Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Industrial Waste Valorization and Sustainable Material Design: A Data-Driven Statistical Approach

STHITADHI DAS

sthdas999@gmail.com

Brainware University https://orcid.org/0009-0001-2484-9530

Research Article

Keywords: Industrial waste upcycling, sustainable engineering materials, statistical modeling, regression analysis, response surface methodology, machine learning, material optimization, waste valorization, circular economy, sustainable engineering

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11633484/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Industrial Waste Valorization and Sustainable Material Design: A Data-Driven Statistical Approach

Sthitadhi Das*
Department of Mathematics, Brainware University, India

Abstract

The increasing generation of industrial waste has created significant environmental and resource-management challenges, while simultaneously providing opportunities for the development of sustainable engineering materials. Industrial waste upcycling offers a promising pathway for converting waste streams into value-added materials for construction, infrastructure, composites, ceramics, and other engineering applications. However, the performance of waste-derived materials is influenced by numerous factors, including waste composition, particle characteristics, replacement level, processing conditions, and curing parameters. This chapter presents a statistical modeling perspective on the development and evaluation of sustainable engineering materials from industrial waste. Established statistical methods, including descriptive analysis, correlation analysis, regression modeling, analysis of variance, response surface methodology, and mixed-effects models, are considered for understanding material behavior and experimental variability. Machine-learning approaches are also discussed as complementary tools for prediction and model comparison. Particular emphasis is placed on using statistical models to identify important factors, predict engineering performance, and optimize material formulations under practical constraints. The chapter further connects statistical prediction with sustainability indicators such as waste utilization, energy consumption, production cost, and environmental impact. A real-data case study will illustrate how statistical modeling can support data-driven material design and contribute to the transition from conventional waste management toward a circular economy.

Keywords: Industrial waste upcycling; sustainable engineering materials; statistical modeling; regression analysis; response surface methodology; machine learning; material optimization; waste valorization; circular economy; sustainable engineering.
2020 Mathematics Subject Classification: Primary 62J05, 62J12; Secondary 62P30, 90C90.

1 Introduction

Industrialization has contributed substantially to economic development, infrastructure expansion, and technological advancement. However, the production, processing, and consumption of industrial materials also generate substantial quantities of residual materials that require appropriate management. These residuals may arise from energy production, metallurgical operations, mining, chemical manufacturing, foundry processes, glass production, and construction activities. When such materials are inadequately managed, they may contribute to environmental contamination, resource depletion, and additional economic costs. The United Nations Environment Programme emphasizes the importance of moving beyond conventional waste management toward waste prevention, resource

*Email ID: sthdas999@gmail.com, ORCID ID: 0009-0001-2484-9530

recovery, and circular-economy practices [1]. Although the Global Waste Management Outlook primarily examines municipal solid waste, its broader resource-management principles are relevant to the development of integrated waste valorization strategies.

The environmental significance of waste management extends beyond the disposal stage. The extraction of virgin raw materials, their transportation, industrial processing, and the manufacture of replacement materials can consume energy and generate environmental burdens. Consequently, sustainable material development requires a systematic assessment of both material performance and resource efficiency. Life-cycle assessment (LCA) provides a recognized framework for evaluating environmental aspects and potential impacts throughout the life cycle of a product or process, including the definition of the goal and scope, inventory analysis, impact assessment, and interpretation [2]. Therefore, the environmental advantages of industrial waste utilization should not be assumed solely from the fact that a waste material has been incorporated into a new product. Such advantages depend on factors including waste processing requirements, transportation distance, substitution of conventional materials, product durability, and end-of-life management.

Industrial waste valorization represents a resource-oriented approach in which residual materials are converted into useful secondary resources or value-added products. In contrast to disposal-oriented practices, valorization considers the possibility of recovering useful physical, chemical, or functional properties from materials that would otherwise have limited economic value. In construction and materials engineering, this approach is particularly relevant because industrial by-products may serve as supplementary cementitious materials, aggregates, mineral fillers, ceramic constituents, or precursors for alkali-activated materials. For example, coal fly ash has been investigated as a resource for several applications, including cementitious and other construction-related products [3]. Similarly, construction and demolition waste can provide secondary aggregates and other recoverable material fractions, although effective implementation depends on appropriate sorting, processing, quality control, and market acceptance [4, 5]. These applications illustrate the potential of waste valorization to reduce the demand for selected virgin resources while creating additional pathways for material utilization.

The range of industrial residues considered for sustainable material design is broad. It includes fly ash, blast-furnace slag, steel-making residues, red mud, foundry sand, construction and demolition waste, waste glass, quarry dust, mining residues, and rice-husk ash. However, these materials cannot be treated as interchangeable ingredients. Their properties may vary substantially according to their sources, production processes, chemical composition, mineralogical structure, particle-size distribution, moisture content, and pretreatment conditions. Even materials bearing the same general name may exhibit different engineering characteristics when obtained from different industrial facilities or production batches. Consequently, a waste material that performs satisfactorily in one formulation may not provide the same performance under different processing conditions or when combined with a different binder or matrix. This variability makes systematic characterization and data-driven evaluation important components of sustainable material development.

The successful incorporation of industrial waste into engineering materials depends on several interacting factors. These may include the proportion of waste used, the water-to-binder ratio, particle fineness, chemical composition, additive dosage, mixing procedure, processing temperature, compaction pressure, curing duration, and environmental exposure. The resulting engineering responses may include compressive strength, tensile strength, flexural strength, density, porosity, water absorption, thermal conductivity, durability, and resistance to chemical or physical degradation. In practice, these responses are often interdependent. For instance, a formulation that improves one mechanical property may simultaneously affect density, water absorption, workability, production cost, or long-term durability. The objective of material design is therefore not necessarily to maximize a single response, but to identify a formulation that achieves an appropriate balance among technical, economic, and environmental requirements.

The circular-economy perspective provides a broader context for this problem. Rather than following a linear pattern of resource extraction, manufacturing, use, and disposal, a circular approach seeks to retain the value of materials through reuse, recovery, recycling, and appropriate reintegration

into production systems. Reviews of construction and demolition waste management have identified material recovery and circular-economy strategies as important opportunities for improving resource efficiency, while also documenting practical barriers to their implementation [4, 5]. However, circularity alone does not guarantee environmental superiority. The suitability of a waste-derived material must be assessed in relation to its intended application, performance requirements, processing demands, and environmental consequences. For this reason, material characterization and quantitative performance evaluation should be integrated with sustainability assessment rather than treated as separate activities.

A major challenge in industrial waste valorization is that material performance is governed by numerous explanatory variables whose effects may be nonlinear and mutually dependent. Traditional experimental comparisons can identify differences between selected formulations, but they may provide an incomplete understanding when many variables change simultaneously. Statistical modeling offers a systematic framework for investigating such relationships. Descriptive statistics summarize the distribution and variability of observed measurements, while graphical exploration and correlation analysis help identify preliminary patterns and possible associations. Regression models can quantify relationships between explanatory variables and engineering responses, whereas analysis of variance (ANOVA) can be used to assess differences among experimental conditions under an appropriate statistical design. These methods provide a foundation for interpreting experimental evidence and identifying factors that may warrant further investigation.

A general representation of the relationship between material properties and explanatory variables is

$$Y_i = f(X_{i1}, X_{i2}, \ldots, X_{ip}) + \varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{1.1}$$

where Y_i denotes an observed engineering-performance response for the ith experimental observation, $X_{i1}, \ldots, X_{ip}$ represent the corresponding waste characteristics and processing variables, $f(\cdot)$ denotes an unknown response function, and ε_i represents unexplained variation. The sample size is n, and p denotes the number of explanatory variables. Depending on the experimental setting, the response may be a continuous measurement, such as compressive strength or thermal conductivity, or another outcome appropriate to the material-design problem.

The form of $f(\cdot)$ depends on the scientific problem and the available data. A linear regression model may be useful when the response can be adequately approximated by additive linear effects. Polynomial regression can accommodate curvature, while interaction terms can represent situations in which the effect of one variable depends on the level of another. Response surface methodology (RSM) provides a systematic collection of experimental-design and modeling techniques for studying relationships between several controllable factors and one or more responses. It is particularly useful when the objective is to identify operating conditions or material formulations that provide desirable responses within a specified experimental region [6, 7]. For industrial waste-derived materials, such methods can help investigate the combined influence of waste replacement level, binder content, curing conditions, and other formulation variables.

A further consideration is the hierarchical structure of many materials experiments. Measurements may be obtained from multiple specimens prepared within the same batch, from several batches produced on different days, or from different laboratories or industrial facilities. Observations within a common batch may share unmeasured characteristics and therefore may not be statistically independent. Treating all measurements as independent without considering the experimental structure can lead to inappropriate uncertainty estimates. Where the design supports such an analysis, mixed-effects models can distinguish systematic effects associated with measured formulation variables from variation attributable to batches, production runs, or other grouping factors. This is relevant to the transition from laboratory experimentation to industrial production, where repeatability and batch-to-batch consistency are important considerations.

Machine-learning methods provide additional tools for analyzing complex relationships between material composition and engineering performance. Algorithms such as regression trees, random forests, support vector regression, and artificial neural networks can be used to represent nonlinear associations without requiring the analyst to specify a complete parametric response function in

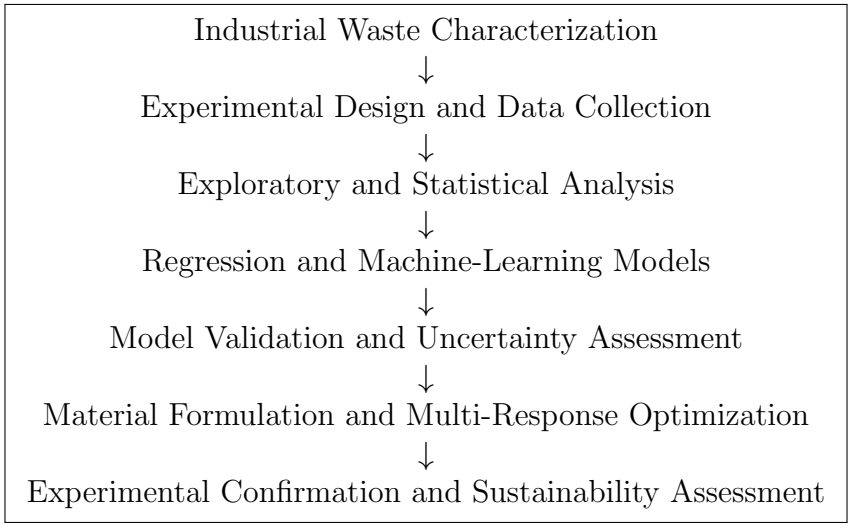
advance. A systematic review by Hassan et al. [8] documents the application of machine-learning methods to the prediction of properties across different types of concrete. Such methods can complement conventional statistical models when the available dataset contains sufficiently informative observations and the evaluation procedure is carefully designed. Nevertheless, predictive accuracy alone does not establish that a model is scientifically appropriate or that its predictions will generalize to new industrial settings. Data leakage, small sample sizes, imbalanced representation of formulations, measurement error, and differences between training and deployment conditions can all affect model performance. Accordingly, model evaluation should consider appropriate data partitioning, cross-validation where suitable, prediction errors, uncertainty, and the relevance of the validation data to the intended application.

The relationship between statistical prediction and material optimization is particularly important. A model may identify formulations associated with favorable experimental responses, but the preferred formulation must also satisfy practical constraints. These may include minimum strength requirements, acceptable water absorption, durability criteria, limits on waste content, manufacturing feasibility, and restrictions on production cost or environmental impact. When several responses are considered simultaneously, the problem becomes a multi-criteria or multi-objective optimization task. A formulation with the highest predicted compressive strength may not be the most appropriate choice if it requires excessive energy, increases cost, or performs poorly in another essential property. Moreover, any model-based optimum should be treated as a candidate for experimental confirmation rather than as conclusive evidence of optimal performance. Design of experiments, regression modeling, and response surface techniques provide established approaches for linking controlled experimentation with process and product improvement [7].

Sustainability assessment adds another dimension to this framework. The quantity of waste incorporated into a formulation is a useful resource-utilization indicator, but it is not sufficient on its own to establish an overall environmental benefit. Other indicators may include energy consumption, greenhouse-gas emissions, water use, production cost, transportation requirements, durability, and the potential for reuse or recycling at the end of the product's service life. Life-cycle assessment can provide a structured basis for evaluating environmental impacts across relevant stages of a product's life cycle, subject to the selected system boundaries, functional unit, inventory data, and assessment assumptions [2]. Consequently, statistical models can contribute to sustainability-oriented material design by estimating engineering responses and supporting comparisons among candidate formulations, while environmental assessment provides a complementary basis for evaluating broader consequences.

The present chapter adopts an application-oriented perspective to demonstrate how established statistical and machine-learning methods can support the evaluation of industrial waste-derived materials. The emphasis is on integrating experimental data, exploratory analysis, regression modeling, ANOVA, response surface methodology, mixed-effects modeling, predictive learning, and optimization within a coherent analytical workflow. The objective is not to introduce new statistical theory, but to explain how established techniques can be selected, interpreted, and applied to material-design problems. Particular attention is given to the quality of the available data, the interpretation of model assumptions, the assessment of predictive performance, and the distinction between statistical association and causal conclusions.

The proposed framework can be summarized as



This workflow emphasizes that sustainable material design is an iterative process rather than a single prediction exercise. Model-based recommendations should be checked against experimental evidence, engineering requirements, and environmental considerations. The resulting integration of statistical analysis and material characterization can support more transparent comparisons among candidate formulations and provide a quantitative basis for identifying promising uses of industrial residuals. Ultimately, the chapter seeks to demonstrate how data-driven methods can help connect industrial waste recovery with the development of technically suitable and environmentally informed engineering materials.

2 Industrial Waste and Sustainable Engineering Materials

2.1 Industrial waste as a secondary resource

Industrial waste should not necessarily be regarded solely as an environmental burden; under appropriate technical, economic, and environmental conditions, it can serve as a secondary resource for manufacturing value-added engineering materials. Industrial waste valorization involves recovering useful constituents or functional properties from residual materials and incorporating them into new products. This approach is consistent with circular-economy principles, which emphasize resource efficiency, material recovery, and the reduction of dependence on virgin raw materials [1]. For example, fly ash can be utilized in cementitious applications, while suitable construction and demolition waste can be processed into secondary aggregates and other construction products [3, 4].

Depending on their industrial origin, waste materials may contain silica, alumina, calcium compounds, iron oxides, carbonaceous matter, glassy phases, metallic constituents, or other potentially useful components. Their engineering functions depend not only on chemical composition but also on mineralogical structure, particle morphology, fineness, porosity, moisture content, and surface characteristics. Consequently, the same waste may be suitable for one application but unsuitable for another. Furthermore, the presence of undesirable constituents, soluble compounds, or potentially hazardous substances may restrict its use and necessitate appropriate treatment and safety assessment.

Waste characterization is therefore a fundamental prerequisite for material development. Relevant measurements may include chemical composition, particle-size distribution, specific gravity, moisture content, loss on ignition, and other application-specific properties. Measurements should also be collected across different batches whenever possible, since variations in the source material or production process may affect the consistency of the final product. From a statistical perspective, characterization provides the explanatory variables required to investigate the association between waste properties and engineering performance. It also helps distinguish genuine formulation effects

from variability attributable to differences in the waste itself.

2.2 Major categories of industrial waste

Industrial waste relevant to sustainable engineering materials can be classified according to its source, composition, and potential application. Important categories include the following:

(i) **Fly ash and bottom ash:** Residues from thermal power generation that may be considered for cementitious materials, bricks, and other construction products, subject to their composition and application requirements [3].

(ii) **Blast-furnace and steel-making slags:** Metallurgical residues that may provide mineral constituents for cementitious binders, aggregates, and other construction materials, depending on their origin, processing, and material properties.

(iii) **Red mud:** A residue from alumina production whose mineral composition may support selected applications in ceramics, construction materials, and other mineral-based products, provided that its alkalinity and potentially problematic constituents are appropriately evaluated.

(iv) **Foundry sand:** Used moulding and core sands that may be assessed as partial replacements for natural sand or as constituents of selected construction products, subject to contamination and quality-control requirements.

(v) **Construction and demolition waste:** Concrete, masonry, brick, and other recoverable fractions that can be processed into secondary aggregates and other construction materials. Their suitability depends on sorting, contamination, particle grading, and processing quality [4, 5].

(vi) **Waste glass:** Glass residues that may be processed for use as aggregate, filler, or a constituent of selected glass-based and cementitious products, depending on particle size and compatibility with the intended application.

(vii) **Quarry dust and stone-processing residues:** Fine mineral fractions that may be investigated as fillers or partial replacements for natural aggregates in appropriate construction applications.

(viii) **Mining and mineral-processing residues:** Tailings and other mineral residues that may contain useful mineral fractions but require careful assessment of chemical stability, contaminant content, and long-term environmental risks.

(ix) **Rice-husk and other biomass-derived ashes:** Ashes that may contain useful mineral constituents, including silica, depending on the feedstock and combustion conditions. Their suitability for cementitious and other applications depends on composition, fineness, and processing history.

(x) **Waste polymers and industrial plastic residues:** Selected polymer wastes that may be incorporated into composites, modified construction materials, or other products after appropriate sorting and processing. Compatibility, durability, and end-of-life considerations are important in determining their suitability.

(xi) **Other metallurgical and process-derived residues:** Sludges, mineral fines, process powders, and related by-products that may offer recovery opportunities when their composition, variability, and environmental safety are adequately established.

These categories illustrate the diversity of potential secondary resources; however, they should not be interpreted as evidence that every waste stream is suitable for direct reuse. The feasibility of valorization depends on the material's characteristics, processing requirements, relevant engineering standards, environmental safety, and the availability of a technically appropriate application. Consequently, a systematic assessment of both material quality and variability is essential before selecting a waste stream for further development.

2.3 Engineering applications

Industrial wastes can be incorporated into engineering materials through different mechanisms, depending on their physical and chemical properties and the characteristics of the receiving material. A waste may function as a partial replacement for cement, natural aggregate, sand, filler, binder, or another constituent. Some materials may contribute to pozzolanic reactions, whereas others may primarily provide particle packing, reinforcement, or mineral filler effects. Suitable aluminosilicate-rich residues may also be investigated as precursors for alkali-activated or geopolymer materials. The actual mechanism and resulting performance must be established for the particular waste, formulation, and processing conditions rather than assumed from the waste category alone.

Consider a composite in which industrial waste partially replaces a conventional constituent. Let

$$Z = \text{percentage of industrial waste incorporated,}$$

and let

$$Y = \text{compressive strength of the resulting material.}$$

The relationship between Z and Y may depend on particle packing, chemical reactivity, porosity, water demand, and the properties of the remaining constituents. A moderate replacement level may improve some aspects of performance, whereas excessive replacement may adversely affect strength or other required properties. The response may therefore be nonlinear and may also depend on variables such as curing time, binder content, and water-to-binder ratio.

A simple quadratic model can represent such curvature:

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 Z_i^2 + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{2.1}$$

where β_0, β_1, and β_2 are unknown parameters and ε_i represents unexplained variation. A negative quadratic coefficient, when supported by the data and model assumptions, may indicate a concave response over the observed range. However, the estimated turning point should be interpreted as a candidate optimum only if it lies within the investigated range and the model provides an adequate representation of the data.

When several formulation variables interact, a more general model is appropriate:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \sum_{j=1}^{p} \beta_{jj} X_{ij}^2 + \sum_{j<k} \beta_{jk} X_{ij} X_{ik} + \varepsilon_i, \tag{2.2}$$

where X_{ij} denotes the jth explanatory variable for the ith observation, β_j represents a linear effect, β_{jj} represents a quadratic effect, and β_{jk} represents an interaction effect. Such a model provides a basis for response surface analysis and formulation optimization when the experimental design contains sufficient information to estimate the required terms [6, 7].

The same modeling framework can be extended to responses such as water absorption, density, flexural strength, thermal conductivity, and durability indicators. Since the formulation that maximizes one response may not optimize the others, practical material development often requires simultaneous consideration of several performance criteria. Statistical analysis can help identify influential variables, quantify uncertainty, compare candidate formulations, and select promising conditions for experimental confirmation.

2.4 Circular economy perspective

The circular economy seeks to retain the value of resources through waste prevention, reuse, recycling, and recovery, thereby reducing dependence on virgin materials and limiting avoidable disposal. Industrial waste valorization can contribute to these objectives by returning suitable residual materials to productive applications. In the construction sector, the recovery of construction and demolition waste illustrates how secondary materials can be incorporated into new production cycles, although implementation remains dependent on material quality, processing infrastructure, economic feasibility, and regulatory requirements [4, 5]. A simplified pathway for industrial waste valorization is

$$\begin{aligned}\text{Industrial Process} &\longrightarrow \text{Waste Generation}\\ &\longrightarrow \text{Characterization and Treatment}\\ &\longrightarrow \text{Statistical Analysis and Material Design}\\ &\longrightarrow \text{Performance Validation}\\ &\longrightarrow \text{Engineering Application.}\end{aligned}$$

Statistical modeling can support this pathway by relating waste characteristics and processing conditions to measurable engineering responses. It can also help quantify batch variability, compare alternative formulations, predict performance within a validated domain, and identify promising conditions for further experimentation. Nevertheless, statistical prediction alone cannot establish the overall sustainability of a proposed material. The environmental benefits of replacing conventional resources must be assessed in relation to factors such as energy consumption, transportation, processing requirements, product service life, and end-of-life management. Life-cycle assessment provides a structured framework for evaluating relevant environmental impacts across a defined product system [2].

Accordingly, successful waste valorization requires the integration of material characterization, statistical analysis, engineering validation, and sustainability assessment. This integrated approach can help distinguish technically feasible applications from those that are merely plausible, while providing a quantitative basis for developing reliable and resource-efficient engineering materials.

3 Data Collection and Experimental Variables

3.1 Data structure

The quality of statistical analysis depends fundamentally on the quality, relevance, and structure of the collected data. In industrial waste valorization, observations may be obtained from laboratory experiments, pilot-scale production, industrial monitoring, or published experimental studies. Each observation should be associated with clearly defined material characteristics, formulation variables, processing conditions, and measured engineering responses. Wherever possible, information about the waste source, production batch, specimen preparation, testing procedure, and measurement units should also be recorded.

A general representation of the dataset is

$$\mathcal{D} = \left\{ \left(Y_i, \boldsymbol{X}_i^\top, \boldsymbol{Z}_i^\top \right) : i = 1, \ldots, n \right\}, \tag{3.1}$$

where Y_i denotes an engineering response, $\boldsymbol{X}_i$ is a vector of p waste characteristics, $\boldsymbol{Z}_i$ is a vector of q processing or formulation variables, and n is the number of experimental observations. When multiple responses are recorded, the response Y_i may be replaced by a vector $\boldsymbol{Y}_i \in \mathbb{R}^m$.

For illustration, consider the partial replacement of a conventional binder with fly ash. The dataset may contain the fly ash replacement percentage, water-to-binder ratio, curing temperature, curing duration, and compressive strength. The primary objective may be to investigate how the formulation and processing variables influence compressive strength while accounting for experimental variability. A hypothetical experimental record is presented below.

Table 1: Illustrative structure of a waste-upcycling dataset.

Obs.	Fly ash (%)	Water/binder ratio	Curing (days)	Temperature (°C)	Strength (MPa)
1	0	0.45	28	25	42
2	10	0.45	28	25	45
3	20	0.45	28	25	48
4	30	0.45	28	25	47
5	40	0.45	28	25	44
6	50	0.45	28	25	39
7	60	0.45	28	25	33
8	70	0.45	28	25	27

Note: The values are hypothetical and are used solely to illustrate the statistical procedures. They do not represent experimentally validated fly ash performance.

The table 1 represents eight formulations evaluated under the same nominal water-to-binder ratio and curing conditions. This controlled comparison illustrates the relationship between replacement level and compressive strength, but it cannot establish the independent effects of curing conditions because those conditions do not vary. A more comprehensive experiment would vary selected factors systematically and include sufficient replication to estimate experimental error.

An important distinction is that specimens prepared from the same batch are not necessarily equivalent to independently produced experimental batches. If several specimens are obtained from each batch, their shared production conditions may induce dependence. Accordingly, the experimental unit, sampling structure, and replication scheme should be established before selecting a statistical model [7].

3.2 Waste characteristics

Waste characteristics describe the composition and physical properties of the material being incorporated into the final product. Depending on the application, the explanatory vector may be written as

$$\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^\top . \tag{3.2}$$

Typical variables include chemical composition, silica and alumina contents, calcium and iron concentrations, particle-size distribution, moisture content, specific gravity, specific surface area, pH, loss on ignition, organic content, and potentially hazardous constituents. Not every variable is required in every study; selection should be guided by material science, the intended application, measurement availability, and relevant safety requirements.

For example, the silica and alumina contents of an aluminosilicate-rich waste may be relevant to its potential use in cementitious or alkali-activated materials. Particle fineness may influence packing, surface area, and reaction rates, whereas moisture content can affect the effective water content of a formulation. Such relationships provide scientific reasons for including these variables in the analysis rather than selecting predictors solely through automated statistical procedures.

Measurements should, whenever feasible, be collected from multiple batches or sources. The resulting variability can then be summarized and incorporated into subsequent models. Where chemical composition is expressed as percentages of total composition, the components are also subject to a sum constraint. Consequently, increases in one component may necessarily correspond to decreases in others, creating dependencies that must be considered during regression modeling.

3.3 Processing and formulation variables

Processing variables describe the conditions under which a material is formulated, manufactured, cured, and tested. They may be represented by

$$\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, \ldots, Z_{iq})^\top . \tag{3.3}$$

Relevant examples include

$$\begin{aligned} Z_{i1} &= \text{waste replacement percentage}, \\ Z_{i2} &= \text{water-to-binder ratio}, \\ Z_{i3} &= \text{curing temperature}, \\ Z_{i4} &= \text{curing duration}, \\ Z_{i5} &= \text{compaction pressure}. \end{aligned}$$

Other variables may include additive dosage, mixing time, particle pretreatment, processing pressure, and drying conditions. The choice depends on the material and production method.

Careful experimental design is essential because formulation variables may interact. For instance, the effect of increasing waste replacement may differ according to curing duration or water-to-binder ratio. Consequently, varying one factor at a time may fail to reveal important interactions. Factorial designs and response surface methodology provide systematic approaches to investigating such effects and identifying promising operating conditions [6, 7].

The experimental range should be selected using scientific knowledge, safety considerations, engineering specifications, and practical manufacturing constraints. Predictions beyond the observed range should be treated cautiously because an apparently adequate model within the experimental domain may behave unrealistically under extrapolation.

3.4 Engineering responses

The response variables quantify the performance of the resulting material. Depending on the application, a multivariate response may be represented by

$$\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{im})^\top , \tag{3.4}$$

where the components may include

$$\begin{aligned} &Y_{i1} = \text{compressive strength}, \quad Y_{i2} = \text{flexural strength}, \quad Y_{i3} = \text{tensile strength}, \\ &Y_{i4} = \text{water absorption}, \quad Y_{i5} = \text{density}, \quad Y_{i6} = \text{thermal conductivity}, \\ &Y_{i7} = \text{durability indicator}. \end{aligned}$$

These responses represent different aspects of material performance and should be interpreted according to their engineering significance. Higher strength may be desirable, whereas lower water absorption may be preferred in some applications. The desirable direction for thermal conductivity depends on whether the intended product is a thermal insulator or a heat-conducting material. Durability also requires a clearly specified measurement procedure, exposure condition, and observation period.

Thus, material selection should not ordinarily be based on a single response in isolation. A formulation that performs well mechanically may be less satisfactory in terms of cost, durability, or environmental impact. The response variables should therefore be chosen to reflect the intended application and the practical criteria against which candidate materials will be evaluated.

4 Exploratory Statistical Analysis

Exploratory statistical analysis is the initial stage of quantitative investigation. Its purpose is to understand the structure of the collected data before fitting predictive or explanatory models. Descriptive summaries, graphical displays, and diagnostic checks can reveal variability, skewness, unusual observations, missing measurements, and preliminary relationships among variables [9, 7]. These procedures do not replace formal modeling; rather, they help determine which models and diagnostic methods are appropriate.

4.1 Descriptive statistics

For observations $X_1, \ldots, X_n$, the sample mean is

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \tag{4.1}$$

and the sample variance is

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2. \tag{4.2}$$

The sample standard deviation is $s_X = \sqrt{s_X^2}$. The mean summarizes the central location of the observations, whereas the standard deviation measures their dispersion around the sample mean.

For a positive variable with a nonzero mean, the coefficient of variation is

$$CV_X = \frac{s_X}{|\bar{X}|} \times 100\%. \tag{4.3}$$

It expresses variability relative to the magnitude of the mean and can assist comparisons among positive measurements recorded on different scales. However, it becomes unstable when the mean is close to zero and is generally inappropriate for variables whose meaningful zero or ratio-scale interpretation is absent.

For the illustrative compressive-strength observations in Table 1, the mean is

$$\bar{Y} = \frac{325}{8} = 40.625 \text{ MPa}.$$

Using the sample-variance formula gives

$$s_Y^2 \approx 53.41 \text{ MPa}^2, \qquad s_Y \approx 7.31 \text{ MPa}.$$

Consequently,

$$CV_Y = \frac{7.31}{40.625} \times 100\% \approx 18.0\%.$$

These calculations summarize the strength values across the eight different formulations. The observed variability combines the differences among replacement levels with any other variation present in the measurements. It should not be interpreted as a pure estimate of measurement error because the formulations themselves differ.

For an experimental study, a descriptive table should normally report the sample size, mean, standard deviation, median, interquartile range, and minimum and maximum values, as appropriate. Where replication is available, summaries should also distinguish variability within a formulation from variability between formulations or production batches.

4.2 Distributional analysis

Histograms, empirical distribution plots, density estimates, and boxplots provide complementary information about the distribution of a variable. Histograms reveal broad distributional features, while boxplots help identify unusual observations relative to the bulk of the data. Quantile–quantile plots can be useful when assessing whether a distribution is approximately consistent with a proposed reference distribution.

For example, compressive-strength observations may exhibit skewness, particularly when several formulations or material sources are combined. A transformation may be useful if it improves model adequacy or stabilizes the variability. Common transformations include

$$Y_i^* = \log(Y_i), \qquad Y_i > 0, \tag{4.4}$$

and

$$Y_i^* = \sqrt{Y_i}, \qquad Y_i \geq 0. \tag{4.5}$$

The logarithmic transformation is suitable only for strictly positive observations, whereas the square-root transformation can be applied to nonnegative observations. The choice should be based on the scientific meaning of the response, the observed distribution, and the assumptions of the intended model. Transformation should not be performed merely because the raw observations fail a normality test; in regression, the relevant distributional assumptions generally concern the model errors rather than requiring every predictor and response to be normally distributed.

4.3 Graphical exploration

Scatterplots provide a useful graphical tool for examining the relationship between an explanatory variable and an engineering response. For the present illustrative example, Figure 1 displays the relationship between fly ash replacement percentage and compressive strength using the eight observed pairs (Z_i, Y_i), $i = 1, \ldots, 8$, where Z_i denotes the replacement percentage and Y_i denotes compressive strength. The scatterplot indicates that compressive strength initially increases and then decreases, with the largest observed value of 48 MPa occurring at a replacement level of 20%.

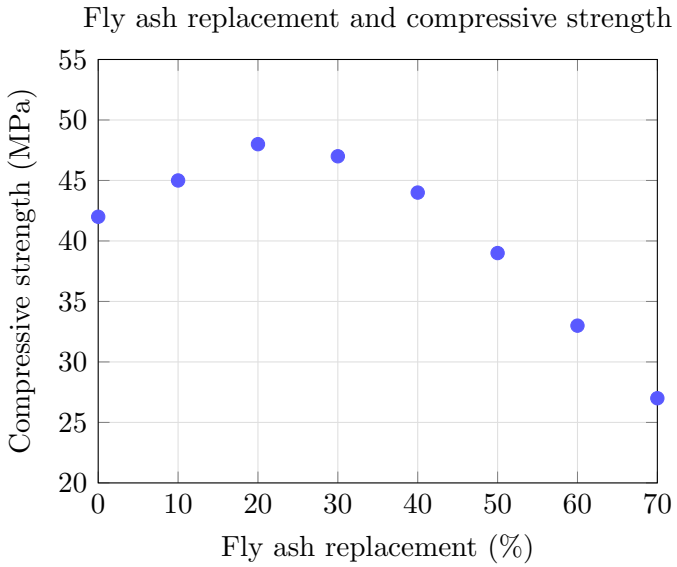


Figure 1: Scatterplot of fly ash replacement percentage against compressive strength for the illustrative dataset.

The observed pattern suggests that a linear model may not adequately represent the relationship between fly ash replacement and compressive strength. However, the largest observed response should not be interpreted as a validated optimum. The dataset contains only eight illustrative

observations, without independent replication or estimates of uncertainty at each replacement level. Consequently, the apparent maximum is descriptive and requires further experimental investigation.

For datasets involving multiple explanatory variables, scatterplot matrices, correlation heatmaps, and plots stratified by waste source or production batch can reveal associations that may not be apparent from individual summaries. Such graphical displays should be interpreted in conjunction with the experimental design, since an apparent association may arise from differences in other explanatory variables or from variations in the composition of the sampled batches.

4.4 Outlier assessment and data quality

An unusual observation may arise from a recording error, instrument malfunction, specimen damage, protocol deviation, or genuine heterogeneity in the waste material. Consequently, an observation should not be removed merely because it is far from the sample mean or appears as an outlier in a boxplot.

Potentially influential observations should be checked against laboratory records, instrument calibration, sample identification, and specimen-preparation procedures. If the observation is valid, it may contain important information about the range of material behavior. If an error is confirmed, the correction or exclusion should be documented transparently.

Data-quality assessment should also examine missing values, inconsistent units, duplicate records, impossible measurements, and differences in testing procedures. Missing observations should be investigated before choosing a deletion or imputation method. In particular, the analysis should preserve the distinction between a genuinely unobserved measurement and a measurement that is absent because a specimen failed or a material did not meet a testing criterion.

5 Correlation and Variable Selection

5.1 Pearson correlation

The Pearson correlation coefficient measures the strength and direction of the linear association between two quantitative variables. For paired observations (X_i, Y_i), it is defined as

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}. \tag{5.1}$$

Its value lies between -1 and 1. Values near 1 indicate strong positive linear association, whereas values near -1 indicate strong negative linear association. A value near zero indicates little linear association, but it does not rule out a nonlinear relationship.

For the eight illustrative observations, let X denote fly ash replacement percentage and Y denote compressive strength. The sample Pearson correlation is approximately $r_{XY} \approx -0.78$. This indicates a negative overall linear association across the observed replacement levels. However, the scatterplot shows that strength first increases and then decreases. The correlation coefficient therefore does not fully describe the response pattern. This example illustrates why correlation should be interpreted alongside graphical exploration and, when appropriate, nonlinear modeling.

5.2 Spearman's Rank correlation

When the relationship is monotonic but may not be linear, Spearman's rank correlation provides a complementary measure. It is the Pearson correlation computed using the ranks of the observations.

When there are no tied ranks, it can be written as

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{5.2}$$

where d_i is the difference between the ranks of X_i and Y_i.

Spearman's coefficient is less dependent on the original measurement scales and can detect monotonic associations that are not linear. However, it does not generally identify a nonmonotonic relationship such as a response that increases and then decreases. Neither Pearson nor Spearman correlation establishes causation; the interpretation must account for the experimental design and potential confounding variables.

5.3 Correlation matrix

For p explanatory variables, the pairwise correlations can be organized into the matrix

$$\boldsymbol{R} = (r_{jk})_{p\times p}, \tag{5.3}$$

where r_{jk} denotes the correlation between predictors X_j and X_k. A heatmap of this matrix can help identify groups of strongly associated variables and possible redundancy in the predictor set.

For example, silica and alumina contents may be associated because they arise from related mineralogical constituents. Likewise, particle size and specific surface area may be strongly related. Such relationships are scientifically plausible but must be assessed using the actual data rather than assumed in advance.

A further issue arises when chemical composition is recorded as percentages that sum to a fixed total. In such cases, the variables are compositionally constrained, and ordinary correlation can be difficult to interpret independently of that constraint. Depending on the research objective, compositional-data methods or scientifically justified transformations may be more appropriate than including every component as an unrestricted predictor.

5.4 Multicollinearity

Multicollinearity occurs when two or more explanatory variables contain overlapping linear information. It can make regression coefficients unstable, increase their standard errors, and complicate the interpretation of individual effects, even when the model retains useful predictive performance.

The variance inflation factor for predictor X_j is

$$VIF_j = \frac{1}{1-R_j^2}, \tag{5.4}$$

where R_j^2 is the coefficient of determination obtained by regressing X_j on the remaining explanatory variables. A value near 1 indicates little linear inflation from the other predictors, whereas larger values indicate increasing redundancy. Values such as 5 or 10 are sometimes used as diagnostic reference points, but neither constitutes a universal threshold for excluding a variable.

In waste-derived material studies, multicollinearity may arise from related chemical components, correlated processing conditions, or variables that are calculated from one another. The appropriate response depends on the objective. For explanatory modeling, scientifically justified simplification or carefully chosen predictor sets may improve interpretation. For prediction, regularization methods such as ridge regression can stabilize estimates without requiring the automatic deletion of scientifically relevant variables [10].

5.5 Variable selection

Variable selection aims to identify a useful set of explanatory variables for the scientific or predictive objective. The process should begin with domain knowledge: variables known to influence material performance should be considered even when their individual correlations with the response are modest. A variable may contribute through an interaction or nonlinear effect that is not captured by a simple pairwise correlation.

Common approaches include scientifically guided selection, comparison of candidate models using information criteria, stepwise procedures, regularization methods such as the lasso, and variable-importance measures from machine-learning algorithms. These approaches differ in their assumptions, stability, and interpretability [7, 10]. Automated selection should not be treated as a substitute for experimental understanding, and stepwise selection in particular can produce unstable models and overstate the apparent certainty of selected coefficients.

For predictive applications, variable selection should be performed within the training and validation procedure. If predictors are selected using the entire dataset before cross-validation, information from the validation observations can leak into the modeling process, leading to an overly optimistic estimate of predictive performance. Where sufficient data are available, the complete selection and model-fitting procedure should be repeated within each training fold.

For the illustrative fly ash example, replacement percentage is an obvious candidate predictor because it is the experimental factor under investigation. However, the eight observations do not support a reliable assessment of a large predictor set, nor can they establish the independent effects of curing, water-to-binder ratio, or waste composition because these variables were held fixed or were not recorded as varying factors. A more informative study would vary the relevant factors systematically, include suitable replication, and evaluate candidate models using diagnostic checks and independent validation where feasible.

Ultimately, the selected variables should provide a defensible balance between scientific relevance, interpretability, data availability, and predictive performance. The purpose is not simply to retain the largest possible number of statistically associated predictors, but to develop a model that adequately represents the material system and supports reliable conclusions within the domain for which data are available.

6 Regression Models for Material Performance

Regression analysis quantifies the relationship between material performance and formulation or processing variables. For waste-derived materials, the response may represent compressive strength, particle size, thermal stability, durability, or pollutant-removal efficiency. The following numerical illustrations demonstrate the methodology; the coefficients and observations are hypothetical and should not be interpreted as results from an actual experiment.

6.1 Multiple linear regression

Suppose the compressive strength Y_i depends on waste replacement percentage W_i, curing temperature T_i, curing duration C_i, and water-to-binder ratio R_i. The multiple linear regression model is

$$Y_i = \beta_0 + \beta_1 W_i + \beta_2 T_i + \beta_3 C_i + \beta_4 R_i + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{6.1}$$

where ε_i represents random variation not explained by the included predictors. The coefficient β_j measures the expected change in strength associated with a one-unit increase in the corresponding predictor, holding the other predictors fixed.

For illustration, consider the hypothetical fitted equation

$$\widehat{Y} = 25 + 0.18W + 0.12T + 0.08C - 18R. \tag{6.2}$$

Here, W is measured in percentage points, T in degrees Celsius, C in days, and R is dimensionless. For $W = 30$, $T = 40$, $C = 28$, and $R = 0.45$, the predicted strength is

$$\begin{aligned}\widehat{Y} &= 25 + 0.18(30) + 0.12(40) + 0.08(28) - 18(0.45) \\ &= 29.34 \text{ MPa}.\end{aligned}$$

Thus, the model predicts a strength of 29.34 MPa for this illustrative combination. This prediction is meaningful only within the context of the assumed equation and the range of experimental conditions used to fit a real model. In practice, the coefficients are estimated by least squares, which minimizes

$$\text{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2. \tag{6.3}$$

The fitted model should also be assessed using residual diagnostics and appropriate measures of predictive performance.

6.2 Polynomial regression

A linear model may be inadequate when increasing the waste content initially improves performance but eventually reduces it. A quadratic model allows curvature:

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 Z_i^2 + \varepsilon_i, \tag{6.4}$$

where Z_i denotes waste replacement percentage. If $\beta_2 < 0$, the fitted curve is concave and may possess an interior maximum. Its stationary point is

$$Z^* = -\frac{\beta_1}{2\beta_2}. \tag{6.5}$$

This point represents a maximum only when $\beta_2 < 0$ and it lies within the feasible experimental range.

For example, consider the illustrative equation

$$\widehat{Y} = 42 + 0.45Z - 0.0125Z^2. \tag{6.6}$$

The estimated stationary point is $Z^* = -\dfrac{0.45}{2(-0.0125)} = 18\%$.

The predicted response there is

$$\widehat{Y}(18) = 42 + 0.45(18) - 0.0125(18)^2 = 46.05 \text{ MPa}.$$

This example illustrates how a quadratic model can describe an interior maximum. The value 18% is not an empirical optimum; it follows solely from the hypothetical coefficients in (6.6).

A cubic model,

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 Z_i^2 + \beta_3 Z_i^3 + \varepsilon_i, \tag{6.7}$$

can accommodate more complex curvature. However, additional polynomial terms increase model flexibility and may cause overfitting, especially when observations are few. Model order should therefore be guided by experimental design, residual analysis, and validation rather than by goodness of fit alone.

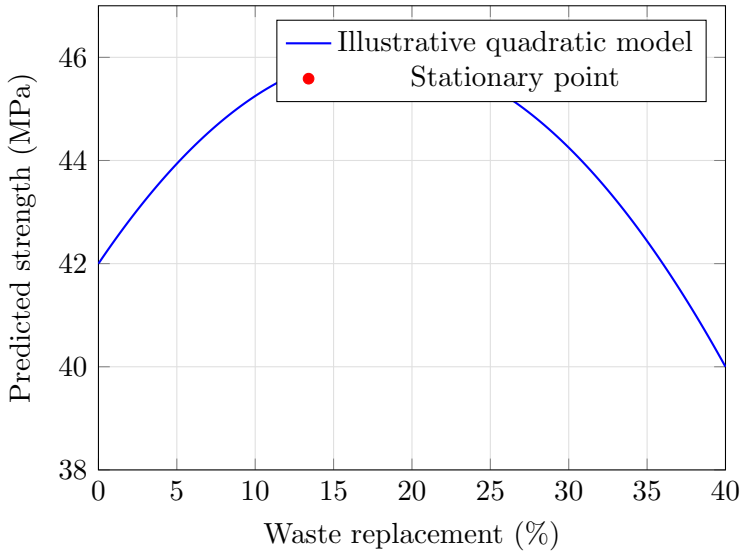


Figure 2: Illustrative quadratic response curve and its stationary point. The curve is generated from hypothetical coefficients.

6.3 Interaction models

The effect of one processing variable may depend on the level of another. For example, the effect of waste replacement may differ between curing temperatures. An interaction model is

$$Y_i = \beta_0 + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \beta_3 Z_{i1} Z_{i2} + \varepsilon_i, \tag{6.8}$$

where Z_{i1} and Z_{i2} denote the two explanatory variables. The interaction coefficient β_3 describes how the effect of one predictor changes with the other, subject to the model's scale and specification.

For a numerical illustration, let $x_1 = (W - 30)/10$ and $x_2 = (T - 40)/10$ be coded waste content and curing temperature, and consider

$$\widehat{Y} = 45 + 2x_1 + 1.5x_2 + 1.2x_1 x_2. \tag{6.9}$$

At $x_2 = -1$, the slope with respect to x_1 is $2 + 1.2(-1) = 0.8$. At $x_2 = 1$, it is $2 + 1.2(1) = 3.2$. Thus, the predicted change per unit of coded waste content is larger at the higher coded temperature. This illustrates an interaction; it does not establish that such an interaction exists in a real material system.

6.4 Model diagnostics

A regression model should be examined for violations of its assumptions and for systematic patterns that it fails to capture. The residuals are

$$e_i = Y_i - \widehat{Y}_i. \tag{6.10}$$

Important diagnostic checks include residual normality when required for inferential procedures, constant error variance, independence, influential observations, and correct functional specification. Residual-versus-fitted plots can reveal curvature or changing variance, while normal Q–Q plots can indicate departures from normality. Leverage and influence measures, such as Cook's distance, can identify observations that exert disproportionate influence.

For example, if residual magnitudes tend to increase with fitted strength, the constant-variance assumption may be questionable. A transformation, a variance model, or a more appropriate regression structure may then be considered. Diagnostic findings should be interpreted jointly rather than using a single plot as a definitive test.

7 ANOVA and Comparative Analysis

Analysis of variance (ANOVA) compares mean engineering responses across material types, replacement levels, or processing conditions. It separates total observed variability into components associated with specified experimental factors and unexplained variation. Reliable inference requires an appropriate design and attention to independence, error variance, and model assumptions.

7.1 One-way ANOVA

Suppose three waste-derived formulations, A, B, and C, are tested, with four independent strength measurements per formulation. The one-way ANOVA model is

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad i = 1, \ldots, k, \quad j = 1, \ldots, n_i, \tag{7.1}$$

where μ is the overall mean and α_i is the effect of formulation i. The hypotheses are

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k, \qquad H_1 : \text{at least one group mean differs.} \tag{7.2}$$

Consider the following hypothetical observations, measured in MPa, given by Table 2:

Formulation	1	2	3	4	Mean
A	42	44	43	45	43.5
B	46	48	47	49	47.5
C	44	45	46	43	44.5

Table 2: Hypothetical compressive-strength observations (MPa) for three formulations, based on four experimental replicates, with the corresponding mean strength.

The overall mean is $\bar{Y}_{\cdot\cdot} = 45.17$ MPa (rounded). The between-group and within-group sums of squares are

$$SS_B = \sum_{i=1}^{k} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \approx 34.67, \tag{7.3}$$

$$SS_W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2 = 15. \tag{7.4}$$

Consequently,

$$SS_T = SS_B + SS_W \approx 49.67.$$

With $k = 3$ groups and $N = 12$ observations, the degrees of freedom are 2 between groups and 9 within groups. Hence,

$$MS_B = \frac{34.67}{2} \approx 17.33, \qquad MS_W = \frac{15}{9} \approx 1.67,$$

and the test statistic is

$$F = \frac{MS_B}{MS_W} \approx 10.40.$$

Under the usual one-way ANOVA assumptions, this value provides evidence against equality of all three population means at the 5% significance level. The calculation is illustrative; actual inference requires genuine independent experimental observations. A significant ANOVA result establishes that not all means are equal, but it does not by itself identify which pairs differ.

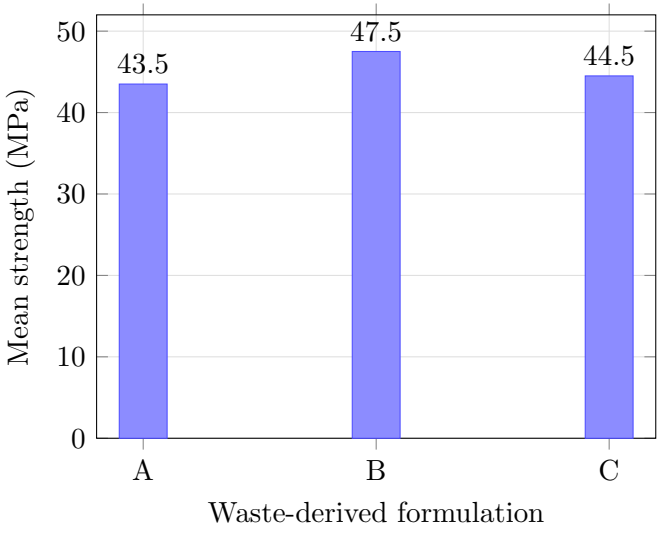


Figure 3: Illustrative group means for three waste-derived formulations.

7.2 Two-way ANOVA

When both waste type and replacement level are studied, a two-way ANOVA model can separate their main effects and their interaction:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \tag{7.5}$$

where α_i is the waste-type effect, β_j is the replacement-level effect, and $(\alpha\beta)_{ij}$ is their interaction. A meaningful interaction indicates that differences among replacement levels are not the same for all waste types. For instance, a replacement level that performs well with fly ash may not perform equally well with slag or rice-husk ash. Replication within factor combinations is generally needed to estimate experimental error and support interaction testing.

7.3 Post-hoc comparisons

When the overall ANOVA test rejects the null hypothesis, a multiple- comparison procedure can be used to investigate specific group differences. Tukey's honestly significant difference procedure is one option for comparing all pairs of means while controlling the family-wise error rate under its assumptions. Comparisons should include uncertainty estimates, such as confidence intervals, rather than reporting only rankings of the sample means. When variances differ substantially, alternative procedures such as Welch's ANOVA and suitable unequal-variance post-hoc comparisons may be more appropriate.

8 Response Surface Methodology

Response Surface Methodology (RSM) combines designed experiments and regression modeling to characterize the effects of controllable variables and identify favorable operating conditions. It is particularly useful when the response depends on several factors and curvature or interactions are expected.

Let $x_1 = \dfrac{W - 30}{10}$, $x_2 = \dfrac{T - 40}{10}$, where W is waste replacement percentage and T is curing temperature in degrees Celsius. These transformations define dimensionless coded variables centred at $W = 30\%$ and $T = 40°\text{C}$. A second-order response surface is

$$\widehat{Y}(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2. \tag{8.1}$$

The linear terms represent local main effects, the squared terms represent curvature, and the product term represents interaction.

8.1 Response surface visualization

For illustration, consider the hypothetical fitted surface

$$\widehat{Y}(x_1, x_2) = 48 + 2x_1 + 1.5x_2 - 2x_1^2 - 1.5x_2^2 + 0.8x_1x_2, \qquad -1 \le x_1, x_2 \le 1. \tag{8.2}$$

The surface is concave because the quadratic terms are negative and the Hessian matrix

$$H = \begin{pmatrix} -4 & 0.8 \\ 0.8 & -3 \end{pmatrix}$$

is negative definite: its first leading entry is negative and $\det(H) = 12 - 0.64 = 11.36 > 0$. Therefore, the stationary point is the unique maximum of this quadratic surface, provided it lies inside the specified feasible region.

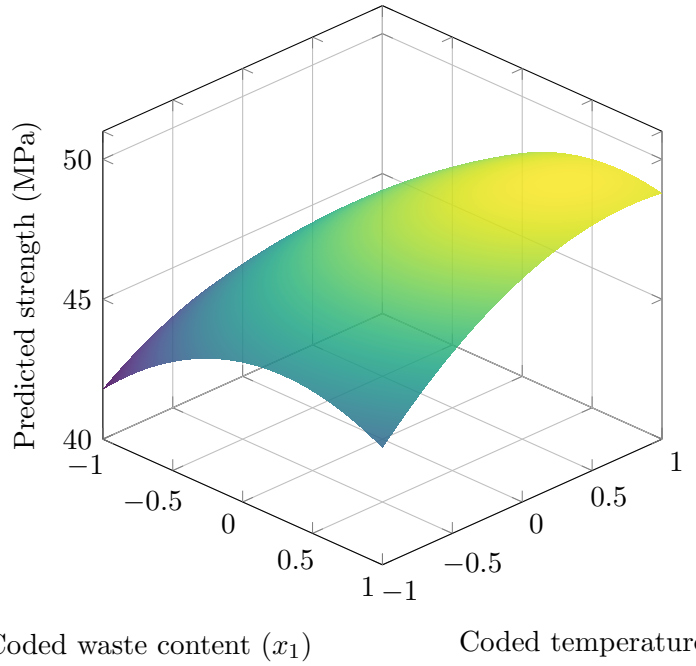


Figure 4: Three-dimensional response surface for the illustrative second-order model in (8.2).

8.2 Optimization through RSM

The estimated optimum is obtained by solving

$$(\widehat{x}_1, \widehat{x}_2) = \arg\max_{(x_1,x_2)\in\mathcal{R}} \widehat{Y}(x_1, x_2), \tag{8.3}$$

where $\mathcal{R}$ denotes the feasible experimental region. For the illustrative surface in (8.2), the stationary point solves

$$\frac{\partial \widehat{Y}}{\partial x_1} = 2 - 4x_1 + 0.8x_2 = 0, \qquad \frac{\partial \widehat{Y}}{\partial x_2} = 1.5 + 0.8x_1 - 3x_2 = 0.$$

Solving these equations gives approximately $\widehat{x}_1 = 0.634$, $\widehat{x}_2 = 0.669$. Returning to the original units,

$$\widehat{W} = 30 + 10(0.634) = 36.34\%, \qquad \widehat{T} = 40 + 10(0.669) = 46.69^{\circ}\text{C}.$$

The predicted response at this point is approximately $\widehat{Y}(0.634, 0.669) \approx 49.14$ MPa.

These values illustrate the optimization procedure, not a validated industrial formulation. In an actual study, the candidate optimum should be checked against operational constraints, prediction uncertainty, and independent confirmation experiments. If the stationary point lies outside the feasible region, optimization must instead be carried out over the permitted region, including its boundaries.

In sustainable material design, maximizing strength alone may not be sufficient. A practical optimization may also impose limits on cost, energy consumption, emissions, durability, or the proportion of waste incorporated. These requirements can be represented as constraints or combined in a multi-response desirability function. The final recommendation should balance predicted performance with environmental and operational feasibility.

9 Mixed-Effects Models for Batch Variability

Industrial waste streams often vary in chemical composition, particle size, moisture content, and contaminant concentration. Consequently, observations from the same production batch may be more similar than observations from different batches. Mixed-effects models account for this hierarchical structure while estimating the effects of formulation and processing variables.

9.1 Random-intercept model

Let Y_{ij} denote the engineering response for observation i from batch j. A random-intercept model is

$$Y_{ij} = \beta_0 + \sum_{r=1}^{p} \beta_r X_{ijr} + b_j + \varepsilon_{ij}, \tag{9.1}$$

where b_j is the batch-specific random intercept and ε_{ij} is the observation-level error. A standard assumption is

$$b_j \sim N(0, \sigma_b^2), \qquad \varepsilon_{ij} \sim N(0, \sigma^2),$$

with independent batch effects and errors. The variance σ_b^2 quantifies between-batch variability, whereas σ^2 represents residual variability within batches after accounting for the predictors.

For example, suppose the estimated variance components are $\widehat{\sigma}_b^2 = 9$ and $\widehat{\sigma}^2 = 4$, in MPa^2. The total variance for observations with the same predictor values is then

$$\widehat{\sigma}^2_{\text{total}} = 9 + 4 = 13.$$

The intraclass correlation coefficient (ICC) is

$$\widehat{\rho} = \frac{\widehat{\sigma}_b^2}{\widehat{\sigma}_b^2 + \widehat{\sigma}^2} = \frac{9}{13} \approx 0.692. \tag{9.2}$$

Thus, approximately 69.2% of the modeled variance is attributable to differences between batches. This hypothetical result would indicate substantial batch heterogeneity and support monitoring batch-level properties during industrial production.

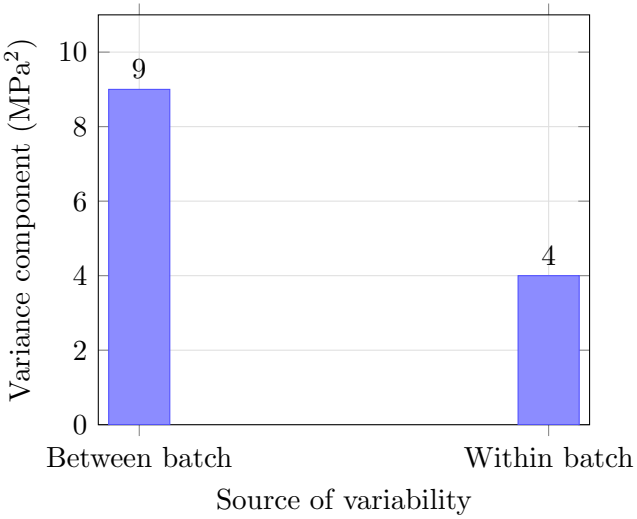


Figure 5: Hypothetical variance components in a random-intercept model for material strength.

9.2 Batch-specific predictions and repeated measurements

The estimated batch effect can be used to obtain batch-specific predictions:

$$\widehat{Y}_{ij} = \widehat{\beta}_0 + \sum_{r=1}^{p} \widehat{\beta}_r X_{ijr} + \widehat{b}_j.$$

The estimates $\widehat{b}_j$ are commonly obtained using empirical best linear unbiased prediction (BLUP) under the fitted model. They are typically shrunk towards zero, especially for batches with limited observations. This avoids treating a noisy batch mean as if it were estimated without uncertainty.

Mixed-effects models can also accommodate measurements taken at multiple curing ages. For example,

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0j} + b_{1j} t_{ij} + \varepsilon_{ij}, \tag{9.3}$$

where t_{ij} is curing age, b_{0j} is a batch-specific intercept, and b_{1j} is a batch-specific slope. The random slope allows strength development to vary between batches. If observations are repeated on the same specimen, their within-specimen dependence must also be considered when specifying the model.

10 Machine-Learning Models for Prediction

Regression models are useful for interpretation and inference, but industrial-material relationships may involve nonlinear effects, thresholds, and complex interactions. Machine-learning methods can provide flexible predictions when sufficient representative data are available. Their usefulness depends on sample size, data quality, validation design, and the intended application.

10.1 Random forest

A random forest averages predictions from an ensemble of decision trees:

$$\widehat{Y}(\boldsymbol{X}) = \frac{1}{B} \sum_{b=1}^{B} T_b(\boldsymbol{X}), \tag{10.1}$$

where T_b denotes tree b and B is the number of trees. Random forests can represent nonlinearities and interactions without specifying a polynomial model in advance. Variable-importance measures can assist exploratory analysis, although importance does not establish causality and may be biased when predictors differ in scale or correlation structure.

10.2 Support vector regression

Support Vector Regression (SVR) estimates a function while balancing prediction errors against model complexity. With a nonlinear kernel, the fitted predictor can be expressed as

$$\widehat{Y}(\boldsymbol{X}) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(\boldsymbol{X}_i, \boldsymbol{X}) + b, \tag{10.2}$$

where K is the chosen kernel and α_i, α_i^* are estimated coefficients. The kernel and tuning parameters should be selected using training data and cross-validation. SVR can be useful for small or moderate datasets, but its performance depends on appropriate scaling and tuning.

10.3 Artificial neural networks

A feed-forward neural network represents a sequence of transformations of the input variables. A single-hidden-layer regression network may be written as

$$\widehat{Y} = \beta_0 + \sum_{h=1}^{H} \beta_h \, g\left(a_{h0} + \boldsymbol{a}_h^\top \boldsymbol{X}\right), \tag{10.3}$$

where g is an activation function and H is the number of hidden units. Neural networks can approximate complex nonlinear relationships, but large numbers of parameters can lead to overfitting when only a limited number of material formulations have been tested.

10.4 Gradient boosting

Gradient boosting builds an ensemble sequentially by adding learners to improve the current fit:

$$\widehat{f}_M(\boldsymbol{X}) = \widehat{f}_0(\boldsymbol{X}) + \sum_{m=1}^{M} \eta \gamma_m h_m(\boldsymbol{X}), \tag{10.4}$$

where h_m is the learner added at iteration m, γ_m is its fitted contribution, and η is a learning-rate parameter. Boosted trees can provide strong predictive performance for structured tabular data, but tree depth, learning rate, and number of iterations require careful tuning.

10.5 Interpretability

Predictive accuracy alone is insufficient when a model will inform material formulation or industrial process control. Partial dependence plots can summarize the average predicted response as one predictor varies, while accumulated local effects can be useful when predictors are correlated. Feature importance and SHAP-type attributions can help describe model behavior, but their results depend on the fitted model and data distribution. These methods should be interpreted as explanations of predictions rather than proof of physical mechanisms.

11 Model Validation and Performance Comparison

Predictive performance must be evaluated using observations that were not used to fit the model. Let $\mathcal{D}_{\text{train}}$ denote the training data and $\mathcal{D}_{\text{test}}$ the held-out test data. The test set should remain separate from model selection and hyperparameter tuning to avoid optimistic performance estimates.

11.1 Prediction-error measures

For n test observations, the root mean squared error (RMSE) and mean absolute error (MAE) are

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}, \tag{11.1}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \widehat{Y}_i|. \tag{11.2}$$

RMSE penalizes large errors more strongly, whereas MAE summarizes the typical absolute error in the response's measurement units. The test-set coefficient of determination is

$$R^2_{\mathrm{test}} = 1 - \frac{\displaystyle\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\displaystyle\sum_{i=1}^{n}(Y_i - \bar{Y}_{\mathrm{test}})^2}, \tag{11.3}$$

where $\bar{Y}_{\mathrm{test}}$ is the mean of the observed test responses. A negative test-set R^2 is possible and indicates that the predictions perform worse, under this metric, than predicting the test-set mean.

Consider five hypothetical test responses: $Y = (40, 45, 50, 55, 60)$ MPa. The following table 3 uses illustrative predictions to demonstrate the calculation. These are not fitted results from a real dataset.

Model	$\widehat{Y}_1$	$\widehat{Y}_2$	$\widehat{Y}_3$	$\widehat{Y}_4$	$\widehat{Y}_5$	RMSE	MAE	R^2
Linear regression	42	46	49	57	58	1.673	1.600	0.944
Random forest	41	46	51	54	61	1.000	1.000	0.980
SVR	43	44	49	56	59	1.612	1.400	0.948
Gradient boosting	40	47	49	56	60	1.095	0.800	0.976

Table 3: Comparison of predictive performance of statistical and machine-learning models for concrete compressive strength.

For the random forest, the squared errors are $1, 1, 1, 1, 1$, giving

$$\mathrm{RMSE} = \sqrt{\frac{5}{5}} = 1 \text{ MPa}, \qquad \mathrm{MAE} = \frac{5}{5} = 1 \text{ MPa}.$$

The test-response mean is 50 MPa and the total sum of squares is 250 MPa^2. Hence,

$$R^2_{\mathrm{test}} = 1 - \frac{5}{250} = 0.98.$$

In this illustration, the random forest has the smallest RMSE, whereas gradient boosting has the smallest MAE. The ranking therefore depends partly on the performance measure. With only five test observations, however, these rankings would be highly uncertain, as seen in Figure 6.

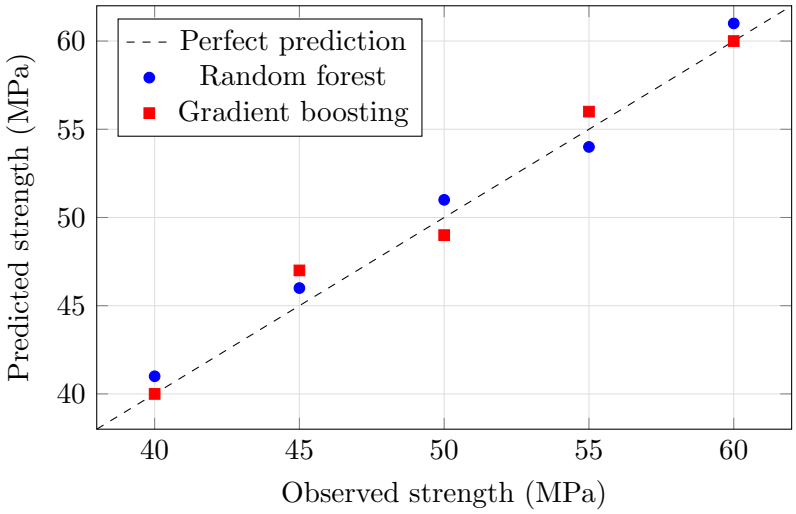


Figure 6: Observed versus predicted values for two illustrative models. Points closer to the dashed line have smaller prediction errors for these observations.

11.2 Cross-validation and fair comparison

In K-fold cross-validation, the training data are partitioned into K folds. Each fold is used once for validation while the remaining folds are used for fitting. The fold-specific errors can be summarized to compare candidate models and tune hyperparameters. Preprocessing, including imputation, scaling, and feature selection, must be learned within each training fold rather than from the full dataset.

For industrial data, random splitting may be inappropriate if observations from the same batch occur in both training and test sets. When the intended application is prediction for a new production batch, group-based validation should hold out entire batches. Where predictions concern future production, a time-ordered split may be more suitable. These choices prevent information leakage and give a more realistic assessment of deployment performance.

11.3 Model selection

The preferred model should combine predictive accuracy with interpretability, robustness, computational cost, and suitability for the intended engineering decision. A model with a slightly larger RMSE may be preferable if its effects are easier to explain, its predictions are more stable across batches, or it supports reliable uncertainty quantification. The table above is a calculation example only; a real model comparison should report metrics calculated from the same validation design.

12 Statistical Optimization of Material Formulation

After model development and validation, the fitted response function can be used to identify promising formulations. Let $\boldsymbol{Z} = (Z_1, \ldots, Z_q)^\top$ denote controllable formulation and processing variables, and let $\widehat{f}(\boldsymbol{Z})$ be the predicted engineering response. A single-response optimization problem is

$$\max_{\boldsymbol{Z} \in \mathcal{R}} \widehat{f}(\boldsymbol{Z}), \tag{12.1}$$

where $\mathcal{R}$ is the feasible region defined by engineering, safety, and operational constraints.

For example, suppose Z is the waste replacement percentage and the feasible range is $0\% \le Z \le 50\%$. Consider the hypothetical prediction function

$$\widehat{Y}(Z) = 42 + 0.45Z - 0.0125Z^2. \tag{12.2}$$

Its stationary point is $Z = 18\%$, which lies within the feasible range and maximizes the fitted quadratic because the coefficient of Z^2 is negative. Nevertheless, this conclusion is conditional on the model being appropriate over the specified range. An independent experiment should be conducted near the proposed optimum before implementation.

12.1 Multiple engineering objectives

A sustainable formulation may need to maximize strength while minimizing water absorption, cost, and carbon footprint. These objectives can conflict, so a single maximum may not exist that simultaneously optimizes every response. One approach is to define desirability functions $d_r(Y_r) \in [0, 1]$ that encode the acceptable range and target for each response. A weighted geometric aggregate is

$$D(\boldsymbol{Z}) = \left\{ \prod_{r=1}^{m} [d_r(\widehat{Y}_r(\boldsymbol{Z}))]^{w_r} \right\}^{1/\sum_{r=1}^{m} w_r}, \qquad w_r \geq 0, \tag{12.3}$$

where w_r expresses the relative importance of response r. The formulation is selected by maximizing $D(\boldsymbol{Z})$ over the feasible region. If any positively weighted desirability is zero, the aggregate is zero. Weights and desirability functions should be justified before examining the final optimum, not chosen solely to favor a preferred formulation.

A second approach is Pareto optimization, which identifies formulations for which no objective can be improved without worsening at least one other objective. The resulting Pareto frontier makes trade-offs explicit and can be particularly useful when engineering performance, cost, and environmental impact are all important.

13 Sustainability and Circular-Economy Assessment

Statistical performance should be connected to material utilization, environmental impact, energy demand, and cost. A formulation with high strength is not necessarily more sustainable if it requires energy-intensive processing, long-distance transport, or hazardous pretreatment. Environmental comparisons should therefore use consistent system boundaries and a clearly specified functional unit, following life-cycle assessment principles.

13.1 Waste utilization

A simple waste-utilization indicator is

$$WU = \frac{m_{\text{waste}}}{m_{\text{material}}} \times 100\%, \tag{13.1}$$

where m_{waste} is the mass of industrial waste incorporated and m_{material} is the total mass of the resulting material, both measured on a consistent basis.

For example, incorporating 30 kg of waste into 100 kg of final material gives

$$WU = \frac{30}{100} \times 100\% = 30\%.$$

This metric measures the waste fraction; it does not by itself demonstrate environmental benefit or product safety.

13.2 Carbon-emission reduction

Let C_{conv} and C_{up} denote emissions per equivalent functional unit for conventional and upcycled materials. The percentage reduction is

$$CR = \frac{C_{\text{conv}} - C_{\text{up}}}{C_{\text{conv}}} \times 100\%. \tag{13.2}$$

If the conventional option emits 500 kg CO_2-equivalent and the upcycled option emits 380 kg CO_2-equivalent per equivalent functional unit, then

$$CR = \frac{500 - 380}{500} \times 100\% = 24\%.$$

A negative value would indicate greater emissions for the upcycled option. Such comparisons should include relevant processing, transportation, pretreatment, and end-of-life stages where these fall within the chosen system boundary.

13.3 Energy consumption and production cost

Energy demand may be modeled as a function of formulation and processing conditions:

$$E = E(Z_1, \ldots, Z_q), \qquad \widehat{E} = \widehat{E}(\boldsymbol{Z}).$$

The total production cost may similarly be decomposed as

$$C_{\text{total}} = C_{\text{raw}} + C_{\text{processing}} + C_{\text{transport}} + C_{\text{energy}}. \tag{13.3}$$

For illustration, if these cost components are respectively 40, 25, 10, and 15 monetary units per functional unit, then the total is 90 monetary units. Such a calculation is meaningful only when the components share the same cost basis and production quantity. Costs associated with waste collection, quality control, emissions treatment, and disposal should also be included when relevant.

13.4 Illustrative sustainability comparison

Consider two hypothetical formulations assessed on the same functional basis. The following values demonstrate how engineering and environmental indicators can be presented together, as in Table 4.

Indicator	Formulation A	Formulation B
Waste utilization (%)	20	30
Strength (MPa)	45	44
Carbon emissions (kg CO_2-eq.)	500	380
Production cost (monetary units)	90	94

Table 4: Hypothetical comparison of engineering, waste-utilization, environmental, and economic indicators for two alternative formulations.

Formulation B incorporates more waste and has lower illustrative carbon emissions, but its predicted strength is slightly lower and its cost is higher. Whether B is preferable depends on minimum strength requirements, cost limits, environmental boundaries, and the relative importance assigned to each outcome.

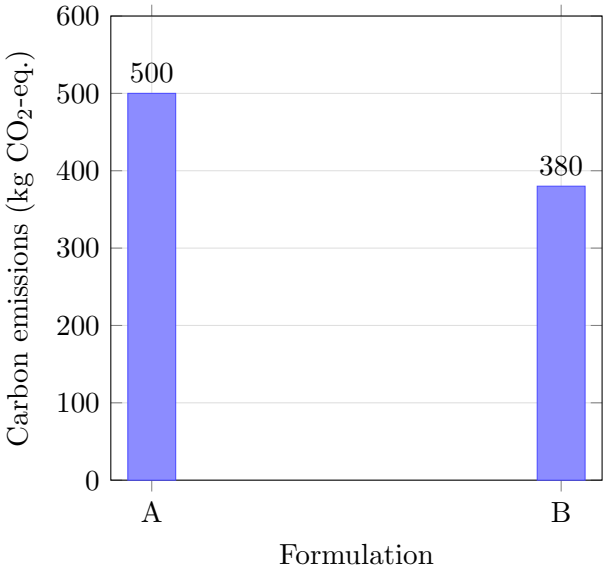


Figure 7: Hypothetical carbon-emission comparison on an equivalent functional basis.

13.5 Integrated sustainability score

For preliminary screening, normalized indicators may be combined into a composite score:

$$S = w_1 S_{\text{engineering}} + w_2 S_{\text{environmental}} + w_3 S_{\text{economic}} + w_4 S_{\text{waste}}, \qquad \sum_{r=1}^{4} w_r = 1, \tag{13.4}$$

where each component is scaled consistently so that a higher value represents better performance. For example, strength may be scored higher when it increases, whereas cost and emissions should be transformed so that lower values correspond to higher desirability.

The weights should reflect transparent engineering priorities or a justified decision-making procedure. Because normalization and weighting can materially affect rankings, sensitivity analysis should examine whether the preferred formulation changes under plausible alternative weights. A composite score is a decision aid, not a substitute for reporting the individual indicators and their uncertainties.

14 Case Study: Statistical Analysis of a Real Industrial-Waste Dataset

14.1 Dataset and research objectives

This case study uses the Concrete Compressive Strength dataset originally developed by Yeh (1998)[11] and distributed through the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength). The dataset contains 1,030 observations, with eight quantitative explanatory variables and one quantitative response variable. The explanatory variables are cement, blast-furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and curing age. The constituent quantities are measured in kg/m^3, curing age is measured in days, and compressive strength is measured in MPa. The accompanying dataset documentation reports no missing attribute values.

The dataset is relevant to industrial-waste valorization because blast-furnace slag and fly ash are industrial by-products that can be incorporated into cementitious materials. Their use in concrete formulations provides a practical setting for examining the relationship between by-product incorporation, mixture composition, and mechanical performance. Nevertheless, the dataset does not

contain production-batch identifiers, detailed waste chemistry, curing temperature, production costs, energy consumption, or life-cycle environmental indicators. Therefore, the analysis investigates the statistical relationships between concrete composition and compressive strength rather than claiming a comprehensive assessment of environmental or economic sustainability.

The principal objectives are to characterize the distribution of compressive strength, investigate its association with mixture constituents, compare alternative regression specifications, examine the observed difference between mixtures with and without fly ash, and explore nonlinear response patterns involving fly ash content and curing age. The numerical results reported below are obtained from the analysis files contained in the accompanying case-study archive.

14.2 Descriptive and graphical analysis

Let Y_i denote the measured compressive strength of the ith concrete mixture. Across the 1,030 observations, the mean compressive strength is 35.818 MPa, with a standard deviation of 16.706 MPa. The observed strengths range from 2.332 to 82.599 MPa. The first and third quartiles are 23.707 and 46.136 MPa, respectively, while the median is 34.443 MPa. Thus, the interquartile range is approximately 22.429 MPa, indicating substantial variation in mechanical performance across the recorded mixtures.

The mean is slightly greater than the median, suggesting some right-sided asymmetry in the empirical distribution. The histogram and boxplot provide complementary views of this variability. In particular, the wide observed range indicates that concrete mixtures with different compositions and curing ages can exhibit markedly different compressive strengths. Consequently, a single average strength would not adequately characterize the dataset.

Variable	Mean	SD	Min.	Q_1	Median	Q_3	Max.
Compressive strength (MPa)	35.818	16.706	2.332	23.707	34.443	46.136	82.599

Table 5: Descriptive statistics for compressive strength in the UCI Concrete Compressive Strength dataset.

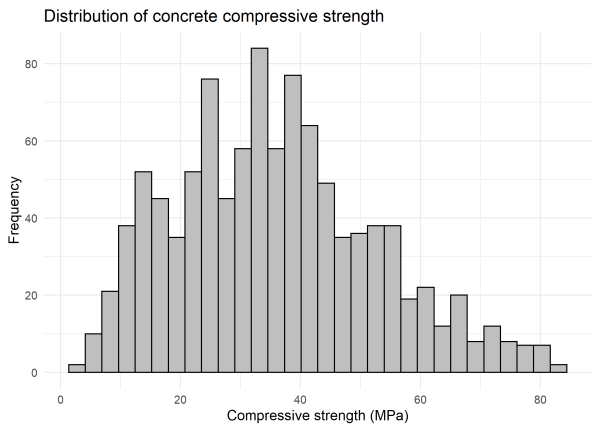


Figure 8: Empirical distribution of concrete compressive strength in the 1,030 observations.

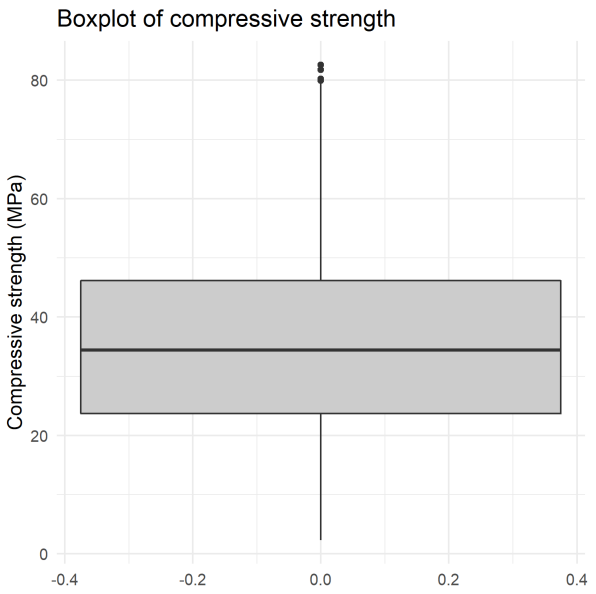


Figure 9: Boxplot of concrete compressive strength, showing the median, interquartile range, and observations beyond the central spread.

Figure 9 presents the histogram and boxplot of compressive strength. These graphics help identify the spread of observations and the presence of relatively low or high strength values. Such observations should not automatically be treated as erroneous, since they may represent genuine differences in mixture composition or curing age.

14.3 Correlation analysis

Pearson correlation coefficients were calculated to summarize the linear associations between the explanatory variables and compressive strength. The largest positive correlation with strength is observed for cement content, with $r = 0.498$. Superplasticizer content has a positive correlation of $r = 0.366$, while curing age has a positive correlation of $r = 0.329$. Blast-furnace slag has a weaker positive correlation with strength ($r = 0.135$).

In contrast, water content has a negative Pearson correlation of $r = -0.290$ with compressive strength. Fine aggregate and coarse aggregate also exhibit negative correlations, equal to -0.167 and -0.165, respectively. Fly ash has a weak negative linear correlation with strength ($r = -0.106$). These coefficients describe the observed mixtures and should not be interpreted as isolated causal effects. For example, the association between fly ash and strength may partly reflect differences in cement content, water content, curing age, and other formulation characteristics.

The correlation matrix also reveals relationships among the explanatory variables. In particular, water and superplasticizer have a Pearson correlation of -0.657, while water and fine aggregate have a correlation of -0.451. Cement and fly ash are negatively correlated ($r = -0.397$), and fly ash and superplasticizer are positively correlated ($r = 0.377$). These relationships indicate that mixture constituents do not vary independently, which is important when interpreting individual regression coefficients.

Variable	Pearson correlation with strength
Cement	0.498
Superplasticizer	0.366
Curing age	0.329
Blast-furnace slag	0.135
Fly ash	−0.106
Coarse aggregate	−0.165
Fine aggregate	−0.167
Water	−0.290

Table 6: Pearson correlations between selected explanatory variables and compressive strength

The Spearman correlation between curing age and strength is 0.596, which is considerably greater than the corresponding Pearson correlation of 0.329. This difference is consistent with the possibility that the association between curing age and strength is not adequately summarized by a simple linear relationship. Spearman's coefficient measures rank-based monotonic association and does not require the relationship to be linear. The Pearson and Spearman correlations between cement and strength are similar, at 0.498 and 0.478, respectively. For fly ash, the corresponding values are -0.106 and -0.078, indicating a weak marginal association with strength under both measures.

The correlation heatmap in Figure 10 provides a compact visualization of the relationships among all nine variables. The results support the use of multivariable models because the observed strength is associated with several mixture characteristics, while the explanatory variables themselves also exhibit appreciable correlations.

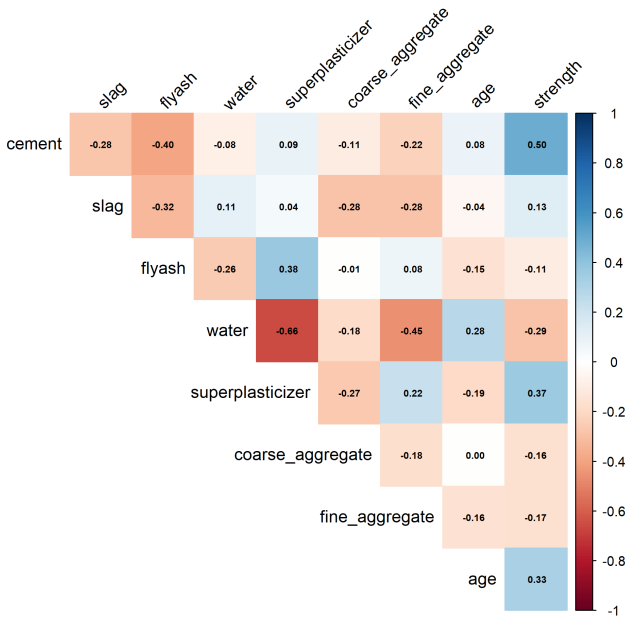


Figure 10: Pearson correlation heatmap for mixture constituents, curing age, and compressive strength.

14.4 Regression and response-surface analysis

The baseline multiple linear regression model is specified as

$$\begin{aligned} Y_i = \beta_0 &+ \beta_1 \mathrm{Cement}_i + \beta_2 \mathrm{Slag}_i + \beta_3 \mathrm{FlyAsh}_i + \beta_4 \mathrm{Water}_i \\ &+ \beta_5 \mathrm{Superplasticizer}_i + \beta_6 \mathrm{CoarseAggregate}_i + \beta_7 \mathrm{FineAggregate}_i + \beta_8 \mathrm{Age}_i + \varepsilon_i, \end{aligned} \tag{14.1}$$

where Y_i is compressive strength and ε_i denotes the unexplained component.

To investigate potential nonlinear relationships, a quadratic regression model adds squared terms for fly ash content and curing age. A response-surface specification further includes an interaction between fly ash and curing age. The corresponding additional terms can be written as

$$\gamma_1 \mathrm{FlyAsh}_i^2 + \gamma_2 \mathrm{Age}_i^2 + \gamma_3 \mathrm{FlyAsh}_i \mathrm{Age}_i.$$

These terms allow the fitted relationship to exhibit curvature and permit the association between fly ash and strength to vary with curing age.

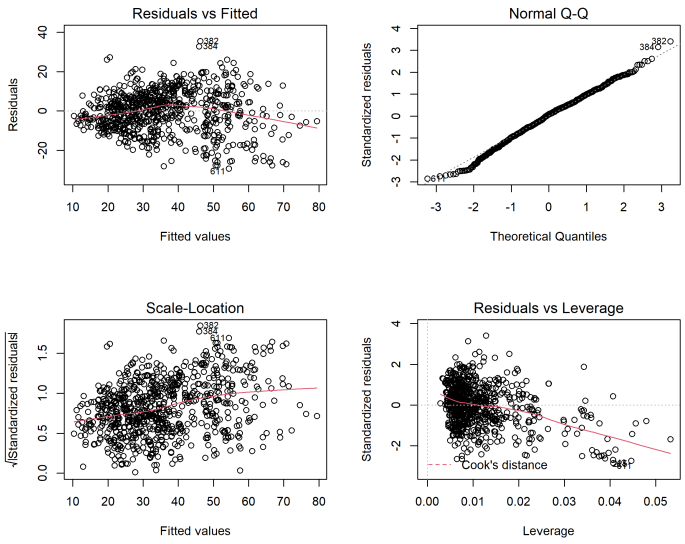


Figure 11: Regression diagnostic plots for the linear model.

Model	Adjusted R^2	AIC	BIC
Multiple linear regression	0.616	6222.697	6269.838
Quadratic regression	0.758	5844.700	5901.270
Response-surface model	0.759	5840.476	5901.761

Table 7: Training-data model-selection criteria for the three regression specifications.

The training-data model-selection results are summarized in Table 7. The adjusted coefficient of determination for the linear regression is 0.616, whereas the quadratic and response-surface models achieve adjusted R^2 values of 0.758 and 0.759, respectively. Thus, the two nonlinear specifications account for a greater proportion of the variation in the training response after adjustment for the number of fitted parameters. The Akaike information criterion (AIC) decreases from 6222.697 for the linear regression to 5844.700 for the quadratic model and 5840.476 for the response-surface model. The Bayesian information criterion (BIC) is 6269.838, 5901.270, and 5901.761, respectively. Accordingly, the response-surface model has the smallest AIC, whereas the quadratic model has a marginally smaller BIC. The difference in BIC between the two nonlinear models is approximately 0.490, so this criterion does not provide strong separation between them. Taken together, the results suggest that allowing curvature improves the in-sample representation of compressive strength, although these criteria alone do not establish superior predictive performance on unseen data.

The diagnostic figure in Figure 11 is used to assess the adequacy of the linear model. Residual-versus-fitted plots help identify systematic patterns or nonconstant residual variance, while normal Q–Q plots assess departures from normality. Such diagnostics are important because a high adjusted R^2 does not, by itself, guarantee that the regression assumptions are appropriate.

For a complete predictive comparison, the candidate models should be evaluated on the same held-out observations using root mean squared error (RMSE), mean absolute error (MAE), and test-set R^2. These quantities should be reported separately from adjusted R^2, AIC, and BIC, which are training-data criteria in the supplied analysis. The latter are useful for comparing fitted regression specifications but cannot replace an independent evaluation of predictive accuracy.

14.5 Exploratory analysis of fly ash incorporation

To examine the observed differences between mixtures with and without fly ash, the data were divided into two groups: mixtures with zero fly ash and mixtures with positive fly ash content. There are 566 observations in the zero-fly-ash group and 464 observations in the positive-fly-ash group.

The mean compressive strength is 36.772 MPa for mixtures without fly ash and 34.654 MPa for mixtures containing positive amounts of fly ash. The corresponding standard deviations are 18.445 and 14.234 MPa, while the medians are 35.301 and 33.743 MPa, respectively. Thus, the zero-fly-ash group has a higher observed mean by approximately 2.117 MPa. The positive-fly-ash group also exhibits a smaller sample standard deviation, indicating less observed dispersion in strength in this particular grouping.

A one-way analysis of variance was used to test $\mathbf{H_0} : \mu_{\text{zero}} = \mu_{\text{positive}}$ against the alternative that the two group means differ. The reported ANOVA gives $F(1, 1028) = 4.108$ with $p = 0.0429$. At the conventional 5% significance level, the null hypothesis is rejected. The observed difference between the two group means is therefore statistically significant under this one-way ANOVA analysis.

Group	n	Mean (MPa)	SD (MPa)	Median (MPa)
Zero fly ash	566	36.772	18.445	35.301
Positive fly ash	464	34.654	14.234	33.743

Table 8: Descriptive comparison of compressive strength by fly ash presence.

Source	df	Sum of squares	F	p-value
Fly ash group	1	1143.1	4.108	0.0429
Residuals	1028	286030.0		

Table 9: One-way ANOVA comparing mean compressive strength between mixtures with zero and positive fly ash content.

This finding, provided by tables 8 and 9 must be interpreted cautiously. The grouping does not control for cement, water, slag, superplasticizer, aggregate proportions, or curing age. The result therefore does not imply that adding fly ash necessarily reduces compressive strength. It indicates only that the average strengths of the two observed groups differ in this dataset. A more informative assessment of the role of fly ash would adjust for the other mixture constituents and examine possible interactions and nonlinear effects. Figure 12 displays the group comparison. The substantial overlap between the two distributions is consistent with the modest difference in their means relative to the observed within-group variation. Hence, fly ash presence alone does not provide a complete explanation of compressive strength.

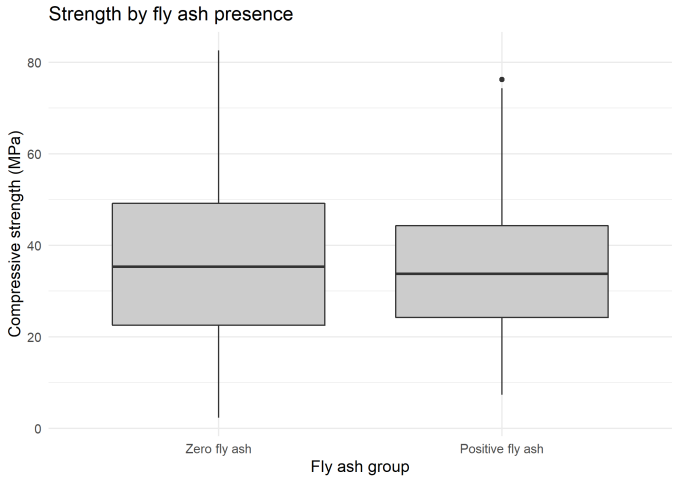


Figure 12: Compressive strength for mixtures with zero and positive fly ash content.

14.6 Exploratory formulation screening

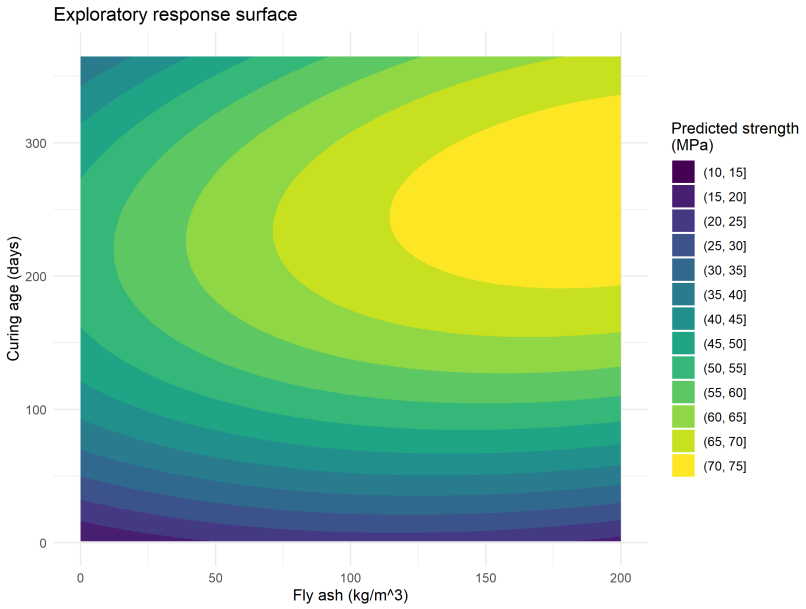


Figure 13: Exploratory response surface for predicted concrete compressive strength as a function of fly ash content and curing age, conditional on the reference values of the remaining constituents.

The response-surface analysis was also used to examine predicted strength over a grid of fly ash contents and curing ages, while holding the other mixture constituents at specified reference values. The resulting contour plot, shown by Figure 13, illustrates the fitted model's conditional predictions and provides a visual summary of the nonlinear response pattern. The supplied grid-search output identifies a candidate with fly ash content of 200 kg/m^3 and curing age of approximately 263.633 days. At this point, the other constituents are fixed at cement content 266 kg/m^3, blast-furnace slag content 24 kg/m^3, water content 185 kg/m^3, superplasticizer content 6.625 kg/m^3, coarse aggregate content 968 kg/m^3, and fine aggregate content 780.09 kg/m^3. The fitted model predicts a compressive strength of approximately 74.092 MPa for this combination.

This candidate is an exploratory model-based result, not a validated optimum. In particular, the prediction is conditional on the chosen reference composition and the fitted response surface. The available grid-search output does not establish that this combination satisfies all practical mixture-design constraints, nor does it demonstrate that an experimental specimen prepared at this composition would achieve the predicted strength. Confirmation would require a feasible mixture-design assessment and independent experimental validation. The result should therefore be interpreted as a candidate for further investigation rather than a final formulation recommendation. The results are presented in Table 10.

Quantity	Value
Cement (kg/m^3)	266.000
Blast-furnace slag (kg/m^3)	24.000
Fly ash (kg/m^3)	200.000
Water (kg/m^3)	185.000
Superplasticizer (kg/m^3)	6.625
Coarse aggregate (kg/m^3)	968.000
Fine aggregate (kg/m^3)	780.090
Curing age (days)	263.633
Predicted strength (MPa)	74.092

Table 10: Exploratory response-surface candidate recorded in the analysis archive, including the selected mixture proportions, curing age, and predicted compressive strength.

Although the response-surface analysis can help screen potential formulations for mechanical performance, it cannot directly identify the most environmentally sustainable formulation. The dataset does not contain carbon emissions, energy consumption, cost, or life-cycle assessment indicators. Accordingly, the candidate cannot be described as the environmental or economic optimum. Such an assessment would require additional measurements and a multi-objective optimization framework that incorporates both mechanical performance and sustainability criteria.

14.7 Reproducibility and limitations

The analysis archive contains the analysis-ready dataset, descriptive statistics, Pearson and Spearman correlation matrices, ANOVA output, regression-model selection criteria, the exploratory response-surface candidate, and the figures used in this case study. These materials provide a basis for reproducing the reported numerical summaries and graphical analyses. The original dataset documentation identifies 1,030 observations, nine variables, and no missing attribute values.

Several limitations should be considered when interpreting the findings. First, the data represent concrete mixture records and do not provide production-batch identifiers; therefore, the analysis cannot quantify between-batch variability or establish performance under a batch-based validation design. Second, correlations and the unadjusted fly ash group comparison are observational and do not establish causal effects. Third, the regression criteria available in the archive describe training-data fit, while a complete numerical table of held-out predictive performance for all candidate models is not included. Consequently, the available results do not support declaring one of the seven candidate prediction methods the definitive best model. Finally, environmental and economic superiority cannot be established without direct sustainability measurements.

Overall, this case study demonstrates how descriptive statistics, correlation analysis, analysis of variance, nonlinear regression, and response-surface screening can be combined to investigate the mechanical performance of concrete containing industrial by-products. The numerical results show substantial variation in compressive strength, a stronger marginal association with cement and curing age than with fly ash alone, and improved training-data fit when nonlinear terms are introduced. These findings illustrate the usefulness of data-driven statistical analysis for preliminary formulation screening while highlighting the need for independent validation and additional environmental and economic data before drawing broader conclusions about sustainable material design.

15 Challenges, Future Directions, and Conclusions

The sustainable utilization of industrial waste in engineering materials presents considerable opportunities for resource conservation, waste minimization, and circular-economy development. However, translating these opportunities into reliable industrial applications requires a systematic understanding of the relationships among waste characteristics, material composition, processing conditions, and final engineering performance. Statistical modeling provides a valuable framework for investigating these relationships, quantifying uncertainty, identifying influential variables, and supporting evidence-based material design. Nevertheless, several methodological and practical challenges must be addressed before data-driven approaches can be applied reliably across different waste streams, material formulations, and industrial environments.

One of the principal challenges is the inherent heterogeneity of industrial waste. Waste materials originating from different industries, production processes, geographical regions, and storage conditions may exhibit substantial differences in chemical composition, particle-size distribution, moisture content, and physical properties. Even materials obtained from the same industrial source may vary considerably between production batches. Such variability can influence both the reproducibility of experimental results and the predictive performance of statistical models. Consequently, future investigations should consider hierarchical and mixed-effects models that explicitly account for batch-to-batch and source-specific variations. Where data are collected from multiple industries or

geographical locations, multilevel modeling may help distinguish general relationships from source-dependent effects. The systematic characterization of waste materials, together with appropriate uncertainty quantification, would further improve the reliability and transferability of the resulting models.

The availability of experimental data represents another important limitation. Materials research frequently involves expensive laboratory procedures, specialized equipment, and time-consuming mechanical or chemical tests, which may restrict the number of observations available for statistical analysis. Although complex machine-learning algorithms can represent highly nonlinear relationships, their use with limited datasets may lead to overfitting and unreliable predictions outside the observed experimental range. Therefore, the complexity of a statistical model should be selected according to the amount, quality, and structure of the available data. Regression models, regularization techniques, and carefully specified response-surface models may provide effective alternatives when sample sizes are relatively small. Cross-validation should be employed to assess predictive performance, while independent datasets should be used whenever possible to evaluate generalizability. Future studies may also investigate optimal experimental-design strategies to select informative formulations and processing conditions, thereby reducing experimental costs without compromising the quality of statistical inference.

Missing observations and measurement uncertainty require equally careful consideration. Certain chemical properties, microstructural characteristics, or mechanical responses may remain unobserved because the corresponding tests are costly, technically difficult, or unavailable for every experimental specimen. Simply deleting incomplete observations may reduce the effective sample size and introduce bias when the probability of missingness is related to observed or unobserved variables. Future research should therefore examine statistically appropriate missing-data procedures, including multiple imputation, likelihood-based methods, and sensitivity analyses under alternative missingness assumptions. The suitability of any procedure should be evaluated in relation to the data-collection mechanism and the scientific objectives of the investigation. In addition, engineering measurements are subject to experimental error, specimen-to-specimen variation, and laboratory-specific conditions. For example, nominally identical specimens may exhibit different compressive strengths because of variations in preparation, curing, composition, or testing. Accordingly, reporting only a fitted mean response or a point prediction may provide an incomplete description of material performance. Prediction intervals, confidence intervals, and, where appropriate, estimates of measurement and between-specimen variability should accompany statistical predictions. Such information is particularly important when predicted material properties are used to assess engineering suitability or compliance with performance requirements.

A further challenge concerns the transferability of laboratory findings to pilot-scale and industrial-scale production. A material formulation that performs satisfactorily under controlled laboratory conditions may behave differently when produced using industrial equipment or under less uniform environmental conditions. Differences in mixing efficiency, temperature regulation, curing procedures, production scale, and raw-material composition may alter the resulting material properties. Therefore, a formulation identified as promising by a laboratory-based statistical model should not automatically be regarded as an industrially validated solution. Future studies should progressively evaluate model performance using laboratory, pilot-scale, and industrial-scale observations. Where datasets from these stages are available, domain-specific effects and changes in the relationships between processing variables and material responses can be investigated explicitly. Periodic model recalibration may also be necessary when production conditions or the characteristics of the incoming waste stream change over time. This approach would help distinguish formulations that perform well under a particular experimental setting from those that remain reliable under practical manufacturing conditions.

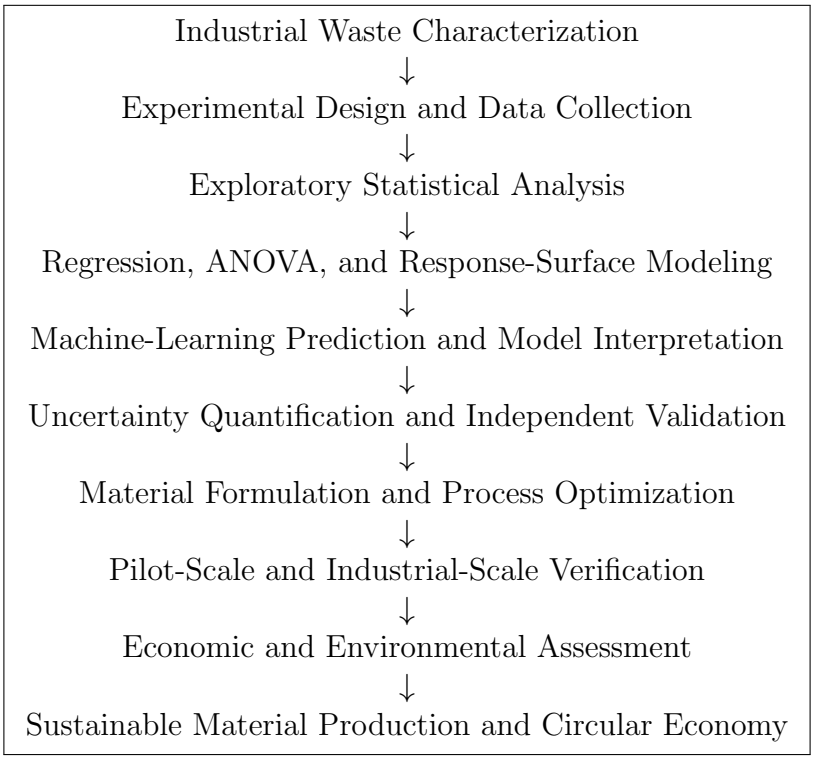
The integration of conventional statistical methods with artificial intelligence offers another important direction for future research. Exploratory data analysis, correlation analysis, regression, analysis of variance (ANOVA), and response-surface methodology provide interpretable tools for understanding experimental relationships and assessing differences among material formulations. How-

ever, complex interactions and nonlinear dependencies may not always be adequately represented by simple parametric models. Machine-learning approaches, including random forests, support vector regression, artificial neural networks, and gradient boosting, can complement conventional methods when sufficient high-quality data are available. Rather than treating statistical modeling and machine learning as competing alternatives, future investigations should evaluate their complementary strengths through systematic model comparison and independent validation. Explainable machine-learning methods are particularly relevant because engineering practitioners need to understand not only the predicted performance of a formulation but also the variables and interactions contributing to that prediction. Feature-importance measures and partial-dependence analyses may assist interpretation, although their results should be examined carefully when predictors are strongly correlated. The combined use of interpretable statistical models and flexible predictive algorithms could provide a more comprehensive basis for material characterization, performance prediction, and formulation development.

Industrial digitalization also creates opportunities to extend statistical modeling beyond the analysis of static experimental datasets. Modern production environments increasingly generate information through sensors, automated monitoring systems, and digital manufacturing platforms. Measurements of temperature, moisture, pressure, material composition, mixing duration, and energy consumption could be integrated with laboratory measurements of mechanical and physical performance. These data may support the development of dynamic predictive models that track changes in production conditions and identify deviations from desired material properties. Future research could investigate online monitoring, sequential model updating, and early-warning procedures for detecting changes in raw-material characteristics or manufacturing performance. However, successful implementation would require reliable sensor measurements, appropriate data preprocessing, robust model validation, and consideration of delays between processing conditions and the availability of final quality measurements. When these requirements are satisfied, data-driven monitoring may support timely adjustments to processing conditions and reduce material waste during production.

Beyond predictive accuracy, future studies should place greater emphasis on the optimization and sustainability assessment of waste-derived materials. A formulation that maximizes a single mechanical property may not necessarily provide the most desirable overall solution. Material selection and formulation optimization may involve several competing objectives, including engineering strength, durability, waste utilization, production cost, energy consumption, and environmental impact. Statistical models can be incorporated into multi-objective optimization procedures to identify candidate formulations that provide an appropriate balance among these objectives. The resulting solutions should then be examined through confirmatory experiments and, where feasible, life-cycle and economic assessments. In particular, the use of industrial waste should not be considered environmentally beneficial solely because it reduces the quantity of waste requiring disposal. The environmental implications of collection, transportation, pretreatment, processing, and long-term material performance must also be considered. Integrating statistical prediction with experimental confirmation and sustainability assessment would therefore provide a stronger basis for determining whether a proposed formulation is technically feasible, economically competitive, and environmentally responsible.

The overall framework for the data-driven development of sustainable materials from industrial waste can be summarized as follows:



In conclusion, the successful upcycling of industrial waste into sustainable engineering materials requires more than identifying a waste material that can be incorporated into a formulation. It requires a systematic investigation of material variability, experimental uncertainty, processing conditions, engineering performance, and environmental consequences. Exploratory statistical analysis can reveal the distributions and variability of experimental variables, while correlation analysis can identify potentially important associations. Regression models provide interpretable descriptions of material performance, and polynomial and response-surface models can represent nonlinear effects and interactions. ANOVA facilitates comparisons among materials and formulations, whereas hierarchical and mixed-effects models can account for structured variability across batches, sources, and production conditions. Machine-learning methods can extend these analyses when more complex relationships are present, provided that model complexity is appropriate for the available data and predictive performance is assessed rigorously.

Importantly, prediction should be regarded as an intermediate step rather than the ultimate objective of statistical modeling. The practical value of a fitted model lies in its ability to support experimentally verifiable formulation decisions, identify promising processing conditions, quantify uncertainty, and guide the efficient use of industrial waste. Such decisions should be supported by independent validation and evaluated against engineering, economic, and environmental criteria. The integration of statistical inference, predictive modeling, optimization, and sustainability assessment can therefore transform experimental observations into actionable knowledge for materials development. Although challenges associated with heterogeneous waste streams, limited datasets, missing observations, measurement uncertainty, and industrial-scale implementation remain substantial, addressing these issues systematically can improve the reliability and practical relevance of data-driven approaches. Ultimately, statistical modeling provides an important link between experimental materials science and sustainable engineering by supporting the development of waste-derived materials that are technically reliable, economically viable, and environmentally responsible.

Declarations

- **Acknowledgement:** The author gratefully acknowledges the support and facilities provided by the Department of Mathematics, Brainware University, Kolkata, India.

- **Funding:** This research received no external funding.
- **Data & Code Availability:** The data and code associated with this study are available at the GitHub repository: https://github.com/sthdas999/Industrial-Waste-Valorization-and-Sustainable-Material-Design.
- **Conflict of Interest:** The author declares no conflict of interest.
- **Ethical Approval:** Ethical approval was not applicable to this study.

References

[1] United Nations Environment Programme, *Global Waste Management Outlook 2024: Beyond an Age of Waste – Turning Rubbish into a Resource*, UNEP, Nairobi, 2024. https://www.unep.org/en/publications/report/global-waste-management-outlook-2024

[2] International Organization for Standardization, *Environmental Management – Life Cycle Assessment – Principles and Framework*, ISO 14040:2006, 2nd ed., Geneva, 2006. https://www.iso.org/standard/37456.html

[3] M. Ahmaruzzaman, A review on the utilization of fly ash, *Progress in Energy and Combustion Science*, **36** (2010), no. 3, 327–363. https://doi.org/10.1016/j.pecs.2009.11.003

[4] C. K. Purchase, D. M. Al Zulayq, B. T. O'Brien, M. J. Kowalewski, A. Berenjian, A. H. Tarighaleslami, and M. Seifan, Circular economy of construction and demolition waste: A literature review on lessons, challenges, and benefits, *Materials*, **15** (2022), no. 1, Article 76. https://doi.org/10.3390/ma15010076

[5] I. Papamichael, I. Voukkali, P. Loizia, and A. A. Zorpas, Construction and demolition waste framework of circular economy: A mini review, *Waste Management & Research*, **41** (2023), no. 12, 1728–1740. https://doi.org/10.1177/0734242X231190804

[6] G. E. P. Box and K. B. Wilson, On the experimental attainment of optimum conditions, *Journal of the Royal Statistical Society. Series B (Methodological)*, **13** (1951), no. 1, 1–45. https://doi.org/10.1111/j.2517-6161.1951.tb00067.x

[7] D. C. Montgomery, *Design and Analysis of Experiments*, 10th ed., John Wiley & Sons, Hoboken, NJ, 2019.

[8] S. I. Hassan, S. A. Syed, S. W. Ali, H. Zahid, S. Tariq, M. M. Su Ud, and M. M. Alam, Systematic literature review on the application of machine learning for the prediction of properties of different types of concrete, *PeerJ Computer Science*, **10** (2024), Article e1853. https://doi.org/10.7717/peerj-cs.1853

[9] NIST/SEMATECH, *e-Handbook of Statistical Methods*, National Institute of Standards and Technology. https://www.itl.nist.gov/div898/handbook/

[10] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, 2nd ed., Springer, New York, 2021. https://doi.org/10.1007/978-1-0716-1418-1

[11] I.-C. Yeh, Modeling of strength of high-performance concrete using artificial neural networks, *Cement and Concrete Research*, **28**(12) (1998), 1797–1808. https://doi.org/10.1016/S0008-8846(98)00165-3.