Robust statistical modelling of the bridge overload risk caused by extreme traffic load hazard

Mattia Cefis, Maurizio Carpita
Department of Economics and Management, University of Brescia, Brescia, Italy.

1. The project of the bridge overload risk

In this paper, we present results of a study conducted thanks to collaboration with the Transportation Engineering research group at the University of Brescia, within the Spoke 7 Project of the Sustainable Mobility Center (MOST), funded by the National Recovery and Resilience Plan (NRRP) 2022-2026. Our contribution focuses on providing expertise in statistical methodology and modelling; in particular, it focuses on assessing the risk of bridge overloads caused by extreme traffic loads.

Prior research in this area has explored the estimation of bridge overload risk, examining various methodologies and risk components. The prediction of load exceedance event severity is a key concern for bridges subjected to high volumes of heavy traffic. A recent study by Ventura et al. (2024) investigated this issue on a heavily trafficked bridge in the south-eastern Brescia, using both a statistical approach (binomial logistic regression model) and a machine learning approach (artificial neural network).

In this analysis, our reference model will be the statistical one for a key binary component of the risk assessment model that, which is the severity ($Y = 1$ if a severe overload event occurs, 0 otherwise) of bridge failure events. This work addresses a significant imbalance problem as, obviously and fortunately, severe overload events are rare, representing approximately 2% of the monitored traffic instances.

We introduce three methodological enhancements compared to the study of Ventura et al. (2024): (i) applying a class-balancing approach to address the rare event category across the entire dataset, (ii) verifying the robustness and stability of results through bootstrap resampling, and (iii) using the rank graduation Box metric to evaluate the predictive power of the predictors of the statistical model. Moreover, we use the full sample of n = 10,824 temporal windows of one hour from January 2022 to March 2023, during which vehicle traffic over the bridge was continuously monitored. Within this sample, a sub-sample (n = 6,184) was identified where at least one overload event occurs (instances where the vehicle loads exceed a certain critical threshold), indicating potential risk for structural damage. In their work, Ventura et al. (2024) focused on this sub-sample, to model the severity: while this approach yielded high model accuracy (nearly 100%), it also introduced potential for estimation bias and overfitting due to the restricted sample and outcome, making the resulting estimators unstable. Additionally, the reported model coefficients exhibited substantial variation, and odds ratios tending towards infinity suggested a lack of robustness, highlighting the need for further validation.
Our research aims to address the challenges identified in the previous study. In the initial phases, adopting a binary regression model, we compared different link functions, including *probit*, *cloglog* and *logit*. The *probit* link (Aldrich and Nelson, 1984) was selected due to its superior stability, evidenced by lower bootstrap standard errors and its better fit, indicated by a smaller AIC value than the others. We employed the significant and not collinear predictors identified by Ventura et al. (2024) and incorporated interaction terms to capture better relationships among risk factors (i.e., mean vehicle speed and the maximum gross vehicle mass -GVM- limit ratio, analyzed separately for standard vehicles and those with more than six axes).

The binary regression model in (1) estimates the conditional probability of a severity event (Y = 1) for a given temporal window $i = 1, 2, \ldots, n$ described by a vector of K explanatory variables (predictors) $\boldsymbol{x}_i^{\top} = (x_{i1}, \ldots, x_{i\mathrm{K}})$:

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Mattia Cefis, Maurizio Carpita, *Robust statistical modelling of the bridge overload risk caused by extreme traffic load hazard*, pp. 141-146, © 2026 Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1062-1.26

$$\hat{p}_i = E(Y_i = 1 \mid \boldsymbol{x}_i) = P(Y_i = 1 \mid \boldsymbol{x}_i) = \Phi(\boldsymbol{x}_i^{\top}\beta), \quad (1)$$

where $\hat{p}_i$ represents the estimated conditional probability that the severity corresponds to the event of interest (bridge failure) given the traffic situation i, $\boldsymbol{\beta}$ is the vector of coefficients to be estimated, representing the influence of each feature on the probability; finally, $\Phi(z)$ is the standard normal cumulative distribution function, which links the linear predictor $\boldsymbol{x}_i^{\top}\boldsymbol{\beta}$ to a probability value in the (0,1) range.

The next section details the method used for balancing the outcome classes and the design developed to test the model stability and robustness; Section 3 presents the results of the model and Section 4 concludes with a discussion about findings and implications for future research.

2. The balanced synthetic samples and the robust design approaches

Given the class imbalance in our target variable (severity Y = 1 accounting for only 2% of observations in the dataset), it is crucial to account for it: as widely reported, class imbalance significantly compromises the training process, leading models to focus on the prevalent class. The limited amount of data for the minority class results in poor model accuracy, particularly in correctly classifying those rare events **(**Menardi and Torelli, 2014). To provide an alternative approach, we explored oversampling techniques to balance the dataset and compared performance against models trained on the original, imbalanced data (i.e., adopting 2% as a class threshold).

While several oversampling techniques exist, such as random oversampling examples (ROSE: Menardi and Torelli, 2014), we focused on implementing the SANSA (synthetic adaptive nearest neighborhood sampling algorithm: Nasir et al., 2022) approach. It is an evolution of the well-known SMOTE (synthetic minority oversampling technique: Chawla et al., 2002) with several advances, improving data balancing by adaptively generating synthetic data focused near decision boundaries, leading to enhanced performance with very noisy datasets; this approach also minimizes overfitting by carefully selecting nearest neighbors and excluding outliers, resulting in more realistic and informative synthetic observations compared to standard SMOTE. While ROSE was also explored, SANSA exhibited superior performance metrics in our analysis, and thus we present only the results obtained using this new approach.

To apply this, we used the *R* package *sansa*, obtaining the best performances by setting its two parameters in this way: $\lambda = 0$, performing uniform oversampling, by generating synthetic samples evenly across all minority instances without giving priority to isolated ones; $ksel = 10$ means each synthetic sample is generated based on the 10 nearest minority neighbors, resulting in smoother and more generalized synthetic data. Note that employing oversampling, it is important to recalibrate the probabilities within the test set to compensate for the effects of the modifications to the training data (Bishop and Nasrabadi, 2006). Following the Bayes' Theorem, posterior probabilities are proportional to prior ones, which can be estimated by each category's relative frequency. Therefore, the estimated probabilities $\hat{p}_i$ in formula (1) obtained with the artificially balanced datasets were corrected obtaining the calibrated severity probability $\hat{p}_i^*$ in formula (2).

$$\hat{p}_i^* = \frac{\frac{0.02}{0.30} \cdot \hat{p}_i}{\frac{0.02}{0.30} \cdot \hat{p}_i + \frac{1-0.02}{1-0.30} \cdot (1-\hat{p}_i)} \quad (2)$$

To further assess stability and robustness of our model, a bootstrap resampling procedure was developed in *R* (with B = 1,000 replications), stratifying by the binary variable severity, to preserve class distributions in each resample. For each iteration, the bootstrap dataset was split into a training set (70%) to which SANSA was applied (the severity events were artificially inflated to 30% of the dataset) and a test set (30%). In each bootstrap replication, all the predictors in both the training and test sets were standardized.

In addition, from an explainable artificial intelligence (xAI) perspective, we also adopted one measure from the rank graduation Box for SAFE AI (Giudici, 2024): the rank graduation explainability (RGE) metric, to assess the predictive importance of the predictors. Note that RGE

allows to rank the variables in terms of their relative importance: it measures how much a variable helps the model distinguish between cases with high predicted severity and those with low predicted severity. The mean RGE (B = 1,000 bootstrap replications) computed on the test set, for the predictor x_j is:

$$\overline{RGE_j} = \frac{1}{B}\sum_{b=1}^{B}\left(\frac{1}{2} - \frac{1}{2}\cdot\frac{cov\left(\hat{p}^{*(b)}, rank\left(\hat{p}_{-j}^{*(b)}\right)\right)}{cov\left(\hat{p}^{*(b)}, rank\left(\hat{p}^{*(b)}\right)\right)}\right), \quad (3)$$

where $\hat{p}^*$ is the estimated probability of the severity in (2) from the model considering all the predictors while $\hat{p}^*_{-j}$ is that one estimated considering $(K-1)$ predictors, excluding x_j; *cov* is the covariance and *rank* assigns a rank to each value in both $\hat{p}^*$ and $\hat{p}^*_{-j}$, based on their order within the respective sets of values.

3. Results

This section presents the comparative results of the analysis using two approaches: imbalance (with the original data) and balanced (with synthetic data generated by SANSA). The estimates of the Probit Regression Model coefficients for both datasets are summarized in Table 1, which also reports the corresponding odds ratios (OR), computed as exp($\hat{\beta}_B$×1.6) to facilitate interpretation in terms of the logit effect sizes (Amemiya, 1981).

It emerges that data balancing substantially affects the estimated relationships, often amplifying the impact of key predictors compared to the original, imbalanced data. In particular, the estimates for *max_GVM_limit_ratio* and its interaction with *mean_speed* show substantial differences: the 90% percentile-bootstrap confidence intervals (CIs) for these parameters do not overlap, indicating that these variations are statistically significant (e.g., 0.942 vs 2.452 for the *max_GVM_limit_ratio* coefficient; OR ≈ 4.51 vs 50.56). The same pattern is observed for the *mean_speed:max_GVM_limit_ratio* interaction, which increases from 0.565 to 1.922 (OR ≈ 2.47 vs 21.65), suggesting a more pronounced moderating effect after balancing. Some features with negative coefficients, such as *mean_speed:max_GVM_limit_ratio_over_6_axes*, translate into OR lower than 1, indicating a strong negative association.

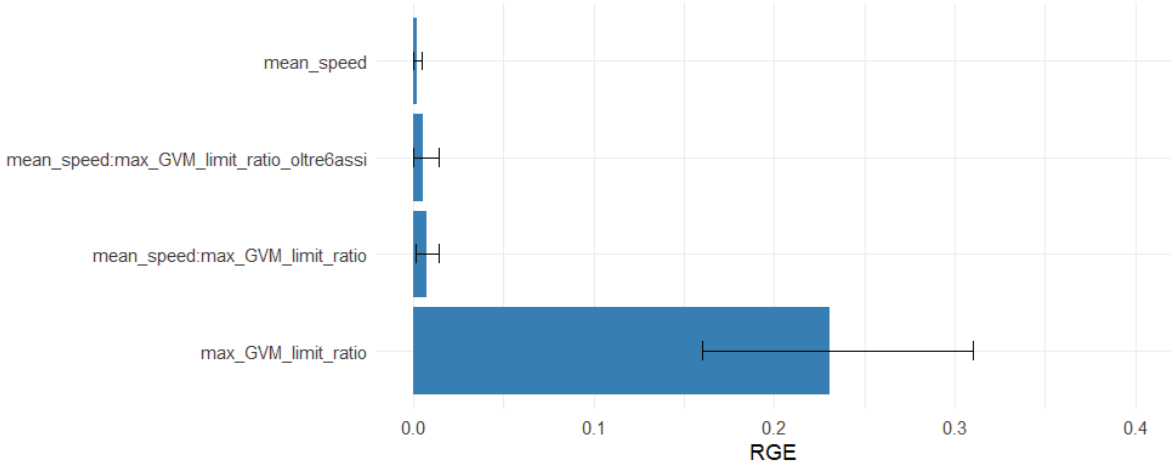
However, despite this amplification of the coefficient estimates and corresponding odds ratios, the bootstrap mean standard errors are similar between the two approaches; the most remarkable difference was observed for the last feature, which has an average standard error of 0.177 in the unbalanced approach compared to 0.153 in the balanced one. This suggests that, although the effect sizes are larger with the balanced approach, the precision of estimates remains comparable.

Table 1. Bootstrap estimates with the training set of the probit model parameters, bootstrap mean standard errors and 90% percentile-bootstrap CIs with imbalance and balanced sampling approaches

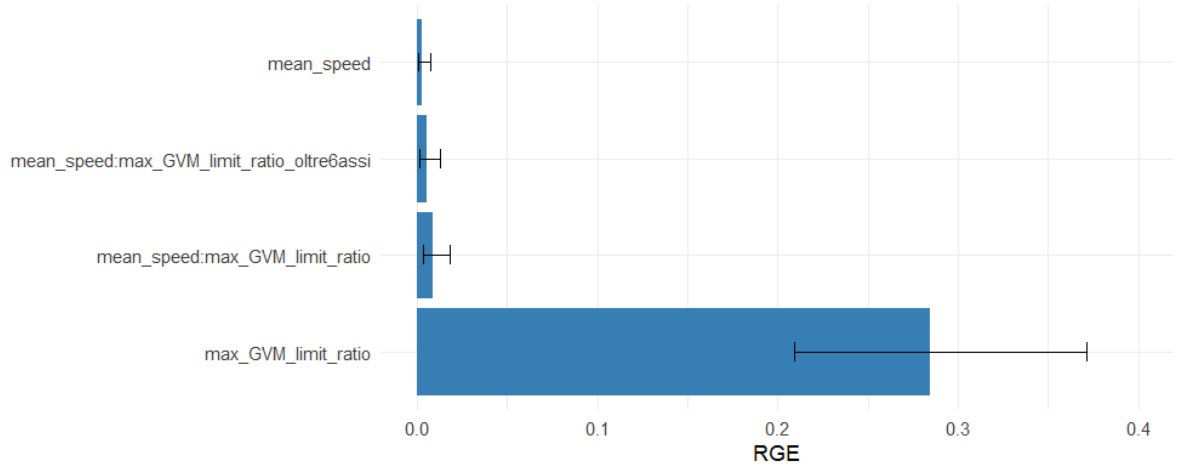
	Imbalanced sample					Balanced Sample				
Regressors	$\hat{\beta}_B$	$\overline{SE}_B$	$\beta_B^{5\%}$	$\beta_B^{95\%}$	$OR(\hat{\beta}_B)$	$\bar{\beta}_B$	$\overline{SE}_B$	$\beta_B^{5\%}$	$\beta_B^{95\%}$	$OR(\hat{\beta}_B)$
Intercept	-2.870	0.091	-3.11	-2.680	-	-3.496	0.095	-4.293	-2.966	-
mean_Speed	0.217	0.096	-0.001	0.458	1.415	0.827	0.089	0.501	1.344	3.755
max_GVM_limit_ratio	0.942	0.063	0.787	1.135	4.514	2.452	0.063	2.116	2.979	50.562
mean_speed : max_GVM_limit_ratio	0.565	0.109	0.230	0.915	2.469	1.922	0.111	1.397	2.755	21.654
mean_speed : max_GVM_limit_ratio_over_6_axes	-0.594	0.177	-1.138	-0.025	0.387	-2.031	0.153	-3.117	-1.355	0.039

Examining the mean RGE index (3) for the predictors after the Bootstrap procedure and taking a look at its 90% percentile-bootstrap confidence interval (Figure 1), it reveals that the ranking of variable importance remained consistent across both the imbalanced and balanced approaches. In both scenarios, the *max_GVM_limit_ratio* emerges as the most crucial variable for evaluating severity. Furthermore, in the balanced approach (Figure 1b), the importance of *max_GVM_limit_ratio* is further amplified, with its mean RGE increasing to nearly 0.3. While the other three variables are

statistically significant, their predictive power, as indicated by their mean RGE values, is substantially lower, with *mean_speed* exhibiting the least predictive capability. Thus, while multiple factors contribute to the model, *max_GVM_limit_ratio* demonstrably dominates in predicting severity.



(a) Imbalanced approach.



(b) Balanced approach.

Figure 1. Mean RGE of the probit model's predictors with 90% percentile-bootstrap CI

Focusing on performance of the probit regression model on the test set (Figure 2), we can see the performance metrics summaries after Bootstrap resampling (B = 1,000) for both imbalance and balanced approaches, along with their 90% confidence intervals. The results highlight that the implementation of SANSA led to a statistically significant improvement in the average performance compared to those obtained from the original data for key metrics. Specifically, the precision scores (the percentage of true positive over all the positive predictions), on average, increased significantly from 8.6% to 14.8% (+72%); a Welch's t-test, accounting for unequal variances, confirmed this improvement to be statistically significant (p-value < 0.05).

Consistent results were also registered on the MCC metric (the Matthews correlation coefficient, that takes into account true and false positives and negatives and it is considered a good metric when dealing with imbalanced datasets, Chicco and Jurman, 2020): the original dataset provides, on average, a MCC score of 0.256, while the oversampled dataset shows a MCC score of 0.359 (+40%).

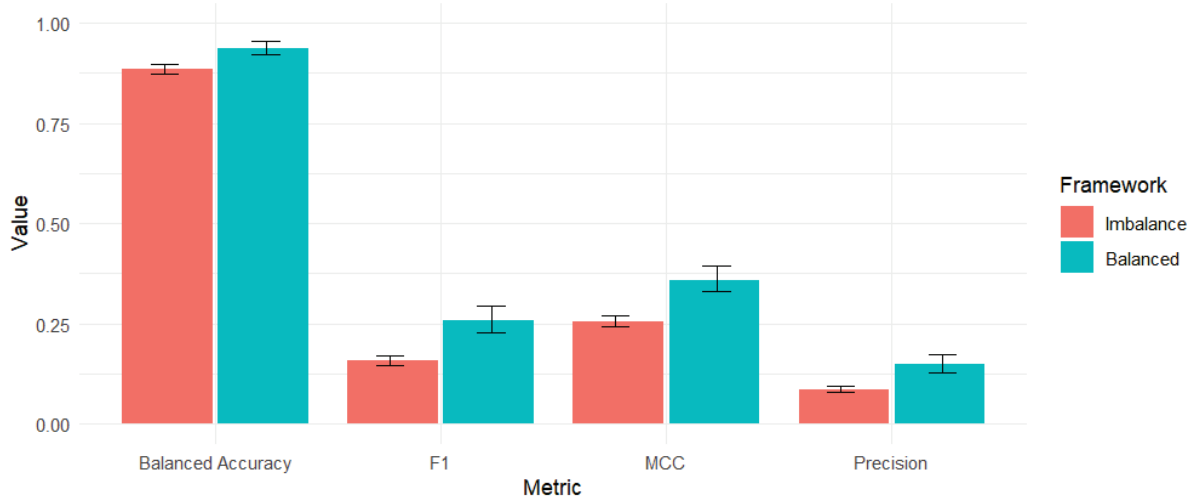


Figure 2. The main performance metrics and their mean and 90% percentile-bootstrap CI for the imbalance and balanced approaches.

In addition, the F1 metric (the harmonic mean between precision and sensitivity) shows a value of 0.257 (+63%) and the balanced accuracy (mean between specificity and sensitivity) shows a value of 0.938 (+6%) for the balanced scenario. Other widely used metrics, such as the area under the curve (AUC, which measures the model's ability to discriminate between classes, 0.98 for both approaches) and the Brier score (mean squared error of probabilistic predictions, 0.02 for both approaches) do not show statistically significant differences between the two frameworks.

4. Conclusion and future directions

This work aimed to contribute to the field of transport engineering by addressing the critical issue of bridge failure risk, a crucial problem for people security. We addressed some challenges identified in previous work of Ventura et al. (2024), which exhibited estimation bias and instability due to the focus on a restricted subsample (frequency failure events ≥ 1) and the lack of robustness of the estimators. To overcome these limitations, we implemented a balanced synthetic sampling approach using SANSA. This technique aimed to generate more robust and reliable results by addressing the class imbalance inherent in the data, improving the model's ability to accurately identify and predict high-severity overload events.

Applying the bootstrap approach, the RGE metric in the field of xAI revealed a consistent ranking of variable importance, with *max_GVM_limit_ratio* emerging as the most critical predictor for assessing severity across both imbalanced and balanced scenarios. Moreover, our findings demonstrate that the SANSA approach yields statistically significant improvements in model performance, particularly in Precision and MCC metrics.

Future research should explore several aspects to further refine and expand this work. One interesting direction is to develop a bivariate *probit* model to consider situations where two binary outcome variables are potentially correlated: it could simultaneously model both frequency and severity, accounting for the potential correlation between these outcomes. Furthermore, the robustness of this result could be enhanced by using updated and more recent data related to the monitoring of vehicles on bridges and others validation designs, for example the out of bag (OoB) validation approach.

Acknowledgements

Thanks to Roberto Ventura and Benedetto Barabino for providing the dataset used in this short paper. This study has been developed for the European Union and Italian Ministry for Universities and Research, National Recovery and Resilience Plan (NRRP), within the Project "Sustainable

Mobility Center (MOST)" 2022-2026, CUP D83C22000690001, Spoke N 7 "CCAM, Connected networks and Smart Infrastructures". Mattia Cefis was supported by the Project "Study of mobile phone siGNals for the evalUation of the interconnections between Mobility and the environment in Lombardia (SIGNUM)" CUP: F53D23010910001 - PRIN 2022 PNRR M4C2 - financed by the European Union - Next Generation EU (DD MUR n. 1409 del 14/09/2022).

References

Aldrich, J.H., Nelson, F.D. (1984). *Linear Probability, Logit, and Probit models*. Number 45. Sage.

Amemiya, T. (1981). Quantitative response models: A survey. *Journal of Economic Literature*, 19(4), pp. 1483-1536.

Bishop, C.M., Nasrabadi, N.M. (2006). *Pattern Recognition and Machine Learning*, Volume 4. Springer.

Chawla, N.V., Bowyer, KV., O'Hall, L., Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, pp. 321-357.

Chicco, D., Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), pp. 1-13.

Giudici, P. (2024). Safe machine learning. *Statistics*, 58(3), pp. 473-477.

Menardi, G., Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1), pp. 92-122.

Nasir, M., Dag, A., Simsek, S., Ivanov, A., Oztekin, A. (2022). Improving imbalanced machine learning with neighborhood-informed synthetic sample placement. *Journal of Management Information Systems*, 39(4), pp. 1116-1145.

Ventura, R., Barabino, B., Maternini, G. (2024). Prediction of the severity of exceeding design traffic loads on highway bridges. *Heliyon*, 10(1), pp. 1-17.