1.5 Distributed error quantification: conditioning the error distribution on the input space

This section corresponds to box labelled "P(E|X)" in Figure 1.17. In section 1.4, the marginalised distribution of the residue $P(e)$ was discussed. This section aims at extracting useful information by conditioning the error distribution to the input variables' space, which mathematically can be denoted by $P(e|X)$. Amongst the motivations for this, we count in the first place that this section will help building our uncertainty model (vid. section 1.7) by refining the methods described in the previous section for uncertainty prediction with new information from $P(e|X)$. Furthermore, model and data boosting (vid. Figure 1.17) can both benefit from the analysis performed in this section. Knowledge of the error distribution conditioned to the input space can help identify those regions where either:

- The dataset is properly representing the region, but model training is deficient.
- Dataset quality is not sufficient for proper training according to the specified performance requirements.

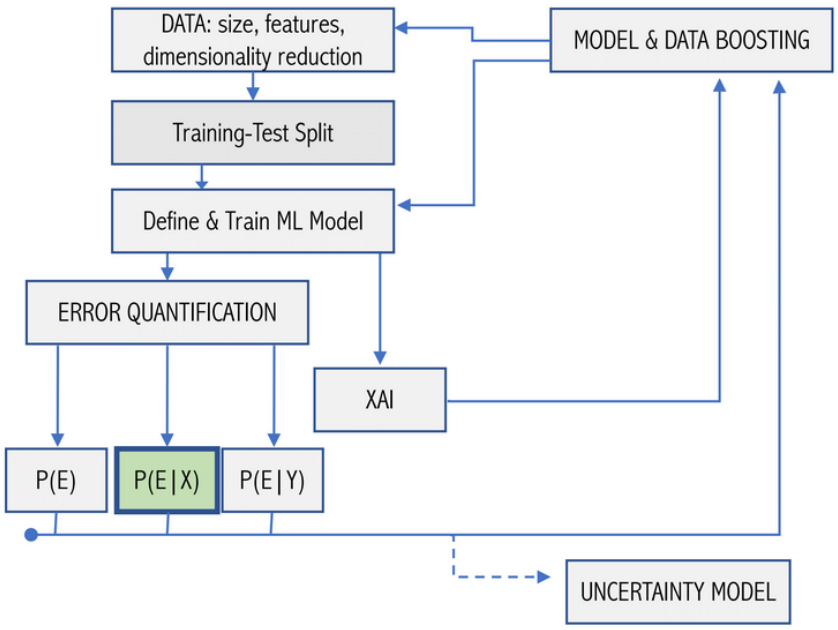


Figure 1.17: Box diagram showing the relative position of section 1.5 in the complete validation pipeline.

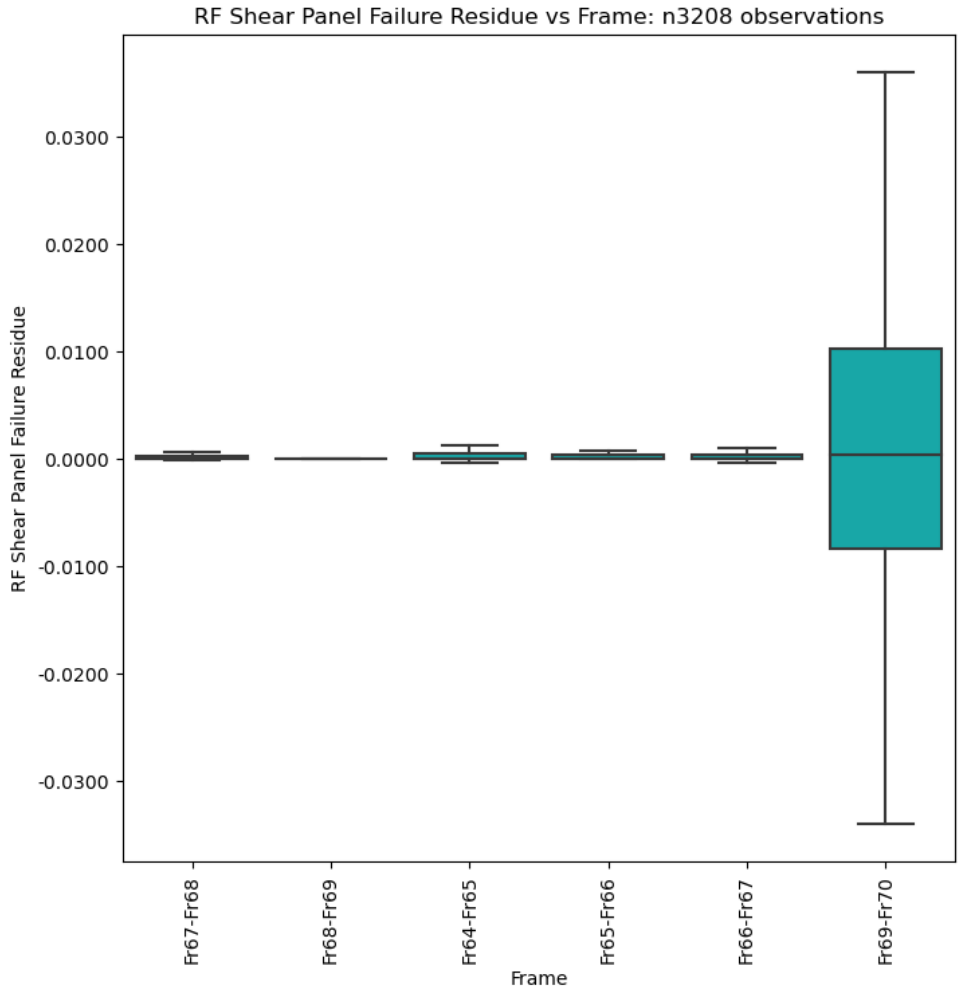
In the first case, training of the model can be reinforced by several means (longer training, hyperparameter tuning, transfer learning, etc.) In the second case, resampling data in deficient regions, or artificial data enhancing techniques such as data augmentation[26] are recommended.

The task of recomposing $P(e|X)$ needs to be addressed in a computationally efficient way, given the input space dimensionality m [recall the input features vector is $\mathbf{X} = (X_1, X_2, \ldots, X_m)$] can be very large. For the case of MS-S18, for instance, the input space's dimensionality is $m = 31$ (28 numerical inputs plus 3 categorical). With industrial sized datasets (typically reaching up to millions of data points), recomposing $P(e|X)$ becomes unaffordable. The easiest way of overcoming this issue is by binning the input space according to a certain criteria. If we denote the binned input space by $X^B = \{X_1^B, X_2^B, X_3^B \ldots X_r^B\}$ where X_j^B represents a certain bin, and there are a total of r bins, then we can substitute the task of recomposing $P(e|X)$ by that of recomposing $P(e|X^B)$. Although simple, this idea allows for an effective computational cost reduction.

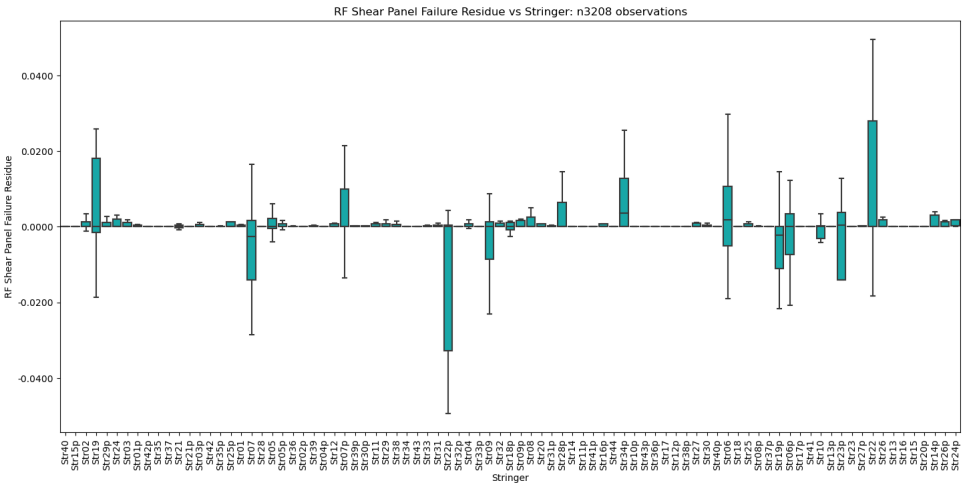
1.5.1 Visualization of error as a function of categorical (discrete) input variables

The immediate criteria for binning the input space is using categorical variables. These variables do in essence bin the input space as they can only take certain (or rather discrete) values. As previously mentioned, for MS-S18 there are three categorical variables enclosing geometrical information: "Frame", "Stringer", and "dp". Figure 1.18 plots, for each categorical input variable, a box plot of the error as a function of the category.

The box plot layout reports the median, Inter-quantile Range and whiskers. Outliers showing anomalous dispersion of the residue could indicate geometrical configurations (*e.g.* the one defined by frames no. 69-70 and stringer no. 26) where training has been inefficient. The engineer in charge could inspect this zone and conclude, for instance, that the physics at play in that region involve exploding gradients which induce strong non-linearities.



(a) Error conditioned to the categorical variable "Frame"



(b) Error conditioned to the categorical variable "Stringer"

Figure 1.18: Box and whisker plots depicting the residue conditioned to two different categorical variables. Figure 1.18(a): "Frame". Figure 1.18(b): "Stringer". Clearly, outliers can be appreciated in both cases ("Fr69-Fr70" in Figure 1.18(a), and various stringers in Figure 1.18(b)).

1.5.2 Bias detection and quantification (1D)

After the visual inspection carried out in Figure 1.18, a statistical test is needed to actually check whether the observed residue for some frames or stringers is unexpected enough to be considered an "outlier". This is, we need to check whether anomalies shown in Figure 1.18 actually are statistically significant or not. To this end, the following procedure is proposed:

A. Detection of error bias in single categories (ANOVA)

The initial method of choice to detect whether a given input categorical variable shows any bias is the one-way ANOVA test[27]. This method consists on the following steps[5]:

- We fix a categorical variable i under analysis.
- The method scatter-plots the residual error versus the category of x^i, for all n samples of the test set.
- ANOVA makes then an hypothesis test, where the null hypothesis $H0$ is that the mean of the (theoretical distribution) error for each of the categories is the same. The ANOVA test rejects $H0$ with a certain confidence if the $p-$value of the test is smaller than a certain level.
- The output of the method is just the $p-$value, if this is smaller than 0.05, then $H0$ is rejected at 95% confidence and we say that the categorical variable x^i shows bias.

In Table 1.12, results of the ANOVA test from MS-S18 are shown. The test shows bias in all the three categorical input variables, but for different output variables each one.

[5]N.B. ANOVA in principle requires data in each category to be normally distributed and homoscedasticity (variances in different categories are similar, cfr.[17, p. 374]). It is important to remark these conditions are sometimes not met.

Table 1.12: P-values results from the one-way ANOVA test. The red labelled cells show that variable "dp" shows bias for the residual distribution of "RF Net Tension", as well as variable "Frame" for the residual distribution of "RF Forced Crippling" and variable"Stringer" for the residual distributions of "RF Net Tension" and "RF Pure Compression". For the rest of the table, p-values greater than 0.05 indicate that $H0$ cannot be rejected with a statistical confidence of at least 95%.

	1-way ANOVA (p-value)					
	RF Forced Crippling	RF Column Buckling	RF In Plane	RF Net Tension	RF Pure Compression	RF Shear Panel Failure
dp	0.60142	0.35218	0.86072	0.00000	0.68189	0.90204
Frame	0.02972	0.69550	0.24502	0.22363	0.39069	0.59135
Stringer	0.32703	0.24055	0.28428	0.00062	0.04613	0.83613

B. Quantification of error bias in single categories test no. 1: based on error mean outlier

This analysis is to be done only if results from the one-way ANOVA test show bias in at least one categorical variable. Under this assumption, the former test has identified a number of categorical variables that show bias. Here we quantify such bias by checking whether there are certain categories which are "outliers", *i.e.* categories where the error mean is substantially different than for the rest of categories. We check this using z-score[28], which consists on the following steps: we initially construct the mean error e in each category. If the categorical variable has s categories, then we have a vector $(e_1, e_2, ..., e_s)$. We then z-score this vector to build

$$\left(\frac{e_1 - \langle e\rangle}{\sigma(e)}, \frac{e_2 - \langle e\rangle}{\sigma(e)}, ..., \frac{e_s - \langle e\rangle}{\sigma(e)}\right),$$

where

$$\langle e\rangle = \frac{1}{s}\sum_{i=1}^{s} e_i;\ \sigma(e) = \sqrt{\frac{1}{s}\sum_{i=1}^{s}(e_i - \langle e\rangle)^2}$$

We make use of the rule of thumb that category i shows **weak bias** if

$$1 < \frac{|e_i - \langle e\rangle|}{\sigma(e)} \leq 3,$$

whereas category i shows **strong bias** if

$$\frac{|e_i - \langle e\rangle|}{\sigma(e)} > 3.$$

The test output is shown in Table 1.13, where for each categorical variable that was previously identified as having bias, a sub-table showing **only** the specific categories that either show weak or strong bias is displayed.

Table 1.13: Z-score results for biased categorical variables. One table is generated for each pair biased categorical input variable-output variable.

Frame with RF Forced Crippling as output

	Fr66-Fr67	Fr69-Fr70
N	4	1
mean	-0.014281	0.008480
Z score	1.174359	1.770030
Bias	Weak	Weak

dp with RF Net Tension as output

	0.000000	0.900000
N	15	2
mean	0.004450	-0.005974
Z score	1.000000	1.000000
Bias	Weak	Weak

Stringer with RF Net Tension as output

	Str06	Str06p	Str31	Str39
N	1	2	1	1
mean	-0.022004	0.028592	-0.026716	0.035710
Z score	1.225013	1.331218	1.463067	1.690886
Bias	Weak	Weak	Weak	Weak

Stringer with RF Pure Compression as output

	Str07
N	1
mean	0.178702
Z score	3.086849
Bias	Strong

C. Quantification of error bias in single categories test no. 2: based on error variance outlier

This analysis is to be done only if results from the one-way ANOVA test show bias is found in at least one categorical variable.

The procedure is similar as before, but here we investigate whether for certain categories the dispersion (not the mean) of the samples in a given category is substantially different than for the rest of categories. For instance, experience in MSP-S18 suggests that the main source of error bias in categorical variables comes from the fact that different categories have different variance. This is quantified by doing a z-score analysis on category variances: if the categorical variable under analysis has s categories, then we have a vector of variances $(v_1, v_2, ..., v_s)$, where

v_i is the variance of the error for all samples in category i. We then z-score this vector to build

$$\left(\frac{v_1 - \langle v\rangle}{\sigma(v)}, \frac{v_2 - \langle v\rangle}{\sigma(v)}, ..., \frac{v_s - \langle v\rangle}{\sigma(v)}\right),$$

where

$$\langle v\rangle = \frac{1}{s}\sum_{i=1}^{s} v_i;\ \sigma(v) = \sqrt{\frac{1}{s}\sum_{i=1}^{s}(v_i - \langle v\rangle)^2}$$

We make use of the rule of thumb that category i shows ***weak variance bias*** if

$$1 < \frac{|v_i - \langle v\rangle|}{\sigma(v)} \leq 3,$$

whereas category i shows ***strong variance bias*** if

$$\frac{|v_i - \langle v\rangle|}{\sigma(v)} > 3$$

The analysis outputs, for each categorical variable that was previously identified as having bias, a table showing only the specific categories that either show weak or strong bias according to this second method.

D. Quantification of error bias in single categories test no. 3: based on identification of linear trend

This test is only applicable to categorical variables for which specific categories have a natural ordering (mathematically, we say there is a canonical geometric embedding for the categorical variable). Such identification needs to be done "by hand" by the engineer in charge.

Such embedding exists when the categorical variable is related for instance to a geometrical location in the plane (for instance the Frames and Stringers are categorical –in so far they are discrete– but they correspond to regions of the plane along a Cartesian and a radial axis, so there is a natural ordering). In those cases, it makes sense to investigate linear trends as these are interpretable.

The method just proceeds to fit a linear model (with just one explanatory variable). If the slope is above a certain threshold and if the fit is good, then we can be confident the linear trend is genuine, and the slope can be used to quantify the linear bias and as a source for uncertainty.

For each categorical variable that has a natural ordering and was previously detected as having bias, the test fits a linear model and outputs the result of this model.

If such model has a slope (statistically) significantly different from zero, then this categorical variable is flagged as having a linear trend.

1.5.3 Bias detection and quantification (2D)

In the previous section we considered the specific effect (bias) of individual input variables on the error. However, it can be the case that such bias is enhanced when groups of input variables are considered together. The method of choice to analyze whether two independent (input) categorical variables have an effect on the error means is the 2-way ANOVA test[29]. In general the method is not very informative except when the combination of input variables is itself interpretable. This usually requires the external feedback of the engineer. For instance, the engineer can know a priori that certain combination of input variables play a synergistic role, go together, etc. For instance, in MSP-S18, the categorical variables **Frame** and **Stringer** are geometric locations and thus, together, provide a location of the specific region of the plane that the configuration analyses. One can thus perform a 2-way ANOVA to analyze the presence of bias accordingly. If such bias exists, then bias quantification tests described in subsection 1.5.2 can be applied.

A. Bias detection (2D)

For the pair of (previously chosen) categorical variables, the test performs 2-way ANOVA as described above and outputs the p-value. If this p-value is smaller than 0.05, we conclude that there exists bias at 95% confidence.

B. Bias quantification (2D)

Suppose we have two categorical variables A and B, where A has q categories $A = (A_1, A_2, \dots, A_q)$ and B has r categories $B = (B_1, B_2, \dots, B_r)$. Suppose also that a 2-way ANOVA concluded that the error is biased on the combined effect of A and B. We then can build all the pairs (A_i, B_j) and interpret each of them as a single category of this "block categorical variable", *i.e.* we have now $q \times r$ categories. We can subsequently apply the bias quantification tests 1 and 2 described above, applied to the "block categorical variable".

- If bias has been detected, the "block categorical variable" with qr categories is built and a table with qr columns and 2 rows is defined (for test no. 1 and test no. 2 results).
- Bias quantification test no. 1 (error mean, see above) is performed on the block categorical variable, and fills up in the table the first row for those columns that show either weak or strong bias.
- Bias quantification test no. 2 (error variance, see above) is performed on the block categorical variable, and fills up in the table the first row for those columns that show either weak or strong bias.

1.5.4 Visualization of error as a function of numerical (continuous) input variables

In this section the same basic analysis is performed as in subsection 1.5.2, although conditioning the error on numerical instead of categorical input variables.

In this section we plot, for each numerical input variable, a scatter plot of the error as a function of the input numerical variable at analysis. From the scatter plot and its linear fit, we expect to (with human intervention) identify bias in the variables which present it, and use visual information for model and data boosting. Illustrative results of two input variables are given in Figure 1.19.

1.5.5 Bias detection and quantification

The goal of this section is to flag statistically significant bias found in output variables, in order to provide mathematical assurance to engineers when interpreting results from Figure 1.19. Similar to categorical variables, we use different tests for bias detection and for bias quantification. As the input variables are now numerical, instead of ANOVA and z-score we use:

A. Detection of linear trends for individual numerical variables

For each input variable, the code fits a linear model that can identify a linear trend of the error as a function of the input variable.
The test outputs (vid. Table 1.14):

- The p-value that accounts for whether such type of bias indeed exists.
- The Pearson correlation coefficient.

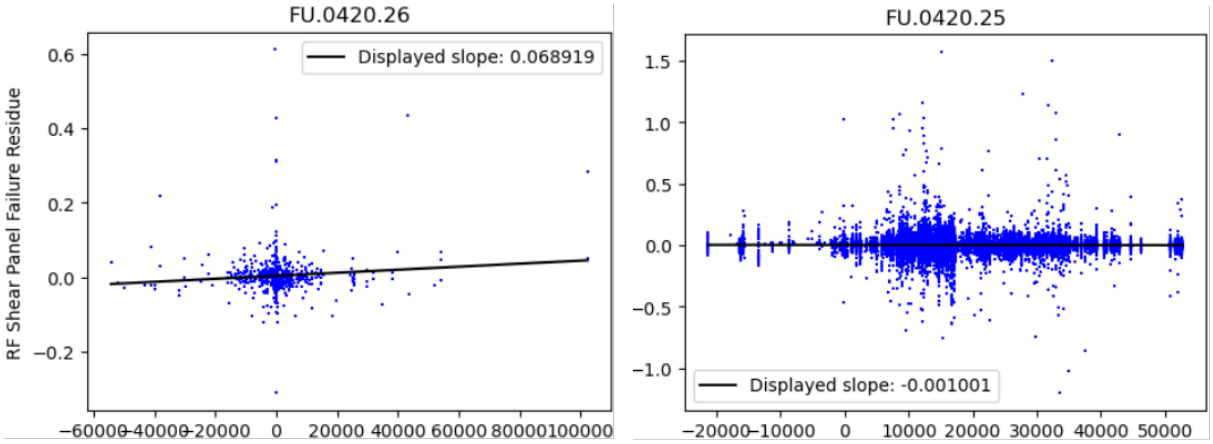


Figure 1.19: Scatter plots of residue against a particular numerical input variable. Left: Input var. FU.0420.26. A linear trend is appreciated, suggesting the model makes worse predictions as input load "FU.0420.26" is larger. Right: input var. FU.0420.25. No linear trend is appreciated, although severe heteroscedasticity can be observed. Note the vertical bar arrangement, showing particular input loads for which dispersion is unusually large. In view of the former results, the engineer may decide to investigate the underlying causes both for the linear trend in the left and for heteroscedasticity on the right. With the information, model and data boosting may be performed. More training data could be decided to be sampled in the range of $[20000, 60000]$ of FU.0420.26, for instance, and some regularization technique could be tried for the loss function in order to penalise inputs triggering high variance observed in the right.

- The slope of the best linear model fit.

Table 1.14: Bias detection and quantification on numerical input variables (the results for just five input variables are shown here – in columns–): P-value (statistical significance of the hypothesis that the variable is biased), Pearson coefficient, and slope of the best linear fit are shown in the first three rows. The last row shows the message "Biased" in case the p-value > 0.05, "NO" otherwise. Data from this table corresponds to residue of the output variable "RF Column Buckling". Analogous tables exist for the rest of the output variables.

factors	FU.0410.14	FU.0410.15	FU.0410.16	FU.0410.24
p-value	0.600663	0.625659	0.037808	0.480859	0.059544
Pearson coeff.	0.029417	0.027417	-0.116347	-0.039608	0.105612
Slope (normalized)	0.013192	0.027099	-0.141826	-0.019688	0.068905
Linear trend	NO	NO	Biased	NO	NO

B. Discretizing continuous variables

The process is to bin each input numerical variable and thus treat them as categorical, so that

one can perform one-way ANOVA followed by bias quantification tests no. 1 (mean outlier) and test no. 2 (variance outlier) former discussed. By default the number of bins is equal to 10.

For each input variable:

- The input variable is binned.
- The steps depicted in sections subsection 1.5.2 are applied.

Results are shown in Table 1.15, Table 1.16 and Table 1.17.

Table 1.15: 1-way ANOVA test results (p-values) for binned numerical input variables. For the output variable "RF Forced Crippling", bias is found in the input variables "FU.0420.25" and "FU.0430.25". Similarly, for the output variable "RF Net Tension", bias is found in the input variable "FU.0430.15".

	1-way ANOVA (p-value)					
	RF Forced Crippling	RF Column Buckling	RF In Plane	RF Net Tension	RF Pure Compression	RF Shear Panel Failure
factors_binned	0.36995	0.57707	0.16276	0.32759	0.72817	0.67178
FU.0410.14_binned	0.14541	0.60928	0.47117	0.86966	0.16037	0.72052
FU.0410.15_binned	0.47235	0.07028	0.23893	0.19848	0.69785	0.36203
FU.0410.16_binned	0.99209	0.29325	0.30693	0.59764	0.86850	0.11857
FU.0410.24_binned	0.28251	0.37036	0.06887	0.30212	0.50172	0.10088
FU.0410.25_binned	0.84519	0.27807	0.35407	0.71585	0.99200	0.10885
FU.0410.26_binned	0.97916	0.23816	0.63145	0.18726	0.59593	0.54611
FU.0420.14_binned	0.11827	0.54753	0.44448	0.69463	0.13006	0.87985
FU.0420.15_binned	0.27700	0.05152	0.19852	0.27864	0.86905	0.30744
FU.0420.16_binned	0.96355	0.29429	0.34697	0.53042	0.88578	0.10234
FU.0420.24_binned	0.28161	0.36135	0.06327	0.32091	0.60557	0.12283
FU.0420.25_binned	0.03686	0.09300	0.46097	0.78027	0.13730	0.83867
FU.0420.26_binned	0.56289	0.15438	0.85899	0.28868	0.77835	0.78597
FU.0430.14_binned	0.20706	0.64043	0.65843	0.16471	0.25147	0.57201
FU.0430.15_binned	0.34776	0.20721	0.70087	0.01776	0.68061	0.87189
FU.0430.16_binned	0.29934	0.05751	0.33328	0.78752	0.76107	0.21693
FU.0430.24_binned	0.50895	0.18564	0.79205	0.09649	0.68746	0.90018
FU.0430.25_binned	0.03176	0.07608	0.30233	0.82106	0.70637	0.73546
FU.0430.26_binned	0.75006	0.18252	0.25337	0.45661	0.99750	0.39317

Table 1.16: Bias quantification in binned FU.0430.15 input variable. Columns represents bins showing bias. The same quantification methods employed for categorical variables (z-score for mean and variance outlier detection) have been used.

FU.0430.15_binned with RF Net Tension as output				
	3	**4**	**8**	**9**
N	269	284	281	299
mean	-0.003908	-0.005275	-0.001238	-0.001153
Z score	1.004073	2.079140	1.096293	1.163220
Bias	Weak	Weak	Weak	Weak

Table 1.17: Summary of binned input variables bias quantification after binning the numerical variables and performing one-way ANOVA and z-score tests to every bin.

RF Forced Crippling		
	FU.0420.25_binned	**FU.0430.25_binned**
1-way ANOVA - ZScore(mean)	Weak	Weak
1-way ANOVA - ZScore(var)	Weak	Weak

RF Net Tension	
	FU.0430.15_binned
1-way ANOVA - ZScore(mean)	Weak
1-way ANOVA - ZScore(var)	Weak

Bibliography

[1] P. Bijlaard. "On the Buckling of Stringer Panels Including Forced Crippling". In: *Journal of the Aeronautical Sciences* 22.7 (1955), pp. 491–501 (cit. on p. 1).

[2] F. P. Preparata and M. I. Shamos. "Convex Hulls: Basic Algorithms". In: *Computational Geometry: An Introduction*. New York, NY: Springer New York, 1985, pp. 95–149. ISBN: 978-1-4612-1098-6. DOI: 10.1007/978-1-4612-1098-6_3. URL: https://doi.org/10.1007/978-1-4612-1098-6_3 (cit. on p. 2).

[3] D. Barrett et al. "Measuring abstract reasoning in neural networks". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 511–520. URL: https://proceedings.mlr.press/v80/barrett18a.html (cit. on p. 3).

[4] B. M. Lake and M. Baroni. "Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks". In: *CoRR* abs/1711.00350 (2017). arXiv: 1711.00350. URL: http://arxiv.org/abs/1711.00350 (cit. on p. 3).

[5] D. Saxton et al. "Analysing Mathematical Reasoning Abilities of Neural Models". In: *CoRR* abs/1904.01557 (2019). arXiv: 1904.01557. URL: http://arxiv.org/abs/1904.01557 (cit. on p. 3).

[6] T. Ebert, J. Belz, and O. Nelles. "Interpolation and extrapolation: Comparison of definitions and survey of algorithms for convex and concave hulls". In: *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE. 2014, pp. 310–314 (cit. on p. 3).

[7] W.-Y. Loh, C.-W. Chen, and W. Zheng. "Extrapolation errors in linear model trees". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.2 (2007), 6–es (cit. on p. 3).

[8] P. Klesk. "Construction of a Neurofuzzy Network Capable of Extrapolating (and Interpolating) With Respect to the Convex Hull of a Set of Input Samples in R". In: *IEEE Transactions on Fuzzy Systems* 16.5 (2008), pp. 1161–1179. DOI: 10.1109/TFUZZ.2008.924337 (cit. on p. 3).

[9] R. Balestriero, J. Pesenti, and Y. LeCun. "Learning in high dimension always amounts to extrapolation". In: *arXiv preprint arXiv:2110.09485* (2021) (cit. on pp. 3, 4, 10).

[10] S. Marsland. *Machine Learning: An Algorithmic Perspective*. 2nd ed. Boca Raton, USA: Chapman & Hall/CRC, 2015 (cit. on pp. 3, 7).

[11] I. Bárány and Z. Füredi. "On the shape of the convex hull of random points". In: *Probability theory and related fields* 77 (1988), pp. 231–240 (cit. on p. 3).

[12] L. Bonnasse-Gahot. "Interpolation, extrapolation, and local generalization in common neural networks". In: *arXiv preprint arXiv:2207.08648* (2022) (cit. on pp. 4, 10, 13).

[13] H. Hotelling. "Analysis of a complex of statistical variables into principal components." In: *Journal of educational psychology* 24.6 (1933), p. 417 (cit. on p. 9).

[14] B. Rosner and D. Grove. "Use of the Mann–Whitney U-test for clustered data". In: *Statistics in medicine* 18.11 (1999), pp. 1387–1400 (cit. on p. 13).

[15] R. Velez Ibarrola and A. Garcia Perez. *Calculo de probabilidades y Estadistica Matematica*. 1st ed. Madrid, Spain: Universidad Nacional de Educacion a Distancia, 1994 (cit. on p. 15).

[16] D. Zhang. "A coefficient of determination for generalized linear models". In: *The American Statistician* 71.4 (2017), pp. 310–316 (cit. on p. 20).

[17] J. D. Jobson. *Applied multivariate data analysis: regression and experimental design*. Springer Science & Business Media, 2012 (cit. on pp. 20, 37).

[18] G. Chen, J. R. Gott, and B. Ratra. "Non-Gaussian Error Distribution of Hubble Constant Measurements". In: *Publications of the Astronomical Society of the Pacific* 115.813 (2003), p. 1269 (cit. on p. 22).

[19] P. Pernot, B. Huang, and A. Savin. "Impact of non-normal error distributions on the benchmarking and ranking of Quantum Machine Learning models". In: *Machine Learning: Science and Technology* 1.3 (2020), p. 035011 (cit. on p. 22).

[20] D. Smyl et al. "Learning and correcting non-Gaussian model errors". In: *Journal of Computational Physics* 432 (2021), p. 110152 (cit. on p. 22).

[21] L. Chai et al. "Using generalized Gaussian distributions to improve regression error modeling for deep learning-based speech enhancement". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.12 (2019), pp. 1919–1931 (cit. on p. 22).

[22] M. C. Jones and A. Pewsey. "Sinh-arcsinh distributions". In: *Biometrika* 96.4 (2009), pp. 761–780 (cit. on pp. 25, 28).

[23] B. Rosner. "Percentage points for a generalized ESD many-outlier procedure". In: *Technometrics* 25.2 (1983), pp. 165–172 (cit. on p. 27).

[24] B. Efron. "Bootstrap methods: another look at the jackknife". In: *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 569–593 (cit. on p. 29).

[25] E. B. Wilson. "Probable inference, the law of succession, and statistical inference". In: *Journal of the American Statistical Association* 22.158 (1927), pp. 209–212 (cit. on p. 31).

[26] L. Taylor and G. Nitschke. "Improving deep learning with generic data augmentation". In: *2018 IEEE symposium series on computational intelligence (SSCI)*. IEEE. 2018, pp. 1542–1547 (cit. on p. 35).

[27] T. K. Kim. "Understanding one-way ANOVA using conceptual figures". In: *Korean journal of anesthesiology* 70.1 (2017), pp. 22–26 (cit. on p. 37).

[28] B. R. Kirkwood and J. A. Sterne. *Essential medical statistics*. John Wiley & Sons, 2010 (cit. on p. 38).

[29] Y. Fujikoshi. "Two-way ANOVA models with unbalanced data". In: *Discrete Mathematics* 116.1-3 (1993), pp. 315–334 (cit. on p. 41).

[30] O. Montenbruck and E. Gill. *Satellite Orbits: Models, Methods and Applications*. 1st ed. Wessling, Germany: Springer, 2001 (cit. on p. 51).

[31] K. Yamanaka and F. Ankersen. "New State Transition Matrix for Relative Motion on an Arbitrary Elliptical Orbit". In: *Journal of Guidance, Control & Dynamics* 25.1 (2002), pp. 60–66. DOI: 10.2514/2.4875 (cit. on p. 52).

[32] D.-W. Gim and K. Alfriend. "State Transition Matrix of Relative Motion for the Perturbed Noncircular Reference Orbit". In: *Journal of Guidance Control and Dynamics* 26 (Nov. 2003), pp. 956–971. DOI: 10.2514/2.6924 (cit. on p. 52).

[33] S.-S. Exchange. *Practical Uses of an STM*. 2019. URL: https://space.stackexchange.com/questions/32916 (cit. on p. 53).

[34] ESA. *Precise Orbit Determination*. 2011. URL: https://gssc.esa.int/navipedia/index.php/Precise_Orbit_Determination (cit. on p. 53).