (fraction of units with scikit-learn ConvergenceWarning). Even when 100% of units emit convergence warnings, the conformal layer consistently maintains nominal coverage, validating its robustness to suboptimal GP fits.

We compare against two classical degradation baselines (drifted Wiener process, exponential trend model) with identical unit-level conformal calibration.

Table 5 shows the GP pipeline achieves better point accuracy and narrower calibrated intervals, especially under short observation prefixes. The drifted Wiener baseline is a strong comparator but trades accuracy for wider intervals, while the exponential trend baseline suffers severe model mismatch under heterogeneous degradation.

Figure 3 visualizes raw vs conformalized intervals at ratio 0.7, highlighting the under-coverage of nominal GP intervals and the coverage-restoring effect of conformal calibration.

4.4. *External Validation on NASA C-MAPSS*

We validate the conformal calibration pipeline on the NASA C-MAPSS FD001 benchmark (100 units). Table 6 shows raw interval coverage varies widely (0.49–0.83), while conformal calibration consistently restores coverage to the nominal 0.90 level. This confirms the cross-domain effectiveness of the calibration component beyond the cryogenic simulation regime. Table 7 summarizes point-prediction performance across all C-MAPSS subsets, contextualizing task difficulty across operating regimes.

5. Discussion

Mode-conditional KM curves provide a lightweight, interpretable heterogeneity-aware survival view for small-sample cryogenic reliability assessment. Even with moderate clustering reproducibility, downstream survival separation is stable under bootstrap refitting, supporting practical stratification when paired with assignment-confidence diagnostics.

Engineering applicability checklist. To help practitioners apply the proposed tools, we recommend four minimal checks:

(1) **Data regime clarity**: Report sample sizes for time-to-event records and degradation units separately, along with the right-censoring rate.
(2) **Stratification support**: Merge strata with fewer than 5–10 failure events into a mixed group, and report per-stratum counts.
(3) **Stability validation**: Report assignment confidence distribution and bootstrap refit stability; treat unstable strata as a risk-screening device only.
(4) **Interval efficiency**: If calibrated 90% interval width exceeds 2× the best-estimate RUL, default to monitoring and additional data collection.

Coverage–width tradeoff. In our simulation at ratio 0.7, calibrated intervals have a median width/true-RUL ratio of 1.7; on NASA FD001, the ratio is 3.0–5.6. This confirms that while calibration can fix under-coverage, interval efficiency depends on the base predictor quality and observed prefix length.

Limitations. Small inferred strata lead to wider bootstrap bands and higher sensitivity to merge rules; we therefore interpret stratified KM as an engineering diagnostic, not a definitive mechanistic taxonomy. Mode-aware routing offers no consistent gain under small per-mode samples, requir-

Table 4. RUL prediction performance across observation ratios (mean $\pm$ std over 3 repeats).

Ratio	MAE	RMSE	Raw Cov.	Conformal Cov.	Raw Width	Conformal Width	GP warning rate
0.5	1.86 ± 0.23	2.98 ± 0.90	0.82 ± 0.16	0.93 ± 0.07	4.40 ± 1.56	10.70 ± 5.28	1.00 ± 0.00
0.7	1.18 ± 0.25	1.64 ± 0.38	0.78 ± 0.16	0.92 ± 0.07	3.41 ± 1.07	5.83 ± 1.06	1.00 ± 0.00
0.9	0.76 ± 0.25	1.11 ± 0.35	0.96 ± 0.06	0.96 ± 0.06	2.24 ± 0.67	2.80 ± 0.91	0.96 ± 0.06

Table 5. Comparison against classical degradation baselines (nominal 90% conformal intervals).

Method	Ratio	MAE	Conformal Cov.	Conformal Width
GP (ours)	0.5	1.86 ± 0.23	0.93 ± 0.07	10.70 ± 5.28
Wiener (drift)	0.5	2.10 ± 0.29	0.88 ± 0.01	10.13 ± 1.67
Exponential trend	0.5	14.86 ± 14.19	0.88 ± 0.01	144.25 ± 118.64
GP (ours)	0.7	1.18 ± 0.25	0.92 ± 0.07	5.83 ± 1.06
Wiener (drift)	0.7	1.62 ± 0.16	0.97 ± 0.06	6.82 ± 0.82
Exponential trend	0.7	5.45 ± 3.62	0.88 ± 0.01	36.36 ± 30.36
GP (ours)	0.9	0.76 ± 0.25	0.96 ± 0.06	2.80 ± 0.91
Wiener (drift)	0.9	0.85 ± 0.17	1.00 ± 0.00	4.16 ± 0.10
Exponential trend	0.9	1.19 ± 0.64	0.89 ± 0.01	6.77 ± 2.85

Table 6. NASA FD001 GP-RUL interval calibration (nominal 90% CI).

Ratio	Raw Cov.	Conformal Cov.	Conformal Width
0.5	0.79	0.90	443.81
0.7	0.83	0.90	301.73
0.9	0.49	0.90	135.55

Table 7. NASA C-MAPSS point-prediction benchmark (RMSE in cycles).

Subset	Ridge	RF	GBRT
FD001	38.8	33.1	31.9
FD002	41.5	32.7	32.4
FD003	55.0	45.4	44.4
FD004	64.2	49.2	46.5

ing larger datasets to realize potential benefits.

Decision-facing operationalization. A simple tiered action rule for calibrated RUL intervals (90% interval $[L, U]$, action threshold T, risk buffer Δ):

- Immediate inspection/replacement if $L \leq T$;
- Defer but monitor if $L > T$ and $U \leq T + \Delta$;
- Continue operation if $L > T + \Delta$.

Calibrated intervals make the risk tradeoff transparent and auditable, avoiding systematic risk underestimation from under-covering raw intervals.

6. Conclusion

This paper presents two complementary, independently deployable tools for small-sample cryogenic reliability assessment of reusable liquid rocket engine components: (A) DP-GMM-based unsupervised stratification with mode-conditional KM curves for heterogeneity-aware reliability estimation under right censoring; (B) GP-based RUL prediction with unit-level conformal calibration for trustworthy prediction intervals under sparse observation. Exploratory analysis clarifies that mode-aware routing is technically feasible but limited by small per-mode samples. External validation on the NASA C-MAPSS benchmark confirms the cross-domain robustness of the calibration component. Future work will integrate physical prior information into the DP-GMM framework to improve mode interpretability, and validate the pipeline on real cryogenic valve test data.

References

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics 2*(6), 1152–1174.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistics 1*(2), 209–230.

Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53*(282), 457–481.

Lei, Y., N. Li, L. Zhang, Z. He, and S. X. Ding (2018). Machinery health prognostics: A systematic review from data acquisition to remaining useful life prediction. *Mechanical Systems and Signal Processing 104*, 799–834.

Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Romano, Y., E. Patterson, and E. J. Candès (2019). Conformalized quantile regression. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Saxena, A., K. Goebel, D. Simon, and N. Eklund (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. In *International Conference on Prognostics and Health Management (PHM)*.

Si, X., W. Wang, C. Hu, and D. Zhou (2011). Remaining useful life estimation—a review on the statistical data-driven approaches. *European Journal of Operational Research 213*(1), 1–14.

Vovk, V., A. Gammerman, and G. Shafer (2005). *Algorithmic Learning in a Random World*. Springer.