THE BALTIC JOURNAL
OF ROAD AND BRIDGE ENGINEERING
2026/21(1)

ISSN 1822-427X/eISSN 1822-4288
2026 Volume 21 Issue 1: 56–108
https://doi.org/10.7250/bjrbe.2026-21.673

PREDICTION OF ROAD MAINTENANCE AND REPAIR COSTS IN TURKEY USING A HYBRID MODELLING APPROACH

HAYDAR GUNDOGDU*, OMER FARUK CANSIZ, MEHMET FATIH CAN

Graduate Education Institute, Iskenderun Technical University, Iskenderun, Turkey

Received 29 October 2025; accepted 18 February 2026

Abstract. Accurate estimation of road maintenance and repair costs is of strategic importance for the efficient management of public resources and the safety of transportation systems. In Turkey, these costs are influenced by a wide range of multidimensional factors, including meteorological and environmental conditions, infrastructure characteristics, traffic intensity, economic indicators, and financial cost components. This study aims to comprehensively examine these factors and to develop a high-accuracy prediction model for road maintenance and repair costs. A national dataset covering a 19-year period (2004–2022) and comprising 21 independent variables identified through the literature and expert judgement was employed. Methodologically, classical statistical approaches – Multiple Linear Regression, Ridge Regression, Least Absolute Shrinkage and Selection Operator, Stepwise Akaike Information Criterion, and Granger Causality Analysis – were integrated with soft computing techniques, including Random Forest, Gradient Boosting, Support Vector Machines, Artificial Neural Networks, Genetic Algorithms, Principal Component Analysis, and Sensitivity Analysis. In total, ten variable-selection techniques were combined with five prediction models, resulting in 50 hybrid model configurations. The results indicate that the RR-ANN hybrid model, constructed using Freight KM, bitumen and salt consumption and minimum wage variables selected via RR, achieves the highest predictive accuracy

* Corresponding author. E-mail: hdmgundogdu@gmail.com

Haydar GUNDOGDU (ORCID ID 0000-0002-2354-8920)
Omer Faruk CANSIZ (ORCID ID 0000-0001-6857-2513)
Mehmet Fatih CAN (ORCID ID 0000-0002-3866-2419)

(MSE = 175 474.92; RMSE = 418.90; R^2 = 0.985; AdjR^2 = 0.980; MAPE = 1.36%). Computational performance analysis further shows that the trade-off between accuracy and execution time is critical in practical applications: RR-based models are the fastest, whereas ANN-based hybrids provide superior accuracy at the cost of higher computational and implementation effort. Overall, the findings demonstrate that hybrid modelling approaches yield more reliable and robust predictions than single-method specifications. The proposed framework contributes methodologically to the literature and offers policymakers a practical and standardised tool to support budget planning and the development of sustainable road maintenance strategies.

Keywords: artificial neural networks, hybrid modelling, ridge regression, road maintenance costs, soft computing.

Introduction

Road transport is the dominant mode of freight and passenger transportation in Turkey and plays a critical role in supporting economic development and enhancing social welfare. Accordingly, the preservation of road infrastructure and the effective implementation of maintenance and repair activities are of fundamental importance. The maintenance and repair operations conducted by the General Directorate of Highways (KGM) are essential for ensuring the continuity, safety, and long-term sustainability of the national road network (KGM, 2019). In its 2019–2023 Strategic Plan, KGM identified key priorities such as improving road maintenance standards, strengthening bridges through seismic retrofitting, and further developing the Pavement Management System. In addition, operational activities including snow and ice control, post-disaster rehabilitation, and the expansion of intelligent transport systems directly influence the scale and complexity of road maintenance and repair costs (MRC) (KGM, 2023a).

Between 2004 and 2022, KGM's budget accounted for an average of 4.01% of the Central Government Budget (KGM, 2023b). Over the same period, MRC constituted 12.35% of KGM's realised expenditure, while their share in the initial annual budget averaged 25.45%. Moreover, the ratio of KGM's initial budget to realised expenditure reached 112.80%, indicating that planned appropriations were frequently insufficient to meet actual expenditure requirements. This persistent gap necessitated substantial supplementary allocations during the fiscal year and underlines the strategic importance of MRC within both KGM's internal budget structure and the broader framework of central government finances.

In this context, accurate forecasting of MRC is vital for improving fiscal performance and operational efficiency within KGM. To overcome the limitations of traditional cost estimation approaches, which often rely on single-method or purely linear frameworks, this study integrates classical statistical models with advanced soft computing techniques, including artificial neural networks (ANN) and genetic algorithms (GA). Using a comprehensive national dataset covering a 19-year period (2004–2022) and incorporating 21 key independent variables identified through the literature and expert judgement, the study aims to develop a high-accuracy prediction framework for MRC. Recent research increasingly emphasises the importance of standardised analytical frameworks that normalise maintenance costs on a per-kilometre basis, thereby ensuring consistency across analyses and enabling objective prioritisation of asphalt road sections. Such standardisation enhances transparency in decision-making processes and supports evidence-based maintenance planning (Akpan & Morimoto, 2022).

Road maintenance and repair costs are shaped by a complex interaction of infrastructural characteristics, economic conditions, traffic intensity, and environmental factors. Consequently, recent studies demonstrate that multivariate and hybrid modelling approaches outperform single-method specifications by effectively capturing non-linear relationships and interdependencies among explanatory variables (Persyn et al., 2020). Analyses based on large-scale and spatially detailed datasets have further been shown to provide more realistic estimates of transport and maintenance costs, thereby improving the reliability of policy and investment evaluations.

Beyond cost prediction, the use of harmonised maintenance cost indicators enables benchmarking across regions and supports life-cycle assessment (LCA) of alternative pavement maintenance strategies. Contemporary research on asphalt pavement systems highlights that integrating economic and environmental impacts within a unified analytical framework – supported by uncertainty analysis – yields more robust and comparable results (Bressi et al., 2022). In this context, innovative techniques such as full-depth reclamation (FDR) have been reported to offer significant advantages over conventional pavement maintenance methods in terms of both cost efficiency and environmental performance (Torres & Evers, 2024).

Against this background, the primary objective of the present study is to identify the key determinants of MRC in Turkey and to develop a high-accuracy hybrid prediction model. Existing evidence indicates that traffic volume, particularly the effects of overloaded heavy vehicles, significantly accelerates pavement deterioration and increases maintenance expenditure (Kumar & Suman, 2025). Conversely, proactive and timely maintenance interventions have been shown to reduce life-cycle costs and improve long-term pavement performance (Pan et al., 2021; Jasim et al., 2024). Moreover, climate-related factors – such as increasing temperature variability and accelerated asphalt ageing – are expected to intensify maintenance requirements under changing climatic conditions (Zhang et al., 2025). Economic uncertainty, particularly fluctuations in construction material and raw material prices, further contributes to volatility in maintenance and repair budgets (Schmitt, 2025).

To address these multidimensional challenges, this study employs a comprehensive dataset covering the period 2004–2022 and analyses 21 independent variables representing infrastructural features, economic indicators, and environmental conditions. In line with recent methodological advances, classical statistical techniques – Multiple Linear Regression (MLR), Ridge Regression (RR), Least Absolute Shrinkage and Selection Operator (LASSO), Stepwise Akaike Information Criterion (Stepwise AIC), and Granger Causality Analysis (GCA) – are integrated with soft computing methods, including Sensitivity Analysis (SA), Random Forest (RF), Gradient Boosting (GB), Support Vector Machines (SVM), ANN, and GA, within a unified hybrid modelling framework. This approach is adopted to enhance predictive accuracy and model robustness, consistent with best practices reported in the recent literature (Elwahsh et al., 2023).

Within this framework, the study systematically evaluates the predictive performance of multiple hybrid model configurations using a 19-year dataset comprising 21 explanatory variables. The results indicate that integrated hybrid approaches achieve forecast accuracy improvements of approximately 2–30% relative to conventional single-method specifications, as reflected in lower prediction errors and improved goodness-of-fit. Beyond the Turkish case, the proposed framework relies on routinely available transport and economic indicators and is therefore applicable to other countries with comparable infrastructure systems and data environments. Accordingly, the study contributes to the transport economics and forecasting literature by offering methodological insights and policy-relevant evidence to support more reliable budget planning and sustainable long-term maintenance strategies.

1. Methods

1.1. Data Sources and variable classification

This study integrates classical statistical techniques with advanced soft computing methods to predict MRC within Turkey's national road network. The analysis is based on 21 fundamental explanatory variables collected over a 19-year period spanning 2004–2022, with a particular focus on their complex and potentially non-linear relationships with MRC. This section describes the data sources, the classification of variables, and the procedures adopted to ensure a comprehensive and systematic data analysis. The data used in this study were compiled from reliable institutional sources to enable a multidimensional examination of the factors affecting road MRC in Turkey. To facilitate analytical clarity and methodological consistency, the variables were organised into thematic categories reflecting their functional roles in maintenance and repair processes.

Physical infrastructure variables were obtained from official reports of the General Directorate of Highways (KGM, 2023b). These sources provided detailed information on road characteristics, including road types, total road and bridge lengths, traffic density indicators, and the number of personnel involved in maintenance activities. Such variables are essential for capturing the structural and operational dimensions of road maintenance processes.

Economic indicators were derived primarily from data published by the Turkish Statistical Institute (TÜİK) and the Ministry of Environment, Urbanisation and Climate Change (ÇŞB) ((TÜİK, 2024); (ÇŞB, 2023)). Key variables included the inflation rate, minimum wage, and material costs – such as bitumen, salt, and fuel – which play a critical role in determining maintenance expenditures (Birim Fiyat, 2023). In addition, exchange rate data obtained from the Central Bank of the Republic of Turkey (TCMB, 2024) were used to support currency conversions and facilitate comparability in cost assessments.

Meteorological variables were analysed using data from the General Directorate of Meteorology (MGM, 2023). Climatic indicators such as precipitation levels, the number of snow-covered days, temperature, humidity, sunshine duration, and evaporation rates were included to capture both short-term operational impacts and long-term sustainability effects on road maintenance requirements. Demographic variables, including population statistics and labour-related indicators, were sourced from TÜİK (2023). These data contributed to assessing the workforce and social dimensions associated with maintenance activities and resource allocation. By integrating infrastructure, economic, meteorological, and demographic variables, the study establishes a comprehensive representation of the multidimensional drivers influencing road MRC.

For analytical consistency, all variables were classified into four main categories: infrastructure variables (road and bridge characteristics and traffic indicators), economic variables (minimum wage, material costs, and demographic indicators), meteorological variables (climatic conditions), and financial variables (exchange rates used for cost standardisation). This classification enabled the integration of heterogeneous data sources and facilitated clearer interpretation of the relative influence of different factor groups on MRC.

Although the number of explanatory variables (21) is relatively high compared to the length of the annual time series, the potential risk of overfitting is explicitly addressed in this study. Regularised regression techniques such as RR and LASSO are employed to stabilise coefficient estimates under multicollinearity, while dimensionality reduction and systematic variable selection procedures, including PCA, Stepwise AIC, SA, and GCA, are applied to control the effective dimensionality of the models. As a result, not all variables are simultaneously or unrestrictedly included in each specification, ensuring model robustness and preventing over-parameterisation.

1.2. Data processing

In this study, a series of data processing steps were carried out in order to ensure the comparability, accuracy, and suitability of the data for prediction models. These procedures aimed to render the analyses more consistent and meaningful at the international level.

First, normalisation was performed. Regional data were divided by the total road lengths of the respective regions, and standard values per kilometre ($/km) were calculated. This method allowed data from different regions to be compared consistently and increased the accuracy of the analyses. The normalisation process minimised the effect of regional differences and enabled the derivation of more meaningful results.

In the second step, currency conversion was applied. Expenditures reported in Turkish Lira (TL) were converted into United States Dollars ($) using the annual average exchange rates provided by the Central Bank of the Republic of Turkey. This conversion took into account the effect of inflation and made international comparisons possible. In this way, cost analyses were evaluated from a global perspective.

Finally, the dataset was structured in a time-series format. The dataset covering the years 2004–2022 was organised in such a way as to allow the identification of temporal trends and the relationships among variables. The time-series structure enabled the analysis of long-term changes in costs and the time-dependent effects of factors. This process constituted the fundamental framework required for the development of prediction models.

These procedures optimised the accuracy and reliability of the data used in the study, while also making possible comparative analyses under different temporal and regional conditions. These pre-processing steps were of critical importance for preserving data integrity, improving the accuracy of prediction models, and ensuring that the analyses met international research standards. The study contained 21 independent variables and one dependent variable, identified through expert opinion and literature review. These variables reflected the objectives of the research and guaranteed methodological robustness. The systematically presented variables in Tables 1 and 2 provided a basis for replicability and clarity in future studies.

In Tables 1 and 2, the independent and dependent variables encompassed various factors affecting the maintenance and repair processes of highways. These variables could be detailed under different thematic groups to explain their relationships. The groups and the effects of the variables were examined in detail below.

Meteorological and Environmental Factors: Meteorological and environmental conditions directly influence pavement performance and, consequently, highway maintenance and repair costs. Snow Covered Days (*X1*) capture winter severity and are closely associated with snow removal, salting, and anti-icing operations, which increase maintenance expenditures. Total Average Precipitation (*X2*) reflects annual rainfall levels; excessive precipitation facilitates water infiltration into pavement layers, accelerating surface deformation. Average Temperature (*X3*) affects asphalt behaviour through thermal expansion and contraction, while extreme temperatures may lead to cracking or rutting. Average Humidity (*X4*) influences pavement ageing and durability, particularly in regions with persistently high moisture levels. Average Sunshine Duration (*X5*) represents exposure to ultraviolet radiation, which contributes to the chemical degradation of asphalt surfaces. Finally, Average Evaporation (*X6*) indicates surface moisture dynamics; rapid evaporation may induce surface drying and cracking. Together, these variables represent climatic stresses affecting road longevity and maintenance demand.

Infrastructure Characteristics: Infrastructure-related variables describe the physical structure and scale of the highway network, which fundamentally determine maintenance requirements. Stabilised Road Length (*X7*), although initially cost-efficient, typically requires more frequent maintenance over time. Asphalt Road Length (*X8*) represents the extent of paved infrastructure that demands regular upkeep to preserve serviceability. Surface-Treated Road Length (*X9*), while economically advantageous, exhibits lower durability compared to asphalt pavements. Bridge Total Length (*X10*) reflects the scale of bridge infrastructure, where maintenance costs depend not only on length but also on structural design and material properties. These variables collectively capture the heterogeneity of road assets and their associated maintenance needs**.**

Traffic and Economic Factors: Traffic intensity and broader economic conditions jointly influence road deterioration and maintenance costs. Vehicle KM travelled (*X11*) measure overall traffic volume, with higher values accelerating pavement wear. Passenger KM (*X12*) indicate usage intensity associated with passenger transport demand, while Freight KM (*X13*) capture heavy-load movements that exert substantial stress on pavement structures. The number of highway personnel (*X14*) reflects workforce capacity; while higher staffing levels enhance maintenance capability, they also increase labour expenditures. Economic demand pressures are further reflected by population size (*X15*) and total vehicle stock (*X16*), both of which contribute to increased road usage. The consumer price index (CPI) (*X17*) serves as a proxy for inflation, affecting the prices of materials and services required for maintenance activities.

Financial and Cost-Related Factors: Financial variables represent labour and material cost components directly associated with maintenance and repair operations. Minimum Wage levels (*X18*) directly influence personnel costs and overall maintenance budgets. Material-related variables include bitumen consumption (*X19*), which affects pavement durability and resurfacing costs; fuel consumption ($X20$), representing operational and equipment-related expenses; and salt consumption ($X21$), which is essential for winter maintenance but may contribute to long-term pavement deterioration. These variables collectively capture the cost structure underlying highway maintenance and repair activities.

Dependent Variable: MRC (Y) expressed the MRC per kilometre and was the dependent variable of this study. This variable was a critical indicator for the sustainable management of road infrastructure. Each of the environmental, infrastructural, traffic intensity, human resources, economic, and functional factors could have varying levels of impact on these costs. In the literature, such analyses were important for optimising maintenance budgets and ensuring the efficient use of resources.

Table 1. Independent and dependent variables (2004–2013)

Codes	Types		Year									
			2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
X1	Meteorological and Environmental Factors	Snow Covered Days	36	25.4	31.3	25.9	31.9	19.5	12.7	20.2	39.9	23.8
X2		Total Average Precipitation, mm	607.4	637.2	607.4	596.7	493.1	793.8	703	642.2	695.2	547
X3		Average Temperature, °C	13.2	13.3	13.3	13.8	13.6	13.7	15.1	12.8	13.8	13.8
X4		Average Humidity, %	62.2	63.2	63.6	61.3	61	63.8	62.9	63.1	62.1	59.6
X5		Average Sunshine Duration, h	6.9	6.7	6.8	6.8	6.9	6.5	6.5	6.7	6.8	6.8
X6		Average Evaporation, mm	6	5.9	6.2	6.6	6.4	6	6	6	6.2	6.1
X7	Infrastructure Characteristics	Stabilised Road Length, km	2236.00	2207.00	2132.00	1796.00	1600.00	1490.00	1314.00	1077.00	1069.00	852
X8		Asphalt Road Length, km	7030.00	7080.00	7204.00	7406.00	8004.00	8681.00	10 197.00	11 561.00	13 150.00	14 870.00
X9		Surface Treated Road Length, km	50461.00	50 302.00	50 159.00	50 619.00	50 305.00	49 782.00	48 929.00	47 912.00	46 462.00	45 294.00
X10		Bridge Total Length, km	227.14	233.32	237.21	247.09	255.65	275.77	296.31	315.18	334.35	350.44
X11		Vehicle KM	57767.00	61 129.00	64 577.00	69 609.00	69 771.00	72 432.00	80 124.00	85 495.00	93 989.00	99 431.00
X12		Passenger KM	174312.00	182 152.00	187 593.00	209 115.00	206 098.00	212 464.00	226 913.00	242 265.00	258 874.00	268 178.00
X13		Freight KM	156853.00	166 831.00	177 399.00	181 330.00	181 935.00	176 455.00	190 365.00	203 072.00	216 123.00	224 048.00

Codes	Types		Year									
			2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
X14	Traffic and Economic Factors	Highway Personnel Count	7360.00	6741.00	6429.00	6349.00	5734.00	5480.00	5222.00	5171.00	5647.00	4830.00
X15		Population	68 010 215.00	68 860 539.00	69 729 967.00	70 586 256.00	71 517 100.00	72 561 312.00	73 722 988.00	74 724 269.00	75 627 384.00	76 667 864.00
X16		Total Vehicle Count	10 236 357.00	11 145 826.00	12 227 393.00	13 022 945.00	13 765 395.00	14 316 700.00	15 095 603.00	16 089 528.00	17 033 413.00	17 939 447.00
X17		Consumer Price Index	113.86	122.65	134.49	145.77	160.44	170.91	181.85	200.85	213.23	229.01
X18	Financial and Cost-Related Factors	Minimum Wage, $	246.58	283.93	293.11	387.12	423.63	389.04	439.3	442.99	448.98	468.96
X19		Bitumen Consumption, kg/km	1492.38	1981.36	2179.17	2661.67	2749.92	2559.82	2422.74	2549.73	2699.08	2109.19
X20		Fuel Consumption, kg/km	141.02	133.75	152.55	134.56	153.84	168.07	155.21	167.36	191.68	82.62
X21		Salt Consumption, kg/km	45.23	34.38	51.24	44.72	28.7	46.51	53.36	60.34	129.53	79.75
Y	Dependent Variable	MRC, $/km	5251.41	7573.13	7950.35	9506.15	10 622.48	10 298.70	14 769.33	15 259.88	16 027.93	17 755.26

Table 2. Independent and dependent variables (2014–2022)

Codes	Types		Year								
			2014	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
X1	Meteorological and Environmental Factors	Snow Covered Days	11.2	25.5	29.5	25.8	9.4	17.6	39.4	18.1	45.7
X2		Total Average Precipitation, mm	641.6	637.8	605.7	536.4	639.2	639.7	507.6	579.7	530.1
X3		Average Temperature, °C	14.5	13.8	14	13.7	15.1	14.4	14.6	14.5	14.1
X4		Average Humidity, %	62.6	62.4	61.1	61.5	64.4	63.4	61.6	60.6	62.2
X5		Average Sunshine Duration, h	6.5	6.5	6.6	6.5	6.1	6.6	6	6.5	6.5
X6		Average Evaporation, mm	5.6	5.7	5.9	6.1	5.9	6.1	6.2	6.5	6
X7	Infrastructure Characteristics	Stabilised Road Length, km	891	744	593	668	564	480	344	325	333
X8		Asphalt Road Length, km	15 922.00	17 095.00	18 646.00	20 793.00	21 923.00	22 680.00	23 707.00	24 774.00	25 545.00
X9		Surface Treated Road Length, km	44 277.00	43 726.00	42 131.00	40 183.00	39 333.00	38 817.00	37 922.00	36 887.00	36 184.00
X10		Bridge Total Length, km	375.73	408.4	424.95	453.4	470.82	501.25	508.82	481.11	490.29
X11		Vehicle KM	102 988.00	113 274.00	119 671.00	127 997.00	131 625.00	135 485.00	126 053.00	142 479.00	140 531.00
X12		Passenger KM	276 073.00	290 734.00	300 852.00	314 734.00	329 363.00	339 601.00	288 992.00	336 188.00	348 489.00
X13		Freight KM	234 492.00	244 329.00	253 139.00	262 739.00	266 502.00	267 579.00	272 913.00	311 818.00	323 512.00

Codes	Types		Year								
			2014	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
X14	Traffic and Economic Factors	Highway Personnel Count	4478.00	5971.00	9341.00	9571.00	9380.00	9187.00	8883.00	8618.00	8416.00
X15		Population	77 695 904.00	78 741 053.00	79 814 871.00	80 810 525.00	82 003 882.00	83 154 997.00	83 614 362.00	84 680 273.00	85 279 553.00
X16		Total Vehicle Count	18 828 721.00	19 994 472.00	21 090 424.00	22 218 945.00	22 865 921.00	23 156 975.00	24 144 857.00	25 249 119.00	26 482 847.00
X17		Consumer Price Index	247.72	269.54	292.54	327.41	393.88	440.5	504.81	686.95	1128.45
X18	Financial and Cost-Related Factors	Minimum Wage, $	458.96	478.31	464.92	440.42	420.15	410	403.7	480.07	514.01
X19		Bitumen Consumption, kg/km	1390.34	2503.63	3267.02	1290.81	990.45	830.79	708.81	1905.84	1699.64
X20		Fuel Consumption, kg/km	50.45	81.49	125.7	127.67	132.87	155	115.14	126	224.42
X21		Salt Consumption, kg/km	49.44	123.31	182.94	160.51	157.93	155.97	153.59	400.36	599.05
Y	Dependent Variable	MRC, $/km	15 417.89	14 429.04	15 115.89	16 923.63	15 252.60	14 685.71	13 487.29	13 468.17	13 610.69

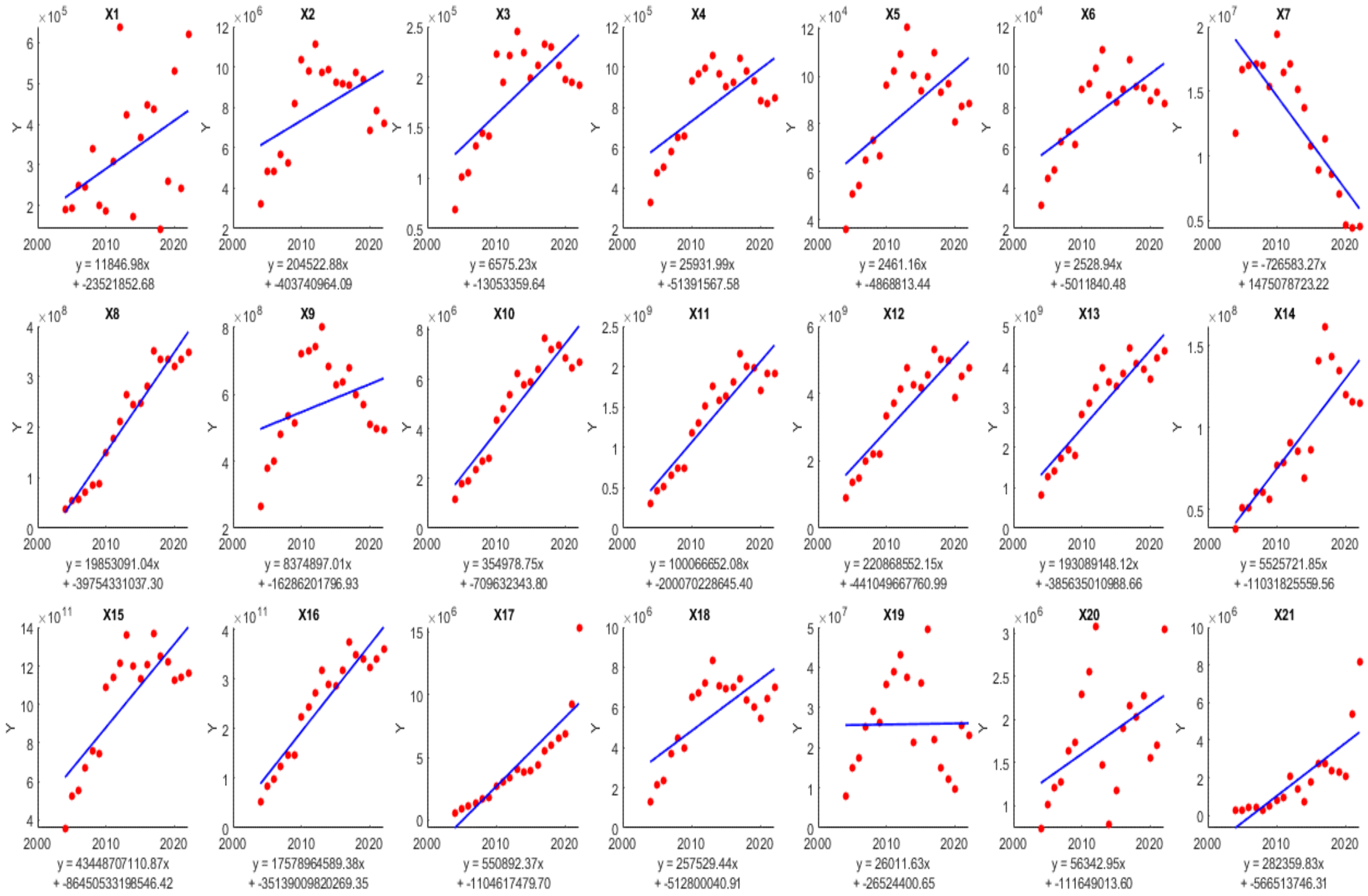

Figure 1. Distribution of the independent variable according to the independent variables

This classification and explanation examined in detail the effects of independent variables on maintenance and repair costs. In the literature, analysing and prioritising the interactions of these factors was of importance for sustainable infrastructure management. Within this framework, the methodology to be employed in the study could reveal the extent of these effects.

1.3. Data evaluation

During the data evaluation stage, the dataset was examined to verify its suitability for statistical analysis using standard diagnostic procedures, including visual inspection, correlation analysis, normality checks, and multicollinearity tests. Scatter plots were used to assess the relationships between the explanatory variables and maintenance and repair costs (MRC). For clarity, the variables were classified into four groups.

- **Meteorological and Environmental Factors:** This group comprises Snow Covered Days, Total Average Precipitation, Average Temperature, humidity, sunshine duration, and evaporation. Snow and precipitation were positively associated with MRC, reflecting increased maintenance requirements, while humidity also contributed to pavement deterioration. The remaining variables exhibited weaker effects.
- **Infrastructure Characteristics:** Infrastructure variables include the lengths of stabilised, asphalt, and surface-dressed roads, together with total bridge length. Greater network extent and bridge infrastructure were associated with higher maintenance costs.
- **Traffic and Economic Factors:** This category covers traffic volumes, population, vehicle ownership, and the consumer price index. Vehicle KM travelled showed the strongest association with MRC, while population and vehicle growth were also positively related. CPI exhibited a non-linear influence.
- **Financial and Cost-Related Factors:** This group includes labour and material inputs, namely highway personnel, minimum wage, and the consumption of bitumen, fuel, and salt. Bitumen and salt consumption were strongly linked to MRC, while fuel consumption also showed a positive relationship. All these evaluations broadened the scope of the study and enabled a clearer understanding of the effects of variables on MRC. The relationships among the data allowed the prediction models to produce more accurate and effective results.

This evaluation process identified the key variables influencing MRC and established a strong foundation for the development of prediction models. The distributions and relationships of the variables were visualised to support the linear modelling approach used in the study (see Figure 1). These visualisations contributed to the systematic analysis of the relationships among variables and to the improvement of model accuracy.

1.3.1. Correlation analysis

Before proceeding to regression analysis, evaluating the relationships among the independent variables was of critical importance to enhance the accuracy and reliability of the model. For this purpose, a correlation analysis was conducted using the Pearson correlation coefficient to assess the linear relationships among the variables.

Although the Pearson correlation coefficient measured the linear relationships between independent variables effectively, it remained insufficient in identifying non-linear relationships. This limitation created the risk of overlooking complex interactions among variables. In particular, for a broader understanding of the effects of the independent variables on MRC, the detection and analysis of non-linear relationships were important.

High correlations among independent variables could cause multicollinearity. Multicollinearity weakened the sensitivity of the coefficients used in regression models and adversely affected predictive accuracy. Such situations could lead to misleading model results and a decline in forecasting performance.

Therefore, it was necessary to employ alternative methods to control the effects of high correlations among independent variables and to improve the reliability of the models. For example, to reduce multicollinearity, variable selection, dimension reduction methods such as PCA, or statistical techniques aimed at improving regression models could be adopted (Li et al., 2023). These approaches contributed to structuring prediction models on a more robust basis.

Figure 2 visualises the Pearson correlations, highlighting the strong relationships among variables. These strong correlations necessitated the implementation of appropriate adjustments to reduce the risk of multicollinearity and enhance model performance. Such adjustments included variable selection or transformation techniques. The results of the correlation analysis made it possible to use the variables more effectively in the modelling process and improve the accuracy of prediction models.

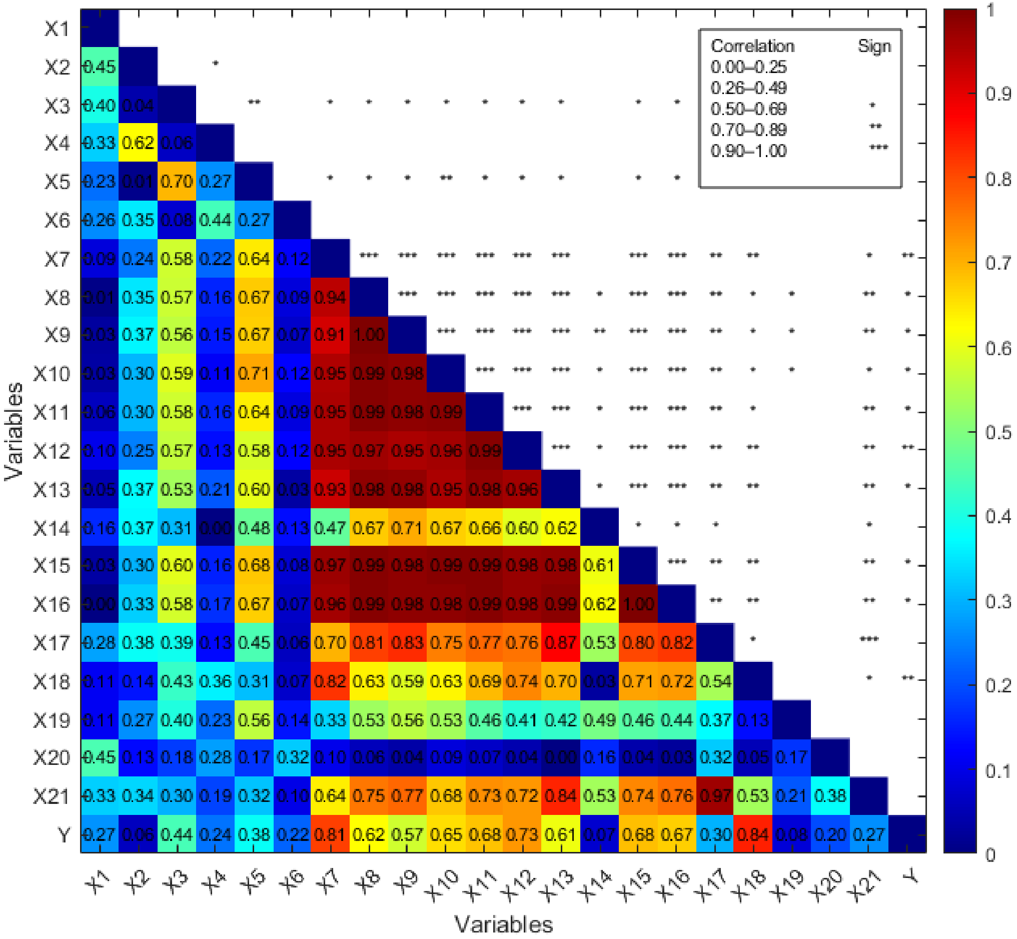

Figure 2. Pearson correlations

Examining the linear relationships between MRC and the independent variables was a critical stage for increasing the accuracy of prediction models and optimising the modelling process. The analysis based on the Pearson correlation coefficient revealed the relationships among the variables and assessed the risk of multicollinearity. The findings from this analysis are presented comprehensively below.

The number of Snow Covered Days showed a significant effect on MRC and exhibited a strong positive correlation. An increase in Snow Covered Days led to intensified functional activities such as snow clearance, salting, and anti-icing. It was observed that, particularly in the winter months, a large proportion of maintenance budgets was allocated to these activities. Similarly, a strong positive relationship was also found between Total Average Precipitation and costs. Precipitation caused deterioration in the physical structure of roads, leading to more frequent maintenance requirements, particularly for stabilised and asphalt roads.

The Average Humidity rate had a moderate positive effect on costs. High humidity weakened the surface materials and infrastructure of roads, increasing structural damage. By contrast, the effects of variables such as Average Temperature and evaporation remained minimal on maintenance costs. Sunshine duration also showed only a weak relationship with maintenance costs, indicating that the indirect effects of these variables were limited.

Infrastructure variables demonstrated notable effects on maintenance and repair costs. The kilometre length of stabilised roads exhibited a significant positive relationship with costs. The low structural durability of stabilised roads required them to undergo maintenance on a regular basis. Asphalt roads and surface-dressed roads also showed positive relationships with costs; however, both road types were identified as requiring less maintenance compared to stabilised roads. The total length of bridges had one of the highest correlation values. The complex engineering structures of bridges and their high maintenance requirements explained this strong relationship. Bridges, in particular, accounted for a considerable share of infrastructure budgets in both regular maintenance and repair activities.

Traffic data emerged as one of the most important determinants of maintenance costs. Vehicle KM exhibited a strong positive correlation with maintenance costs. Traffic density accelerated the wear of roads, leading to frequent maintenance needs. Passenger KM and Freight KM also showed positive relationships with costs, but these effects were weaker compared to Vehicle KM. These findings clearly demonstrated the impact of traffic density on the durability of road infrastructure and emphasised that maintenance planning should be carefully conducted in regions with high traffic density.

Economic indicators exhibited both direct and indirect effects on maintenance costs. The population and the total number of vehicles showed strong positive relationships with costs. More densely populated areas and higher vehicle numbers increased maintenance costs due to more frequent road use. CPI showed a non-linear relationship with costs. CPI was observed to affect budget allocations and resource use depending on economic conditions. The effect of the minimum wage on maintenance costs was found to be weak. This indicated that labour costs constituted only a small share of total maintenance expenditure.

Bitumen Consumption showed a strong positive correlation with maintenance and repair costs. As bitumen was widely used as a road surfacing material, it directly influenced costs. Salt consumption also showed a significant effect on maintenance costs. It was found that the salt used for snow clearance and anti-icing activities accounted for a large portion of costs during the winter months. Fuel consumption showed a moderate positive relationship. Fuel was a fundamental cost component in operating maintenance equipment.

The correlation analysis clearly revealed the critical variables affecting maintenance and repair costs. Variables such as the number of Snow Covered Days, Total Average Precipitation, Vehicle KM, and Bridge Total Length exhibited strong positive relationships with maintenance costs and should be prioritised in the modelling process. However, in order to reduce the risk of multicollinearity, it was recommended that a selection be made among highly correlated variables or that these variables be restructured using transformation methods.

This analysis provided an important foundation for increasing the accuracy of prediction models and for optimising maintenance budgets. The detailed examination of variables deepened the understanding of the key factors influencing maintenance costs and offered a strategic guide for future planning efforts.

1.3.2. Normality assessment of dependent and independent variables

Normality assessment is a fundamental aspect of regression analysis, ensuring the residuals of the dependent variable align with a normal distribution. This assumption underpins the validity and reliability of predictive models, enhancing their utility for inference and decision-making. In this study, normality was rigorously evaluated using statistical tests and graphical methods, offering a detailed understanding of variable distributions (Mohd Razali & Yap, 2011).

- Shapiro-Wilk Test (KS): Particularly effective for small to medium-sized datasets, this test is highly sensitive to deviations from normality, evaluating whether sample distributions significantly diverge from normality (Shapiro & Wilk, 1965).
- Kolmogorov-Smirnov Test (SW): Ideal for larger datasets, it compares sample distributions against theoretical normal distributions, detecting subtle deviations and ensuring data integrity (Massey, 1951).

Table 3. Normality assumption results

Variables	Kolmogorov-Smirnov Test			Shapiro-Wilk Test			Skewness/ Std. Error	Kurtosis/Std. Error	Results
	Test Statistic	df	*p*-value	Test Statistic	df	*p*-value			
Snow Covered Days	0.125	19	0.200*	0.975	19	0.864	0.454	−0.510	Close to Normal
Total Average Precipitation, mm	0.185	19	0.086	0.950	19	0.389	0.963	0.843	Close to Normal
Average Temperature, °C	0.176	19	0.125	0.960	19	0.571	0.546	−0.287	Close to Normal
Average Humidity, %	0.085	19	0.200*	0.987	19	0.994	−0.523	−0.246	Normal
Average Sunshine Duration, h	0.249	19	0.003	0.872	19	0.016	–1.945	1.203	Significant Deviation
Average Evaporation, mm	0.147	19	0.200*	0.953	19	0.443	0.693	0.487	Close to Normal
Stabilised Road Length, km	0.147	19	0.200*	0.904	19	0.057	1.142	−0.915	Slight Deviation
Asphalt Road Length, km	0.145	19	0.200*	0.898	19	0.045	0.380	–1.510	Slight Deviation
Surface Treated Road Length, km	0.159	19	0.200*	0.885	19	0.027	−0.623	–1.459	Deviation from Normality
Bridge Total Length, km	0.129	19	0.200*	0.903	19	0.056	0.167	–1.576	Close to Normal
Vehicle KM	0.139	19	0.200*	0.917	19	0.101	0.077	–1.537	Close to Normal
Passenger KM	0.126	19	0.200*	0.940	19	0.264	−0.095	–1.329	Close to Normal
Freight KM	0.139	19	0.200*	0.944	19	0.310	0.739	−0.802	Close to Normal
Highway Personnel Count	0.159	19	0.200*	0.896	19	0.042	0.468	–1.553	Slight Deviation
Population	0.090	19	0.200*	0.949	19	0.377	0.001	–1.288	Close to Normal
Total Vehicle Count	0.107	19	0.200*	0.958	19	0.539	0.109	–1.228	Normal
Consumer Price Index	0.219	19	0.017	0.737	19	–	4.442	6.107	Significant Deviation
Minimum Wage, $	0.187	19	0.080	0.872	19	0.016	–2.339	0.960	Significant Deviation
Bitumen Consumption, kg/km	0.141	19	0.200*	0.956	19	0.493	−0.581	−0.826	Close to Normal
Fuel Consumption, kg/km	0.169	19	0.161	0.961	19	0.594	−0.302	1.021	Normal
Salt Consumption, kg/km	0.270	19	0.001	0.684	19	–	4.614	6.166	Significant Deviation
MRC, $/km	0.235	19	0.007	0.903	19	0.055	–1.668	−0.160	Slight Deviation

Skewness and Kurtosis metrics were also analysed to assess distribution symmetry and peakiness. Variables exceeding twice the standard error were deemed non-normal and transformed accordingly (Kim, 2013). Together, these metrics provide nuanced insights into variable behaviour, preparing the data for robust regression analysis.

By combining statistical tests with graphical evaluations, this study ensures a thorough assessment of normality. Non-normal variables were transformed, ensuring adherence to regression assumptions and enhancing the model's predictive power.

Table 3 summarises the normality assessment for MRC and 21 independent variables, evaluated through KS and SW tests, alongside skewness and kurtosis metrics.

Significant deviations were identified in variables such as the CPI, Salt Consumption, Minimum Wage, and Average Sunshine Duration, with low p-values ($p < 0.05$) and high skewness/kurtosis. For instance, CPI had a skewness of 4.442 and kurtosis of 6.107, indicating a highly skewed distribution, necessitating transformations like logarithmic or Box-Cox methods.

In contrast, variables like Snow Covered Days, Total Average Precipitation, Average Temperature, Vehicle KM, and Freight KM aligned well with normality ($p > 0.05$, skewness/kurtosis near zero). For example, Average Temperature showed a skewness of 0.546 and kurtosis of −0.287, requiring no adjustment.

Variables such as Stabilised Road Length, Asphalt Road Length, and MRC showed slight deviations, with borderline *p*-values and moderate skewness/kurtosis. For example, MRC had a KS *p*-value of 0.007, an SW *p*-value of 0.055, and a skewness of −1.668, suggesting minor transformations may improve suitability.

Highly normal variables, including Humidity, Vehicle Count, and Fuel Consumption, exhibited *p*-values > 0.05 and skewness/kurtosis near zero, requiring no further adjustments.

In summary, variables like CPI, Salt Consumption, and Minimum Wage need transformation due to significant deviations, while most others align well with normality or require minimal adjustment, ensuring robust regression analysis.

1.3.3. Multicollinearity detection

Multicollinearity was assessed using Variance Inflation Factor (VIF) and tolerance values. As shown in Table 4, VIF and tolerance metrics were calculated, with variables exhibiting VIF > 10 or Tolerance < 0.1 flagged as having severe multicollinearity (Kutner et al., 2005). These calculations identified redundancies among the 21 independent variables and their relationships with the dependent variable, MRC. Resolving multicollinearity improves variable selection, enhances model reliability, and ensures valid regression coefficients, contributing to a robust and interpretable regression model.

The analysis of Tolerance and VIF values revealed significant multicollinearity among independent variables, potentially undermining regression model reliability. Variables like Asphalt Road Length (VIF = 6943.531, Tolerance = 0.000), Vehicle KM (VIF = 3849.151, Tolerance = 0.000), and Freight KM (VIF = 1699.578, Tolerance = 0.001) exhibited severe redundancy. Similarly, Surface-Treated Road Length and Population showed very low Tolerance values and high VIF scores, indicating dependency on other predictors.

Moderate multicollinearity was noted in variables like Average Humidity (VIF = 125.301, Tolerance = 0.008) and Average Sunshine Duration (VIF = 35.279, Tolerance = 0.028), while Average Temperature (VIF = 4.097, Tolerance = 0.244) posed minimal concerns. In contrast, variables such as Bitumen, Fuel, and Salt Consumption, with zero Tolerance and VIF values, indicated no variability and unsuitability for inclusion. Addressing these multicollinearity issues is critical for enhancing model stability and reliability.

Table 4. VIF and tolerances

No	Variable	Tolerance	VIF
X1	Snow Covered Days	0.089	11 286
X2	Total Average Precipitation, mm	0.065	15 405
X3	Average Temperature, °C	0.244	4097
X4	Average Humidity, %	0.008	125 301
X5	Average Sunshine Duration, h	0.028	35 279
X6	Average Evaporation, mm	0.029	34 590
X7	Stabilised Road Length, km	0.003	374 557
X8	Asphalt Road Length, km	–	6 943 531
X9	Surface Treated Road Length, km	0.002	635 880
X10	Bridge Total Length, km	0.003	325 152
X11	Vehicle KM	–	3 849 151
X12	Passenger KM	0.018	56 447
X13	Freight KM	0.001	1 699 578
X14	Highway Personnel Count	0.006	177 829
X15	Population	0.002	510 875
X16	Total Vehicle Count	0.028	36 134
X17	Consumer Price Index	0.013	78 770
X18	Minimum Wage, $	0.002	458 042
X19	Bitumen Consumption, kg/km	–	–
X20	Fuel Consumption, kg/km	–	–
X21	Salt Consumption, kg/km	–	–
Y	MRC, $/km		

1.3.4. Descriptive statistics

Descriptive statistics provided insights into variable distributions, guiding necessary data transformations. As shown in Table 5, key metrics such as minimum, maximum, mean, standard deviation, skewness, and kurtosis were analysed to understand the characteristics of the data. Notable findings from these metrics highlighted critical patterns and outliers, informing subsequent analysis and model adjustments.

- Skewness: Salt Consumption displayed significant positive skewness, while other variables were within acceptable limits.
- Kurtosis: A leptokurtic distribution was observed in Salt Consumption (kurtosis > 4), indicating heavy tails.
- Variability: High standard deviations in variables like Bitumen Consumption highlighted substantial interannual variability.

These findings underscored the need for data transformations, supporting robust regression model development.

The normality of variables was assessed by comparing skewness and kurtosis values against their standard errors, flagging deviations exceedingly twice the standard error for transformation (Tabachnick & Fidell, 2007). The CPI and Salt Consumption required transformations to improve model accuracy.

High correlations, such as those between Asphalt Road Length and Total Vehicle Count, indicated potential outliers that were addressed through transformations and dimensionality reduction techniques like PCA.

Multicollinearity was managed by removing redundant variables, consolidating them into composite measures, or using regularisation techniques such as RR and Lasso. Variables with low Tolerance (<0.10) and high VIF values were iteratively adjusted to ensure stability.

Advanced reduction methods, including Stepwise AIC, RF, and GA, optimised predictors while maintaining interpretability. Combined with scatter plots, normality tests, and VIF evaluations, these approaches produced a robust, generalisable regression model for MRC from 2004 to 2022.

Table 5. Descriptive statistics of variables

Variable	*N*	Min	Max	Mean	Std. Dev	Skewness	Skew Std. Error	Kurtosis	Kurtosis Std. Error
Snow Covered Days	19	9.40	45.70	27.18	10.92	0.21	0.52	–1.25	1.01
Total Average Precipitation, mm	19	493.10	793.80	609.28	86.34	0.52	0.52	–0.48	1.01
Average Temperature, °C	19	12.70	15.10	13.80	0.71	0.56	0.52	–1.17	1.01
Average Humidity, %	19	59.60	64.40	62.13	1.40	0.10	0.52	–0.92	1.01
Average Sunshine Duration, h	19	6.00	6.90	6.60	0.27	–0.64	0.52	–0.87	1.01
Average Evaporation, mm	19	5.60	6.60	6.09	0.31	0.31	0.52	–0.99	1.01
Stabilised Road Length, km	19	325.00	2236.00	1066.78	664.76	0.52	0.52	–1.18	1.01
Asphalt Road Length, km	19	7030.00	25 545.00	15 094.32	5725.56	0.45	0.52	–1.08	1.01
Surface Treated Road Length, km	19	36 184.00	50 461.00	44 207.26	4192.98	0.58	0.52	–0.90	1.01
Bridge Total Length, km	19	227.10	490.30	352.41	83.58	0.26	0.52	–1.14	1.01
Vehicle KM	19	57 767.00	140 531.00	97 395.95	28 190.44	0.25	0.52	–1.31	1.01
Passenger KM	19	174 312.00	348 489.00	263 257.16	59 675.82	0.03	0.52	–1.55	1.01
Freight KM	19	156 853.00	323 512.00	232 276.26	48 286.21	0.11	0.52	–1.49	1.01
Highway Personnel Count	19	3250.00	9571.00	6902.79	1997.11	0.40	0.52	–1.12	1.01
Population	19	68 010 215.00	85 279 553.00	76 611 005.10	5 925 043.71	0.22	0.52	–1.45	1.01
Total Vehicle Count	19	10 236 357.00	26 482 847.00	18 539 145.70	5 272 267.74	0.17	0.52	–1.49	1.01
Consumer Price Index	19	113.90	1128.50	338.33	311.65	1.18	0.52	0.34	1.01
Minimum Wage, $	19	246.60	514.00	414.99	71.73	–0.06	0.52	–1.14	1.01
Bitumen Consumption, kg/km	19	708.80	3267.00	1923.91	849.78	0.56	0.52	–0.82	1.01
Fuel Consumption, kg/km	19	50.40	224.40	141.15	49.52	0.05	0.52	–1.14	1.01
Salt Consumption, kg/km	19	28.70	599.10	144.81	157.88	2.04	0.52	4.39	1.01
MRC, $/km	19	5251.40	17 755.30	12 271.20	3800.95	0.23	0.52	–1.10	1.01

1.4. Variable reduction methods

Variable reduction procedures were carried out through various methods that were widely accepted in the literature and had strong analytical foundations. In this context, the ability of each method to determine the importance of independent variables was carefully evaluated, and LASSO, RR, RF, SA, GA, PCA, GCA, GA, SVM, and Stepwise AIC methods were applied. In this process, each method was meticulously designed to minimise the effect of multicollinearity among the independent variables in the dataset and to include only the most significant variables in the modelling process.

As a result of the analysis, each method was independently evaluated, and groups of variables containing up to four independent variables were formed. These groups were compared in detail according to the analytical performance and selection criteria of the methods. Figures 3–12 present these groups of up to four variables identified as the most important by the selection methods.

1.5. Modelling methods

Following the identification of variable groups, prediction models were constructed using these groups. These models were implemented through LASSO, RR, RF, MLR, and ANN algorithms. However, during the analysis process, reapplication of variable reduction within LASSO, RR, and RF methods during modelling was prevented, and the variables identified by each method were kept fixed. Each model was combined with different groups of variables, and in total, fifty hybrid models were designed and analysed using MATLAB software.

This study utilised various analysis methods to predict the dependent variable (Y) using groups of independent variables (X), each containing up to four variables selected through different techniques. These methods reduced dimensionality and ensured that relevant predictors were chosen for effective modelling and forecasting. The primary analysis methods were:

- **LASSO:** Optimised the penalty parameter (λ) using 10-fold cross-validation, minimised the Mean Squared Error (MSE), and eliminated irrelevant predictors. The selected variables addressed multicollinearity and simplified the model.
- **RR:** Applied an L2 penalty to shrink coefficients without excluding variables, ensuring stability with highly correlated predictors through a grid search for the optimal λ.
- **RF:** Modelled non-linear interactions and ranked variable importance using 200 decision trees and Out-of-Bag (OOB) error estimates, effectively capturing dynamic relationships.

- **MLR:** Served as a baseline but faced challenges with linearity and residual normality.
- **ANN:** Analysed each group with 1–10 hidden neurons and four various activation functions such as compet, tansig, purelin, and logsig and using seven different training algorithms (trainlm, trainbfg, trainoss, trainrp, traingd, trainscg, trainbr), ANN minimised MSE and excelled in capturing non-linear patterns despite higher computational costs.

As a result, the effectiveness of different variable reduction and prediction methods was evaluated within the scope of the study, and the impact of addressing multicollinearity problems on model performance was examined in detail. This section presented comparative analyses of the findings and highlighted the prominent aspects of different methods. Figures 13–22 show the performance graphs of prediction analyses, comparing the predicted values of each method with the actual values of the dependent variable.

2. Results and discussion

2.1. Variable reduction

This study employed multiple variable selection methods, which were widely accepted in the literature and had strong analytical foundations, to identify the most important independent variables and improve model performance, while also addressing issues of overfitting and multicollinearity.

In this study, ten different statistical and machine learning-based variable selection methods were applied to reveal the relationships between the independent variables (X) and the dependent variable (Y) in the most meaningful way. All methods were executed on normalised datasets (X_norm, Y_norm), and at the end of each method, up to four variables with the highest importance were reported. In cases of ties, ranking rules within the methods were taken into account.

First, the LASSO (L1 penalty) method was applied using 10-fold cross-validation, and coefficient selection was carried out at the IndexMinMSE point, whereby irrelevant variables were shrunk to zero, and strong linear relationships were highlighted. The RR (L2 penalty) approach selected variables based on coefficient magnitudes, identifying stable and widely effective predictors under multicollinearity conditions. The Stepwise AIC method formed more parsimonious linear sub-models by forward and backward selection steps to minimise the Akaike Information Criterion. SA selected variables with the highest individual linear correlations with Y, providing a simple and explanatory pre-screening.

Among machine learning-based methods, RF employed an ensemble model of 100 trees, ranking variable importance through out-of-bag permutation error contributions and capturing non-linear interactions. GBB (LSBoost) were trained incrementally with 60 weak learners (learning rate = 0.1), and variables contributing most to error reduction were identified using predictor importance metrics. In the SVM + ReliefF approach, a Gaussian kernel SVM model was applied, followed by the ReliefF algorithm with parameter k = 10 neighbours, which identified the variables with the highest neighbourhood-based discriminative power.

For dimension reduction, PCA was employed, and variable contributions were calculated based on the absolute values of the loadings in components explaining at least 95% of the cumulative variance, with the highest-contributing variables selected. The GCA, used in a time-series context, was established with a maximum of two lags, and variables meeting the $p < 0.05$ significance level were identified as causal candidates contributing to the prediction of the dependent variable. Finally, the heuristic optimisation-based GA method was used, with the objective function minimising the MSE of the linear model. Under the constraint of a maximum of four variables, the algorithm was executed with 100 generations, 50 individuals, 80% crossover rate, and elitism selection. As a result, cooperative subsets of variables that most reduced the MSE were identified.

The outputs of all methods were consolidated on a method-by-method basis, and the selected variables were recorded in Excel tables. In addition, a bar chart of PCA contributions was produced, and the results obtained from coefficient magnitudes, correlations, importance scores, and causality tests were compared in detail. In this way, the linear and non-linear, individual and interactive, as well as static and time-dependent effects of the variables were comprehensively evaluated.

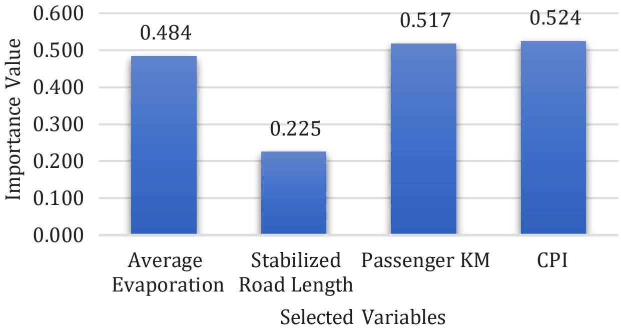

Figure 3. Selected independent variables with LASSO

Figure 3 shows that in the LASSO model, the most important variables were the CPI (0.524) and Passenger KM (0.517). These were followed by Average Evaporation (0.484). Stabilised Road Length (0.225) had a comparatively low importance value. These results showed that the model emphasised both economic indicators

(CPI, Passenger KM) and climatic factors (Average Evaporation), while road infrastructure played a secondary role.

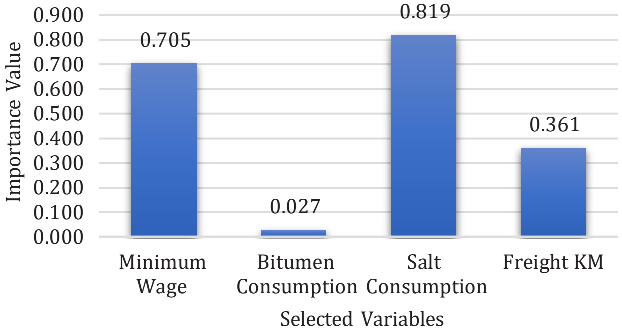

Figure 4. Selected independent variables with RR

Figure 4 shows that in the RR model, the most important variables were Salt Consumption (0.819) and Minimum Wage (0.705). Freight KM (0.361) had a moderate effect, while Bitumen Consumption (0.027) was found to be almost negligible. These results indicated that the model was particularly sensitive to winter maintenance materials and economic indicators, while the use of road materials remained insignificant.

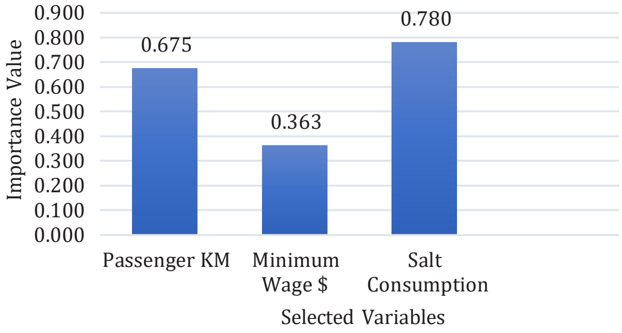

Figure 5. Selected independent variables with Stepwise AIC

Figure 5 shows that in the Stepwise AIC model, the most important variable was Salt Consumption (0.780). This was followed by Passenger KM (0.675), while Minimum Wage (0.363) had a lower importance level. These results showed that the model particularly highlighted winter maintenance materials and passenger transportation factors, while economic indicators had only secondary effects.

Figure 6 shows that in the SA, the most important variable was Stabilised Road Length (0.724). This was followed by Population (0.420) and Passenger KM (0.334). Minimum Wage (0.163) had low importance. The results showed that the model was particularly sensitive to road infrastructure and demographic factors, while economic indicators had a limited effect.

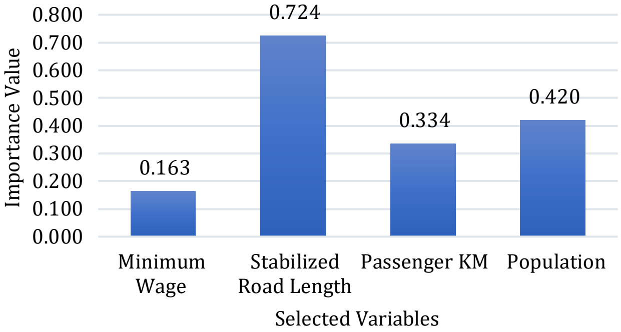

Figure 6. Selected independent variables with SA

Figure 7 shows that in the RF model, the most important variable was Stabilised Road Length (0.803). This was followed by Passenger KM (0.337) and the CPI (0.233). Surface Treated Road Length (0.061) had a very low importance. The results showed that the model was strongly sensitive to road infrastructure, with passenger transportation and economic indicators playing a supporting role, while the length of surface-dressed roads was almost ineffective.

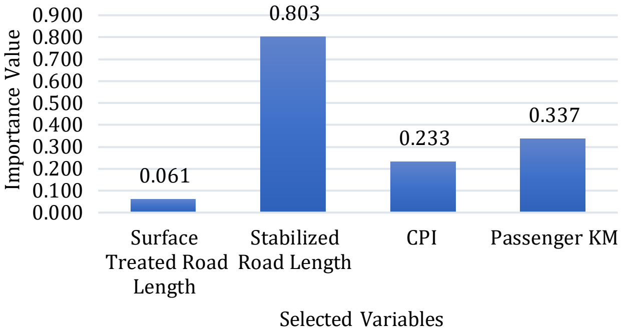

Figure 7. Selected independent variables with RF

Figure 8 shows that in the SVM model, the most important variable was Total Vehicle Count (0.981). This was followed by Total Average Precipitation (0.348), Fuel Consumption (0.275), and Salt Consumption (0.255). The results showed that the strongest predictor in the model was vehicle density, while climatic and consumption factors played supporting roles.

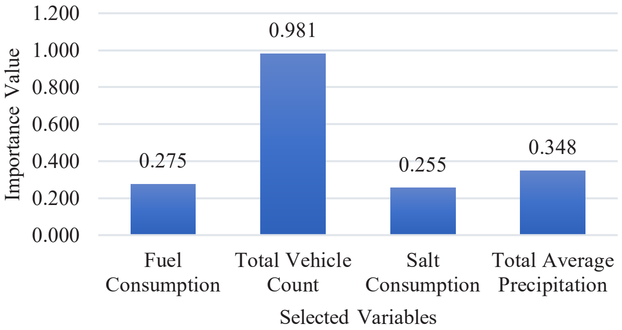

Figure 8. Selected independent variables with SVM

Figure 9 shows that in the GB model, the most important variables were Stabilised Road Length (0.865) and Asphalt Road Length (0.809). These were followed by Freight KM (0.725). Minimum Wage (0.046) had a very low importance value. The results showed that the model regarded road infrastructure and transport variables as strong predictors, while economic variables were not decisive.

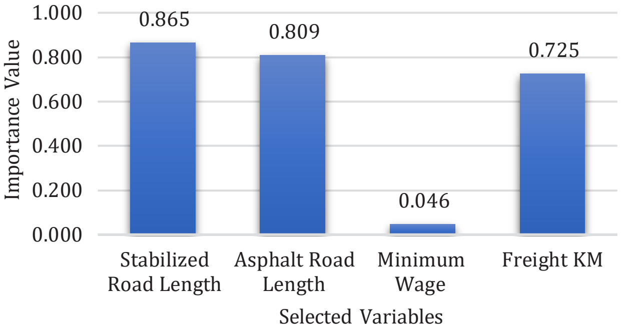

Figure 9. Selected independent variables with GB

Figure 10 shows that in the PCA results, the most important variable was Snow Covered Days (0.802). This was followed by Highway Personnel Count (0.433) and Bitumen Consumption (0.255). Average Sunshine Duration (0.071) had low importance. This indicated that climatic conditions (especially snow covered days) played a decisive role in the model, while personnel and material usage were of secondary importance.

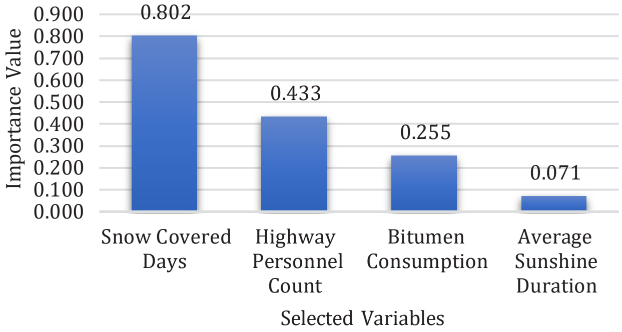

Figure 10. Selected independent variables with PCA

Figure 11 shows that in the GCA, the most important variable was Salt Consumption (0.861). This was followed by Bitumen Consumption (0.574) and Total Average Precipitation (0.251). Highway Personnel Count (0.132) had very low importance. The results showed that winter maintenance materials (salt and bitumen) were critical predictors for the model, while precipitation and personnel factors had only secondary effects.

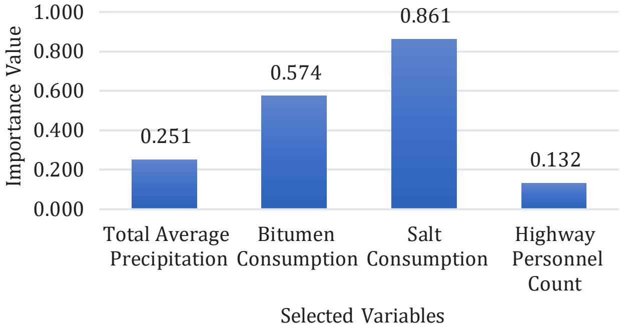

Figure 11. Selected independent variables with GCA

Figure 12 shows that in terms of GA results, the most influential variable was the CPI (0.671), indicating that economic factors played a decisive role in the model. This was followed by Surface Treated Road Length (0.322) and Bitumen Consumption (0.316), suggesting that infrastructure and material usage contributed as secondary but relevant factors. In contrast, Minimum Wage (0.132) exhibited relatively low importance, implying that labour costs had only a minor effect on the model. Overall, these findings highlight that macroeconomic indicators, particularly the CPI, were the dominant drivers, whereas infrastructure and resource variables provided supportive influence, and wage levels had limited explanatory power.

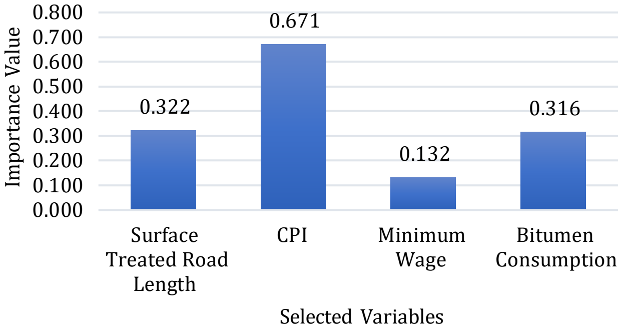

Figure 12. Selected independent variables with GA

The integration of these ten techniques significantly improved the accuracy and stability of the model by revealing critical interactions among economic, traffic-related, and infrastructural factors. As shown in Figures 3–12, the selected key variables emphasised the importance of these interactions in predicting road maintenance costs. While the LASSO and RR methods addressed multicollinearity issues, RF and GA successfully captured non-linear relationships, thus providing a reliable and interpretable model.

2.2. Prediction of highway maintenance and repair expenses

This study applied various analysis methods using groups of independent variables (X), each consisting of up to four variables selected through different techniques, in order to predict the dependent variable (Y). These methods reduced dimensionality and ensured the selection of the most appropriate variables for effective modelling and forecasting. The fundamental analysis methods are listed below.

After variable selection, analyses for each subset were carried out on non-normalised features (X_selected = X(:, selected_indices)). All models were trained to predict the dependent variable Y; following each setup, predictions were obtained, and MSE, RMSE, R^2, AdjR^2, and MAPE indicators were calculated using the compute metrics procedure. In addition, the prediction series of each model was stored, and comparative tables and figures were created in later reporting stages. The five analysis techniques used are presented below with their parameterisations in the code base and their methodological rationale.

- **LASSO (L1 penalty) analysis:** On the matrix of selected features, the LASSO model was established with 10-fold cross-validation (lasso (X_selected, Y, ‘CV’, 10)). The optimal penalty parameter was determined at FitInfo.IndexMinMSE, which gave the lowest MSE on the cross-validation error curve. Predictions were then computed as y_pred_lasso = X_selected * B(:, lasso_lambda) + FitInfo.Intercept(lasso_lambda) using the final coefficients and intercept. The L1 penalty shrank coefficients towards zero, eliminating irrelevant features; thus, under multicollinearity it provided parsimony and stability. This setup aimed to capture the dominant linear effects within the selected subset.
- **RR (L2 penalty) analysis:** For Ridge regularisation, a logarithmic penalty grid was defined (lambdas = logspace(−0.01, 0.01, 100)). For each λ, coefficients were computed with ridge (Y, X_selected, lambdas(i), 0), and predictions were obtained as y_pred_temp = X_selected * B_ridge(2:end) + B_ridge(1), with MSE recorded. The optimal λ was selected with [~, best_idx] = min(mse_vals), and the final model was built using this λ to compute y_pred_ridge. The L2 penalty shrank coefficient magnitudes and reduced variance, capturing distributed/linear effects stably under multicollinearity, and revealed predictors that contributed consistently but were not eliminated, unlike in LASSO.
- **RF (ensemble tree-based regression):** To model non-linear relationships and feature interactions, TreeBagger was used: TreeBagger(200, X_selected, Y, ‘Method’,’regression’,’MinLeafSize’,5,’OOBPredictorImportance’,’on’). In this way, an ensemble of 200 trees was established with a minimum leaf size of 5 and OOB importance calculations. Predictions were produced using predict (rf_model, X_selected). OOB-based evaluation internally monitored the model’s generalisation error, while the tree structure captured high-order interactions and non-linear boundaries, revealing complex patterns in the selected subset.
- **MLR** and diagnostic tests: For comparison, a classical MLR model was established with fitlm(X_selected, Y); predictions were obtained with predict(mdl_mlr, X_selected). The residuals of the model were tested for normality using the Kolmogorov–Smirnov test (kstest), and the homoscedasticity assumption was tested with Levene’s test on absolute deviations (vartestn(...,’LeveneAbsolute’)). In the code flow, if $p < 0.05$ in Levene’s test, the violation of homoscedasticity was reported as a warning; otherwise, the test was considered passed. This diagnostic framework documented the statistical validity of MLR results, verifying the robustness of findings obtained on a linear basis.

– **Enhanced ANN:** To capture non-linear patterns and high-dimensional interactions, feed-forward ANN architectures were scanned within model selection using perform_optimized_ysa(P, T) (P = X_selected', T = Y'). The number of neurons was varied within the range n ∈ [1,10]. The procedure searched across seven different training algorithms (trainlm, trainbfg, trainoss, trainrp, traingd, trainscg, trainbr) and pairs of nine activation functions (compet, tansig, purelin, logsig) selecting the best model according to the lowest MSE. The best model's weights (W1, W2), biases (b1, b2), number of hidden neurons, activation functions, training MSE curve, prediction/actual series, and residuals were systematically returned. This approach enabled the selected subsets to produce highly sensitive predictions across flexible and non-linear decision surfaces. For each selected variable group, 280 ANN hybrid models were developed separately, resulting in a total of 2800 ANN hybrid models.

For each analysis technique, MSE, RMSE, R^2, AdjR^2, and MAPE metrics and prediction series were consolidated across selection–analysis combinations and exported in structured tables. In addition, on the time axis (2004–2022), the actual series and the corresponding prediction series were compared in the same subplots, visually illustrating differences across methods. For ANN, the performance and architectural details of the best models, as well as (when available) closed-form expressions and parameter listings, were exported into Word reports. This setup ensured that all five analysis pipelines were tested under identical conditions on the same selected feature subsets, enabling consistent quantitative (error and fit metrics) and qualitative (graphical evidence) comparisons of linear/non-linear and individual/interactive effects.

Performance evaluation was carried out using MSE, RMSE, R^2, AdjR^2, and MAPE criteria. ANN produced the most consistent and accurate results, outperforming all other methods by successfully capturing complex relationships and long-term trends, despite higher computational costs. RF and Ridge Regression also provided robust predictions, but could not reach the sensitivity of ANN. Regularisation techniques such as LASSO and RR effectively addressed multicollinearity, though in some cases they showed deviations at later stages.

As shown in Figures 13–22, ANN provided predictions closest to the actual values and demonstrated its superiority in capturing complex interactions. These results confirmed the effectiveness of the selected variable groups and demonstrated that ANN was the most reliable method, offering powerful predictions by balancing complexity and interpretability.

In this study, ten different variable selection methods (LASSO, RR, Stepwise-AIC, SA, RF, GB, SVM, PCA, GCA, GA) and five different prediction methods (LASSO, RR, RF, MLR, ANN) were combined, and a total of 50 models were evaluated. Model performance was analysed using Mean Absolute Error (MAE), Root Mean Square

Error (RMSE), Mean Squared Error (MSE), Coefficient of Determination (R^2), Adjusted Coefficient of Determination ($AdjR^2$), and Mean Absolute Percentage Error (MAPE).

LASSO reduced multicollinearity problems through variable selection. However, LASSO–LASSO and LASSO–RR combinations produced predictions close to the actual values only in specific years. According to overall error measures, the LASSO–ANN model achieved higher accuracy compared to other LASSO-based combinations, though it did not rank among the top across the full table.

The RR method regularised multicollinearity by shrinking coefficients. In particular, the RR–ANN model produced the lowest error values in the periods 2004–2008 and 2014–2018. This result showed that RR's stabilisation of variables, when combined with ANN's capacity to learn non-linear relationships, generated a strong synergy. The RR–RR and RR–MLR models performed well in linear effects but suffered performance losses in years dominated by non-linear relationships.

Stepwise-AIC provided a classical approach by selecting variables based on statistical criteria. The Stepwise-AIC–ANN model stood out with the lowest error values in the 2019–2022 period. By contrast, the Stepwise-AIC–RR and Stepwise-AIC–MLR models exhibited relatively lower accuracy.

Selection methods based on SA generally produced weak performance. The SA–ANN model performed better compared to other SA-based combinations, but error rates remained high. The SA–RF and SA–MLR combinations ranked among the weakest results.

RF offered a strong selection strategy based on variable importance. However, when used as a selection method, its predictive performance remained limited. The RF–ANN model generated lower error values than other combinations, but did not rank among the top overall.

GB-based selection methods achieved the highest accuracy levels in the period 2009–2013 with the GB–ANN model. The GB–LASSO and GB–RR combinations produced moderate performance. This result demonstrated that GB was particularly effective when combined with strong learning algorithms such as ANN.

SVM-based selection methods achieved limited success despite their capacity to capture non-linear separations. The SVM–ANN model was the strongest among SVM-based combinations but produced higher error values compared to the RR–ANN and GB–ANN models.

PCA, although possessing dimensionality reduction capability, did not significantly improve predictive performance. All PCA combinations, including PCA–ANN and PCA–RR, exhibited low accuracy in the overall ranking.

GCA-based selection methods, although built on a strong theory of detecting inter-variable relationships, remained limited in improving predictive accuracy. The GCA–ANN model produced better results than other GCA combinations, but overall error values were found to be high.

GA stood out with their ability to determine optimal variable combinations in the solution space. The GA–ANN model achieved strong performance, ranking third overall. GA–RR and GA–MLR models remained moderate, while GA–LASSO exhibited low accuracy.

The findings obtained from the table showed that ANN-based hybrid models systematically produced the most successful results. RR–ANN (2004–2008, 2014–2018), GB–ANN (2009–2013), and Stepwise-AIC–ANN (2019–2022) provided the best performance in their respective periods.

Among the moderately successful combinations were GA–ANN and LASSO–ANN, while SA-, PCA-, and GCA-based combinations exhibited the weakest performance.

Overall, while classical methods (LASSO, RR, MLR) were useful for modelling linear relationships, soft computing-based hybrids (particularly ANN combinations) produced more reliable, generalisable, and higher-accuracy predictions in complex and non-linear structures.

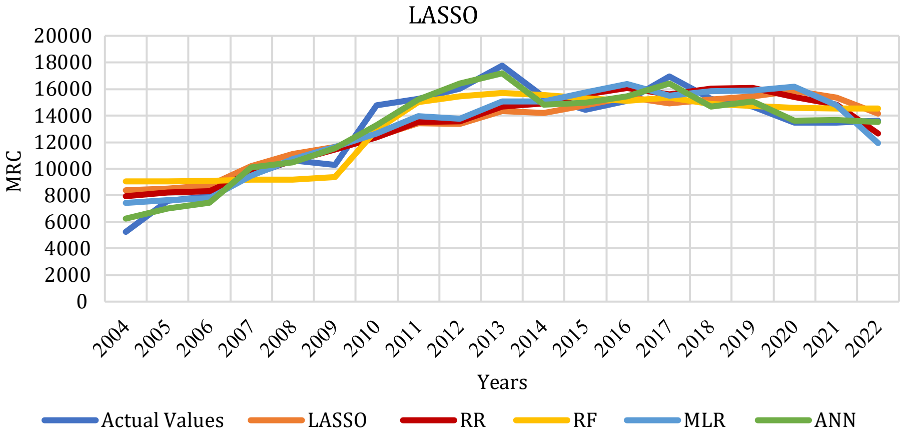

Figure 13. Model predictions and actual values with independent variables selected with LASSO

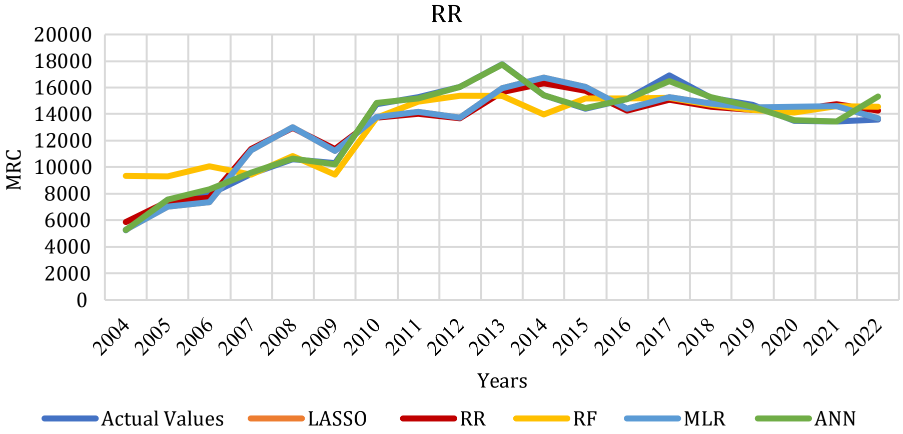

Figure 14. Model predictions and actual values with independent variables selected with RR

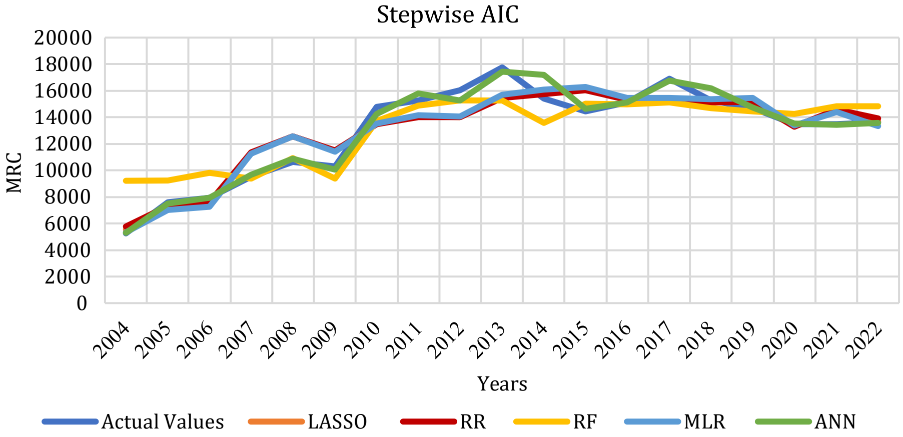

Figure 15. Model predictions and actual values with independent variables selected with Stepwise AIC

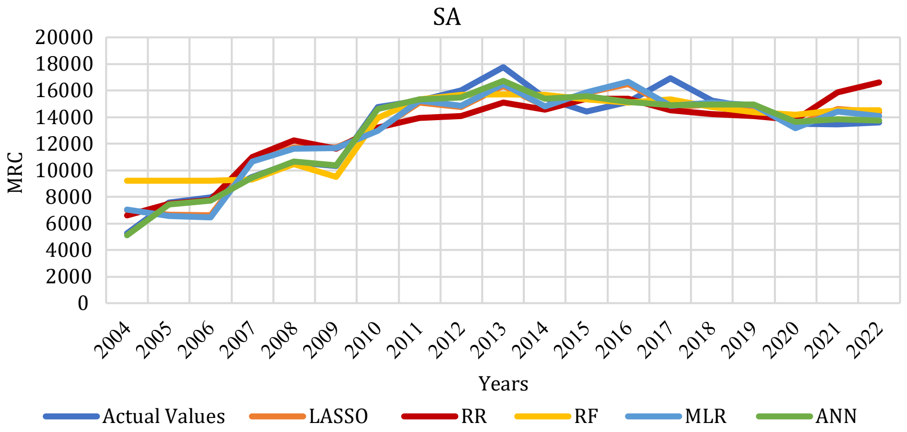

Figure 16. Model predictions and actual values with independent variables selected with SA

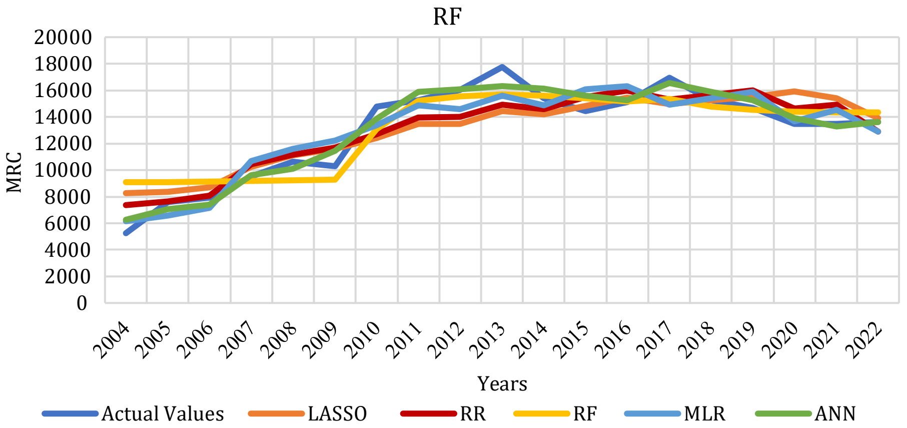

Figure 17. Model predictions and actual values with independent variables selected with RF

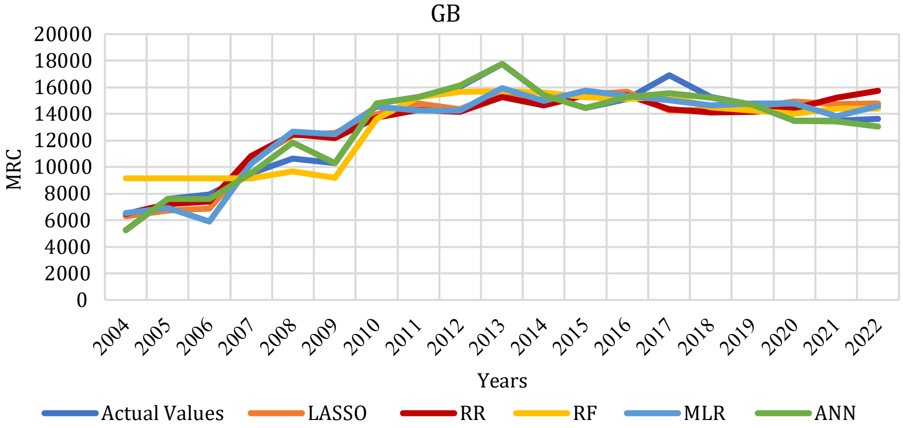

Figure 18. Model predictions and actual values with independent variables selected with GB

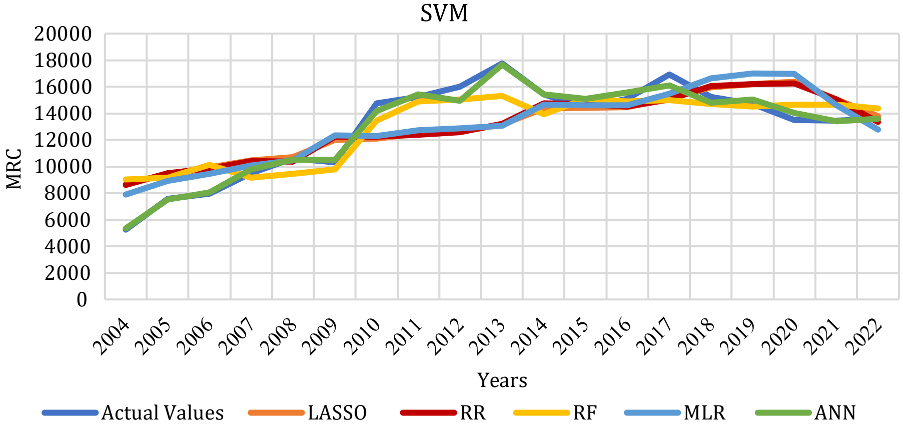

Figure 19. Model predictions and actual values with independent variables selected with SVM

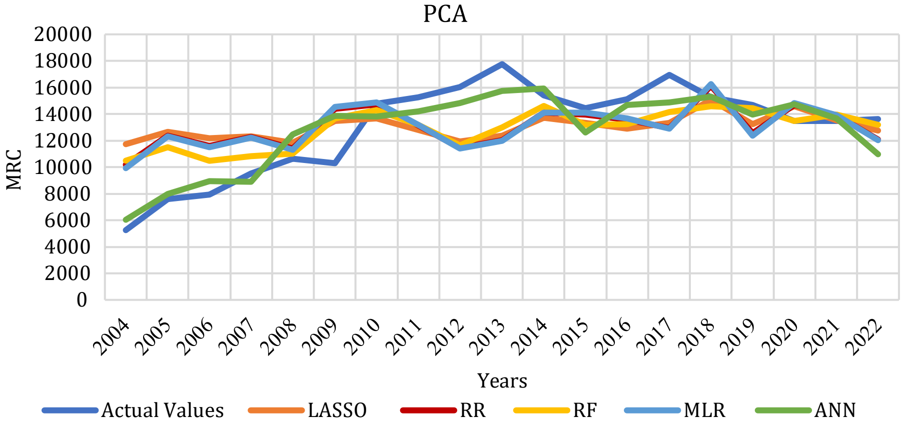

Figure 20. Model predictions and actual values with independent variables selected with PCA

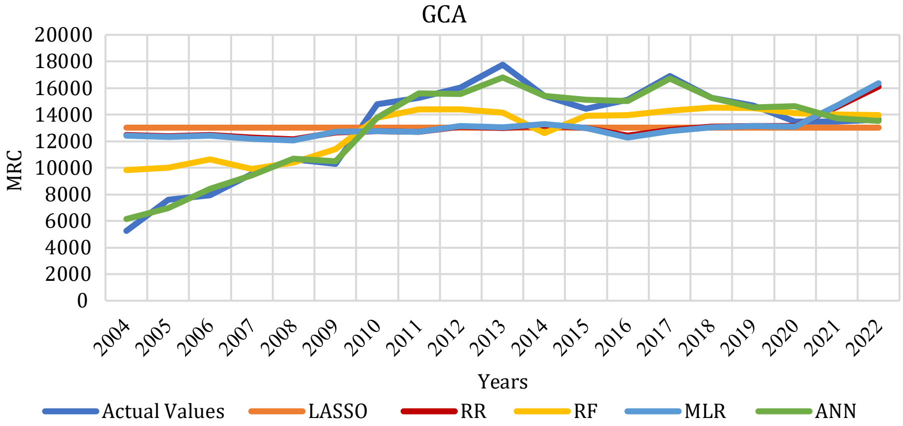

Figure 21. Model predictions and actual values with independent variables selected with GCA

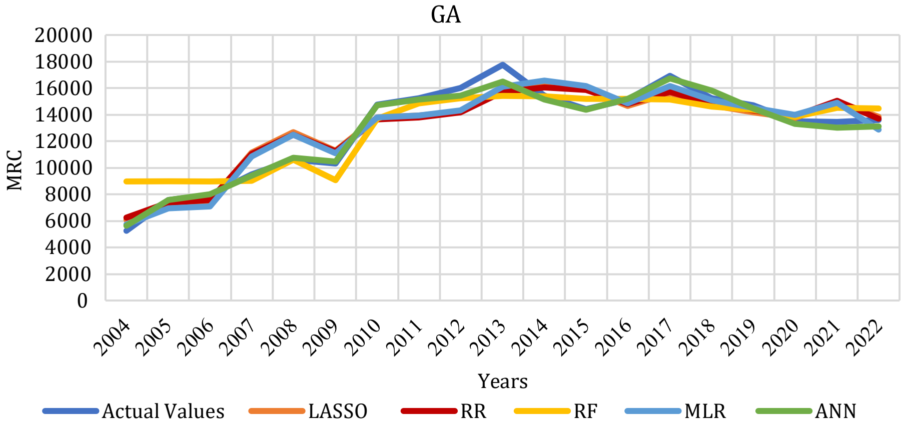

Figure 22. Model predictions and actual values with independent variables selected with GA

The performance of the 50 models constructed in this study was evaluated according to the MSE, RMSE, R^2, AdjR^2, and MAPE criteria, as presented in Table 6. The findings revealed that ANN-based combinations systematically achieved the highest accuracy. Across all selection methods where ANN was used, MSE and RMSE values fell to their lowest levels, while R^2 and AdjR^2 values exceeded 0.95. In particular, the RR–ANN (MSE = 175 475; R^2 = 0.985; MAPE = 1.36%) and GB–ANN (MSE = 197 392; R^2 = 0.983; MAPE = 1.56%) combinations showed a clear superiority in model performance. These results strongly confirmed ANN's success in capturing complex and non-linear relationships.

The results obtained with classical methods lagged behind ANN-based hybrids. For example, combinations such as LASSO–LASSO (MSE = 3 012 972; R^2 = 0.736; MAPE = 12.39%) and SVM–MLR (MSE = 4 428 012; R^2 = 0.612; MAPE = 14.58%) drew attention with low accuracy and high error values. This finding showed that classical regression techniques used in isolation were inadequate for complex data structures.

RR-based selection methods produced the most successful results when used together with ANN, whereas they remained at a moderate level with other prediction algorithms. For example, RR–MLR (MSE = 1 651 776; R^2 = 0.855; MAPE = 8.37%) and RR–RR (MSE = 1 734 248; R^2 = 0.848; MAPE = 8.88%) offered acceptable performance, but deviations were observed in situations where non-linear effects were dominant.

Stepwise-AIC–based selection methods achieved high accuracy with ANN (MSE = 291 233; R^2 = 0.974; MAPE = 2.29%) and also showed strong performance when combined with MLR and RR (e.g., Stepwise-AIC–MLR: MSE = 1 447 957; R^2 = 0.873;

MAPE = 7.81%). This result showed that statistically based variable selection worked compatibly with modern prediction methods.

Selection methods based on SA generally showed limited success. Although SA–ANN (MSE = 377 872; R^2 = 0.967; MAPE = 2.56%) produced strong results, marked performance losses were observed in combinations such as SA–RR (MSE = 2 494 895; R^2 = 0.781; MAPE = 10.57%).

RF and GB selection methods produced high accuracy when used with ANN. For example, the RF–ANN (MSE = 494 294; R^2 = 0.957; MAPE = 5.12%) and GB–ANN (MSE = 197 392; R^2 = 0.983; MAPE = 1.56%) models ranked among the top. However, when the same selection methods were combined with MLR or RR, a decrease in performance was observed.

SVM-based selection methods produced strong results with ANN (MSE = 194 298; R^2 = 0.983; MAPE = 2.35%), yet showed quite weak performance with other prediction methods. In particular, the SVM–LASSO and SVM–RR combinations had the lowest R^2 values (around 0.59) and drew attention with high error rates.

PCA- and GCA-based selection methods were generally found to be unsuccessful. Although PCA–ANN (MSE = 2 253 053; R^2 = 0.802; MAPE = 9.96%) was partially acceptable, R^2 values fell to levels as low as 0.20 in the combinations of PCA with other prediction methods. GCA-based combinations mostly produced low accuracy, although GCA–ANN (MSE = 297 434; R^2 = 0.974; MAPE = 3.72%) yielded relatively good results.

GA–based selection methods exhibited strong overall performance. In particular, GA–ANN (MSE = 573 104; R^2 = 0.950; MAPE = 5.05%) provided high accuracy, while GA–MLR (MSE = 1 239 868; R^2 = 0.891; MAPE = 7.93%) was also found to be statistically satisfactory.

Overall, ANN-based models clearly outperformed other methods across all combinations. RR–ANN and GB–ANN achieved the highest accuracy, while Stepwise-AIC–ANN and GCA–ANN also ranked among the top. By contrast, PCA- and GCA-based classical combinations showed the lowest performance. These findings confirmed that ANN-centred hybrid approaches offered the most reliable and generalisable results in modelling complex and non-linear relationships.

Table 6. Analysis methods metrics

Selection Method	Analysis Method	MSE	RMSE	R^2	AdjR^2	MAPE, %
LASSO	LASSO	3 012 972	1736	0.736	0.720	12.398
LASSO	RR	2 455 905	1567	0.785	0.772	10.998
LASSO	RF	1 869 491	1367	0.836	0.826	10.683
LASSO	MLR	2 309 141	1520	0.798	0.786	10.118
LASSO	ANN	409 193	640	0.964	0.954	4.792
RR	LASSO	1 651 778	1285	0.855	0.847	8.376
RR	RR	1 734 248	1317	0.848	0.839	8.880
RR	RF	2 151 977	1467	0.811	0.800	11.396
RR	MLR	1 651 776	1285	0.855	0.847	8.372
RR	ANN	175 475	419	0.985	0.980	1.357
Stepwise AIC	LASSO	1 467 980	1212	0.871	0.864	7.740
Stepwise AIC	RR	1 511 633	1229	0.867	0.860	7.865
Stepwise AIC	RF	2 233 163	1494	0.804	0.793	11.572
Stepwise AIC	MLR	1 447 957	1203	0.873	0.866	7.811
Stepwise AIC	ANN	291 233	540	0.974	0.969	2.291
SA	LASSO	1 458 449	1208	0.872	0.865	9.461
SA	RR	2 494 895	1580	0.781	0.768	10.573
SA	RF	1 667 476	1291	0.854	0.845	9.777
SA	MLR	1 437 970	1199	0.874	0.866	9.451
SA	ANN	377 872	615	0.967	0.957	2.560
RF	LASSO	2 907 368	1705	0.745	0.730	12.068
RF	RR	1 938 077	1392	0.830	0.820	9.847
RF	RF	1 831 149	1353	0.839	0.830	10.617
RF	MLR	1 514 577	1231	0.867	0.859	9.039
RF	ANN	494 294	703	0.957	0.944	5.116
GB	LASSO	1 867 694	1367	0.836	0.827	10.124
GB	RR	2 128 550	1459	0.813	0.802	10.448
GB	RF	1 727 916	1315	0.848	0.840	10.341
GB	MLR	1 670 368	1292	0.854	0.845	9.730
GB	ANN	197 392	444	0.983	0.978	1.565
SVM	LASSO	4 665 746	2160	0.591	0.567	15.613
SVM	RR	4 616 075	2149	0.595	0.571	15.685
SVM	RF	2 203 956	1485	0.807	0.795	11.825

Selection Method	Analysis Method	MSE	RMSE	R^2	AdjR^2	MAPE, %
SVM	MLR	4 428 012	2104	0.612	0.589	14.584
SVM	ANN	194 298	441	0.983	0.978	2.351
PCA	LASSO	9 704 111	3115	0.149	0.099	25.048
PCA	RR	9 101 476	3017	0.202	0.155	23.317
PCA	RF	6 474 098	2544	0.432	0.399	18.969
PCA	MLR	9 088 394	3015	0.203	0.156	23.111
PCA	ANN	2 253 053	1501	0.802	0.746	9.966
GCA	LASSO	11 403 292	3377	–	–0.059	28.468
GCA	RR	10 471 950	3236	0.082	0.028	27.384
GCA	RF	3 705 192	1925	0.675	0.656	14.448
GCA	MLR	10 464 501	3235	0.082	0.028	27.367
GCA	ANN	297 434	545	0.974	0.966	3.728
GA	LASSO	1 368 571	1170	0.880	0.873	8.248
GA	RR	1 358 270	1165	0.881	0.874	8.039
GA	RF	1 704 924	1306	0.850	0.842	9.867
GA	MLR	1 239 868	1113	0.891	0.885	7.933
GA	ANN	573 104	757	0.950	0.935	5.045

The results in Table 7 provide a detailed evaluation of the performance of different prediction models in terms of processing time. These data offered an important perspective not only for assessing the accuracy of the models but also for evaluating their suitability in practical applications. Processing time was a critical factor influencing model selection, particularly when dealing with large datasets or real-time applications.

The LASSO method, with a processing time of 17.34 seconds, exhibited slower performance compared to other linear regression models. Although the regularisation technique was effective in resolving multicollinearity problems, this increase in computation time could limit the applicability of LASSO, especially when working with larger datasets. This indicated that while the model was advantageous in accuracy-focused analyses, it was less likely to be preferred in situations where speed requirements were prioritised.

Table 7. Method estimation periods

Methods	Elapsed Time, s
LASSO	17.340
RR	0.033
RF	1.272
MLR	2.716
ANN	2 453.99

The RR method, with a processing time of only 0.03 s, stood out as the fastest model. The extremely low processing time of RR made it possible to conduct rapid analyses even with large datasets. Furthermore, its high accuracy rates rendered RR both a fast and reliable option. RR was, therefore, an ideal solution in situations where speed was critical.

The RF method, with a processing time of 1.27 s, provided impressive speed while also being capable of modelling non-linear relationships. This method stood out for achieving high accuracy with complex data structures while keeping processing time at reasonable levels. This demonstrated that RF offered balanced performance, making it a preferred choice across a wide range of applications.

MLR, with a processing time of 2.71 s, showed relatively slower performance among classical computational methods. Model selection procedures and optimisation based on information criteria increased the processing time for this method. However, as its accuracy results were limited, the additional processing time was not sufficiently justified. Therefore, MLR appeared more suitable for smaller datasets or cases where linear relationships were dominant.

ANN, with a processing time of 2453.99 s, had by far the longest computation time. This high processing cost was related to the architectural complexity of the model (e.g., increasing the number of neurons and multi-layered structures) and the computational load of the activation and training functions employed. Although ANN provided superior performance in terms of accuracy, this lengthy processing time limited its use in large datasets or real-time applications. Nevertheless, in projects where high accuracy was required, the predictive performance offered by ANN could justify its time cost.

Overall, fast methods such as RR were more prominent in situations with time constraints. RF provided balanced performance by ensuring high accuracy with reasonable processing times. LASSO and MLR could be used with smaller-scale datasets, while ANN, despite offering superior accuracy, stood out as the costliest method in terms of processing time. These results clearly demonstrated that attention needed to be paid not only to the accuracy of models but also to functional factors such as processing time.

2.3. The best predictive model: RR–ANN

In this study, the combination of the RR method and the ANN model was identified as the best-performing model in terms of performance metrics and configuration parameters. The regularisation effect of RR, combined with the capacity of ANN to model non-linear relationships, enabled the model to achieve superior performance by providing low error rates and high generalisation capacity. The performance of the model was notable, with MSE = 175 474.9241, RMSE = 418.8973, R^2 = 0.9846, MAPE = 1.3570, and AdjR^2 = 0.9802. These values demonstrated that the model explained a large proportion of the variance in the dependent variable while maintaining error rates at a minimum level. Furthermore, the MAPE value, calculated as 1.2367%, showed that the model produced predictions extremely close to the actual values.

The configuration parameters of the model were also optimised to achieve the best performance. The optimal number of neurons in the hidden layer was determined as nine, a configuration that enhanced the ANN's capacity to learn non-linear relationships. The activation function logsig was used, enabling the neural network to successfully model complex transformations. Taken together, these configuration settings and performance metrics indicated that the RR–ANN model stood out as the best model, delivering superior performance in prediction tasks with high accuracy and generalisation capacity. The regularisation capacity of RR, combined with the ability of ANN to model non-linear structures, placed this model ahead of other combinations. These findings clearly revealed that RR–ANN offered a strong modelling strategy for complex data structures.

Based on the literature and the requirements for determining MRC, the RR–ANN model was presented and formulated in graphical form in this study (Figure 23). This formula provided a methodological foundation for the prediction process by combini.ng the regularisation capacity of RR with the modelling ability of ANN for non-linear relationships.

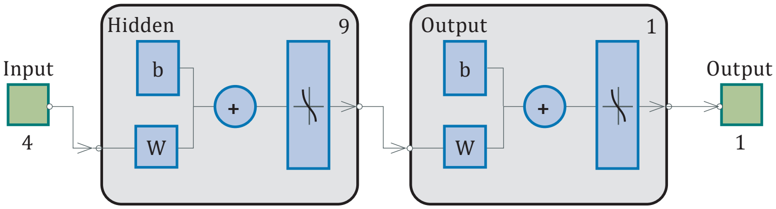

Figure 23. RR–ANN model

According to the RR–ANN predictive analysis modelling values, this formula was arranged as in Equation (1).

Equation (1) MRC prediction formula:

$$Y = \mu Y + \sigma Y \times \operatorname{logsig} \left(\mathrm{W}_2 \times \operatorname{logsig}\left(\mathrm{W}_1 \times \begin{bmatrix} (X18-\mu X18)/\sigma X18, (X19-\mu X19)/\sigma X19, \\ (X21-\mu X21)/\sigma X21, (X13-\mu X13)/\sigma X13 \end{bmatrix} + \boldsymbol{B}_1 \right) + \boldsymbol{B}_2 \right) \tag{1}$$

where:
X21, *X18*, *X19* and *X13* are independent variables: Minimum Wage, Bitumen Consumption, Salt Consumption, and Freight KM;
W_1 and W_2 are the weight matrices of the hidden and output layers of the ANN;
μ is average of variables between 2004 and 2022 (recalculated each year by including the latest available data);
σ are standard deviations of variables between 2004 and 2022 (recalculated each year by including the latest available data);
$\boldsymbol{B_1}$ and $\boldsymbol{B_2}$ are the bias vectors of the hidden and output layers;
logsig is the activation function:

$$\operatorname{logsig}(x) = \frac{1}{1+e^{-\chi}};$$

Y denotes the predicted Maintenance and Repair Cost, $/km (MRC).

This formula effectively combined the ability of RR to minimise multicollinearity problems with the capacity of ANN to learn complex and non-linear relationships. The regularisation effect of RR enhanced the generalisation capacity of the model, while the highly accurate learning mechanism of ANN strengthened the model’s success in predicting maintenance and repair costs. This formula provided a critical tool for accurately and generalizable forecasting MRC.

$$\mathrm{MRC} = 13021.34 + 3376.88 \times \operatorname{log\,sig}\left(\operatorname{log\,sig}\left(\begin{bmatrix} 0.1816 & -0.4518 & 0.6277 & 0.01638 \\ 0.1816 & 0.4518 & 0.6277 & 0.01638 \\ 2.028 & -0.5886 & -0.4444 & 1.964 \\ -0.2773 & -3.59 & -0.6423 & -1.253 \\ -0.1902 & -0.477 & -0.6622 & 0.0284 \\ -0.19 & -0.4768 & -0.6621 & 0.02858 \\ 0.3588 & 1.245 & 0.7126 & -0.09981 \\ -0.19 & -0.4764 & -0.6616 & 0.02814 \\ -2.371 & 0.4036 & -0.1373 & -3.508 \end{bmatrix} \times \begin{bmatrix} \dfrac{\text{Minimum Wage \$} - 415.483}{69.076} \\ \dfrac{\text{Bitumen Consumption} - 1999.599}{706.041} \\ \dfrac{\text{Salt Consumption} - 134.572}{138.654} \\ \dfrac{\text{Freight KM} - 226917.579}{48043.188} \end{bmatrix} + \begin{bmatrix} 0.2341 \\ 0.2341 \\ -1.606 \\ 0.7311 \\ -0.2705 \\ -0.2707 \\ 1.014 \\ -0.2697 \\ -1.344 \end{bmatrix} \right) \times \begin{bmatrix} 0.9292 \\ 0.9292 \\ -2.557 \\ 2.613 \\ -1.009 \\ -1.01 \\ 1.777 \\ -1.008 \\ -2.86 \end{bmatrix}^{T} + \begin{bmatrix} -0.0341 \end{bmatrix} \right)$$

2.4. Discussion

This study presented a large-scale hybrid modelling framework for the prediction of MRC, encompassing both classical statistical methods and soft computing techniques. Most previous studies were either limited to single methods or based on restricted datasets (Han et al., 2023; Göksal, 2024). By contrast, this research systematically tested a total of 50 hybrid models, combining ten variable selection techniques and five different prediction methods using 21 independent variables over a 19-year time series. In doing so, it not only filled the methodological gap in the literature but also, for the first time at the national scale, provided such a comprehensive evaluation.

The results demonstrated that classical methods used in isolation (MLR, LASSO, RR) delivered acceptable performance in modelling linear relationships but remained insufficient when non-linear interactions dominated. Soft computing methods (SA, RF, GB, SVM, ANN, GA) captured non-linear relationships effectively but lost stability in the presence of multicollinearity and high-dimensionality issues. The coexistence of these two limitations highlighted the importance of hybrid models.

The best model was obtained with the RR–ANN hybrid formulation. This model, which combined the regularisation effect of RR with the non-linear learning capacity of ANN, was mathematically expressed as in Equation (1). Here, X represented the four selected critical independent variables (Freight KM, Bitumen Consumption, Minimum Wage, Salt Consumption). With this configuration (9 hidden neurons, logsig activation function), the model achieved MSE = 175 474.92; RMSE = 418.90; R^2 = 0.985; AdjR^2 = 0.980; and MAPE = 1.36%. This performance was superior to that reported in similar studies in the literature (Jasim et al., 2024; Elwahsh et al., 2023) and clearly demonstrated the effectiveness of hybrid modelling in terms of accuracy.

Variable importance analyses revealed that the strongest determinants of costs were traffic intensity (Passenger KM, Freight KM), material consumption (Bitumen, Salt), and economic indicators (CPI, Minimum Wage). Environmental factors (Snow Covered Days, Total Average Precipitation, and evaporation) provided secondary but meaningful contributions. These results were consistent with studies reporting that heavy vehicle loads shortened road lifespan (Pais et al., 2019), that despite signs of deceleration since 2022, volatility in highway cost indices left the trend uncertain (FHWA, 2024), that fluctuations in material prices directly affected budgets (U.S DOT, 2025), and that climate change increased maintenance frequency (Zhang et al., 2025).

In terms of processing times, significant differences were observed among the methods. RR, with 0.03 s, was the fastest method, while RF, with 1.27 s, offered a balanced solution between speed and accuracy. ANN-based models, however, required 2453.99 s, bringing the highest computational cost. This finding revealed the necessity for decision-makers to make strategic choices with respect to the accuracy-speed trade-off: while ANN-based hybrids should be preferred in long-term budget planning where accuracy was critical, RR or RF methods were more suitable for real-time applications where speed was prioritised.

The limitations of the study included the use of a dataset specific to Turkey and the high computational cost of ANN. Nevertheless, the high accuracy and methodological contributions provided by the proposed hybrid formulation mitigated these limitations and positioned the study as a strong reference in the literature.

3. Limitations

Several limitations of this study should be acknowledged. First, the analysis relies on annual national-level data covering the period 2004–2022. While this long-term perspective is suitable for strategic budget planning, the use of annual aggregates may mask short-term fluctuations and regional heterogeneity in maintenance and repair costs. Future research could address this limitation by employing higher-frequency data (e.g., monthly or quarterly) and regionally disaggregated datasets to capture spatial and temporal variations more explicitly.

Second, although the study integrates multiple regularisations, variable-selection, and hybrid modelling techniques to mitigate overfitting risks, the relatively limited number of observations inherent in annual time-series data may still constrain model generalisability. Expanding the dataset with additional temporal coverage or complementary cross-sectional information could further strengthen the robustness of future analyses.

Third, potential structural breaks arising from extraordinary events – such as economic crises, policy reforms, or the COVID-19 pandemic – are not modelled explicitly in the current framework. While hybrid and non-linear methods partially accommodate such dynamics, future studies could incorporate regime-switching models or structural break tests to better account for abrupt changes in cost behaviour.

Finally, the analysis focuses primarily on financial, infrastructural, economic, and meteorological determinants of MRC. Institutional, managerial, and contractual factors – such as procurement practices, maintenance scheduling efficiency, and contractor performance – are not explicitly considered due to data limitations. Incorporating such variables in future research may provide a more comprehensive understanding of cost dynamics and further enhance the policy relevance of the modelling framework.

Conclusions

The research presented a comprehensive approach to the prediction of road MRC in Turkey by combining classical statistical methods with soft computing techniques within a hybrid framework. The RR–ANN hybrid model achieved the highest accuracy among all the tested models (R^2 = 0.985; MAPE = 1.36%), with Freight KM, bitumen consumption, minimum wage, and salt consumption identified as the most critical variables. Whilst RR was the fastest and ANN the most accurate yet computationally costly method, the hybrid RR–ANN formulation successfully balanced accuracy with efficiency.

From a scientific perspective, the study demonstrated that integrating RR regularisation with ANN effectively addresses both multicollinearity and non-linear relationships, whilst the use of a large 19-year dataset with 21 independent variables provided more reliable and generalisable findings compared to previous limited-scale studies. From a practical standpoint, the hybrid framework enables more accurate budget planning, reduces long-term costs through proactive maintenance strategies, and strengthens data-driven decision-support systems in infrastructure management.

Future research should focus on integrating IoT-based real-time data, optimising ANN processing times through GPU or distributed computing techniques, and incorporating deep learning and ensemble methods into the hybrid framework, alongside comparative analyses with international datasets to assess its generalisability.

In conclusion, the comparative results clearly demonstrate that hybrid model configurations substantially outperform conventional single-method approaches in forecasting highway maintenance and repair costs. In particular, hybrid formulations combining structured variable selection methods with artificial neural networks consistently yield the lowest error measures (MSE and RMSE) and the highest goodness-of-fit values across all model classes. Relative to traditional linear and standalone machine learning specifications, these integrated models achieve improvements in forecast accuracy on the order of 2–30%, as evidenced by pronounced reductions in forecast errors and marked increases in explanatory power. These findings confirm the effectiveness of hybrid modelling in capturing complex cost dynamics and provide a robust, transferable framework to support more reliable budget planning and sustainable road asset management in countries with comparable infrastructure systems.

Acknowledgements

The first author, Haydar GUNDOGDU, would like to express my deepest gratitude to my esteemed advisors, whose invaluable guidance and support were instrumental in the successful completion of this work.

The Authors would also like to thank the General Directorate of Highways, the Ministry of Environment, Urbanization and Climate Change, the Turkish Statistical Institute (TÜİK), the Central Bank of the Republic of Turkey (CBRT), and the General Directorate of Meteorology, whose websites provided the data used in the study.

This study is part of the Doctoral Thesis "Determination of Highway Maintenance and Repair Costs with Soft Computing and Classical Techniques" of Haydar GUNDOGDU, which is being developed at the Institute of Graduate Studies, Civil Engineering Program, at Iskenderun Technical University, Turkey.

Statement of the Use of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this manuscript, AI-assisted tools were used solely to improve grammar and clarity of expression in certain sections of the text. The scientific content, data analysis, interpretation of the results, and the conclusions are entirely the responsibility of the authors. The authors take full responsibility for the content of this manuscript.

Disclosure Statement

The authors hereby declare that there are no financial, professional, or personal conflicts of interest that could have influenced the design, execution, or presentation of the work detailed in this article.

REFERENCES

Akpan, U., & Morimoto, R. (2022). An application of Multi-Attribute Utility Theory (MAUT) to the prioritization of rural roads to improve rural accessibility in Nigeria. *Socio-Economic Planning Sciences*, *82*, Article 101256. https://doi.org/10.1016/J.SEPS.2022.101256

BİRİM FİYAT. (2023). Poz Detay Bilgileri. https://www.birimfiyat.net/10.330.5423-dokme-asfalt-batman-rafineri

Bressi, S., Primavera, M., & Santos, J. (2022). A comparative life cycle assessment study with uncertainty analysis of cement treated base (CTB) pavement layers containing recycled asphalt pavement (RAP) materials. *Resources, Conservation and Recycling*, *180*, Article 106160. https://doi.org/10.1016/j.resconrec.2022.106160

ÇŞB. (2023). Asgari Ücret. https://www.csgb.gov.tr/poco-pages/asgari-ucret/

Elwahsh, H., Allakany, A., Alsabaan, M., Ibrahem, M. I., & El-Shafeiy, E. (2023). A deep learning technique to improve road maintenance systems based on climate change. *Applied Sciences (Switzerland)*, *13*(15), Article 8899. https://doi.org/10.3390/app13158899

FHWA. (2024). *National Highway Construction Cost Index 2024 Q2*. Office of Transportation Policy Studies, https://www.fhwa.dot.gov/policy/otps/nhcci/NHCCI_Narrative_Article_2024_Q2.pdf

Göksal, F. P. (2024). Türkiye Ölçeğinde Karayolu Yol- Bakım Maliyetleri ve Gelecek Tahmini. *Black Sea Journal of Engineering and Science*, *7*(3), 392–400. https://doi.org/10.34248/bsengineering.1414038

Han, C., Huang, J., Yang, X., Chen, L., & Chen, T. (2023). Long-term maintenance planning method of rural roads under limited budget: A case study of road network. *Applied Sciences (Switzerland)*, *13*(23), Article 12661. https://doi.org/10.3390/app132312661

Jasim, A. F., Ali, Z. K., & Al-Saadi, I. F. (2024). A comprehensive review of life cycle cost assessment of recycled materials in asphalt pavements rehabilitation. *Advances in Civil Engineering*, 2024, Article 2004803. https://doi.org/10.1155/2024/2004803

KGM. (2019). *KGM 2019–2023 STRATEJİK PLANI*. https://www.kgm.gov.tr/SiteCollectionDocuments/KGMdocuments/MerkezBirimler/Kurumsal/StratejikPlan/strateji(2019-2023).pdf

KGM. (2023a). *KGM 2023 YILI DEVLET VE İL YOLLARI BAKIM-İŞLETME MALİYETLERİ*. https://www.kgm.gov.tr/Sayfalar/KGM/SiteTr/Istatistikler/YapimBakimIsletmeMaliyet.aspx

KGM. (2023b). *KGM YillaraGoreDevletVeIlYollari*. https://www.kgm.gov.tr/Sayfalar/KGM/SiteTr/Istatistikler/DevletveIlYolEnvanteri.aspx

Kim, H.-Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics*, *38*(1), Article 52. https://doi.org/10.5395/rde.2013.38.1.52

Kumar, A., & Suman, S. K. (2025). Effects of overloaded commercial traffic on pavement surface layer. *Intelligent Transportation Infrastructure*, *4*. https://doi.org/10.1093/iti/liaf005

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill Irwin. https://users.stat.ufl.edu/~winner/sta4211/ALSM_5Ed_Kutner.pdf

Li, Z., Lang, L., Sun, G., Cai, Z., & Luo, Z. (2023). Enhancing multiple linear regression for price prediction: A PCA-integrated approach. *2023 4th International Conference on Computer, Big Data and Artificial Intelligence (ICCBD+AI)*, Guiyang, China, 337–341. https://doi.org/10.1109/ICCBD-AI62252.2023.00063

Massey, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, *46*(253), 68–78. https://doi.org/10.1080/01621459.1951.10500769

MGM. (2023). *Resmi İklim İstatistikleri*. https://www.mgm.gov.tr/veridegerlendirme/il-ve-ilceler-istatistik.aspx?k=parametrelerinTurkiyeAnalizi.

Mohd Razali, N., & Yap, B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J. Stat. Model. Analytics*, *2*. https://www.nrc.gov/docs/ml1714/ml17143a100.pdf

Pais, J. C., Figueiras, H., Pereira, P., & Kaloush, K. (2019). The pavements cost due to traffic overloads. *International Journal of Pavement Engineering*, *20*(12), 1463–1473. https://doi.org/10.1080/10298436.2018.1435876

Pan, Y., Shang, Y., Liu, G., Xie, Y., Zhang, C., & Zhao, Y. (2021). Cost-effectiveness evaluation of pavement maintenance treatments using multiple regression and life-cycle cost analysis. *Construction and Building Materials*, *292*, Article 123461. https://doi.org/10.1016/j.conbuildmat.2021.123461

Persyn, D., Diaz-Lanchas, J., Barbero, J., Conte, A., & Salotti, S. (2020). *A new dataset of distance and time related transport costs for EU regions*. https://publications.jrc.ec.europa.eu/repository/handle/JRC119412

Schmitt, R. R. (2025). *Bureau of Transportation Statistics Style Manual*, 2025 Edition. https://doi.org/10.21949/x1at-2e34

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Tabachnick, B. G., & Fidell, L. S., (2007). *Experimental designs using ANOVA*. Thomson/Brooks/Cole.

TCMB. (2024). *Gösterge Niteliğindeki Merkez Bankası Kurları*. https://www.tcmb.gov.tr/kurlar/kurlar_tr.html

Torres-Machi, C., & Evers Jonathan Schmidt Anneliese Crayton, E. (2024). *Pavement rehabilitation analysis: A life-cycle cost and long-term performance comparison of full depth reclamation and overlays*.

TÜİK. (2023). *Genel Nüfus Sayımları Metaverisi*. https://data.tuik.gov.tr/Kategori/GetKategori?p=Nufus-ve-Demografi-109

TÜİK. (2024). *Tüketici fiyat endeks rakamları (2003=100)*. https://data.tuik.gov.tr/Bulten/DownloadIstatistikselTablo?p=KMExlm5AVU2ln21dc2evQ2SnPKPmGEBqV6H8CcJSjNNBzZZT2CJNzYtIqx1WGQK8

U.S DOT. (2025). *Understanding construction change orders A U.S. DOT project delivery center of excellence report*. https://www.transportation.gov/procurement-office-info/volpe-volpe-national-transportation-systems-center

Zhang, R., Sun, L., Qiao, Y., Sias, J. E., & Dave, E. V. (2025). Multidimensional comparative analysis of future climate change impacts on pavement infrastructure aging. *Transportation Research Part D: Transport and Environment*, 142. https://doi.org/10.1016/j.trd.2025.104702