ARD: large yacht sustainability metric analysis

Pop-up ARD 2023

Report No.	:	70138-2-RD
Date	:	January 2024
Version	:	1.0
		Final Report

ARD: large yacht sustainability metric analysis

Pop-up ARD 2023

MARIN order No.	:	70138
MARIN Project Manager	:	Kirsten Odendaal
Number of pages	:	47
Ordered by	:	MARIN ARD
Reported by	:	Kirsten Odendaal
Reviewed by	:	Christian Veldhuis, Enrico Della Valentina, Bart Mak

Version	Date	Version Description
1.0	10/01/2024	Final report

CONTENTS PAGE

REVIEW OF TABLES AND FIGURES

REVIEW OF DELIVERABLES

The following table indicates all deliverables written in the course of the present project:

MARIN Report No.	Contains	Date
70169.301	Final ARD Report	December 2023

MANAGEMENT SUMMARY

In the evolving landscape of maritime sustainability, the emergence of the Yacht Environmental Transparency Index (YETI score) stands as a novel endeavour aiming to assess and promote environmental practices within the yacht industry. Therefore, this report embarks on a comprehensive exploration of the Yacht Environmental Translational Index (YETI) score, paralleling its principles and implications with the well-established Energy Efficiency Design Index (EEDI) utilised in the shipping domain.

The core objectives encompass a comparative analysis between the YETI score and EEDI, leveraging exploratory data analysis methods to delve into existing datasets and quantification of parameter importance. Additionally, this report investigates advanced synthetic data generation techniques strategically applied to fill gaps within the existing dataset. Through these multifaceted approaches, this study endeavours to offer insights into the effectiveness, limitations, and potential advancements of these environmental indices, thereby contributing significantly to the discourse on fostering sustainable practices within maritime industries. The following conclusions summarise the findings of the present project:

Exploratory Data Analysis and Sensitivity Study

- Comparison between Eco-points, Eco-score, and EEDI indicated a high degree of noise and scatter in the Eco-score metric due to the corresponding noise in the normalisation metric, GT.
- Eco-score exhibits similar trends to EEDI and globally penalises smaller yachts, consistent with initial global sensitivity study observations.
- The sensitivity analysis promptly identified a critical bug within the generator database, highlighting the immediate utility of such analyses in detecting intricate issues that might be challenging to discern manually.
- Among the eleven parameters, three parameters - Length, GT, and Avg. Electric Load emerged as the most influential on the eco-point metric. Surprisingly, main engines, generators, and associated parameters exhibited marginal impacts, contrary to initial presumptions.
- Further delving into these three parameters revealed that GT and Avg. Electric Load significantly contributes to approximately 90% of the total variance, showcasing strong and distinctive underlying trends.
- The impact of the speed-power profile on eco-point scores indicated intriguing trends, particularly observing that with increasing GT, the global impact of the speed-power profile diminishes, leading to globally similar eco-point evaluations irrespective of the speed-power profiles.

This study has effectively provided concrete and mathematical evidence regarding parameter impacts on the YETI eco-point metric, dispelling intuitions that were primarily experience-based and challenging to quantify without empirical evidence.

Synthetic Data Generation

- Regarding synthetic data generation, two methodologies were evaluated: quasi-random model-based generation (QMG) and Gaussian Copula (GC). While the former offers transparency and modularity but lacks capturing intricate parameter interactions, the latter presents more realistic data by capturing complex parameter relationships but operates as a black-box approach.
- Additionally, the extraction of realistic speed-power profiles into synthetic data points via exponential curve-fitting and nearest neighbour approaches was achieved. However, challenges were encountered in the matching approach when using the YETI tool as a back-calculator, leading to multiple minima and favouring unrealistic top speeds, necessitating the implementation of practical upper bounds.

- Nevertheless, the methodologies devised in the synthetic data generation phase have enabled a reliable population of sparse data regions, facilitating a deeper understanding of governing trends within the YETI fleet dataset and thereby allowing for further insight into the critical decision-making processes related to YETI metrics.

Collectively, these conclusions underscore the significance of empirical analysis and systematic approaches in unravelling complexities within the YETI fleet domain, providing a robust foundation for future enhancements and strategic decision-making. This comprehensive analysis contributes substantially to the ongoing evolution of YETI fleet management practices, offering actionable insights derived from rigorous analysis and methodologies.

1 INTRODUCTION

Recently, a new work item has been proposed to the ISO [1]. This proposal introduces a new standard aimed at evaluating the environmental footprint of large yachts through a metric known as Eco-points. Unlike the prevalent Energy Efficiency Design Index (EEDI), primarily tailored for conventional vessels, Eco-points concentrate on the distinct ecological aspects pertinent to luxury yachts. This novel metric promises a comprehensive assessment framework enabling comparisons across the fleet and the establishment of environmental ratings. The Eco-points methodology, akin to EEDI, encompasses a series of formulas designed to gauge various components contributing to the environmental impact of large yachts. Its current scope focuses predominantly on operational efficiency, yet offers a platform that can be expanded, potentially including a comprehensive life cycle assessment of a yacht's environmental impact. Nevertheless, the overarching objective revolves around setting precise benchmarks, ensuring the forthcoming generation of yachts embodies enhanced eco-friendliness.

However, the existing ECO-points metric grapples with significant limitations rooted in a relatively limited and incomplete existing yacht database. These constraints impede the accuracy and robustness of sustainability evaluations for large yachts. Addressing this challenge, this project adopts an innovative approach by employing advanced statistical analysis techniques. The aim is to ascertain the overall model sensitivity and robustness, thereby augmenting the coverage and effectiveness of the ECO-points metric across a wider spectrum of vessels. Concurrently, this initiative aligns with the imperative focus on zero-emission research within the maritime sector.

Thus, the primary goal of this project is twofold: comprehending the current ECO-point methodology and enriching the sustainability database by leveraging sophisticated statistical methods. Through this pioneering approach, the endeavour strives to bridge the existing gaps in the database, culminating in a more inclusive and precise portrayal of yacht sustainability, encompassing a broader range of realistic vessels.

Observed Importance and Benefits:

1. Understanding: Delving into the composition of current ECO-point metrics and contrasting them with established metrics like EEDI, underscoring the significance of sustainability and ecological responsibility in the Large Yacht sector.
2. Exploration: Application of contemporary statistical techniques to address a real-world case study carrying substantial sustainability implications.
3. Transparent Methodology: A comprehensive comparison, evaluation, and detailed reporting of both general and advanced methodologies for generating synthetic data.

1.1 Report Structure

The report will be structured into three main sections to address the primary objectives above:

1. Data pre-processing and analysis:
   a. Begins by thoroughly analysing the existing YETI fleet dataset and identifying the gaps and regions with limited information. Furthermore, investigation of the correlations and relationships between variables within the data is considered.
   b. Existing ECO points and ECO-score formulations are shown and compared. These are further analysed and reflected upon on the overall link between the non-yacht EEDI metric.
2. Model sensitivity analysis:
   a. A detailed look into how the input parameters influence the outcomes using a global sensitivity metric (Sobol Indices). This approach quantifies the importance of the input parameters and their interactions with the output eco score metric.

3. Realistic data generation (space-filling methodologies):
    a. To populate sparse data regions within the limited fleet dataset with realistic data points, two methodologies to generate synthetic data are explored:
    b. Quasi-Random Model-based Generation: A manual look into conventional modelling methodologies to create potentially realistic data points using a combination of interpolation, added noise, and domain knowledge.
    c. Statistical Generation Technique: Application of the Gaussian Copula Model (GC) algorithm to the dataset to uncover underlying clusters by leveraging inherent existing data distributions.

2 ENVIRONMENTAL INDICES

2.1 YETI (Yacht Environmental Transparency Index)

The yacht environmental transparency index (YETI), is a new metric which measures a yacht's environmental impact based on its operational profile. This profile is determined using a collection of yacht AIS details, which revealed, on average, that superyachts are in port 60 per cent of the year, at anchor 30 per cent, and at sea 10 per cent. Using this generalised operational profile, YETI considers a yacht's general parameters, along with the speed-power profile, engines, gensets, battery banks, and more. The index then converts calculated emissions into Eco-points. This metric is the quantity of pollutants emitted per kilogram of analysed fuel and the corresponding exhaust gases [2]. Finally, the yacht's environmental impact score (Eco score) is obtained by dividing Eco points by the vessel's gross tonnage.

The International Maritime organisation (IMO) has introduced the Energy Efficiency Design Index (EEDI) to limit the CO2 emission from shipping. Unlike the IMO's rating system for ships, which rates CO2 output against the vessel's utility, YETI focuses on the characteristics of yachting operation – namely, the utility of providing their owners with enjoyment.

Eco-points Formulation: The eco-points metric is a comprehensive evaluation system that considers both upstream (production) and downstream (operational emissions) impacts within its assessment framework. It utilises a summation formulation incorporating specific impact factors (see Table 2-1) to precisely quantify the environmental effects stemming from emissions like CO2, CO, HC, NOx, Ammonia, and PM. Additionally, the metric encompasses upstream utilities such as fuel and urea production, as well as shore power impacts, integrating these factors by multiplying them with the yearly produced equivalent. This approach relies on established operational profiles and corresponding system configurations to accurately gauge the cumulative environmental impact across the entire lifecycle of a process or product.

$$\mathrm{Eco-points} = f_{CO_2} \cdot \overline{CO_2} + f_{CO} \cdot \overline{CO} + f_{HC} \cdot \overline{HC} + f_{NO_x} \cdot \overline{NO_x} + f_{Am} \cdot \overline{Ammonia} + f_{PM} \cdot \overline{PM} + f_{Dies} \cdot \overline{Fuel_{Dies}} + f_{Urea} \cdot \overline{Urea} + f_{shore} \cdot \overline{Shore}$$

Table 2-1: Specific impact factors for emissions upstream and downstream

Specific Impact Downstream (Emission to Air)		
Emission	f	units
CO2	0.013284	points/kg
CO*	0.0084	points/kg
HC*	0.0333	points/kg
NOx	1.000129	points/kg
Ammonia	2.139424	points/kg
PM 2.5	7.94427	points/kg
Specific Impact Upstream		
Utility	f	units
Diesel Production (WTT)	30.23	points/ton
Shore Power	1.17	points/MWh
Urea Solution	40.94	points/ton

** TNO 2007*

2.2 EEDI (Energy Efficiency Design Index):

The Energy Efficiency Design Index (EEDI) is an international measurement tool established by the International Maritime Organization (IMO). It aims to assess the energy efficiency of new ships based on their design parameters and performance characteristics. EEDI sets a standard for reducing greenhouse gas emissions from ships by evaluating their CO2 emissions per ton of cargo transported per nautical mile travelled. Ships are required to meet specific EEDI requirements based on their type, size, and operational functions. Compliance with EEDI is mandatory for new ships built after set deadlines, encouraging the maritime industry to develop and adopt more energy-efficient vessel designs and technologies. However, the EEDI is not enforced for yachts yet, so only time can tell whether or not it will be applicable to yachts in the future. [3].

3 EXPLORATORY DATA ANALYSIS (EDA):

Exploratory Data Analysis (EDA) is a crucial first step in the data analysis process, where you explore and understand your data before conducting more advanced statistical or machine learning analyses. It involves a series of techniques and methods to summarise, visualise, and interpret data sets to gain insights into their underlying patterns, structure, and characteristics. Typically, an EDA process can be decomposed into the following key components:

1. Data Cleaning: During EDA, you identify and deal with missing values, outliers, and inconsistencies in the data. Cleaning the data ensures that your analyses are based on high-quality information.in
2. Data Exploration: EDA involves summarising the main characteristics of your data, such as its size, data types, and basic statistics (mean, median, standard deviation, etc.). This step helps you get a general sense of the data.
3. Patterns and Relationships: EDA enables the discovery of patterns and relationships between variables. It can uncover correlations, trends, and dependencies, which can guide subsequent analyses and decision-making.
4. Visualisation: EDA includes creating visual representations of the data, such as histograms, scatter plots, box plots, and more. Visualisation helps you grasp the data's distribution, relationships between variables and potential patterns.

Using these four steps, an exploratory analysis is conducted on the existing dataset with the hope of understanding the initial trends and comparisons between the various sustainability metrics.

3.1 Data Exploration and Cleaning

The first step in the exploratory data analysis is to ensure that we have a dataset that does not contain any irregularities. This is conventionally known as data cleaning. This step traditionally focuses on eliminating data points that are either duplicated or not found within the dataset. After applying this approach and restructuring the dataset for future analysis the following dataset comparison can be shown as:

1. Initial YETI yacht database: (rows: 54, columns: 127)
2. Cleaned YETI yacht database: (rows: 42, columns: 19)

It can be seen that the initial data set has reduced from 54 vessels to approximately 42 points. This reduction is due to the fact that 12 vessels did not contain any indication of speeds for the associated speed-power profiles. This parameter is a critical component in the analysis and cannot be ignored. When looking at the feature columns we can see a drastic reduction. This is mainly due to the re-structuring of the dataset to retain the minimum critical features that are required for both the EEDI score and the YETI score to be extracted. The list of parameters can be seen in the table below,

Table 3-1: List of retained parameters for the exploratory data analysis

	Parameters	
1	*ship*	Ship ID
2	*L*	Length
3	*GT*	GT
4	*avg_elec*	Avg. Electric Load
5	*nom_me*	Nominal Main Engine Power
6	*num_me*	Number Main Engine
7	*sfc_me*	SFC Main Engine
8	*pto_me*	PTO Main Engine
9	*nom_gen1*	Nominal Generator 1 Power
10	*num_gen1*	Number Generator 1
11	*sfc_gen1*	SFC Generator 1
12	*nom_gen2*	Nominal Generator 2 Power
13	*num_gen2*	Number Generator 2
14	*sfc_gen2*	SFC Generator 2
15	*me_total*	Main Engine Total Power
16	*ae_total*	Auxiliary Total Power
17	*eco*	Eco-points
18	*eco_gt*	Eco-score
19	*EEDI*	EEDI

The second step of the data cleaning process is to identify if there is any irregular or outlier data points. These points typically have the power to negatively bias the overall trends and data analysis. Appendix A presents the data statistics of the cleaned dataset Additionally, the data structure histograms for each parameter are presented in Appendix B. From both of these, two obvious outliers are identified:

- Fully electric (No installed ME – only generators are considered) – data point is removed from the dataset.
- Large installed power (One vessel has an extremely large installed power as compared to the rest of the vessels in the data set) – data point is removed from the dataset to ensure distributions are not extremely biased.

Further inspection of the parameter histograms also presents an indication of data gaps or sparse regions within the feature parameters:

- Length: vessels > 65 m show a quick drop in the number of vessels
- GT: 2000 – 3000 GT indicates limited data
- Avg Electric: 300 kW – 400 kW indicates limited data in this region

3.2 Data Correlations

After the data cleaning process, a correlation investigation is continued. This allows for the discovery of patterns and relationships between variables in order to uncover underlying trends and dependencies. A heatmap of all associated pair-wise correlations found within the dataset is presented as follows,

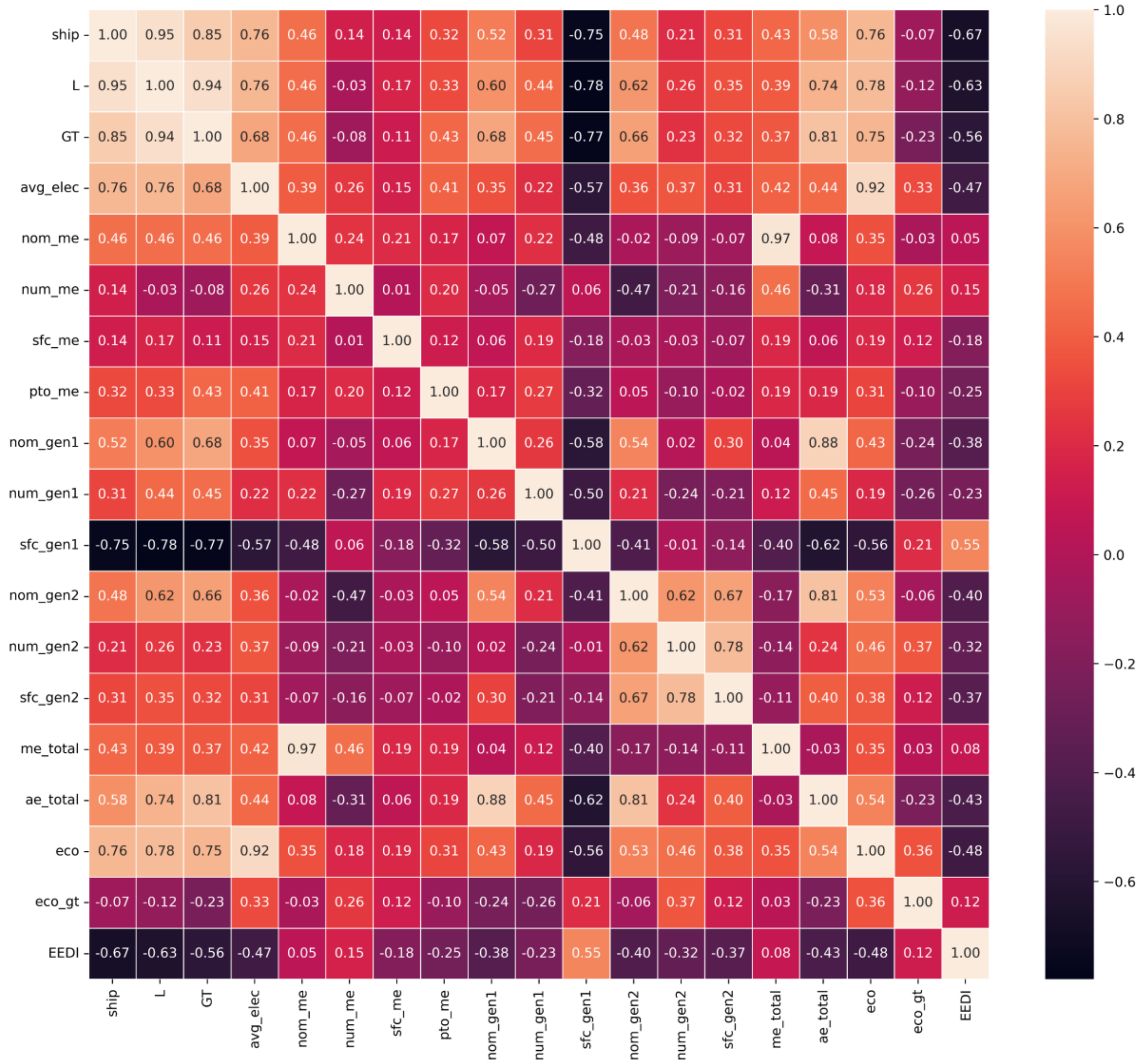


Figure 3-1: YETI dataset correlation heatmap

The Pearson correlation metric is applied. This metric accounts for the linear correlation component which ranges between extremes [-1.0, 1.0]. Unfortunately, more complex non-linear interactions are not thoroughly expressed. Nevertheless, such an analysis can give a quick and accurate indication of the parameters which are most strongly correlated to one another. Immediately it can be seen that the Average Electric Load parameter correlates $r = 0.92$, with the Eco-points metric. A summary of the top 15 (absolute) correlators are shown in Table 3-2.

Table 3-2: Top 15 absolute feature pair-wise correlations

Number	Parameter 1	Parameter 2	Absolute Pearson Correlation (r)
1	Nominal Main Engine	Main Engine Total	0.969
2	Length	GT	0.938
3	Avg. Electric Load	Eco-points	0.916
4	Nominal Gen1	Aux. Engine Total	0.883
5	Nominal Gen2	Aux. Engine Total	0.808
6	GT	Aux. Engine Total	0.806
7	Num. Gen2	SFC Gen2	0.784
8	Length	Eco-points	0.781
9	Length	SFC Gen1	0.779
10	GT	SFC Gen1	0.765
11	L	Avg. Electric Load	0.761
12	GT	Eco-points	0.752
13	L	Aux. Engine Total	0.740
14	GT	Avg. Electric Load	0.685
15	GT	Nominal Gen1	0.683

From this initial analysis, it is evident that parameters: Length, GT, and Avg. Electric Load shows the largest correlation to the raw eco-point metric.

3.3 Data visualisation and comparison

To further grasp the data's relationships and potential patterns, a direct visual inspection is carried out. The complete pair-wise scatter plot visualisations for all parameters can be found in Appendix C. However, having identified that Length, GT, and Average Electric Load are influential parameters, a direct comparison analysis is conducted by comparing each parameter with the corresponding environment metric; Eco-points, Eco-score (Eco-points/GT), and EEDI.

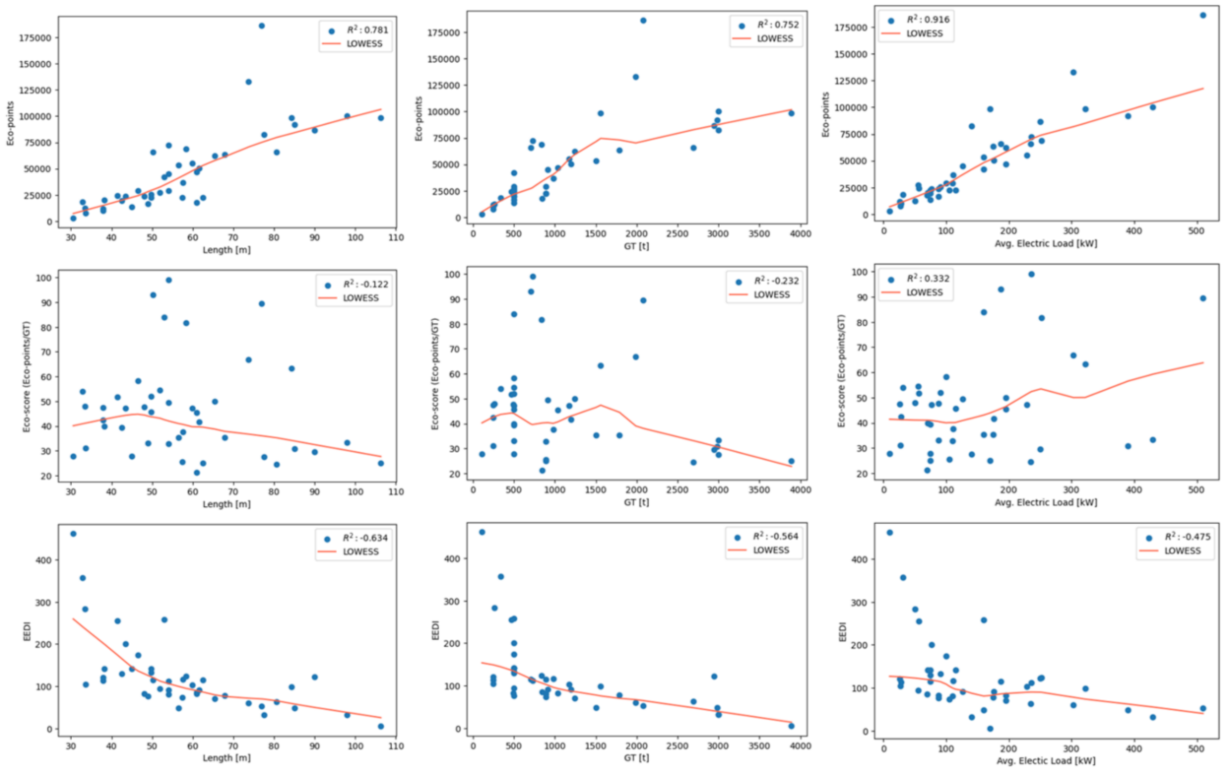


Figure 3-2: Scatter plot visualisation of the top correlating parameters

Based on a visual inspection of the associated scatterplots (BLUE) and the local moving average trend (RED) a few general observations can be made:

1. The eco-points metric demonstrates a very linear trend for all three parameters.
2. The eco-score exhibits an opposite trend when considering the Length and GT parameters. From an initial inspection Eco-score inherently penalises smaller yachts. Additionally, a large amount of scatter is observed. The likelihood is that the extreme scatter is a result of the normalising parameter GT which propagates its noise to the overall eco-points score.
3. The EEDI score demonstrates a very similar trend to the eco-score metric when looking at the Length and GT parameters. However, it exhibits much less scattering.

In summary, the following key takeaways are observed from the exploratory data analysis:

- Length, GT, and Avg. Electric Load is the strongest correlator with respect to the YETI eco-points metric.
- As Length increases; eco-points increases, eco-score decreases, and EEDI decreases .
- As GT increases; eco-points increases, eco-score decreases, and EEDI decreases.
- As Average Electric Load increases; eco-points increases, eco-score increases, EEDI decreases.

While initial trends can be observed, statistical significance and/or conclusions can be challenging to extract with a high degree of confidence due to the small size of the dataset. Additionally, such correlation studies, while often very important to investigate immediate trends, only consider the pair-wise connections and do not consider interaction terms. In other words, the global importance of features is not fundamentally captured. Therefore, a novel approach to evaluate the YETI tool's global sensitivity to the input parameters is explored.

4 GLOBAL SENSITIVITY ANALYSIS

4.1 General Sensitivity Background

Sensitivity analysis, often referred to as a sensitivity study or what-if analysis, is a critical technique used in engineering to assess the impact of changes in input variables or parameters on the output or performance of a system, model, or process. It involves systematically varying these input parameters within a defined range to understand how variations in these factors affect the overall outcomes, predictions, or results. Sensitivity analysis is used to gain insights into the robustness and reliability of engineering designs and models, making it an essential tool for decision-making and risk assessment.

Some general purposes and applications include:

- Model Validation and Verification: Sensitivity analysis helps validate and verify mathematical models by examining the effects of parameter changes on model outcomes.
- Optimisation: It aids in identifying critical input parameters that significantly influence the desired output, facilitating optimisation and efficient resource allocation.
- Risk Assessment: Engineers can use sensitivity analysis to assess the risks associated with uncertainties in input parameters, improving risk management strategies.
- Decision Making: Sensitivity analysis provides valuable insights for informed decision-making by demonstrating how varying parameters impact outcomes and support optimal decisions.
- Quality Control: In manufacturing, sensitivity analysis helps identify influential factors affecting product quality, leading to improved manufacturing processes and product quality.

The main importance of such methods are:

- In engineering, where precise and efficient designs are critical, understanding how variations in input parameters affect outputs is paramount.
- It enhances reliability and robustness by identifying parameters that could lead to system failures or suboptimal performance.
- Engineers can use sensitivity analysis to optimise designs, reduce costs, and improve efficiency by focusing on critical input parameters.

Many sensitivity analyses exist. However, the most common approaches are,

1. **One-at-a-time (OATS) Sensitivity Analysis:** In OATS, one input parameter is varied at a time while keeping all other parameters constant, allowing researchers to assess the impact of each parameter individually on the model’s output. Here are some pros and cons of the OATS approach:

Pros of the OATS Approach:

- Simplicity: OATS is easy to implement and understand, making it accessible to individuals without extensive expertise in sensitivity analysis.
- Interpretability: The results are easy to interpret because they directly show how each parameter influences the output, providing clear insights into the system’s behaviour.
- Low Cost: Since only one parameter is varied at a time, OATS typically requires fewer model evaluations compared to more complex sensitivity analysis methods, making it computationally efficient.

Cons of the OATS Approach:

- Misses Interaction Effects: OATS only considers the individual impact of each parameter, neglecting potential interactions between parameters. Parameters can interact complexly in many systems, and OATS doesn't capture these interactions.
- Inefficient for High-Dimensional Spaces: In systems with many parameters, OATS can be impractical as it requires running multiple simulations, each time varying a single parameter. This can be time-consuming and computationally expensive.
- Limited Insights: While OATS provides insights into the sensitivity of parameters individually, it may not provide a comprehensive view of the overall system behaviour. Some parameters may have minimal individual sensitivity but contribute significantly when combined with others.
- Risk of Oversimplification: Relying solely on OATS can lead to an oversimplified understanding of a complex system, potentially missing critical factors that affect the output.

2. **Global Sensitivity Analysis with Sobol Indices:** GSA is a more comprehensive approach to sensitivity analysis that examines how variations in input parameters collectively affect the output of a model. One widely used method in GSA is the Sobol indices, which provide insights into the relative importance of individual parameters and their interactions. Sobol indices are a set of sensitivity indices used to quantify the contribution of individual parameters and parameter interactions to the variance in the model's output. They provide a detailed breakdown of the importance of each parameter and various orders of interactions, allowing for a deeper understanding of a system's behaviour.

Pros of Global Sensitivity Analysis with Sobol Indices:

- Captures Parameter Interactions: Unlike the one-at-a-time approach, Sobol indices consider the impact of individual parameters and their interactions. This is particularly important for systems with complex, nonlinear, or non-additive relationships between parameters.
- Robust Insights: Sobol indices provide robust insights into the system's behaviour by quantifying the relative importance of parameters, including those that may have a small individual effect but a significant impact when combined with others.
- Risk Assessment: They are helpful in risk assessment by highlighting parameters and interactions that contribute the most to output variability, aiding in risk management and mitigation.
- Comprehensive Understanding: Sobol indices offer a more comprehensive understanding of a system, making them valuable for research, design, and decision-making processes.

Cons of Global Sensitivity Analysis with Sobol Indices:

- Computational Intensity: Sobol indices can be computationally intensive, especially when dealing with complex models or high-dimensional parameter spaces. The method requires running many model simulations to estimate the indices accurately.
- Data Requirements: Accurate estimation of Sobol indices may require a significant amount of data, which could be a limitation in cases where data is scarce or expensive.
- Complex Interpretation: Interpreting Sobol indices can be challenging, especially for high-dimensional models with numerous parameters and interactions. Visualising and understanding the results may require advanced techniques.

Due to the low computational cost of the YETI tool, the more complex global sensitivity using the Sobol indices approach was adopted to extract as much information as possible.

4.2 Sobol Methodology:

Applying a Sobol Indices approach for modelling sensitivity involves a specific methodology to evaluate how input parameters influence the model's output. The general steps are as follows:

1. Define the Model and Parameters:
   - Clearly define the mathematical or computational model you want to analyse.
   - Identify all the input parameters (factors) and their respective ranges or distributions.
2. Select Output Metrics:
   - Determine the output metrics or performance measures that you want to evaluate for sensitivity.
3. Generate Input Samples:
   - Generate a set of input parameter samples from their respective distributions. The number of samples should be sufficient for reliable statistical analysis, considering the desired level of precision.
4. Model Evaluation:
   - Evaluate the model for each combination of input parameter samples to generate the corresponding output values. This step may involve running simulations or solving equations.
5. Compute Sobol Indices:
   - Calculate the Sobol indices using the collected output data. The Sobol indices are typically calculated in a systematic way, breaking down the variance of the output into components attributed to individual parameters and their interactions.
   - The total Sobol index (S_{total}) quantifies the total variance in the output explained by all parameters.
   - The first-order Sobol indices (S_i) represent the variance attributable to each individual parameter.
   - Higher-order Sobol indices, $(S_{ij}, S_{ijk}, etc.)$ measure interactions between parameters.
6. Rank and Interpret Sobol Indices:
   - Interpret the results to understand which parameters have the most significant effect on the model's output. Parameters with high Sobol indices are the most influential.
7. Iterate and Refine:
   - If the analysis reveals important parameter interactions or dependencies, consider refining the model or conducting further investigations to account for these relationships.

4.3 Case Study: YETI Tool Eco-score

An initial sensitivity case focusing on conventional propulsion system configurations (diesel-direct) was explored by varying 11 YETI tool input parameters and using three different (scaled) speed-power profiles. The design bounds of the input parameters are as follows:

Table 4-1: Parameters and associated bounds used in the sensitivity analysis

Input Parameters	Minimum	Maximum	Units
Length	10	150	[m]
Functional Unit: GT	500	1500	[t]
Avg. Elec Load	0	500	[kW]
Main Nominal Power	0	5000	[kW]
Main Num. Engines	1	3	[-]
Gen 1 Nominal Power	50	1050	[kW]
Gen 1 Num. Engines	1	3	[-]
Gen 2 Nominal Power	50	1050	[kW]
Gen 2 Num. Engines	1	3	[-]
Waste Heat Recov. 1	0 (False)	1 (True)	[-]
Waste Heat Recov. 2	0 (False)	1 (True)	[-]
	Output Parameter		
Eco-score	-	-	[Eco-points/GT]

The corresponding speed-power profiles used are,

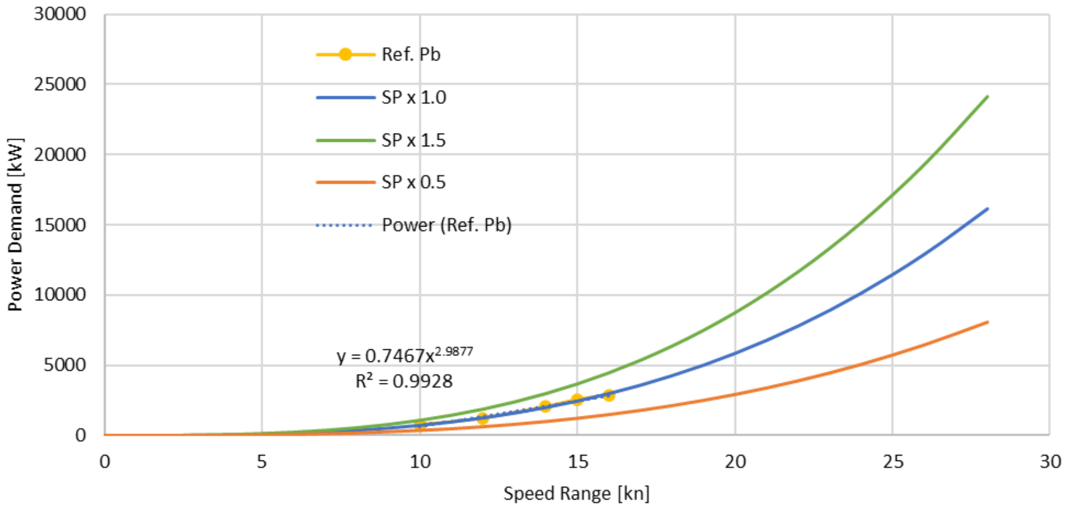


Figure 4-1: Manually scaled speed-power profiles (0.5, 1.0, 1.5) used in the sensitivity analysis

The speed-power profiles are based on an initial reference curve. This curve is then fitted using a cubic expression to extrapolate across the entirety of the speed range. From there, two additional curves are generated by considering a $\pm$50% scaling factor. While these profiles are not explicitly accounted for in the sensitivity analysis, the results are compared to each case to see the impact of the input speed-power curve on the overall metric importance and model sensitivity. It should be noted that the speed-power selection has been arbitrarily selected.

4.3.1 Design of Experiments (Data Sampling):

The systematic generation and collection of many evaluations are required to perform the sensitivity analysis. The Sobol sequence is a quasi-random low-discrepancy sequence used to generate uniform samples of parameter space.

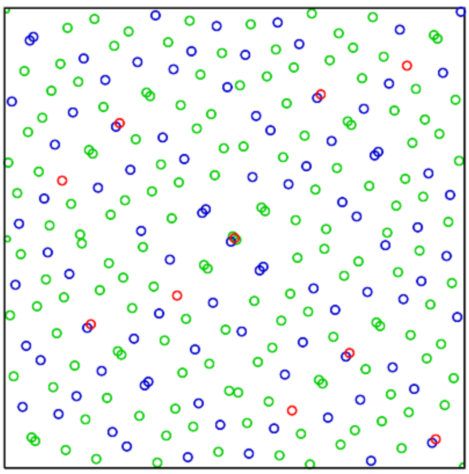


Figure 4-2: Example visualisation of Sobol sequence sampling using Saltelli's extension in 2 dimensions

Saltelli's scheme extends the Sobol sequence in a way to reduce the error rates in the resulting sensitivity index calculations. In order to confidently evaluate the second-order interaction terms, the resulting sampling size should have N * (2D + 2) points. Where N is the initial population size (typically set to values of base 2), and D is the number of parameters in the sensitivity study. Through conventional experience, sample sizes of 5,000 – 10,000 points are usually enough to produce suitable results. However, this was doubled due to the YETI tool's low computational cost.

Thus,

- D = 11 parameters
- N = 1024 initial sample

This resulted in a total sample size of 24576 x 3 (speed-power) = 73728 total evaluations. This approach was applied using the Python package SALib [4]. Once the design space has been generated, the corresponding model is evaluated for each sample. In the context of model sensitivity, the input parameter combinations may not exhibit physically genuine relationships. This is not a concern using this methodology as this approach is attempting to quantify the model robustness by understanding both the individual and interaction of individual variables on the final output, eco-scores.

4.3.2 Case evaluation: Initial Results:

Using the methodology described in the Section 4.2, the Sobol indices were evaluated and compared for each corresponding parameter and speed profile. The results are visualised in the format of stacked bar plots in Figure 4-3. Here, The first-order Sobol indices $(S_1\text{: Blue bar})$ represent the variance attributable to each individual parameter. The higher-order Sobol indices $(S_{ij}\text{: Orange bar})$ measure interactions between parameters. The stack of these columns represents the total Sobol index (S_{total}), which quantifies the total variance in the output explained by all parameters. Therefore, the larger the combined columns, the more sensitive the output is to the associated input parameter.

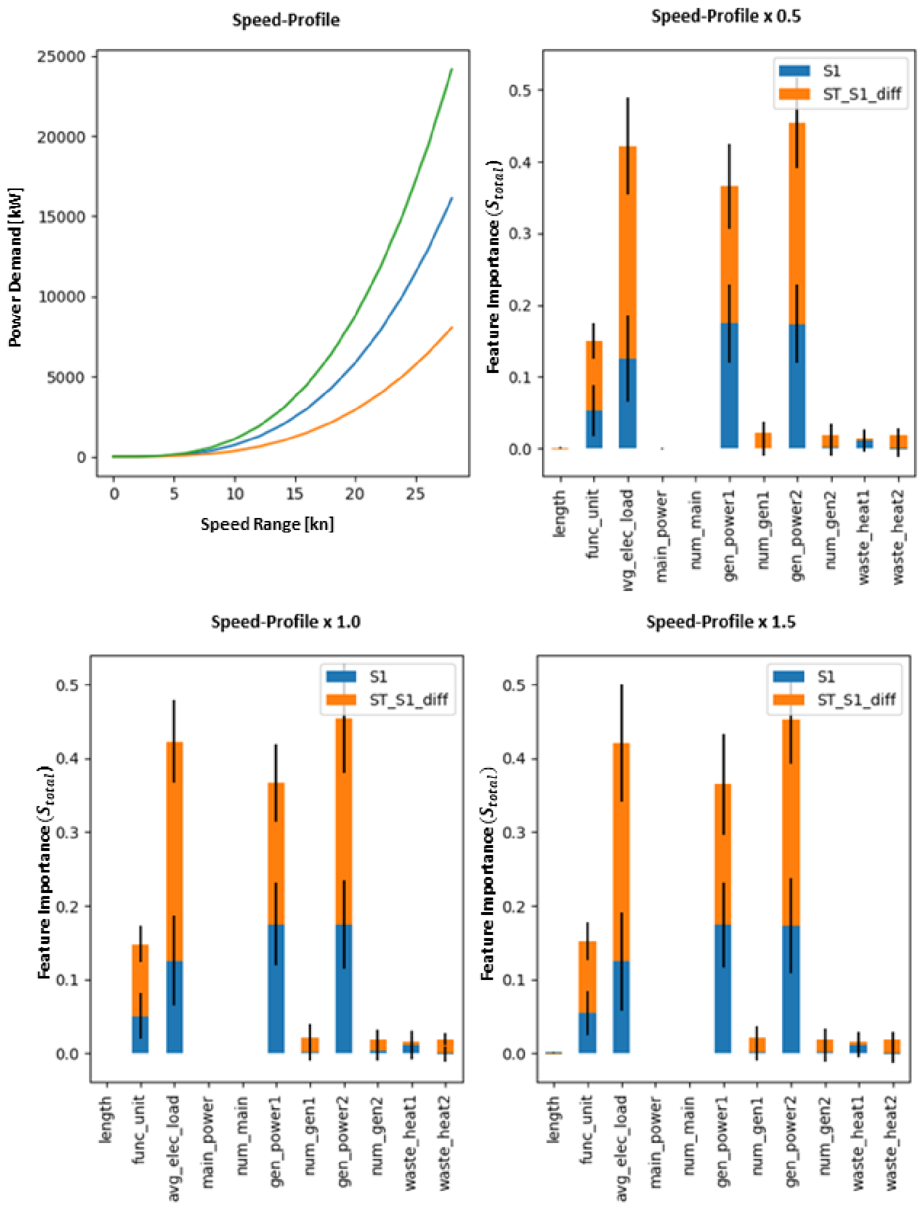


Figure 4-3: Sobol Global Sensitivity (S_{total}) evaluation for 3 speed-profile cases – Top-right: SP0.5 Bottom-left: SP1.0 and Bottom-right: SP1.5

Based on the results, we can visually observe that four parameters are the most impactful parameters:

- Func. Unit: GT (func_unit)
- Average Electric Load (avg_elec_load)
- Generator Power 1 (gen_power1)
- Generator Power 2 (gen_power2)

In this case, changing the speed profile does not change the results for the three different profiles. Ultimately, these initial results yielded a few “yellow” flags that required further investigation.

- Why is the length parameter not very impactful?
- Why Is the effect of the speed profile not very impactful?
- Why do the generator inputs show greater importance than the main installed power?

To understand these questions, additional individual analysis was conducted by looking further into the specific parameters. Ultimately, it can be observed that some of the evaluated results massively inflate the eco-score metric. This phenomenon can be observed in the following figures:

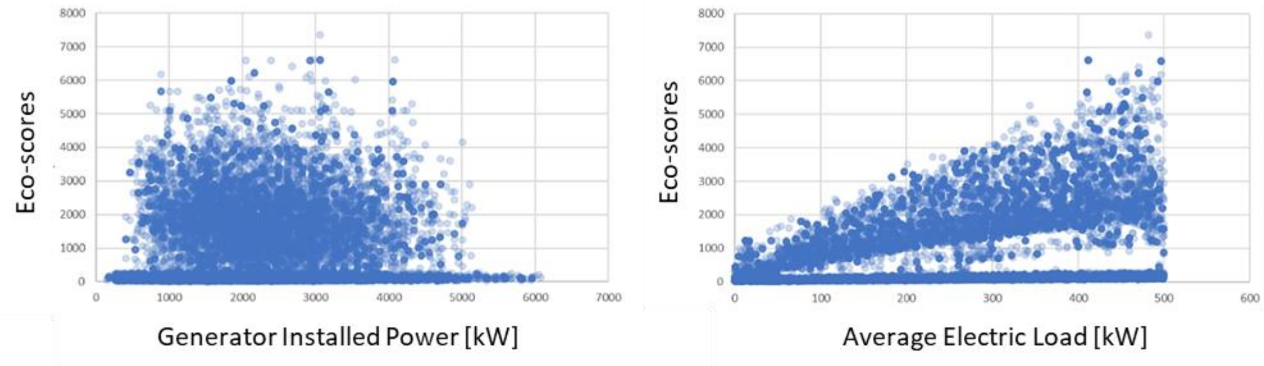


Figure 4-4: *Total generator power (left) and average electric load (right) to the eco-score metric visualisation*

Ultimately, eco-scores should fall within the realistic bounds of ~50 – 200 (based on the extreme bounds of the existing fleet dataset). Therefore, seeing eco-scores in the 1000's is extremely alarming. This initial identification does not immediately indicate an error but raises a degree of doubt. Looking further into the data, a single case evaluation was explored, which fell in this extreme range. The example parameters are shown as:

Table 4-2: *Single extracted example from evaluated samples which shows extreme eco-score score*

Length	Func Unit (GT)	Avg. Elec Load	Main Power	Num Main	Gen Power 1	Num Gen 1	Gen Power 2	Num Gen 2	Waste Heat 1	Easte Heat 2	Eco points1
138.994	1206.543	223.877	661.621	3	900	1	575	2	1	0	1311.4

In this case, while not necessarily optimal, the evaluated parameters could be considered a real yacht. To identify the extreme eco-score inflation, each parameter was systematically varied. Ultimately, the source was related to the generator inputs. As the power of the generators varied, an extremely large discontinuity presented itself within the YETI tool. For instance, the resulting Eco-score outputs by changing ONLY the Generator #2 Power parameter is observed in Table 4-3 and Figure 4-5, respectively.

Table 4-3: *Generator power to Eco-score direct comparison*

Power [kW]	Eco-score [-]
65	85.6
99	82.6
150	75.0
200	70.9
325	68.2
345	72.9
498	1384.9
900	83.7

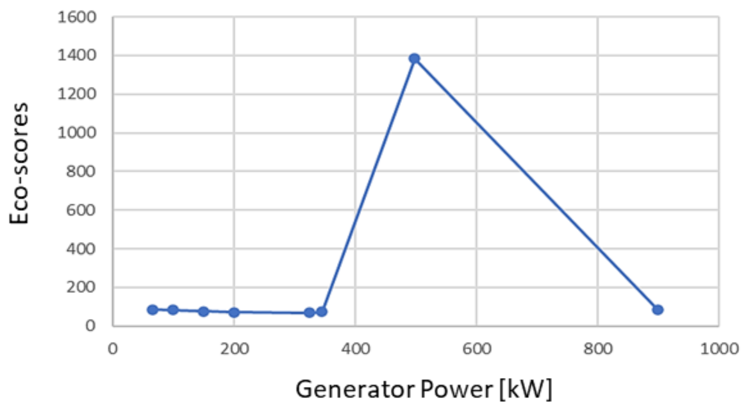


Figure 4-5: *Observed eco-score abnormality as generator power increases*

Therefore, the initial sensitivity study quantified and identified this extreme response as a reaction due to an error in the YETI Tool. The direct cause was associated with an incorrect inputting in the existing engine database. Unfortunately, the current status of the tool does not meet the requirements to fulfill a global sensitivity study for YETI tool insights. Therefore, to properly quantify a sensitivity response, a temporary solution was applied by removing the incorrect engines from the database and re-running the simulation. Of course, it is recommended to perform a dedicated review of the databases to ensure such irregularities are no longer present.

4.3.3 Case Evaluation: Second Iteration

Having fixed the corresponding database irregularity, using the same approach previously demonstrated, the Sobol indices were again evaluated and compared for each corresponding parameter and speed profile. The results are visualised in stacked bar plots for the total sensitivity and interaction terms.

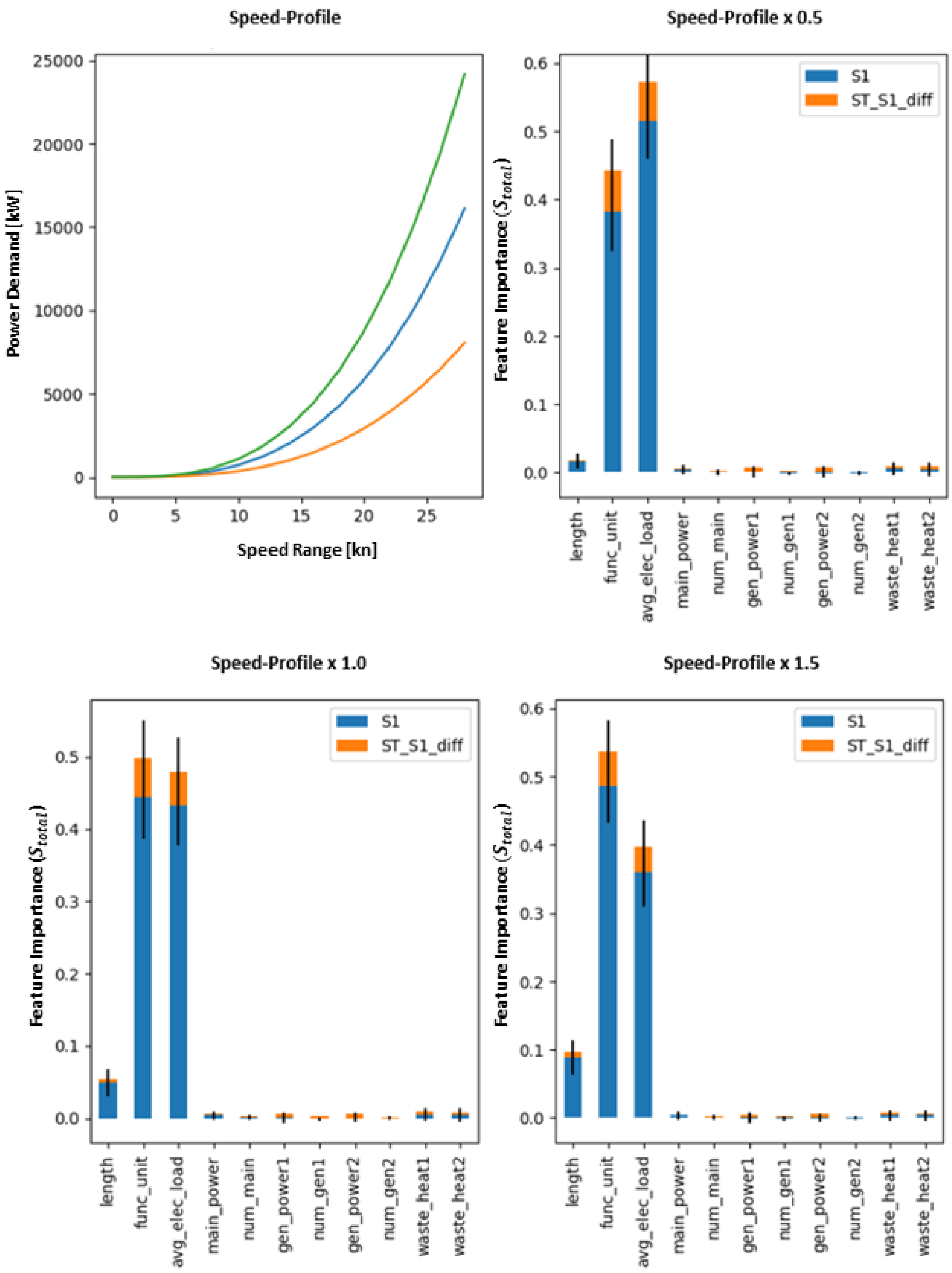


Figure 4-6: Sobol Global Sensitivity (S_{total}) evaluation 2nd for 3 speed-profile cases -Top-right: 0.5 Bottom-left: 1.0 and Bottom-right: 1.5

Based on the results, we can visually observe that now three parameters are the most impactful parameters.

- Length (length)
- Func. Unit: GT (func_unit)
- Average Electric Load (avg_elec_load)

In this case, specifically, it can be observed that approximately 90% of the total modelling eco-score variance is due to only GT and Average electric load. Additionally, it can be observed that while there are interaction effects, first-order effects dominate. This means that the impact of changing multiple parameters at time has marginal consequences on the overall eco-scores. Nevertheless, if we look into the interaction terms, it can be seen that the parameter pair that exhibits the biggest contribution to the output variance is GT and Average Electric Load. This finding enforces the overall importance of these two parameters.

The following general observations can be additionally seen from the corresponding results:

1. As the speed-power profile increases (power demand increases), Length's importance increases. This effect currently appears linear across the varying profiles.
2. As the speed-power profile increases (power demand increases), GT's importance increases. The corresponding trend shows some form of convergence as the 2nd (x1.0) and 3rd (x1.5) speed-power cases don't deviate much.
3. As the power profile increases (power demand increases), the Average Electric Load's importance decreases. Much like the Length case, an approximate linear trending is observed.

Based on these observations, it can also be concluded the speed profile additionally provides a great deal of importance to the output eco-score. This idea can be further confirmed by investigating the evaluated correlations between all parameters for each speed profile. A summary of these visualisations can be seen in Figure 4-7.

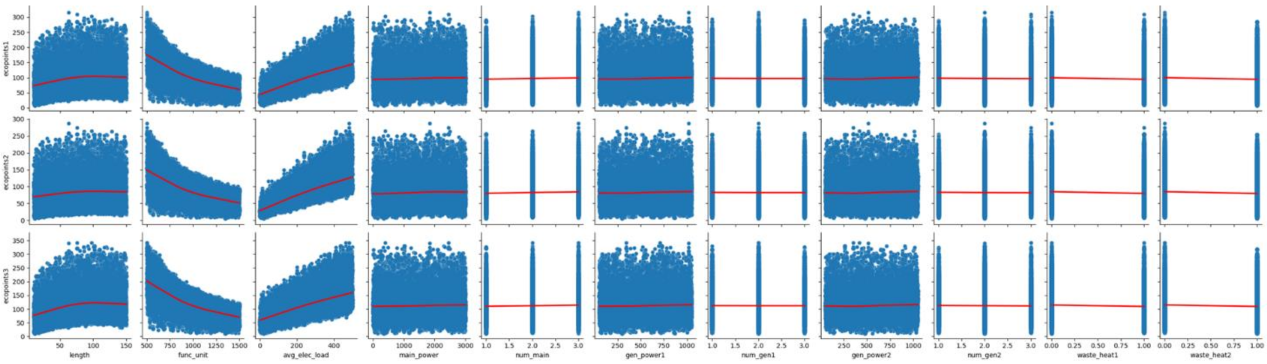


Figure 4-7: Pair-wise correlation plots of all parameters for three speed-power profiles

The 11 columns represent each of the input parameters. Each row represents the associated speed profile in this case, where 1, 2 and 3 are SP0.5, SP1.0, and SP1.5, respectively. Visually, it can be seen that only the first three parameters exhibit any clear trending (i.e. trends of the last 8 columns are very nearly horizontal). This visual confirmation is no coincidence, as these are the identified importance parameters found via the Sobol Sensitivity approach. Looking further at these parameters, all three speed-power profiles can be directly compared. The resulting comparisons between the two most impactful parameters, GT and Average Electric Load and their observed linear trends are shown below.

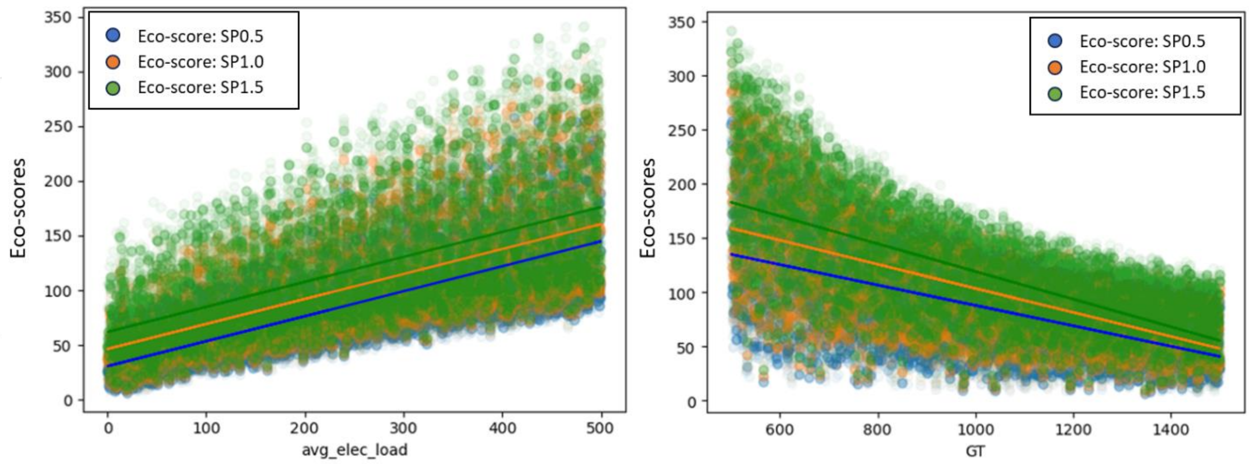


Figure 4-8: Trend observations for varying speed-power profiles for Average Electric Load (left) and GT (right)

Visually, it can be confirmed that changing the speed profile can impact the overall magnitude of the eco-score metrics. A nearly constant offset is seen in the Average Electric load case. However, when considering GT, the larger cases demonstrate a marginal impact due to power demand. However, as the GT decreases, the importance of the speed-power profile becomes much more impactful. These findings are truly interesting because even though a quasi-random sampling is performed, underlying modelling trends are clearly found.

4.4 Sensitivity Analysis Initial Conclusions

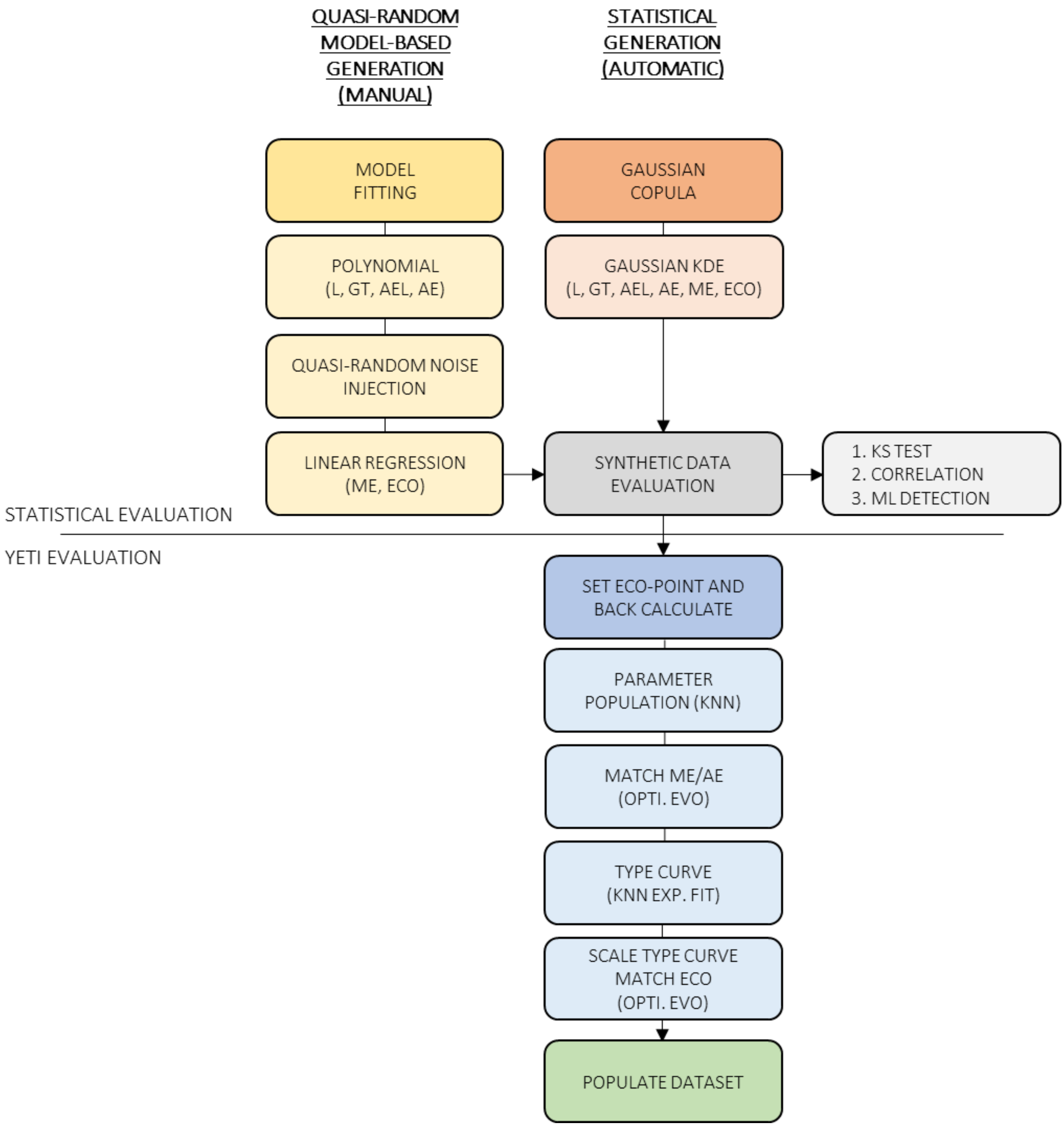
The initial conclusions and key takeaways are as follows:

1. The sensitivity analysis demonstrated its immediate usefulness by identifying a critical bug in the generator database that would otherwise have been difficult to identify manually.
2. Of the 11 parameters, 3 were identified as the most impactful to eco-score. These include Length, GT, Average Electric Load.
3. It is interesting to note that main engines, generators, and associated parameters have a marginal relative impact on the eco-score metric.
4. Of these 3 parameters, 2 constitute approximately 90% of the total variance. These include GT and Avg. Electric Load.
5. Even with a quasi-random sampling approach, these 2 parameters demonstrate strong and unique underlying trends. The impact of the speed-power profile also demonstrated strong impacts on eco-scores. In general, the increasing load presented an increase in the eco-score metric. However, the most interesting trend observation is that as the GT increases, the global impact of the speed-power profile becomes negligible. Thus, irrespective of the individual speed-power profile, similar eco-scores are observed.

While many of the outcomes may seem logical from an engineering perspective. These intuitions are generally experience-based and challenging to quantify without the need to “take my word on it.” This study has now provided clear evidence of the parameter impacts on the YETI eco-score metric in a concrete and mathematical format.

5 DATA POINT (VESSEL) GENERATION:

Two methods are explored in the general approach to generating realistic (statistically reliable) data points: Manual and Automatic approaches. These are further described in the overall methodology and workflow as follows:



The workflow consists of two different evaluation phases. The statistical evaluation contains the workflow to generate new synthetic data points. However, in order to generate a completely new vessel data point, a YETI tool evaluation is required to back-calculate and fully populate the new synthetic dataset using a mixture of real data and optimisation processes. The existing dataset contains a mixture of ship types but is predominantly displacement vessels. We can see that extreme outliers include high-speed yachts and sailing yachts. Semi-displacement yachts are also contained more often.

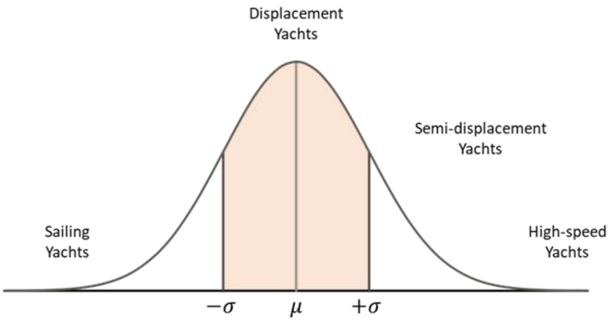


Figure 5-1: General sample distribution when considering the vessel types

Therefore, throughout the workflow, the following assumptions are made:

1. Predominantly displacement-type only yachts: The type of vessels are based on approximately exponential speed-power curves. Semi-displacement characteristics may be semi-captured, but not guaranteed because these are not explicitly considered when describing the speed-power curves.
2. Sailing Yachts excluded: Due the extremely small sample size of sailing yachts within the existing dataset these points are neglected. A similar methodology could be used to account for these specific vessel types only when sufficient data points are collected.
3. Fully electric excluded: Currently only one data sample contains a fully electric solution. As indicated with the sailing vessels, more data is required to make statistically viable approximations.

5.1 Quasi-random model-based generation (QMG):

The first synthetic tabular point generation method involves a manual statistical curve fitting approach. More formally, the method is known as, Quasi-random model-based generation. The approach considers and accounts for the uncertainty and uniqueness in the dataset via selective noise injections. This randomly generated noise is imposed on the ideally fitted curve in order to capture the real-world uncertainties exhibited in the original dataset.

The noise parameter bounds is considered via the dataset's local standard deviation observed in the true fleet dataset. In other words, the established noise variance is determined via the nearest neighbouring points within a certain binning region. The initial bin region was set at $\pm 10\%$; however, manual tuning was conducted via curve inspection. Additional research can be conducted to determine the most optimal or realistic noise bandwidths which should be applied. A schematic indicating how the observed variance changes across a data set can be seen in Figure 5-2.

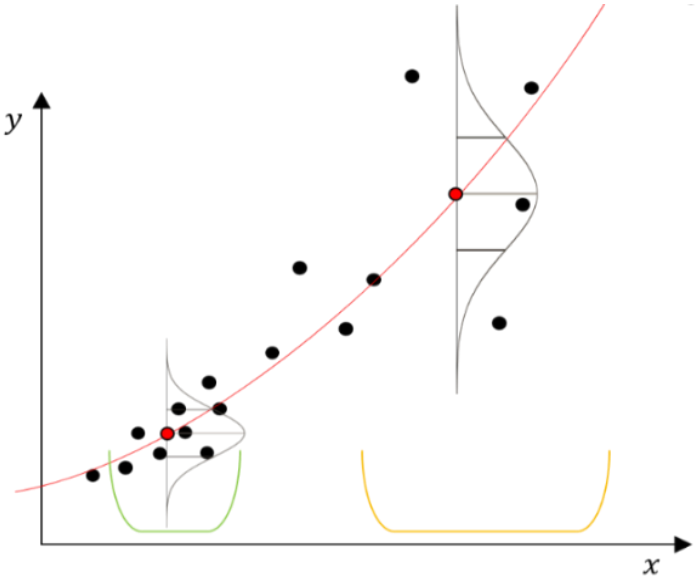


Figure 5-2: Schematic of varying data variance across the parameter regions

For each point along the fitted curve an observed natural variance in the dataset can be observed. Thus, by quantifying this randomness across the dataset, we establish the quasi-random characteristic.

Using the information obtained in the exploratory data analysis, an initial model fitting for three parameters using Length as the governing geometrical parameter was conducted. The corresponding ideal fit and injected noise (70 new points), can be observed in Figure 5-3.

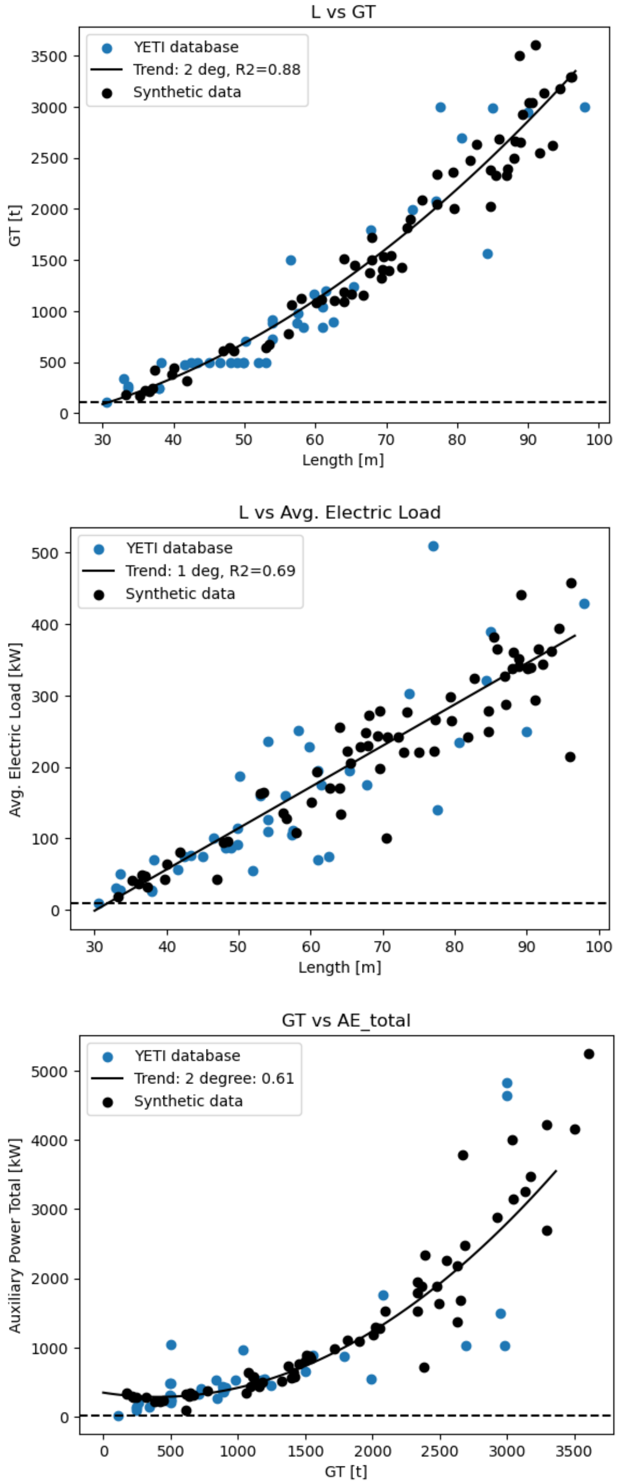
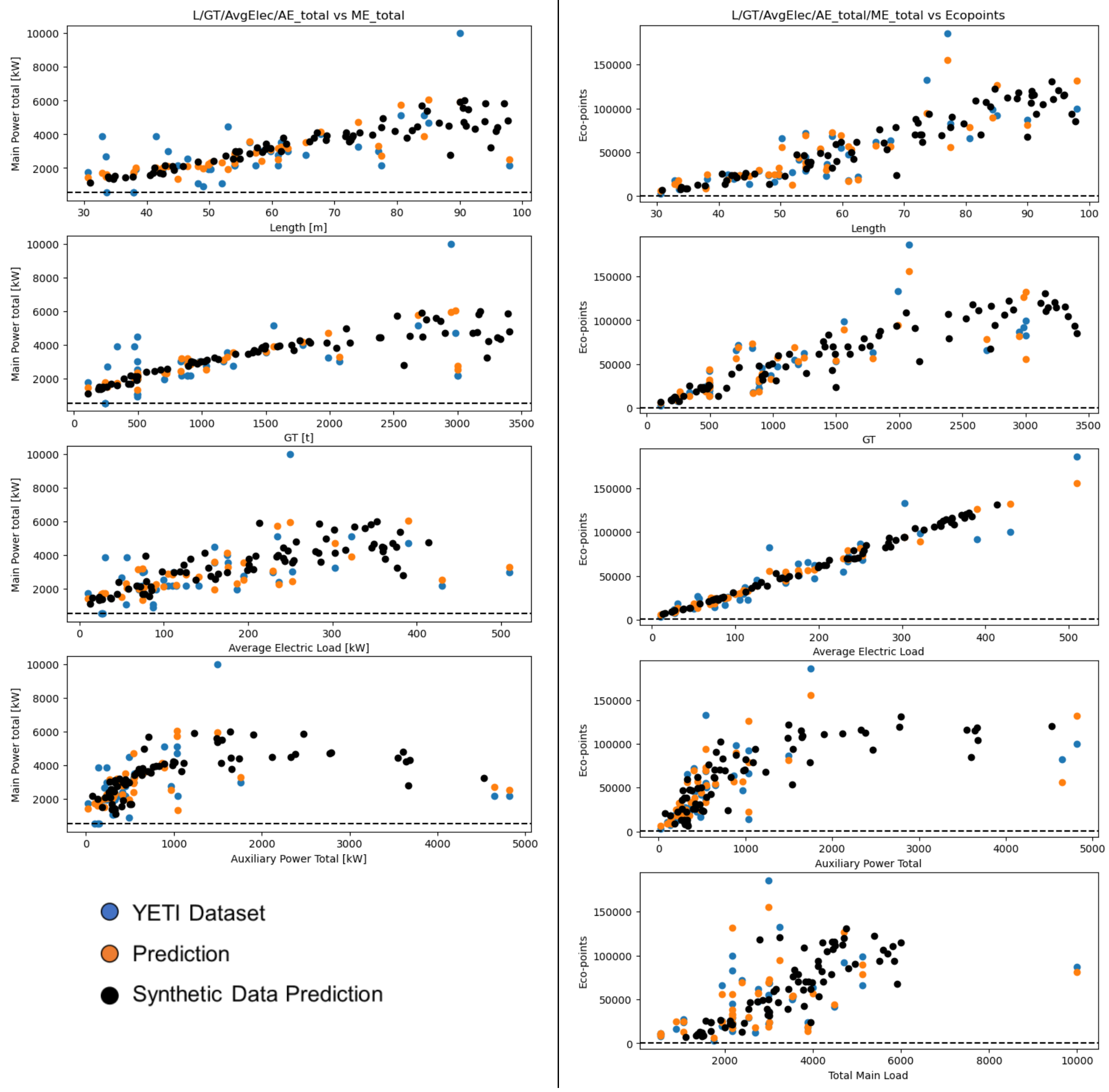


Figure 5-3: Fitted trends and associated noise injections for (top) L vs GT, (middle) L vs AEL, and (bottom) GT vs AET

In this case, four parameters are initially described using this method: Length, Gross Tonnage, Average Electric Load, and Total Installed Auxiliary Load. It can be observed that a generally good fitting can be approximated using conventional polynomial expressions. Furthermore, based purely on visual inspection, the degree of injected noise seems appropriate across all explored parameters. Thus, allowing for a “realistic” generated sample. The next step is to complete the synthetic approximation by estimating the total installed main engine power and the observed Eco-points. To achieve this objective, a basic multiple linear regression methodology was applied to individually predict these parameters from the newly generated samples.

Total installed main engine power ($R^2 = 0.52$)	**Eco-points** ($R^2 = 0.88$)



Based on the results it can be observed that the exact prediction of the total installed propulsion power is much more challenging than the Eco-points metric. This is likely due to non-linear interactions, thus a linear model is not sufficient to fully capture all extremities. Nevertheless, it can be visually observed that while we are not able to obtain perfect predictions for some of the extreme points (orange), the cluster of predictions actually fall neatly within the extreme bandwidths of the actual dataset (blue). It is also noticed that the induced randomness is inherently injected and transferred within the predicted metrics (black). This helps to further visualise the like trends in regions with sparse data. While the total installed main power is challenging, the opposite is true for the Eco-points metric. In this case we can very accurately quantify this metric using a small subset of variables. This is mainly attributed to the fact that a near linear relationship is exhibited between the Eco-points metric and the average electric load. While the prediction are not perfect, they don’t have to be as the induced noise presents a natural degree of randomness that is smartly captured and retained in the synthetic data. This fact is further enforced through a direct visual comparison which shows all resulting distributions fall within a likely and logical location.

5.2 Gaussian Copula Generation (GC):

Gaussian Copula is a statistical modelling technique for automatic data synthesis. They serve as a powerful tool for modelling dependencies among multiple variables. Its fundamental principle lies in the separation of the marginal distributions from the joint distribution, allowing for the representation of complex interdependencies while assuming Gaussian (normal) marginal distributions for individual variables. This separation enables users to define a correlation structure through a correlation matrix, describing the strength and direction of relationships between variables. By simulating data from this joint distribution, Gaussian copulas generate synthetic datasets that preserve the specified dependencies. This approach provides flexibility in modelling various types of dependencies, accommodating linear and nonlinear relationships.

However, its reliance on assuming Gaussian distributions for marginal variables can be a limitation. Real-world data often deviates from perfect normality, potentially leading to inaccuracies in the generated synthetic datasets. Additionally, Gaussian copulas may not adequately capture extreme events or tail dependencies, as they assume a constant relationship structure across all quantiles, which may not align with actual observations in highly volatile scenarios such as financial crashes. The open-source SDV python package [5] is used to create, generate, and evaluate the automatic synthetic data method.

5.3 Performance Evaluation and Comparison

To determine which approach is the most suitable method to populate the dataset, a comparison between the QMG and the GC is performed using three commonly applied performance metrics when evaluating synthetic data.

1. KS test (Quality metric): One way to evaluate the quality of synthetic data is to use the KS metric, which stands for Kolmogorov-Smirnov metric. This metric measures the distance between the cumulative distribution functions (CDFs) of the real and synthetic data for each variable. The smaller the distance, the more similar the distributions are. Most often, the KS compliment is used (1 – KS), Thus, a score of 1 indicates 100% replication.

2. Correlation comparison: Correlations between two or more columns are extremely important for ML applications, which help uncover relationships between features and the target variable and help create a well-trained model. If the synthetic data and real data demonstrate similar correlations across all features, the underlaying patterns have been successfully captured and replicated.

3. ML Detection analysis: This is a detection analysis method in which a ML model is trained on a mixture of both the synthetic and real data. The purpose is then to identify how hard it is to distinguish between real and synthetic instances with flags indicating whether the data is real or synthetic. The easier it is to predict the flag, the more distinguishable between real and synthetic data. If the model with less than or equal to 50% accuracy, indicates that the ML cannot clearly distinguish between the two thus indicating a close match between real and synthetic data points.

5.3.1 KS Quality Test

Gaussian Copula (Automatic)	Quasi-Random Model-Based (Manual)
Overall KS Quality Score**: 88.76%** SUCCESS: ✓ The synthetic data covers over 90% of the numerical ranges present in the real data ✓ The synthetic data follows over 90% of the min/max boundaries set by the real data ✓ Over 90% of the synthetic rows are not copies of the real data	Overall KS Quality Score: **74.06%** SUCCESS: ✓ The synthetic data follows over 90% of the min/max boundaries set by the real data ✓ Over 90% of the synthetic rows are not copies of the real data WARNING: ! The synthetic data is missing more than 10% of the numerical ranges present in the real data

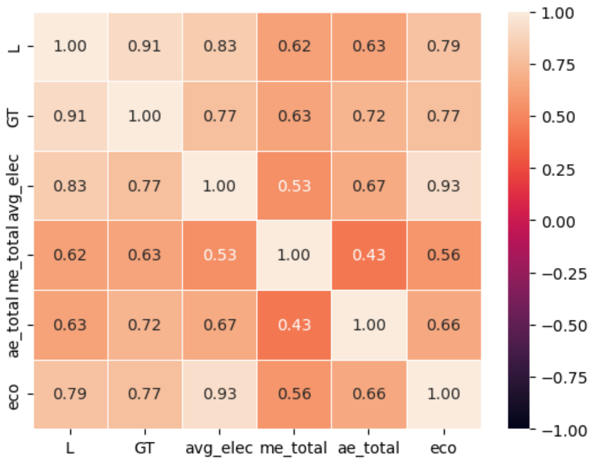
Based on the KS quality analysis, it is observed that the Gaussian copula performs much better than the QMG approach. There is even an indicated warning that the numerical ranges are not fully covered.

5.3.2 Correlation Analysis:

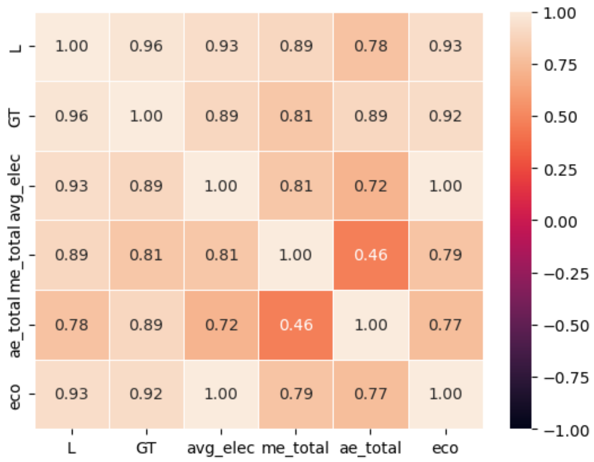
A correlation analysis is conducted by comparing the real dataset with both synthetic approaches. In this case a direct visual comparison can be perfomred as the bounds are maintained across all comparisons. The resulting heatmaps can be seen as follows:

REAL DATA CORRELATIONS

Gaussian Copula (Automatic)



Quasi-random model-based (Manual)



Based on the analysis, it can be seen that, again, the Gaussian Copula seemingly captures the correlations between the pair-wise parameter comparisons more clearly. To quantify the total degree difference between the real and synthetic datasets, a mean average per cent error (MAPE) metric is used to compare both approaches.

- **Quasi-random model-based (manual) MAPE = 51%**
- **Gaussian Copula (automatic) MAPE = 21%**

It can be seen that in a global correlation comparision, the Gaussian Copula captures approximately 50% more correlations than the Quasi-random model-based prediction approach. It is interesting to note that the Total installed auxiliary power versus the Total installed main engine power is challenging to capture in both cases. Nevertheless, all other parameters seem to exhibit generally similar trends.

5.3.3 ML Detection Approach:

The final comparison method applies a ML detection approach. In this case a vanilla (default) classifier ensemble model is trained and evaluated using a hold-out dataset made-up of synthetic data from both methodologies. In this case a total of 500 random evaluations were conducted. The training procedure is as follows:

1. Trained: 42 Real points and 42 random synthetic points (QMG & GC)
2. Tested: 30 generated synthetic data points (QMG & GC)
3. Repeat: Sample new 72 synthetic data points and repeat evaluation model accuracy

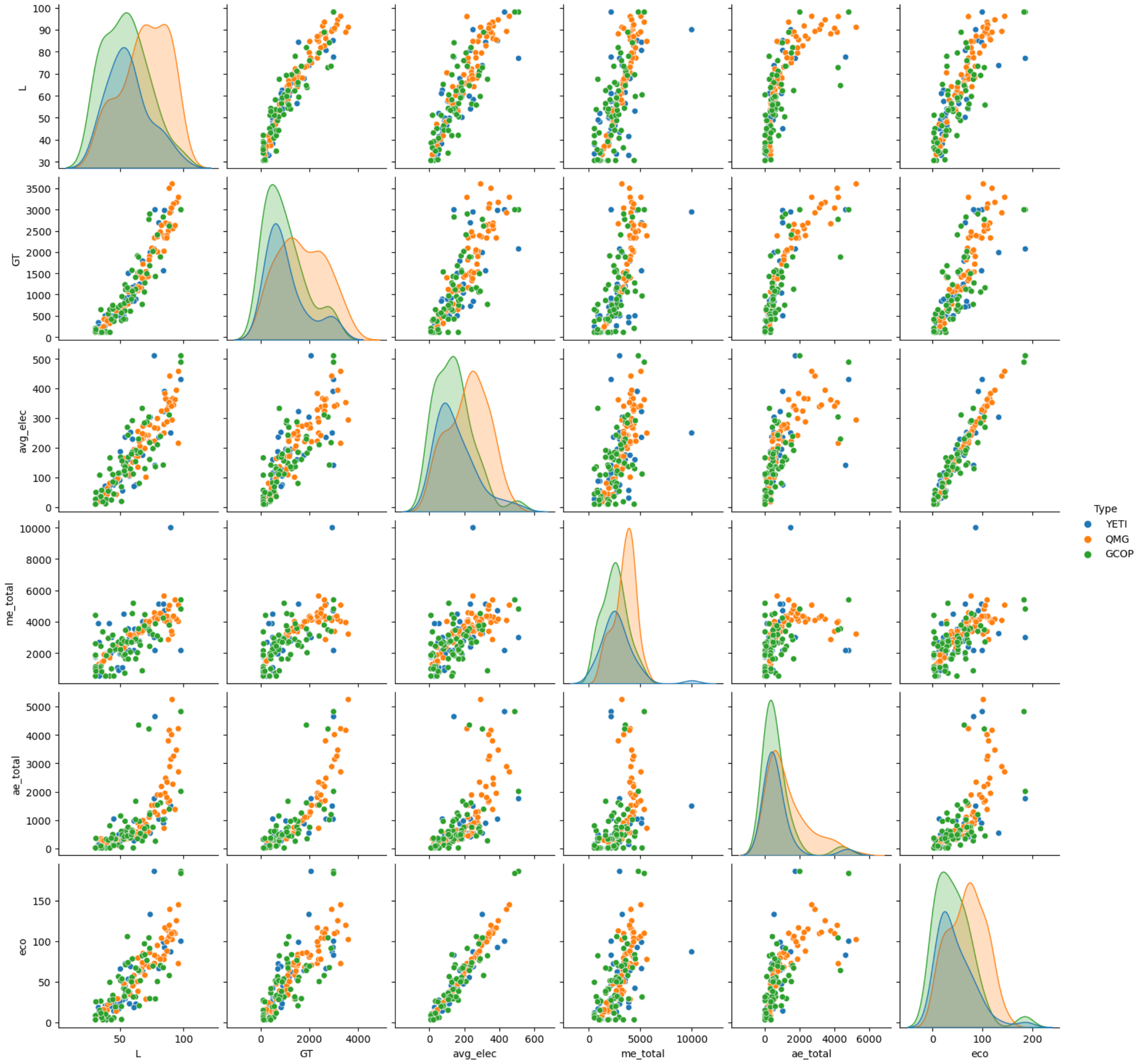
The corresponding results after 500 evaluations are:

- **Quasi-random model-based (manual): 75.9%**
- **Gaussian Copula (Automatic): 66.3%**

Over the course of 500 random evaluations the model could predict 76% of the time correctly on the QMG data generator. Whereas, for the GC, the model was only able to predict 66%. Thus, the GC was able to produce more realistic vessels as compared to the QMG and fool the model 34% as compared to only 23%. Thus indicating again that the Gaussian Copula presents a high degree of realistic data points as compared to the manual approach.

5.4 Synthetic Data Visual Comparison

A direct visual inspection of the three data sources: YETI Fleet dataset, QMG, and GC can be seen below:



Based on the generated data points visualisation, a few interesting observations can be found. The automatic approach (Gaussian Copula) seemingly replicates the data distributions very well. Due to the inherent statistical sampling, the selected points are distributed much like the existing dataset. Fortunately, the does seem to be a broad distribution but extreme points are limited. Looking at the manual points, the distributions do not following the underlying distributions. This is due to the fact that only trends/relationships were used to develop the underlying predictions. Thus, a broader spread is possible. There are pros and cons of both these approaches;

- Pro Automatic: The data is very realistic and follows the inherent data distributions very closely, thus expanding the existing database in a very realistic manner.
- Con Automatic: Completely sparse regions are not likely to be abundantly populated as the method considers the current dataset distributions as ground truth. Therefore, distributions are not likely to contain much information in sparse regions.

- Pro Manual: The generated data points can naturally populate the sparse regions of the existing dataset as an regression formulation is developed for each individual input parameter.
- Con Manual: It is unclear if the generated data truly follows the underlying distribution of the real dataset.

It should be noted that since the automatic approach is a sampling method based on the generated probability distribution functions – it is possible to conditionally sample sparse regions in the dataset. In other words, if I would like to generate more points in the upper length regions or upper average electric loading conditions, this is possible by fixing the parameters and sub-sampling from the observed partial distributions.

6 POINT-WISE VESSEL ESTIMATION

While we now have a method to generate reliable synthetic data points, a follow-up approach is required to convert this information into full vessel input data points. As indicated in Chapter 5, the second phase of the methodology can be conducted; the YETI Re-evaluation. This portion can be decomposed into six critical steps:

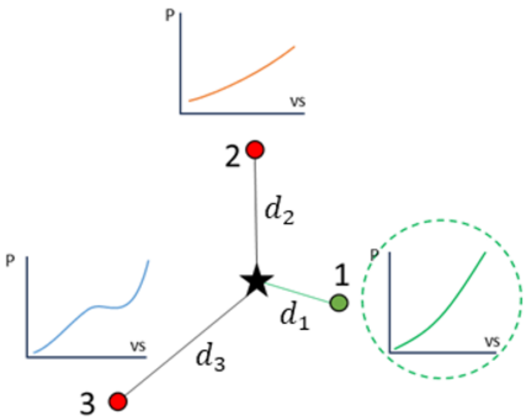
1) Set synthetic Eco-points as the target parameter.
2) Populate remaining unknown parameters using a nearest neighbour approach.
3) Match the synthetic total auxiliary and main engine power with existing engines from the YETI database.
4) Establish the appropriate speed-power curve using nearest neighbour and curve fitting methodology.
5) Scale the associated speed-power curve to match the target synthetic eco-points score.
6) Populate the existing synthetic dataset and combine with the real data set.

6.1 Methodology Challenges and Assumptions:

The largest difficulty in the proposed approach is obtaining realistic speed-power curves that are unique to the associated vessel type. The reason for this is because there only exists a few geometrical parameters which is not nearly enough inputs for most low-fidelity power prediction approaches. Thus a solution is required to account for the natural variation in the existing fleet data base. Assuming that similar geometrical vessels (Length and GT) have similar speed-power profiles, the nearest neighbouring vessels characteristics can be used to establish and retain baseline information such as:

- Speed-power type curve
- Number of engines
- Avg. hotel heat load
- Bow thruster
- Shore power
- Etc.

The nearest neighbour approach uses an Euclidean distance metric to establish which points in a N-dimensional space is nearest to the point of interest (synthetic data point). For the purposes of this study, only the closest neighbouring point is selected but in actuality k amount of points can be used in conjunction with weighted averaging based on the associated distances.



$$1\text{NN} = \min D; \quad D \in \mathbb{R}: \{d_1, d_2, .. d_n\}$$

Figure 6-1: K-nearest neighbour schematic where K=1

Once the closest neighbouring vessel speed-power is extracted a type curve is predicted using an exponential curve fitting approach ($y = a * x^b$). In this case the coefficients a and b are optimised for each nearest neighbouring speed-power curve thus allowing for a unique representation of for each vessel type. Figure 6-2, demonstrates an example of both the fitted and generated datapoints.

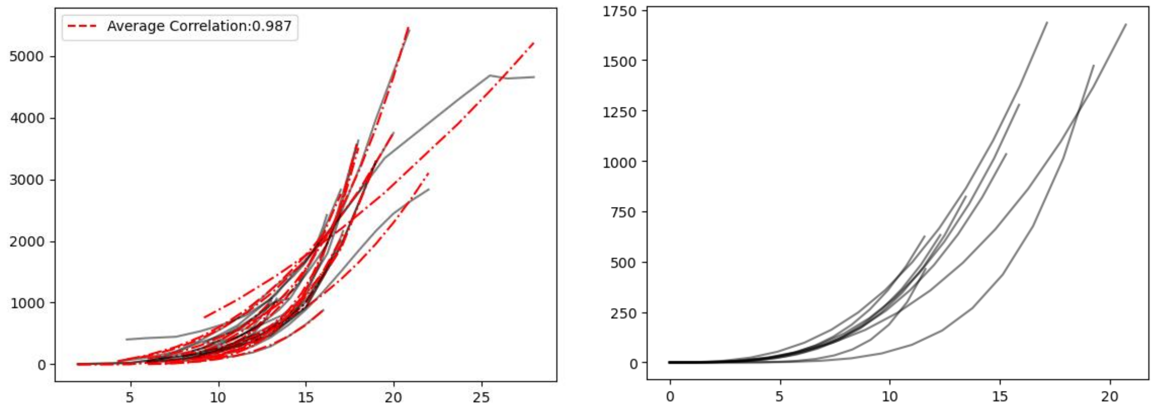
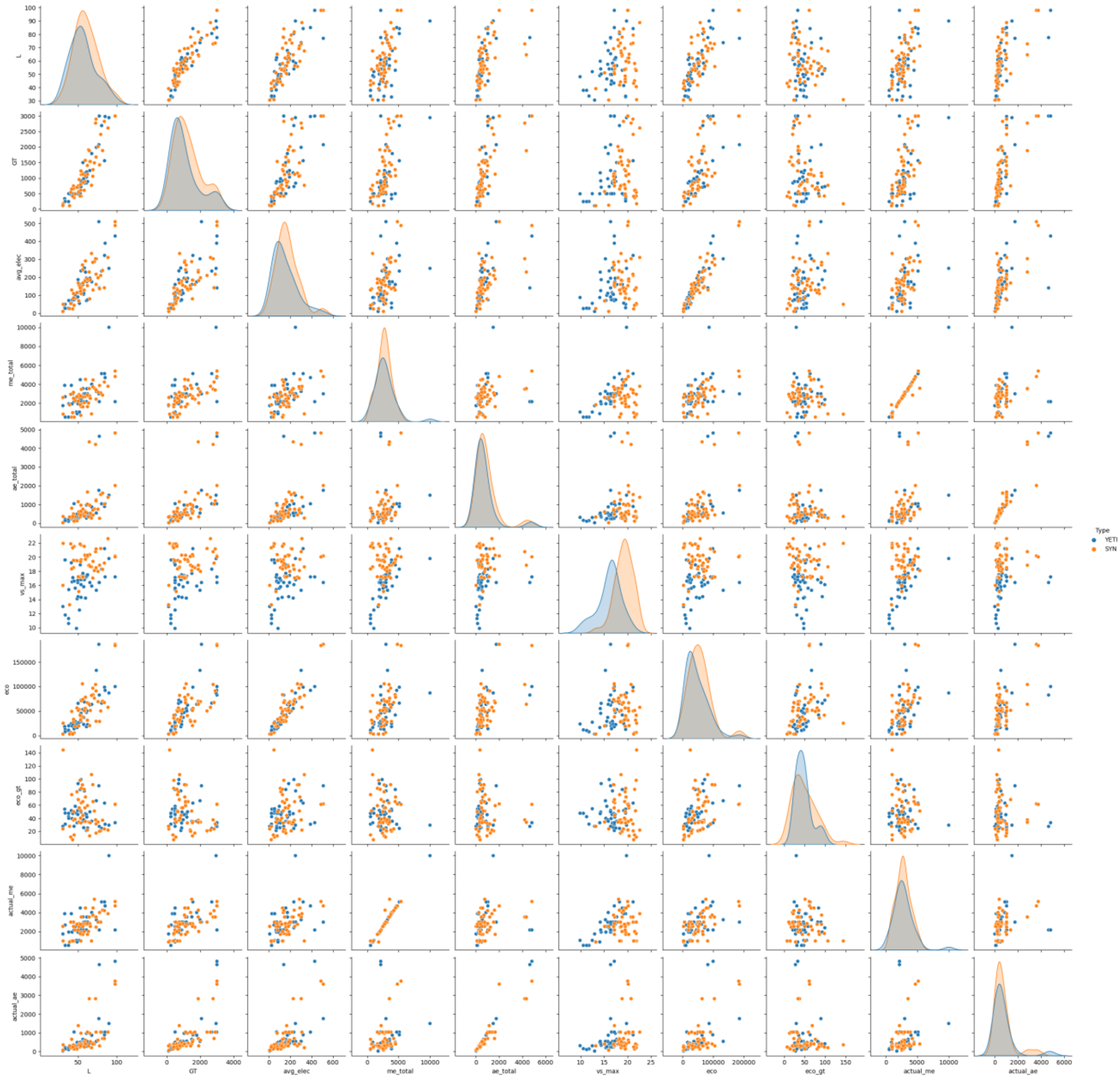


Figure 6-2: (left) Exponential curve-fitting on a subset of real speed-power points (right) Generated speed-profiles

In this case, we clearly see the uniqueness inherently found in actual speed-power curves is equally generated in the synthetic approximation. Additionally, we not only capture the inherent uniqueness but the overall fitting accuracy of the entire dataset exhibits an average $R^2 = 0.99$.

The above logic assumes that the speed-power profile exhibits a linear trend on the overall eco-score evaluation. This assumption was extracted from the general sensitivity study, which seemed to indicate such a trend. However, there are interesting cases where non-linear impacts are observed. Thus, when globally scaling the type curve to match the eco-points metric, the optimiser finds alternative extreme minima that greatly reduce the speed-power curve, thus increasing the top achievable speeds well beyond the expected ranges (as observed from the initial fleet database and personal intuition). Thus, optimisation constraints were applied to counteract these strange phenomena to maintain a realistic speed range. The corresponding upper search bound is set as $V_{s,max}$ + 3 knots using the most similar vessel as the governing baseline.

6.2 Final Results Comparison:



6.3 Conclusions for Realistic Vessel Automatic Generation

The methodology for generating synthetic fleet points is developed and compared. The initial conclusions and key takeaways are as follows:

1) Quasi-random model-based generation provides a transparent and modular framework which can be easily understood and modified but currently lacks capturing the complex parameter interactions.
2) Gaussian Copula presents more realistic data and better captures the complex nature of most parameters but is not easily understood (black-box approach).
3) Realistic speed-power profiles can be generated for each sample using exponential curve-fitting and nearest neighbour approach on the actual dataset. Thus incorporating physical world knowledge into the synthetic data points.
4) Unfortunately, the matching approach using a scaled type curve can find multiple minima within the YETI tool. Consequently, this phenomenon typically favours reducing the speed-power curves

to achieve a higher top speed. Therefore, a practical upper bound was set to avoid solutions that are well beyond the existing fleet database.

Using the proposed synthetic data generation approaches, sparse data regions can now be reliably populated to observe governing trends on the YETI fleet dataset to help aid in important YETI metric decision-making.

Future steps:

1) **Add more variables to the generation model:** Introduce the top speed as an input generation variable, aiming to predict the top-speed relationship. Establishing this relationship can serve as an upper boundary constraint, enhancing the accuracy of speed-power profile evaluations and further refining the realism of synthetic data points.
2) **Explore alternative models:** While two different synthetic data generation methodologies were lightly explored, alternative methods exist. For instance, the more complex algorithms, such as Generative Adversarial Networks (GAN), are also commonly used for such problems. Specifically, the CTGAN (Conditional Tabular) approach is a new and novel framework which focuses on structured datasets. Therefore, it would be an interesting inclusion to evaluate how these more complex algorithms deal with such datasets. Additionally, a deeper look into the existing methods is also worthwhile. While they show great promise, there is still some lingering uncertainties.
3) **Further verification of synthetic data:** While a systematic evaluation process has been implemented to assess the feasibility of estimated vessel points, consider an additional layer of evaluation. Engage domain experts to meticulously analyse each generated point, validating the authenticity and feasibility of synthetic data before integration into the existing dataset. Expert analysis will ensure the credibility and reliability of the synthetic data points, strengthening the dataset's overall integrity and usefulness for decision-making processes.

These steps aim to refine the data generation process, leveraging expert insights to ensure the accuracy and relevance of synthetic data points, ultimately enhancing the quality and applicability of the dataset for future analyses and strategic decision-making within the YETI fleet domain.

7 CONCLUSIONS AND RECOMMENDATIONS

The comprehensive assessment and analysis conducted in the report have yielded significant insights into enhancing the understanding and decision-making process regarding the YETI fleet metrics, specifically focussing on the tool sensitivity and existing dataset limitations. The following conclusions summarise the findings of the present project:

Exploratory Data Analysis and Sensitivity Study

- Comparison between Eco-points, Eco-score, and EEDI indicated a high degree of noise and scatter in the Eco-score metric due to the corresponding noise in the normalisation metric, GT.
- Eco-score exhibits similar trends to EEDI and globally penalises smaller yachts, consistent with initial global sensitivity study observations.
- The sensitivity analysis promptly identified a critical bug within the generator database, highlighting the immediate utility of such analyses in detecting intricate issues that might be challenging to discern manually.
- Among the eleven parameters, three parameters - Length, GT, and Avg. Electric Load emerged as the most influential on the eco-point metric. Surprisingly, main engines, generators, and associated parameters exhibited marginal impacts, contrary to initial presumptions.
- Further delving into these three parameters revealed that GT and Avg. Electric Load significantly contributes to approximately 90% of the total variance, showcasing strong and distinctive underlying trends.
- The impact of the speed-power profile on eco-point scores indicated intriguing trends, particularly observing that with increasing GT, the global impact of the speed-power profile diminishes, leading to globally similar eco-point evaluations irrespective of the speed-power profiles.

This study has effectively provided concrete and mathematical evidence regarding parameter impacts on the YETI eco-point metric, dispelling intuitions that were primarily experience-based and challenging to quantify without empirical evidence.

Synthetic Data Generation

- Regarding synthetic data generation, two methodologies were evaluated: quasi-random model-based generation (QMG) and Gaussian Copula (GC). While the former offers transparency and modularity but lacks capturing intricate parameter interactions, the latter presents more realistic data by capturing complex parameter relationships but operates as a black-box approach.
- Additionally, the extraction of realistic speed-power profiles into synthetic data points via exponential curve-fitting and nearest neighbour approaches was achieved. However, challenges were encountered in the matching approach when using the YETI tool as a back-calculator, leading to multiple minima and favouring unrealistic top speeds, necessitating the implementation of practical upper bounds.
- Nevertheless, the methodologies devised in the synthetic data generation phase have enabled a reliable population of sparse data regions, facilitating a deeper understanding of governing trends within the YETI fleet dataset and thereby allowing for further insight into the critical decision-making processes related to YETI metrics.

Collectively, these conclusions underscore the significance of empirical analysis and systematic approaches in unravelling complexities within the YETI fleet domain, providing a robust foundation for future enhancements and strategic decision-making. This comprehensive analysis contributes substantially to the ongoing evolution of YETI fleet management practices, offering actionable insights derived from rigorous analysis and methodologies.

Wageningen, January 2024
MARITIME RESEARCH INSTITUTE NETHERLANDS

Dr. Ir. H.J. Prins
Manager Research & Development

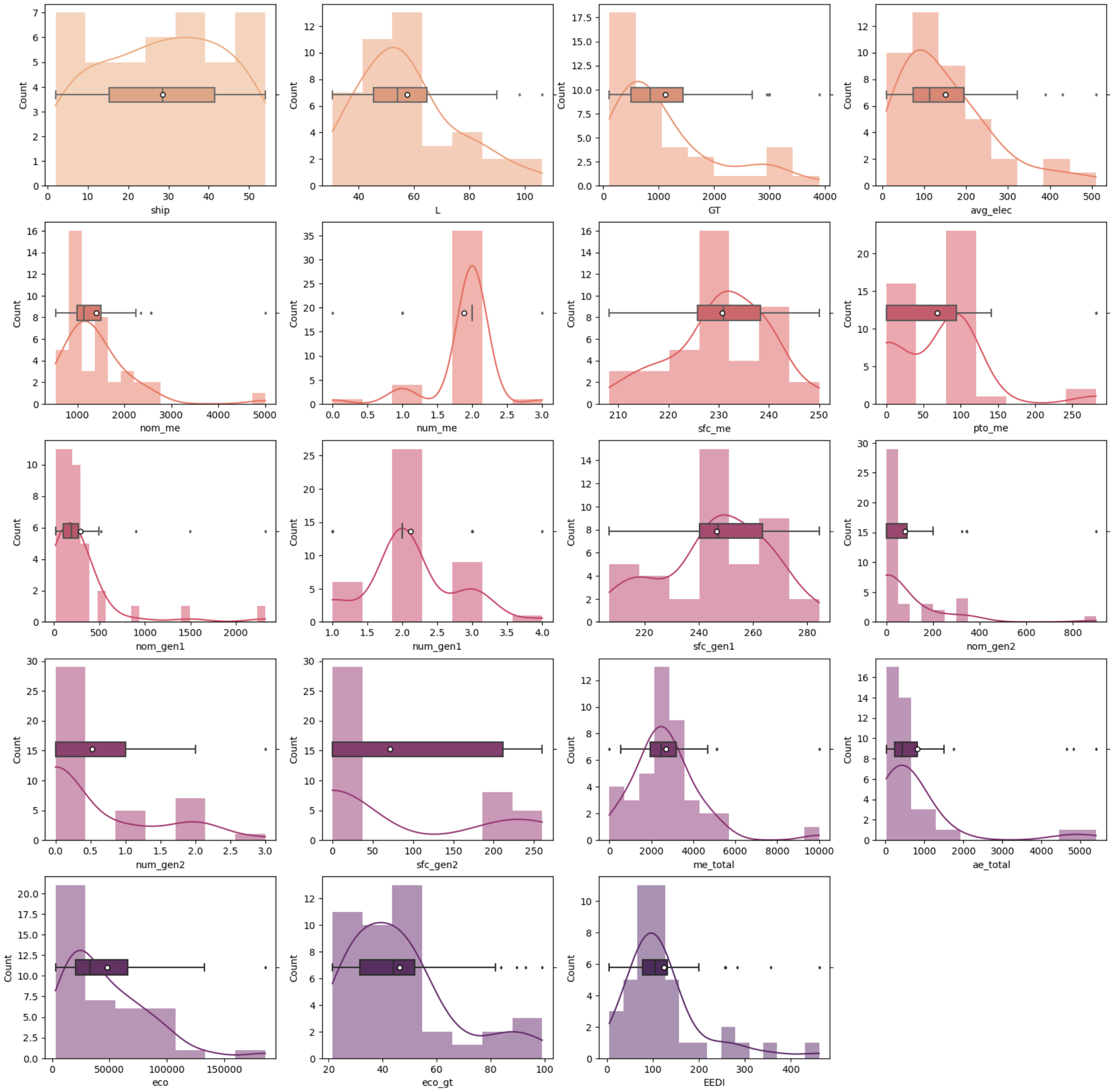
REFERENCES

1. ISO Form 4 – New Work Item Proposal, ISO/NP TS 23099, 2023-06-24
2. Water Revolution Foundation (WRF). url: https://waterrevolutionfoundation.org/
3. International Maritime Organization (IMO), (2022) Improving the energy efficiency of ships. url: https://www.imo.org/en/OurWork/Environment/Pages/Improving%20the%20energy%20efficiency%20of%20ships.aspx
4. Herman, J. and Usher, W. (2017) SALib: An open-source Python library for sensitivity analysis. Journal of Open Source Software, 2(9). doi:10.21105/joss.00097
5. Patki, N., Wedge, R. and Veeramachaneni, K. (2016) The Synthetic Data Vault. IEEE DSAA. doi:10.1109/DSAA.2016.49

APPENDIX A REDUCED YETI DATASET MAIN STATISTICAL PROPERTIES

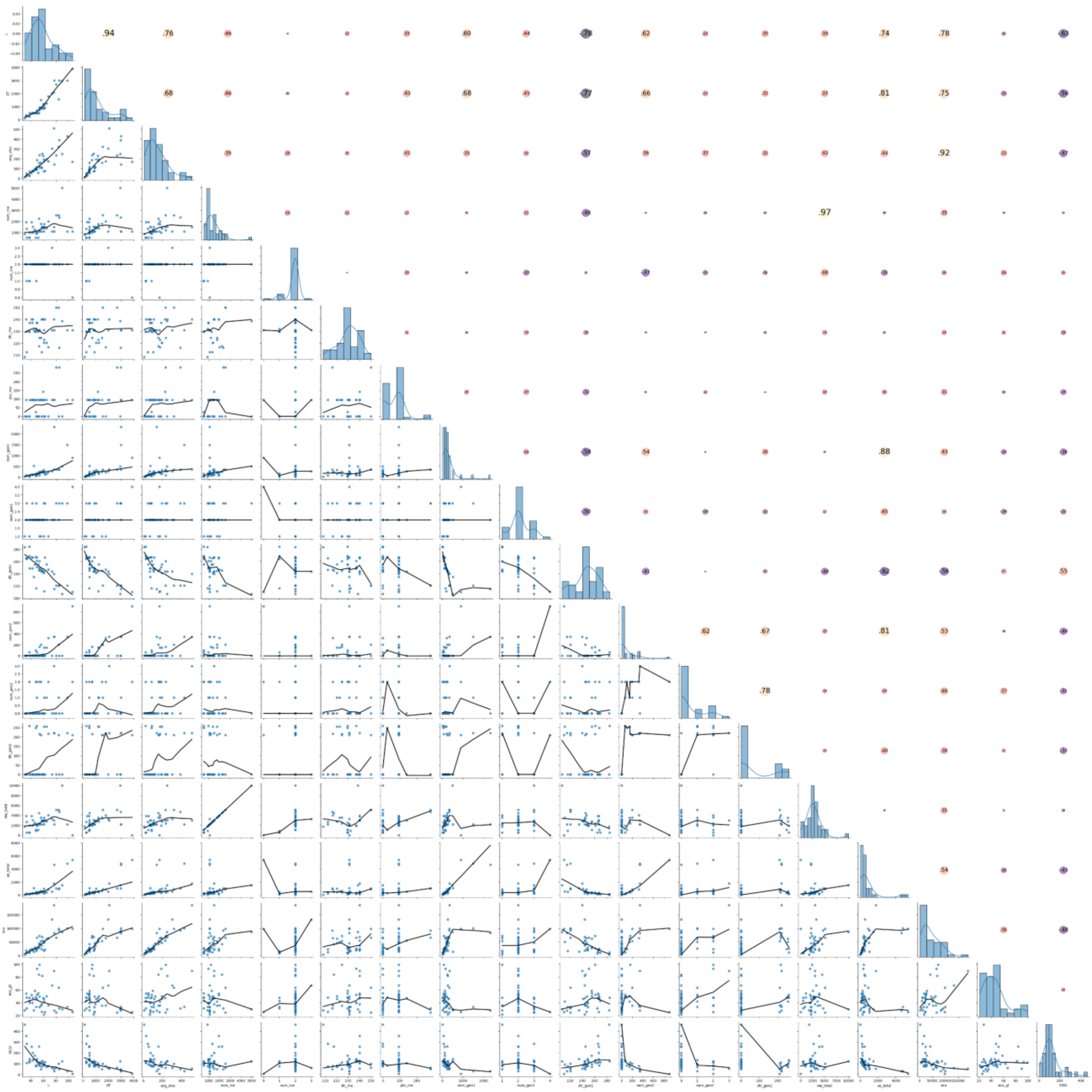
	L	GT	Avg. Elec.	Nom ME	Num ME	SFC ME	PTO ME	Nom Gen1	Num Gen1	SFC Gen1	Nom Gen2	Num Gen2	SFC Gen2	ME Total	AE Total	Eco point	Eco score	EEDI
count	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42
mean	57.6	1122.4	151.8	1399.5	1.9	230.7	68.9	292.1	2.1	246.7	79.8	0.5	71.7	2713.9	818.3	47796.5	46.3	123.2
std	18.1	944.1	113.7	775.3	0.5	9.6	66.9	409.6	0.7	19.7	169.3	0.9	108.9	1671.2	1220.9	38092.6	19.6	87.7
min	30.6	109.0	10.0	533.0	0.0	208.2	0.0	19.0	1.0	207.0	0.0	0.0	0.0	0.0	20.0	3032.0	21.4	5.1
25%	45.4	499.0	75.0	996.8	2.0	225.8	0.0	103.8	2.0	240.2	0.0	0.0	0.0	1938.0	239.5	20582.5	31.5	77.3
50%	54.0	841.0	113.0	1137.0	2.0	231.0	94.0	192.5	2.0	247.2	0.0	0.0	0.0	2461.0	430.0	33031.5	44.0	104.1
75%	64.7	1436.0	195.0	1500.0	2.0	238.3	94.0	270.0	2.0	263.5	90.5	1.0	211.5	3182.3	813.5	66016.8	52.0	132.0
max	106.3	3891.0	510.0	5000.0	3.0	250.0	282.0	2322	4.0	284.4	900.0	3.0	260.2	10000.0	5400.0	186006.0	99.1	461.7

APPENDIX B REDUCED YETI DATASET DISTRIBUTIONS



APPENDIX C REDUCED YETI DATASET PAIR-WISE SCATTER AND CORRELATION ANALYSIS

The following figure presents all the corresponding correlations and pair-wise comparisons of each feature explored in the corresponding report. This dataset contains a mixture of continuous and integer variables. Scatter plot visualisations include a locally average trendline (black) which shows the observed relationship between the paired features. The upper triangle presents the bubble representation for the absolute correlation strength. The larger the bubble, the more strongly correlated the two parameters are.



M A R I N
P.O. Box 28

6700 AA Wageningen
The Netherlands

T +31 317 49 39 11
E info@marin.nl

I www.marin.nl