

Certificate in Maritime Data Analytics

Statistical Methods for Maritime Data Analysis

ANOVA (Analysis of Variance) – Related terms: F-test, between-group variance, within-group variance. ANOVA tests whether the means of three or more groups differ significantly by comparing the ratio of between-group variance to within-group variance. In maritime data, ANOVA can assess whether fuel consumption varies across vessel types (e.g., tankers, bulk carriers, container ships). Example: comparing average voyage speed for ships operating in different sea zones. Practical application includes optimizing route planning by identifying zones where speed differences are statistically significant. Challenges involve meeting assumptions of normality and homoscedasticity; maritime data often exhibit skewed distributions and heteroscedasticity, requiring transformations or robust alternatives.

ARIMA (AutoRegressive Integrated Moving Average) – Related terms: time-series forecasting, differencing, p-d-q parameters. ARIMA models predict future values of a series by combining autoregressive (AR) terms, integration (I) for differencing, and moving-average (MA) components. For ship arrival times at a port, an ARIMA model can forecast daily inbound traffic, aiding berth allocation. Example: fitting an ARIMA(2,1,1) model to weekly container volume data. Practical application includes capacity planning for terminal operators. Challenges include selecting appropriate orders, handling seasonal patterns (often requiring SARIMA extensions), and dealing with missing or irregularly spaced observations common in AIS datasets.

Bayesian Inference – Related terms: prior distribution, posterior distribution, Markov Chain Monte Carlo (MCMC). Bayesian inference updates prior beliefs with observed data to obtain a posterior distribution, allowing probabilistic statements about parameters. In maritime risk assessment, Bayesian methods can estimate the probability of collision given historical near-miss events. Example: using a beta prior for the collision probability and updating with new AIS incident reports. Practical application involves real-time decision support where data arrive incrementally. Challenges include choosing appropriate priors, computational intensity of MCMC for high-dimensional models, and communicating probabilistic results to non-technical stakeholders.

Chi-Square Test – Related terms: independence test, goodness-of-fit, contingency table. The chi-square test evaluates whether observed frequencies differ from expected frequencies under a null hypothesis. In maritime contexts, it can test whether the distribution of vessel types across different port regions deviates from a uniform expectation. Example: constructing a 3×4 contingency table of ship class versus port zone and computing the chi-square statistic. Practical application includes monitoring compliance with port allocation policies. Challenges arise from small expected counts, which violate test assumptions; maritime datasets with rare vessel categories may require Fisher's exact test as an alternative.

Cluster Analysis – Related terms: k-means, hierarchical clustering, silhouette score. Cluster analysis groups observations into homogeneous subsets based on similarity metrics. For AIS trajectory data, clustering can

identify typical shipping lanes or anomalous routes. Example: applying k-means to latitude-longitude points to extract major maritime corridors. Practical application includes maritime traffic management and environmental monitoring of high-traffic zones. Challenges involve selecting the number of clusters, handling high dimensionality, and accounting for temporal dynamics that static clustering ignores.

Confidence Interval – Related terms: margin of error, 95 % CI, bootstrapping. A confidence interval provides a range of plausible values for an unknown parameter with a specified confidence level. In ship fuel efficiency studies, a 95 % CI around the mean fuel consumption per nautical mile quantifies estimation uncertainty. Example: using t-distribution to compute the interval for a sample of 30 voyages. Practical application includes reporting performance metrics to regulators. Challenges include ensuring the interval's coverage when data are non-normal or autocorrelated, which may require bootstrap or Bayesian credible intervals.

Cross-Validation – Related terms: k-fold, hold-out, overfitting. Cross-validation assesses model generalizability by partitioning data into training and testing subsets repeatedly. For predicting vessel arrival times, a 5-fold cross-validation can evaluate the predictive accuracy of a regression model. Example: splitting a year's worth of AIS records into five blocks and rotating the test block. Practical application ensures that deployed models perform reliably on unseen voyages. Challenges include temporal dependence; random folds may break time order, so blocked or rolling validation schemes are often preferred for maritime time-series data.

Data Imputation – Related terms: missing data, multiple imputation, mean substitution. Imputation fills gaps in incomplete datasets, preserving analytical power. In AIS streams, gaps arise from satellite coverage loss; linear interpolation or model-based imputation can reconstruct missing positions. Example: applying Kalman filtering to estimate ship tracks during signal outages. Practical application enables continuous monitoring of vessel routes. Challenges include avoiding bias; naïve methods can distort speed estimates, while multiple imputation adds computational complexity.

Decision Tree – Related terms: CART, random forest, feature importance. Decision trees partition data based on predictor variables to predict an outcome. For classifying high-risk voyages, a CART model can split on variables such as vessel age, cargo type, and weather severity. Example: a tree that first splits on cargo hazard level, then on vessel length. Practical application includes rule-based alerts for maritime safety officers. Challenges involve pruning to prevent overfitting, handling unbalanced classes (e.g., few collision events), and ensuring interpretability when trees become deep.

Empirical Bayes – Related terms: shrinkage estimator, hierarchical modeling. Empirical Bayes estimates prior hyperparameters from the data itself, then performs Bayesian updating. In port throughput analysis, empirical Bayes can shrink noisy monthly volume estimates toward a global mean, reducing variance. Example: estimating ship arrival rates for small ports with limited observations. Practical application improves forecasting for under-sampled locations. Challenges include assuming exchangeability across groups; heterogeneity in maritime operations may violate this assumption.

Exponential Smoothing – Related terms: Holt-Winters, smoothing parameter, forecast horizon. Exponential smoothing generates forecasts by weighting recent observations more heavily. For daily cargo handling volumes, simple exponential smoothing can produce short-term forecasts with minimal computational effort. Example: setting smoothing factor $\alpha = 0.3$ to predict next-day throughput. Practical application assists terminal operators in staffing decisions. Challenges include selecting appropriate smoothing parameters, handling trend and seasonality (requiring Holt or Holt-Winters extensions), and sensitivity to outliers common in episodic shipping events.

Generalized Linear Model (GLM) – Related terms: logistic regression, Poisson regression, link function. GLMs extend linear regression to response variables that follow exponential family distributions. For modeling the count of port calls per vessel, a Poisson GLM relates explanatory variables (e.g., vessel size, route length) to the expected count. Example: using a log link to predict expected arrivals. Practical application supports capacity planning and tariff design. Challenges include overdispersion (variance > mean) requiring negative binomial models, and zero-inflation when many vessels have no calls in a period.

Geostatistics – Related terms: variogram, kriging, spatial autocorrelation. Geostatistical methods model spatially correlated data, providing predictions at unsampled locations. In maritime pollution monitoring, kriging can estimate oil concentration across a sea area based on sparse sensor measurements. Example: constructing an empirical variogram of concentration versus distance and applying ordinary kriging. Practical application aids environmental response planning. Challenges include non-stationarity due to ocean currents, anisotropy, and the need for high-quality spatial covariance estimates.

Granger Causality – Related terms: vector autoregression (VAR), lagged variables, causality test. Granger causality assesses whether one time series provides predictive information about another. In maritime logistics, testing whether weather index series Granger-cause changes in vessel speed can inform dynamic routing. Example: fitting a VAR(2) model to speed and wind speed series and performing an F-test. Practical application improves predictive models for voyage duration. Challenges include requiring stationary series, selecting appropriate lag length, and interpreting causality as predictive rather than true causal influence.

Hierarchical Clustering – Related terms: dendrogram, agglomerative, linkage criteria. Hierarchical clustering builds a tree of nested clusters without pre-specifying the number of groups. For AIS data, it can reveal nested shipping routes, from broad oceanic corridors to local port approaches. Example: using Ward's linkage on Euclidean distances of vessel trajectories. Practical application assists in multi-scale traffic analysis. Challenges include computational cost for large datasets, sensitivity to distance metric, and difficulty in cutting the dendrogram at an appropriate level.

K-Means Clustering – Related terms: centroid, within-cluster sum of squares, elbow method. K-means partitions observations into K clusters by minimizing intra-cluster variance. For classifying vessels based on speed, draft, and cargo, K-means can produce distinct operational profiles. Example: selecting $K = 4$ based on the elbow plot of total within-cluster sum of squares. Practical application supports targeted maintenance regimes. Challenges include sensitivity to initial centroids, requirement of spherical cluster

shapes, and inability to handle outliers without preprocessing.

Kaplan-Meier Estimator – Related terms: survival function, censoring, hazard rate. The Kaplan-Meier estimator computes the probability of survival beyond a given time, accommodating right-censored observations. In maritime reliability, it can estimate the time-to-failure distribution of engine components. Example: plotting the survival curve for a fleet of engines with some still operating at study end. Practical application guides preventive maintenance scheduling. Challenges include handling left-censoring (late entry), small sample sizes, and assuming independence between censored and uncensored observations.

Kruskal-Wallis Test – Related terms: non-parametric ANOVA, rank-based test, median comparison. The Kruskal-Wallis test compares medians across three or more groups without assuming normality. For comparing median fuel consumption across vessel categories when data are skewed, this test is appropriate. Example: ranking fuel consumption values and computing the H statistic. Practical application provides robust inference when parametric assumptions fail. Challenges include reduced power relative to ANOVA, and post-hoc analysis requiring pairwise comparisons with adjusted significance levels.

Logistic Regression – Related terms: binary outcome, odds ratio, maximum likelihood. Logistic regression models the log-odds of a binary event as a linear combination of predictors. In maritime safety, it can predict the probability of a collision given ship speed, traffic density, and visibility. Example: estimating odds ratios for each predictor and interpreting the increase in collision risk per unit speed increase. Practical application feeds into real-time alert systems. Challenges include multicollinearity among predictors, class imbalance (few collisions), and need for regularization to avoid overfitting.

Markov Chain Monte Carlo (MCMC) – Related terms: Gibbs sampler, Metropolis-Hastings, posterior sampling. MCMC generates dependent samples from a target posterior distribution when analytic forms are intractable. For Bayesian vessel trajectory models, MCMC can sample latent positions conditioned on noisy AIS observations. Example: running a Gibbs sampler for 10,000 iterations to estimate posterior means of speed parameters. Practical application enables full uncertainty quantification. Challenges include diagnosing convergence, choosing appropriate proposal distributions, and computational burden for large maritime datasets.

Monte Carlo Simulation – Related terms: random sampling, stochastic modeling, variance reduction. Monte Carlo simulation propagates uncertainty through a model by repeatedly sampling input distributions. In port capacity analysis, it can simulate arrival and service times to estimate berth occupancy distributions. Example: drawing 5,000 random vessel inter-arrival times from an exponential distribution and computing utilization. Practical application supports risk-aware planning. Challenges involve selecting realistic input distributions, handling correlated inputs, and ensuring sufficient sample size for stable results.

Multicollinearity – Related terms: variance inflation factor (VIF), ridge regression, predictor redundancy. Multicollinearity occurs when explanatory variables are highly correlated, inflating coefficient variance and destabilizing estimates. In maritime fuel consumption models, diesel consumption may be correlated with

engine load and ship speed, leading to multicollinearity. Example: calculating VIFs and finding values > 10 for speed and load variables. Practical application prompts variable selection or regularization techniques. Challenges include detecting hidden correlations, especially when variables are transformed, and balancing model interpretability with statistical stability.

Non-Parametric Regression – Related terms: kernel smoothing, spline regression, LOESS. Non-parametric regression estimates relationships without assuming a specific functional form. For ship speed versus sea state, LOESS can capture nonlinear effects of wave height on speed. Example: fitting a locally weighted scatterplot smoothing curve with $\text{span} = 0.3$. Practical application yields flexible predictive models adaptable to complex maritime phenomena. Challenges include selecting smoothing parameters, boundary bias, and computational load for large AIS datasets.

Principal Component Analysis (PCA) – Related terms: dimensionality reduction, eigenvectors, scree plot. PCA transforms correlated variables into orthogonal components that capture maximal variance. In maritime sensor networks, PCA can reduce a high-dimensional set of environmental variables (temperature, salinity, current) to a few principal components for downstream modeling. Example: retaining components that explain 85 % of total variance. Practical application simplifies models and visualizes multivariate patterns. Challenges include interpreting components, handling missing data, and ensuring that variance captured aligns with operational relevance.

Propensity Score Matching – Related terms: causal inference, treatment effect, covariate balance. Propensity score matching creates comparable groups based on the probability of receiving a treatment, reducing selection bias. In evaluating the impact of a new navigation aid, vessels using the aid (treated) are matched to similar vessels not using it (control) based on size, route, and cargo. Example: estimating the average treatment effect on voyage duration. Practical application provides quasi-experimental evidence for policy decisions. Challenges include achieving adequate overlap, choosing matching algorithms, and assessing residual imbalance.

Quantile Regression – Related terms: conditional quantile, asymmetric loss, τ -level. Quantile regression models specific percentiles of the response distribution, offering insight beyond the mean. For maritime delay analysis, estimating the 90th percentile of arrival delays helps design robust schedules. Example: fitting a $\tau = 0.9$ quantile regression with ship length and weather index as predictors. Practical application supports risk-averse planning. Challenges include selecting appropriate quantiles, interpreting coefficients that vary across quantiles, and handling heteroscedasticity.

Random Forest – Related terms: ensemble learning, bagging, feature importance. Random forest builds many decision trees on bootstrap samples and aggregates predictions, improving accuracy and robustness. In classifying vessel behavior (e.g., fishing vs. transport), a random forest can capture complex interactions among speed, heading, and AIS message frequency. Example: training 500 trees and extracting the top five important variables. Practical application yields high-performing classifiers with built-in out-of-bag error estimates. Challenges include interpretability compared to single trees, hyperparameter tuning (number of

variables tried at each split), and computational cost for large maritime datasets.

Regression Diagnostics – Related terms: residual plot, Cook’s distance, heteroscedasticity test. Diagnostics assess the validity of regression assumptions and identify influential observations. For a linear model of fuel consumption, residual plots can reveal non-linearity, while Cook’s distance flags voyages that disproportionately affect coefficient estimates. Example: detecting a single long-duration voyage with high leverage. Practical application ensures model reliability before deployment. Challenges include handling autocorrelated residuals common in time-ordered maritime data and distinguishing genuine outliers from legitimate extreme events.

Seasonal Decomposition – Related terms: STL (Seasonal-Trend decomposition using Loess), additive model, multiplicative model. Seasonal decomposition separates a time series into trend, seasonal, and residual components. For monthly cargo throughput, STL can isolate a recurring summer peak from the underlying growth trend. Example: applying STL with seasonal window = 12 months. Practical application informs capacity adjustments aligned with seasonal demand. Challenges include selecting appropriate window lengths, handling irregular seasonality due to holidays or weather anomalies, and ensuring that residuals are approximately white noise.

Spearman Rank Correlation – Related terms: monotonic relationship, non-parametric correlation, rho. Spearman’s rho measures the strength of a monotonic association between two variables using ranks. In maritime analytics, it can assess the relationship between vessel draft and reported fuel consumption when the relationship is non-linear. Example: computing $\rho = 0.68$ indicating a strong positive monotonic trend. Practical application provides a robust correlation metric when data contain outliers. Challenges include reduced power compared to Pearson’s correlation for linear relationships and the need for adequate sample size.

Spatial Autocorrelation – Related terms: Moran’s I, Geary’s C, spatial lag. Spatial autocorrelation quantifies the degree to which nearby observations exhibit similar values. In analyzing piracy incidents, a high positive Moran’s I suggests clustering in certain maritime zones. Example: calculating Moran’s I = 0.42 for incident counts across grid cells. Practical application guides resource allocation for patrols. Challenges include defining appropriate spatial weights, accounting for edge effects, and distinguishing true clustering from underlying covariate patterns.

Survival Analysis – Related terms: hazard function, censoring, Cox proportional hazards model. Survival analysis models time-to-event data, accommodating censored observations. In maritime maintenance, it can estimate the time until a propeller requires overhaul. Example: fitting a Cox model with covariates ship age and operating hours to obtain hazard ratios. Practical application informs predictive maintenance schedules. Challenges include proportional hazards assumption violations, handling time-varying covariates, and sparse failure events leading to unstable estimates.

Time-Series Cross-Correlation – Related terms: lagged correlation, cross-covariance, pre-whitening.

Cross-correlation evaluates the similarity between two time series at different lags. For ship speed and wave height, a peak at lag = 2 hours may indicate that speed reacts to wave changes after a short delay. Example: plotting the cross-correlation function and identifying significant lags beyond the confidence bounds. Practical application improves lag selection for predictive models. Challenges include spurious correlations due to shared trends, requiring detrending or pre-whitening before analysis.

Time-Series Decomposition (STL) – Related terms: trend component, seasonal component, remainder. STL decomposes a series into additive components using Loess smoothing, handling non-linear trends and changing seasonality. For weekly AIS-derived traffic density, STL can separate a long-term increase in traffic from periodic weekly patterns. Example: specifying seasonal window = 7 days. Practical application facilitates separate modeling of trend and seasonality. Challenges include choosing smoothing parameters that balance over-smoothness against noise retention, and ensuring residuals are random.

Time-Series Forecasting (ARIMAX) – Related terms: exogenous variables, dynamic regression, lagged predictors. ARIMAX extends ARIMA by incorporating external regressors. For forecasting port call volume, weather indices (e.g., wind speed) can be included as exogenous variables. Example: fitting an ARIMAX(1,1,1) model with wind speed lagged by one day. Practical application yields more accurate forecasts by leveraging auxiliary information. Challenges involve parameter selection for both ARIMA and exogenous parts, multicollinearity among exogenous variables, and ensuring stationarity after differencing.

U-Statistics – Related terms: unbiased estimator, kernel function, Hoeffding's decomposition. U-statistics provide unbiased estimators for population parameters defined as expectations of symmetric kernels. In maritime data, the Gini coefficient of cargo distribution across vessels can be estimated via a U-statistic. Example: using pairwise differences of cargo loads to compute the Gini index. Practical application offers robust measures of inequality. Challenges include computational intensity for large samples (quadratic complexity) and handling ties in rank-based kernels.

Variance Inflation Factor (VIF) – Related terms: multicollinearity diagnostic, tolerance, eigenvalue analysis. VIF quantifies how much the variance of a regression coefficient is inflated due to linear dependence among predictors. In a model predicting emissions, a VIF of 12 for ship speed indicates severe multicollinearity with engine load. Example: computing VIFs and deciding to drop or combine correlated variables. Practical application improves model stability. Challenges include interpreting VIF thresholds (commonly 5 or 10) in the context of domain knowledge, and addressing collinearity without sacrificing important predictors.

Wavelet Transform – Related terms: multi-resolution analysis, discrete wavelet transform (DWT), time-frequency localization. Wavelet transforms decompose signals into localized time-frequency components, useful for detecting transient features. In analyzing ship vibration signals, wavelets can isolate fault-related bursts from background noise. Example: applying a Daubechies-4 DWT to a 10-minute vibration record and reconstructing high-frequency coefficients. Practical application supports condition monitoring. Challenges include selecting appropriate mother wavelet, determining decomposition level, and interpreting coefficients in a physical context.

Wilcoxon Signed-Rank Test – Related terms: non-parametric paired test, median difference, rank-based statistic. The Wilcoxon signed-rank test evaluates whether the median difference between paired observations differs from zero. For assessing the effect of a speed-reduction policy, compare vessel speeds before and after implementation for the same ships. Example: calculating the test statistic and obtaining a p-value

Zero-Inflated Poisson (ZIP) Model – Related terms: excess zeros, mixture model, overdispersion. ZIP models address count data with more zeros than expected under a standard Poisson distribution. In modeling daily port calls, many days may have zero arrivals for certain vessel types, necessitating a ZIP approach. Example: estimating separate processes for the zero-inflation component and the Poisson count component. Practical application improves forecast accuracy for low-traffic routes. Challenges include model identifiability, selecting appropriate covariates for each component, and convergence issues during maximum likelihood estimation.