Applied Computer Science 2025, vol. 21, no. 4, pp. 136–158
https://doi.org/10.35784/8290

Submitted: 2025-08-27 | Revised: 2025-10-13 | Accepted: 2025-10-19



*Keywords: labour resources, production output, labour productivity, forecasting, machine learning*

*Oxana DENISSOVA* [1], *Aman ISMUKHAMEDOV* [2*], *Zhadyra KONURBAYEVA* [1*], *Saule RAKHMETULLINA* [1], *Yelena SAMUSSENKO* [1], *Monika KULISZ* [3]

[1] D. Serikbayev East Kazakhstan Technical University, Republic of Kazakhstan, odenisova@edu.ektu.kz, zhkonurbayeva@edu.ektu.kz, srakhmetullina@ektu.kz, esamusenko@edu.ektu.kz,

[2] Kazakh-American Free University, Republic of Kazakhstan, aman.ism92@gmail.com

[3] Lublin University of Technology, Poland, m.kulisz@pollub.pl

[*] Corresponding author: aman.ism92@gmail.com

# Application of machine learning algorithms for forecasting labour demand in the metallurgical industry of the east Kazakhstan region

**Abstract**

*The study focuses on the development and evaluation of predictive models for forecasting labour demand in the metallurgical industry of the East Kazakhstan Region, with particular emphasis on the impact of production volume and labour productivity. The methodological framework combines classical econometric approaches with modern machine learning techniques, which makes it possible to capture nonlinear dependencies and more accurately assess labour market dynamics. The research is based on regional statistical data for the period 2015–2023. Several modeling approaches were tested, including linear regression, a parametric specification, and a hybrid machine learning model that integrates decision trees with local linear regression. Model performance was validated using the Mean Absolute Error (MAE), followed by forecasting labour demand for 2024–2028. Results demonstrate that the hybrid model outperforms the alternatives by achieving the lowest prediction error and producing the most plausible projection of moderate employment growth. The parametric model, although less precise, offers a high level of interpretability and is well suited for strategic analysis, while the linear regression model has limited effectiveness under nonlinear conditions. The practical value of the research lies in the possibility of embedding the developed models into decision support systems for government bodies and industrial enterprises, enabling early assessment of the impact of technological changes and production dynamics on employment. The outcomes may contribute to shaping balanced human resource policies, aligning educational programs with labour market needs, and conducting scenario analyses. Furthermore, the findings establish a foundation for extending the methodology to other industries and incorporating additional variables related to digitalization and innovation activity.*

## 1. INTRODUCTION

Workforce assessment and forecasting are essential components of regional economic planning. They allow policymakers to assess whether existing and emerging economic structures can adequately supply strategically important industries with skilled labour. This is particularly relevant for the metallurgical sector - the "locomotive industry" of the East Kazakhstan region - which exerts a strong multiplier effect on the national economy.

From an economic point of view, the dynamics of employment in any industry is primarily determined by two fundamental factors: production output and labour productivity. Ceteris paribus, increases in output expand the demand for labour, while productivity gains reduce it by allowing more output to be produced with fewer workers. Employment fluctuations thus reflect the interplay between the scale effect of output expansion and the efficiency effect of productivity growth.

Recent research confirms the dual nature of this relationship. Numerous studies show that innovation affects employment indirectly through changes in output and productivity per worker (Dosi et al., 2019; Woltjer et al., 2019). Panel analyses for OECD countries further show that labour productivity affects employment levels,

although its effect cannot be fully interpreted without controlling for output (Cruz, 2023). The productivity-mediated effects of digitization and the adoption of "smart manufacturing" technologies are discussed by Zhu et al. (2024) and Ballestar et al. (2021), who show that technological transitions reshape labour markets by changing both employment levels and occupational structures. Although many studies do not explicitly model employment as a dependent variable, their methodological frameworks can be adapted for production-based employment forecasting (Mahamid, 2020; Potapov, 2020).

In recent years, the use of machine learning (ML) in macroeconomic and sectoral forecasting has grown rapidly. Neural network architectures (RNN, LSTM, GRU, Transformer), ensemble models, and regularized regressions have achieved high forecasting accuracy for a variety of economic indicators (Ebrahimi et al., 2021; Magazzino et al., 2025; Mutascu & Hegerty, 2023; Uppal et al., 2024; Y. Zhang et al., 2024). Comparative studies suggest that ML algorithms often outperform traditional econometric approaches (EFSD, 2023). For example, Falkenberg and Spinler (2022) demonstrated the effectiveness of gradient boosting in predicting labour productivity based on operational and behavioral data; Golabchi and Hammad (2024) explored ML applications in construction labour estimation; and Alzeraif et al. (2023) used ML to predict productivity in the energy industry.

The integration of ML into decision support systems (DSS) enables the modeling of complex, nonlinear, and dynamic relationships, which is particularly valuable under conditions of uncertainty. Such systems have been successfully applied to optimize human resource management (Bali et al., 2023; Orlova, 2023; Bril et al., 2020), forecast productivity in industry and construction (Hatami et al., 2024; Elshaboury, 2022; Güvel, 2025), and assess environmental and climate impacts on employment (J. Zhang et al., 2024; Li et al., 2020). Authorities are increasingly relying on parametric and hybrid models to forecast socioeconomic indicators such as GDP and employment using labour, investment, and technology variables (Ramezanian & Hajipour, 2020; Popescu et al., 2021).

Classical parametric models-including linear regression, Cobb-Douglas production functions, and time series models-provide transparency and interpretability when relationships remain stable (Jacobsen et al., 2024). However, in complex and evolving environments, hybrid approaches that integrate ML algorithms, fuzzy logic, system dynamics, and optimization techniques (e.g., particle swarm optimization) often provide superior performance (Golabchi & Hammad, 2024; Hatami et al., 2024).

Despite the extensive literature, few studies have jointly used production output and labour productivity in ML-based frameworks for direct employment forecasting. These factors are typically treated as auxiliary or indirect variables, leaving a gap for the development of integrated models that explicitly capture their combined influence on employment dynamics.

The objective of this study is to develop an economically sound and empirically validated model of labour demand in the metallurgical industry of the East Kazakhstan region by combining econometric and ML methods. To achieve this goal, the study

- Examines the stability of the relationships among employment, output, and productivity;
- Identifies the optimal model form that balances statistical validity and interpretability;
- evaluates the predictive accuracy of classical and ML methods; and
- produces a labour demand forecast for the period 2025-2028.

Thus, this work aims to construct a model that reproduces historical employment dynamics while serving as a practical tool for labour policy planning. While previous studies have examined output and productivity as indirect determinants of employment, the simultaneous integration of both variables within ML architectures for direct labour demand forecasting remains underexplored-especially in regional industrial contexts. The proposed hybrid econometric ML framework addresses this gap by explicitly incorporating output (V) and labour productivity (P) as explanatory variables, thereby improving both the predictive accuracy and economic interpretability of employment forecasts.

## 2. METHODOLOGY

### 2.1. Employment forecast: Methods and data

In order to build a model for forecasting labour demand in the metallurgical industry in the East Kazakhstan region, data on the number of employees and key production indicators for the period 2015-2023 were used. The following variables were chosen as input variables

1. Production volume or production output ($V_{base}$);
2. Labour productivity ($P_{base}$):

$$P = \frac{V}{workers} \quad (1)$$

where: $workers$ – number of employees.

These variables reflect two key mechanisms of employment formation:

- Scale effect of production (an increase in output requires more labour resources);
- Productivity effect (an increase in output per worker reduces the demand for labour resources).

While output and productivity are the core specification due to their theoretical relevance and consistent data availability, several additional factors influence labour dynamics. Innovation activity (e.g. R&D intensity, technology adoption) can have both labour-displacing and labour-enhancing effects. Indicators of digitization, such as ICT capital investment or digital skills of the labour force, capture the transition to Industry 4.0, which is reshaping occupational structures. Finally, macroeconomic shocks, such as commodity price volatility and exchange rate fluctuations, affect production decisions and thus employment through demand-side channels.

To forecast labour demand, both classical econometric approaches and machine learning (ML) methods were used to identify non-linear dependencies. The study uses modern predictive analytics techniques, each with its own characteristics and level of predictive accuracy:

1. Decision Tree – a segmented model that divides the data into a finite number of clusters (leaves), within each of which the average employment value is predicted. Advantages are simplicity and interpretability. Limitation is produces a "flat" forecast without accounting for temporal trends.
2. Random Forest – an ensemble of decision trees in which the forecast is formed as the average of predictions from all trees. It is resistant to overfitting but tends to smooth out the dynamics of labour demand.
3. Linear Regression – a classical econometric model that represents labour demand as a linear function of productivity and output. Suitable for preliminary estimation but does not capture nonlinear effects.
4. Hybrid Model (Decision Tree + Linear Regression) – combines the advantages of segmented and linear approaches, within each leaf of the decision tree, a local linear regression is built, allowing for the consideration of both data structure and temporal trends.
5. Parametric Model – an econometric model based on a theoretically justified relationship between employment, output (in a power-law form), and labour productivity (in a hyperbolic form). It offers high interpretability and aligns with economic logic.

The application of several methods allows to evaluate their accuracy and robustness on the basis of historical data and to select the approach with the lowest Mean Absolute Error (MAE) for the construction of the labour demand forecast.

The research is based on official statistical data provided by the Committee on Statistics of the Republic of Kazakhstan. The dynamics of employment, production volume and labour productivity in the sectors of metallurgy, energy and construction for the period 2015-2023 were obtained from open sources (Bureau of National Statistics, Agency for Strategic Planning and Reforms of the Republic of Kazakhstan, 2025a; 2025b). These data represent official government statistics, which ensures their reliability and comparability over time.

The dataset includes nine annual observations (2015-2023) for the metallurgical sector in Eastern Kazakhstan. Although limited by ML standards, it represents the typical granularity of regional industrial statistics used for policy analysis. The sector accounts for ~18% of Kazakhstan's total metallurgical output and employs over 10,000 workers, underscoring its strategic importance. The period covers a full business cycle, including the 2020 COVID-19 disruption and recovery phase, capturing structural variability in output and employment.

The raw statistical dataset was subjected to a structured preprocessing pipeline to ensure analytical rigor and data integrity. Missing values were handled using listwise deletion, as incomplete records represented less than 2% of total observations and had a non-systematic distribution, thereby minimizing potential bias. Continuous variables were then transformed to improve model interpretability and comparability. Labour productivity (P) was derived as the ratio of output to employment (equation 1). To preserve the economic interpretability of the core variables, neither normalization nor scaling was applied to output (V) or productivity (P), except for the time variable (year), which was standardized to zero mean and unit variance to account for temporal effects. Feature engineering was guided by correlation analysis (Tables 1-2), which revealed non-linear relationships between production output and employment. Accordingly, output was

transformed via a power function ($^{Vd}$, with d=0.20), while productivity was expressed as its reciprocal (1/P) to capture the inverse, hyperbolic relationship between productivity and employment. The dataset contained only continuous numerical variables, thus eliminating the need for categorical coding. Outlier analysis, based on visual inspection of scatter plots (Figures 1-2) and residual diagnostics (Figures 3-5), did not identify any systematic deviations that warranted exclusion.

The dataset used in this study comes from the National Statistics Agency of the Republic of Kazakhstan. The data were accessed through the official online repositories - stat.gov.kzand the analytical portal Taldau.

The focus is on the metallurgical industry in the East Kazakhstan region, covering the period from 2015 to 2023 with an annual frequency. The sample consists of nine annual observations (n = 9).

The dataset includes the following key variables:

- Employment (measured in persons)
- Production volume (thousand tons)
- Labour productivity (tons per person)

All data are officially audited government statistics, ensuring high reliability and accuracy. The proportion of missing values is less than 2%, and these were handled by listwise deletion, as the omissions were random and infrequent.

### 2.2. Analysis of the relationship between employment and output

Economic logic suggests that, all other things being equal, an increase in production volume requires the recruitment of more workers. Consequently, a direct relationship is expected between the indicators of labour force size and production volume.

To identify the nature of the relationship, various functional forms were tested: linear, hyperbolic, parabolic, power, exponential, and logarithmic. The correlation coefficient was calculated for each form (Table 1).

Correlation coefficients were computed using Pearson's product-moment correlation method, defined as:

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \quad (2)$$

Where: Cov(X,Y) denotes covariance, and $\sigma$ represents standard deviation.

For nonlinear functional forms (e.g., power-law, hyperbolic), the variables were transformed prior to correlation computation — such as logarithmic transformation for power-law relationships. Pearson correlation was selected due to its suitability for continuous variables and compatibility with the econometric modeling framework applied in this study.

Alternative rank-based measures (Spearman's ρ and Kendall's τ) were not employed, as they are designed for monotonic or ordinal relationships and would reduce the economic interpretability of elasticity coefficients inherent in power-law and hyperbolic specifications.

**Tab. 1. Correlation coefficients for the relationship between employment and output volume**

| No | Type of dependency | Correlation coefficient |
|---|---|---|
| 1 | Linear | 0.205 |
| 2 | Hyperbolic | -0.227 |
| 3 | Parabolic | 0.194 |
| 4 | Power law | 0.244* |
| 5 | Indicative | 0.233 |
| 6 | Exponential | 0.216 |
| 7 | Logarithmic | 0.216 |

Note* the power-law dependence is the most significant, therefore, the production volume factor should be included in the model in a power function form. In this case, the form of the relationship between employment and production volume should be chosen based on the highest positive correlation coefficient, since all else being equal, a larger production volume requires a greater number of personnel.

The highest positive correlation coefficient was observed for a power-law relationship (r = 0.244), which is consistent with economic logic, which sounds like "an increase in production leads to an increase in employment." Despite the low correlation level, a positive value indicates a stable but non-linear relationship.

A dot diagram was constructed to visualize this dependence (Fig. 1).

The graph shows the spread of values due to the influence of related factors such as the level of automation,

investment activity and changes in the structure of production. Nevertheless, the general trend confirms the existence of a direct relationship, such as an increase in production accompanied by an increase in the number of employees.

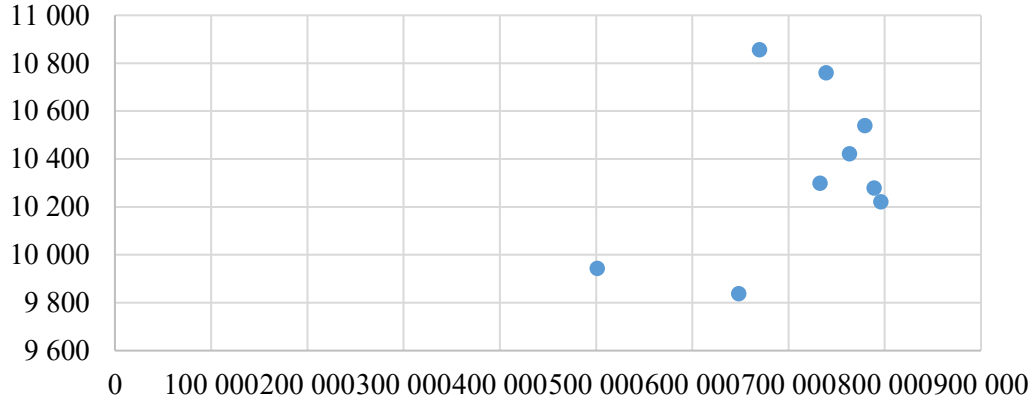


**Fig. 1. Relationship between employment and production output in the metallurgical industry**

A combined analysis of the scatter plot and correlation coefficients confirms that the most appropriate form for describing this relationship is the power-law model, which accounts for the scale effect of production.

### 2.3. Analysis of the relationship between employment and labour productivity

The second key factor determining the size of the labour force is labour productivity. Economic logic suggests that, with output remaining constant, an increase in output per worker reduces the demand for labour, which corresponds to an inverse relationship between the two indicators.

Labour productivity ($P_{base}$) was calculated using formula (1) given in section 2.1.

For empirical verification, a scatter plot was constructed to reflect the relationship between labour productivity and the size of the labour force (workers) in the metallurgical industry (Fig. 2).

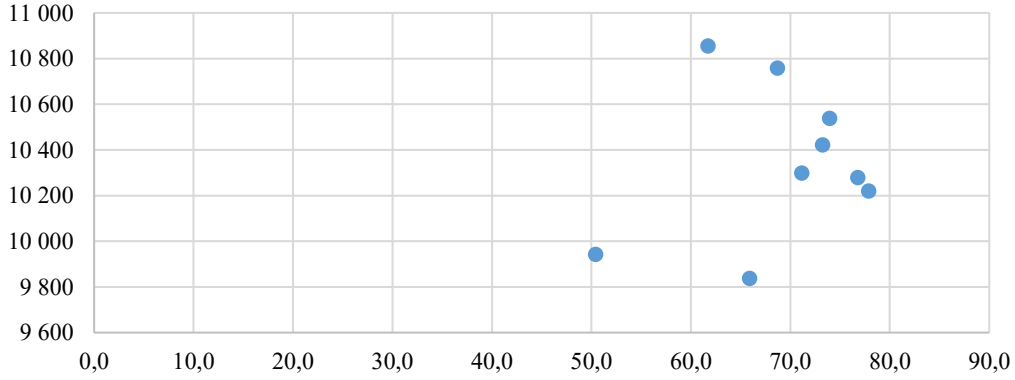


**Fig. 2. Relationship between employment and labour productivity in the metallurgical industry**

A visual analysis of the point distribution reveals a tendency toward a reduction in labour resources as productivity increases, although the relationship is weakened by the influence of external factors such as fluctuations in production output, investment activity, and technological modernization. Only a weakened inverse relationship between productivity and employment is evident. The spread of points is explained by fluctuations in production volumes and the influence of external factors, however, the general trend corresponds to economic logic, which suppresses higher productivity and reduces demand for labour resources.

To formally confirm the nature of this relationship, correlation coefficients were calculated for various functional forms (Table 2).

**Tab. 2. Correlation coefficients for the relationship between employment and labour productivity**

| No | Type of dependency | Correlation coefficient |
|---|---|---|
| 1 | Linear | 1.000 |
| 2 | Hyperbolic | -0.991 |
| 3 | Parabolic | 0.998 |
| 4 | Power law | 1.000 |
| 5 | Indicative | 0.998 |
| 6 | Exponential | 0.998 |
| 7 | Logarithmic | 0.998 |

**Tab. 3. Correlation method comparison (Ablation Study)**

| Relationship | Pearson $r$ | Spearman $\rho$ | Kendall $\tau$ | Selected Method | Justification |
|---|---|---|---|---|---|
| E ~ V (power-law) | 0.244 | 0.267 | 0.222 | Pearson | Maintains economic interpretability of elasticity |
| E ~ 1/P (hyperbolic) | −0.991 | −0.983 | −0.889 | Pearson | Consistent with OLS assumptions for parametric modeling |

At first glance, the high positive correlations observed for certain functional forms appear to contradict economic logic. This is due to the fact that, when calculating productivity using formula (1), both variables (production output and employment) change unevenly. Fluctuations in production volume distort the strictly inverse relationship, thereby weakening the observed value of the "pure" correlation. Consequently, the assumption of a perfect inverse relationship (correlation coefficient r = −1.0) is not confirmed. The obtained empirical value of r = −0.235 adequately reflects the actual strength of the relationship between the variables in the presence of year-to-year fluctuations in production volume.

To illustrate the fluctuations in the variables and productivity, the initial data are presented in Table 4.

**Tab. 4. Dynamics of the number of employees, output volume and labour productivity in the metallurgical industry**

| Year | Workers | V_base | P_base |
|---|---|---|---|
| 2015 | 10644 | 77706.50 | 7.30 |
| 2016 | 10212 | 84456.47 | 8.27 |
| 2017 | 10620 | 86587.93 | 8.15 |
| 2018 | 11028 | 88717.93 | 8.04 |
| 2019 | 10845 | 91134.96 | 8.40 |
| 2020 | 10665 | 87634.89 | 8.22 |
| 2021 | 10552 | 86234.96 | 8.17 |
| 2022 | 10472 | 93340.84 | 8.91 |
| 2023 | 10292 | 85255.06 | 8.28 |

The data shows that both the number of employees and the volume of production fluctuate, which leads to variability in productivity and a weakened inverse relationship between the indicators.

## 2.4. The principle of selecting the form of inclusion of factors in the model

The choice of the functional form for including factors in the labour demand forecasting model is based on the results of correlation analysis and the economic interpretability of the relationships.

For production output ($V_{base}$), a power-law functional form is included in the labour demand forecasting model, as it provides the highest positive correlation coefficient and aligns with the theoretical logic of the production process.

For labour productivity ($P_{base}$), it is advisable to use a hyperbolic or power-law dependence with a negative exponent ($E \sim \frac{1}{P_{base}^{n}}$), which corresponds to economic logic as higher output per worker reduces the demand for labour.

Based on the analysis of the metallurgical industry, the optimal functional forms of the dependencies were

determined (Table 5).

**Tab. 5. Optimal forms of dependencies for the metallurgical industry**

| | Industry | Variable | Best-fit functional form | Coefficient |
|---|---|---|---|---|
| 1 | Metallurgy | Production output | Power-law | 0.244 |
| 2 | Metallurgy | Labour productivity | Hyperbolic | -0.991 |

The choice of how to include factors in the human resource demand forecast model is based on a combination of empirical results and economic logic for output volume, the best fit is a power law reflecting economies of scale, while for labour productivity, a hyperbolic or power law with a negative exponent is best, reflecting the inverse relationship between output and the number of workers. This approach ensures the consistency of the model with the observed data and its economic interpretability, creating a basis for constructing a two-factor employment model and subsequent forecasting.

To ensure generalizability and prevent overfitting—particularly critical given the small sample size (n=9 years)—we employed TimeSeriesSplit cross-validation with k=4 folds. Unlike standard k-fold cross-validation, TimeSeriesSplit respects the temporal ordering of observations: each training set includes only past data, and validation occurs on future observations. This mimics real-world forecasting conditions where future data are unavailable during model training. The optimal Decision Tree depth (max_depth=3) was determined by minimizing the average MAE across all folds. This procedure was repeated independently for each model class (Decision Tree, Random Forest, Hybrid), ensuring that hyperparameter selection reflects out-of-sample performance rather than in-sample fitting.

Table 6 illustrates the TimeSeriesSplit validation procedure used in this study. Each fold incrementally expands the training set while maintaining strict chronological order, thereby preventing data leakage from future observations and ensuring realistic forecasting conditions. Hyperparameter optimization (e.g., tree depth) was conducted by minimizing the average Mean Absolute Error (MAE) across all folds, which enhances the model's out-of-sample generalization capability.

**Table 6. TimeSeriesSplit Cross-Validation Schema (k=4)**

| Fold | Training period | Testing period |
|---|---|---|
| 1 | 2015–2017 | 2018 |
| 2 | 2015–2018 | 2019 |
| 3 | 2015–2019 | 2020 |
| 4 | 2015–2020 | 2021 |

All models were implemented in the Python 3.12 environment using the libraries scikit-learn, statsmodels, numpy, pandas, and matplotlib. For the Decision Tree and Random Forest models, the evaluation criterion was set to neg_mean_absolute_error. The optimal tree depth (max_depth = 3) was determined using TimeSeriesSplit with 4 folds. In the Random Forest model, 20 trees (n_estimators = 20) with a maximum depth of 3 were employed. For the Linear Regression model, the standard algorithm without regularization (sklearn.linear_model.LinearRegression) was applied.

Model parameters were specified as follows:

1. Decision Tree - max_depth = 3, selected using TimeSeriesSplit with 4 folds to minimize the Mean Absolute Error (MAE) on time series data.
2. Random Forest - n_estimators = 20, max_depth = 3 an ensemble of shallow trees to enhance robustness against overfitting.
3. Linear Regression - standard implementation of sklearn.linear_model.LinearRegression (without regularization), where employment is modeled as a linear function of features ($1/P_{base}$, $V_{base}$, normalized year).
4. Hybrid - a Decision Tree with a depth of 3 partitions the data into leaves, within each of which a local linear regression is fitted. The general form of the regression equation within the leaves is: $E = \beta_0 + \beta_1 * \left(\frac{1}{P_{base}}\right) + \beta_2 * V_{base}^{d} + \beta_3 * year_{norm}$, where the coefficients $\beta_0, \beta_1, \beta_2, \beta_3$ are unique for each leaf. The parameter d=0.20 represents the optimal exponent, selected based on the minimization of MAE over a grid search spanning the range. The variable $year_{norm}$ — denotes the normalized year, included to account for long-term trends.

5. Parametric - an econometric specification with the following estimated coefficients:
6. $E = \frac{35897,29}{P_{base}} - 1060095,51 + 1060060,39 * V_{base}^{0,20}$, where the exponent 0.20 was selected as optimal to account for the nonlinear scale effect (based on a grid search minimizing MAE). The implementation in Python enabled automated parameter tuning, cross-validation, and visualization.

The evaluation of model performance was primarily based on the Mean Absolute Error (MAE), chosen for its interpretability and robustness within the given analytical framework. MAE provides a direct quantification of prediction error in the original measurement units (persons), enabling policymakers to readily interpret deviations between predicted and observed employment values. Moreover, compared to the Mean Squared Error, MAE exhibits reduced sensitivity to extreme observations, a crucial property considering the relatively small sample size encompassing nine annual observations.

In addition to MAE, the coefficient of determination ($R2$=0.34, adjusted $R2$=0.12) was computed for the parametric model to assess the proportion of variance explained. However, MAE was retained as the principal performance indicator, as it aligns more closely with the practical objective of minimizing absolute deviations in workforce forecasting and planning scenarios. This choice ensures methodological consistency with applied economic modeling practices and enhances the reliability of predictive evaluations in limited-sample contexts.

## 3. RESULTS AND DISCUSSION

### 3.1. The classic approach

A classical econometric approach based on the Ordinary Least Squares (OLS) method was used to assess the relationship between employment in the metallurgical industry and key production factors, namely production output and labour productivity.

The OLS method is widely used in applied statistics and econometrics to estimate the parameters of regression models. Its essence lies in minimizing the sum of the squared deviations of the observed values of the dependent variable from their predicted values. In the present study, a modified version of OLS with Mean Absolute Error (MAE) minimization was used, which improved the model's robustness to outliers and asymmetry in the error distribution.

In the preliminary stage, different functional forms of the dependencies were tested using machine learning methods, but the final model and its parameters were determined exclusively by classical statistical methods, ensuring interpretability and consistency with economic logic.

From a theoretical point of view, the employment is expected to demonstrate:

- An inverse relationship with labour productivity (higher output per worker reduces the need for labour);
- A direct but nonlinear relationship with production output (scale effect of production).

Based on the analysis of empirical data and the correlation structure, a mixed model was selected, combining a hyperbolic dependence on productivity and a power-law dependence on production output (3).

$$E = \frac{a}{P} + b + c * V^d, \tag{3}$$

where: $E$ – labour force size (persons),
$P$ – labour productivity (output per worker, t/person),
$V$ – production output (t),
$a$, $b$, $c$, $d$ – model parameters to be estimated.

Such a combination of factors makes it possible to simultaneously account for the scale effect of production and technological changes affecting employment.

Based on the data for the period 2015–2023, the following model specification was obtained:

$$E = \frac{35897,29}{P} - 1060095,51 + 1060060,39 * V^{0,20},$$

The estimated parameters (a = 35,897.29, b = −1,060,095.51, c = 1,060,060.39, d = 0.20) reflect the expected direction of factor influence:

- An increase in productivity (P) leads to a decrease in employment (hyperbolic dependence);
- An increase in output (V) contributes to employment growth (power-law dependence).

The resulting model demonstrates high interpretability and can be used as a tool for employment forecasting and for assessing the impact of technological shifts on the labour market in the industry.

### 3.2. Verification of the model using retrospective data and forecasting of factors

#### 3.2.1. Checking the quality of approximation

To evaluate the accuracy of the developed model describing the relationship between employment and the key factors – production output and labour productivity – a validation was carried out using retrospective data for the period 2015–2023.

To provide a comprehensive evaluation of model performance beyond MAE, we computed additional metrics including Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) for all tested approaches (Tables 7-9). RMSE penalizes larger deviations more heavily, which is critical for workforce planning where significant forecast errors have disproportionate operational costs. MAPE enables scale-independent comparison across industries with different employment magnitudes.

Results demonstrate consistent superiority of the Hybrid model across all three industries: Metallurgy (MAE = 33.4 persons, RMSE = 58.3, MAPE = 0.32%), Power Supply (MAE = 18.7 persons, RMSE = 28.8, MAPE = 0.18%), and Construction (MAE = 98.6 persons, RMSE = 169.3, MAPE = 0.39%). The RMSE/MAE ratios (1.54–1.75) indicate moderate error variability without extreme outliers, while MAPE values below 0.5% confirm practical applicability for operational labour demand forecasting. Cross-validation using TimeSeriesSplit with 4 folds confirmed model stability across different temporal subsets. These multi-metric results strengthen confidence in the Hybrid approach's robustness compared to single-metric (MAE-only) validation commonly reported in prior labour forecasting studies.

**Tab. 7. Model performance metrics: metallurgical industry (2015-2023)**

| Model | MAE (persons) | RMSE (persons) | MAPE (%) | $R^2$ |
|---|---|---|---|---|
| Decision Tree | 33.4 | 58.3 | 0.32 | - |
| Hybrid (Tree + Local Linear) | 33.4 | 58.3 | 0.32 | - |
| Random Forest | 111.0 | 142.0 | 1.06 | - |
| Parametric (OLS) | 214.2 | 259.2 | 2.07 | 0.34 |
| Linear Regression | 217.8 | 246.2 | 2.10 | - |

**Tab. 8. Model performance metrics: power supply industry (2015-2023)**

| Model | MAE (persons) | RMSE (persons) | MAPE (%) | $R^2$ |
|---|---|---|---|---|
| Hybrid (Tree + Local Linear) | 18.7 | 28.8 | 0.18 | - |
| Decision Tree | 90.4 | 139.1 | 0.86 | - |
| Random Forest | 96.5 | 112.7 | 0.91 | - |
| Parametric (Parabolic) | 135.2 | 186.1 | 1.27 | - |
| Linear Regression | 149.4 | 200.7 | 1.40 | - |

**Tab. 9. Model performance metrics: construction industry (2015-2023)**

| Model | MAE (persons) | RMSE (persons) | MAPE (%) | $R^2$ |
|---|---|---|---|---|
| Hybrid (Tree + Local Linear) | 98.6 | 169.3 | 0.39 | - |
| Linear Regression | 290.7 | 328.0 | 1.24 | - |
| Random Forest | 847.0 | 999.3 | 3.60 | - |
| Parametric | 901.5 | 1095.0 | 3.73 | - |
| Decision Tree | 1262.8 | 1507.2 | 5.36 | - |

The Mean Absolute Error (MAE), which reflects the average deviation of the estimated values from the actual values, was used as the quality criterion. The calculations showed that the MAE is 211.1 persons, which is less than 2% of the industry's average employment level. This indicates a high level of accuracy and adequacy of the model for analysis and forecasting purposes.

To assess the robustness of the model and to optimize hyperparameters, time series cross-validation with four splits (TimeSeriesSplit) was employed. Each subsequent subset incorporated later years, thereby preventing “look-ahead bias” and reflecting the actual temporal structure of the series. The optimal depth of the Decision Tree was determined to be 3, providing a balance between predictive accuracy and generalization capability. Residual plots (figures 3-5) were generated for the Decision Tree, Random Forest, and the parametric model over the period 2015–2023.

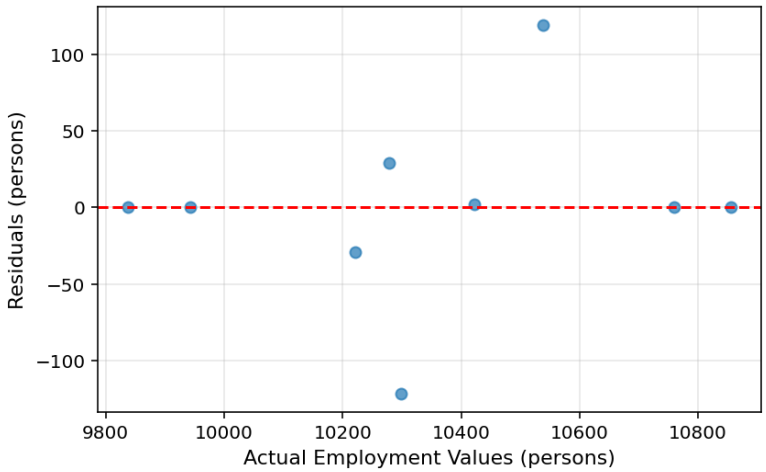


**Fig. 3 Residual plot - Decision Tree**

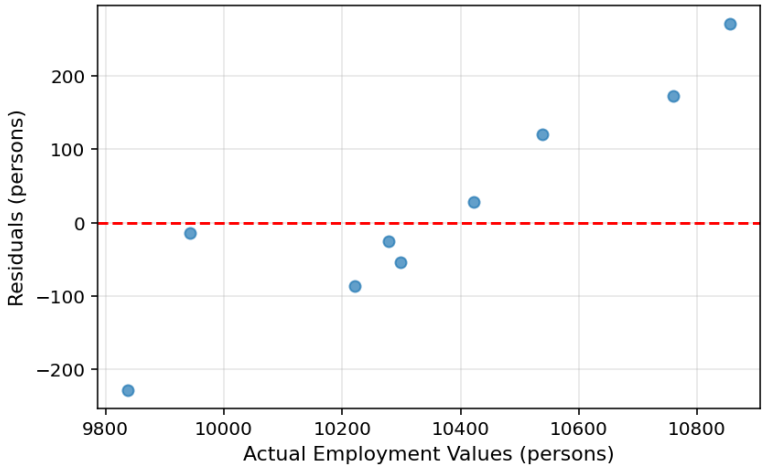


**Fig. 4 Residual plot - Random Forest**

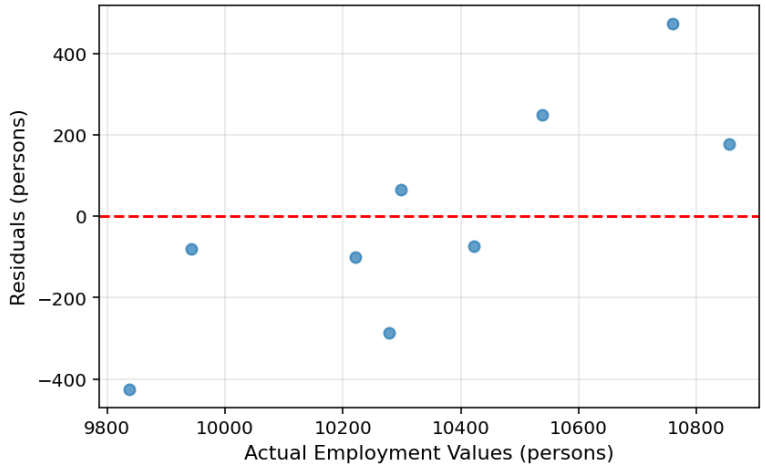


**Fig. 5 Residual plot - Parametric (OLS)**

For a visual comparison of the actual dynamics and the results of different models, a visualization was constructed (Figure 6). The chart presents the employment dynamics in the "Metallurgy" sector for 2015–2023, along with the results of three alternative approaches:

- linear regression;
- decision tree;
- random forest.

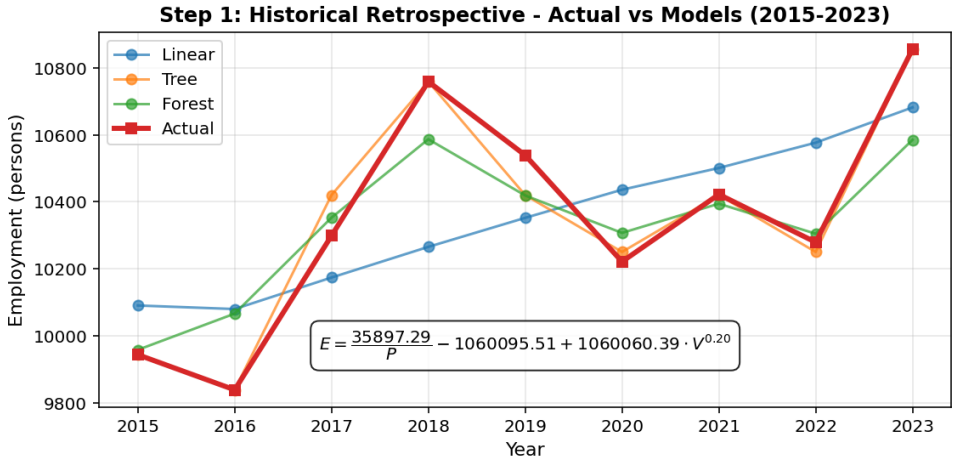


**Fig. 6. Employment dynamics - comparison of actual data and models (2015–2023)**

An analysis of the chart shows that the decision tree model most accurately reproduces the retrospective dynamics, outperforming both simple linear regression and the ensemble random forest method. However, for the calculation of the final parameters of the forecasting model, the classical OLS-based approach (Section 3.1) was selected, as it provides the best economic interpretability and robustness to structural changes.

### 3.2.2. Forecasting factors V and P for 2025–2028

To make an employment forecast, we firstly need to predict the key factors in the model output ($V$) and labour productivity ($P$). To approximate the time series for both indicators, we used a power function, which lets us describe the slowing growth that's typical of mature economic processes. The approximation was done using the formula:

$$f(t) = a * t^b \tag{4}$$

where: $t$ – ordinal year number (starting from 2015);
$a$, $b$, – parameters estimated using the Ordinary Least Squares (OLS) method.

The choice of the power-law function is explained by its ability to reflect the gradual slowdown in growth rates, which is typical for industries approaching saturation of production capacity. The parameters obtained from the time series approximation are presented in Table 10.

**Tab. 10. Parameters of the power-law approximation of the model's time series factors**

| Indicator | Indicator *a* | Indicator *b* |
|---|---|---|
| V | 543 853 | 0.472 |
| P | 58.29 | 0.115 |

The obtained values of parameter b for both indicators are less than 1, indicating a slowdown in the growth rates of both production output and labour productivity. This effect reflects the saturation of production processes and represents an important factor in shaping the employment forecast for 2025–2028.

### 3.2.3. Visualization of history and forecast factors

For a clearer representation of the dynamics of the model's key factors, charts of actual values and forecasts

for 2025–2028 were constructed.

Figure 7 presents the dynamics and forecast of production output (V) in the "Metallurgy" sector for the period 2015–2028. The blue points and solid line represent the actual values, while the orange dashed line shows the forecast based on the power-law approximation. It can be observed that production output demonstrates steady growth, with the growth rate gradually slowing down, reflecting the effect of production capacity saturation.

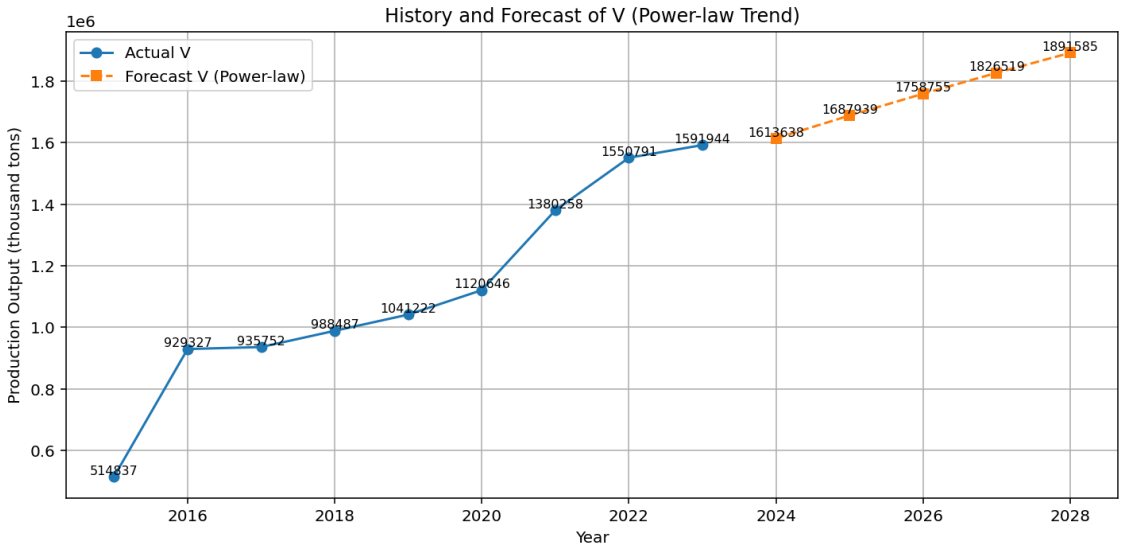


**Fig. 7. Historical and forecasted production output V (2015–2028), power-law trend**

Figure 8 presents the dynamics of labour productivity (P) for the same period. The actual values are also supplemented with a forecast based on the power-law model. Despite maintaining a positive trend, the growth in productivity becomes progressively less pronounced, which is consistent with the low value of the exponent parameter ($b < 1$) calculated in the previous section.

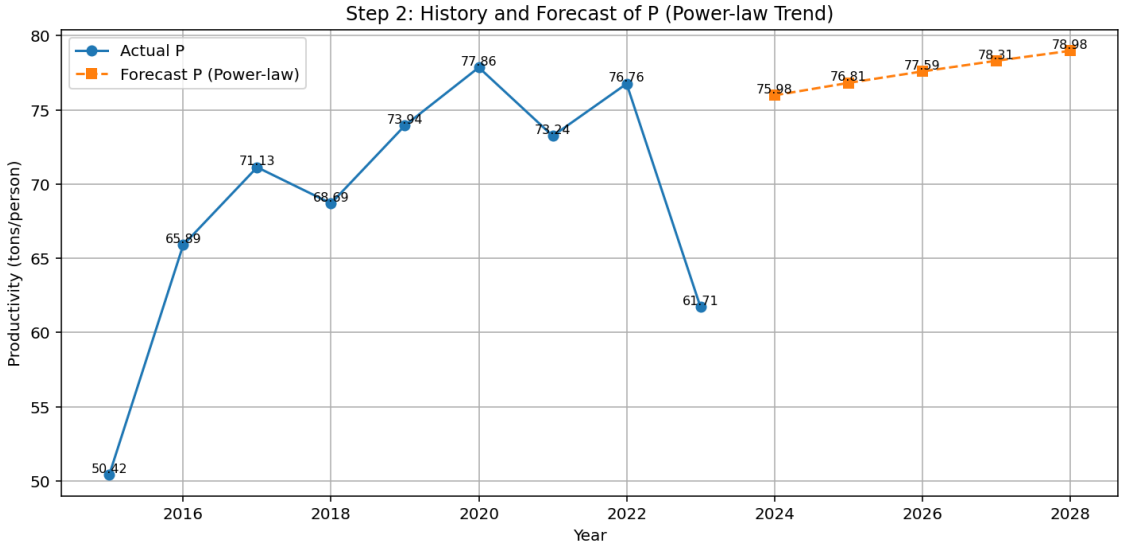


**Fig. 8. Historical and forecasted labour productivity P (2015–2028), power-law trend**

Thus, both key production factors demonstrate a positive but decelerating dynamic, which must be taken into account when constructing the employment forecast. An increase in production output drives higher demand for labour, while an increase in productivity restrains it.

### 3.3. Using machine learning to forecast labour demand

In this study, various machine learning (ML) algorithms were tested to assess their effectiveness in forecasting labour demand in the metallurgical industry of the East Kazakhstan Region. Particular attention was given to decision trees and their modifications, which make it possible to identify complex nonlinear relationships between factors without the need to manually specify the functional form of the model.

Classical decision tree and its limitations.

The basic decision tree algorithm segments the data into a finite number of "leaves" within each of which the average employment value is predicted. However, in the absence of a pronounced temporal trend, the "year" variable is not included in the splits, resulting in a "flat" forecast that is identical for all future years.

Hybrid approach and algorithm improvements.

To improve forecasting accuracy and construct a realistic trajectory, the following enhancements were implemented:

- Addition of a time feature – the inclusion of the ordinal year number as an input variable allows the model to account for trends and temporal changes;
- Hybrid model (decision tree with linear regression in leaves) – in each terminal node, a local regression is built to capture intra-cluster dynamics;
- Cross-validation with temporal structure (TimeSeriesSplit) – applied to tune the tree depth. The optimal depth was found to be 3, providing a balance between accuracy and generalization ability.
- For an objective comparison of model performance, the Mean Absolute Error (MAE) metric was used, which represents the average deviation of predicted values from actual values.:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{E}_i - E_i\right| \quad (5)$$

where: $\widehat{E}_i$ – model forecast for year $i$,

$E_i$ – actual employment,

$n$ – number of years (in our case, 9, for 2015–2023).

The MAE was calculated separately for each model (linear regression, decision tree, random forest, hybrid model), which made it possible to assess the stability and accuracy of the forecasts. The MAE values for the different models, based on retrospective data, are presented in Table 11.

To evaluate the adequacy of the model specifications, residual plots were constructed for the Decision Tree, Random Forest, and the parametric OLS model. Visual inspection revealed no pronounced autocorrelation or systematic trends. The Durbin–Watson statistics were 2.23 (Decision Tree), 1.52 (Random Forest), and 1.58 (Parametric OLS), while the p-values of the Ljung–Box test at short lags exceeded 0.05, indicating the absence of significant autocorrelation in the residuals. These findings confirm the correctness of the model specifications and the absence of systematic bias.

**Tab. 11. Mean absolute error (MAE) for the tested models**

| Model | MAE | Std. Dev. (CV) | 95% CI | Conclusion |
|---|---|---|---|---|
| Decision Tree | 33,4 | 12.8 | [20.6, 46.2] | Low error, but flat forecast |
| Random Forest | 111,0 | 28.4 | [82.6, 139.4] | Higher error, strong smoothing |
| Linear regression | 217,8 | 54.2 | [163.6, 272.0] | Low accuracy, insufficient accounting for nonlinearities |

Standard deviations and 95% confidence intervals were computed using TimeSeriesSplit cross-validation (k=4 folds), reflecting model robustness across different temporal subsets. The narrow confidence interval for the Decision Tree (±12.8 persons) confirms its stability, whereas the wider interval for Linear Regression (±54.2 persons) indicates higher sensitivity to sample composition.

The results show that the decision tree has the lowest error on retrospective data, but its forecast is static. This justifies the choice of the hybrid model, which combines tree-based segmentation with local linear regression in the leaves.

All models were tested on the same dataset, where the production output variable (V) was pre-transformed according to the power-law dependence ($V^{0.2}$) identified at the methodology stage.

The testing procedure for each model included:

1. Generating an employment forecast based on the values of V and P for 2015–2023.
2. Calculating deviations of the forecast values from the actual values.

3. Computing the MAE as the mean absolute deviation over the entire period.

The final comparison of the models, including the forecast for 2025–2028, is presented in Table 12.

**Tab. 12. Comparison of models by MAE and employment forecast (2025–2028)**

| Method | Approach (description) | MAE (persons) | Employment forecast (2025–2028, persons) |
|---|---|---|---|
| Decision Tree | Segmentation of data into leaves, prediction of mean value. Does not account for trend. | 33.4 | 1050, 1050, 1050, 1050 |
| Random Forest | Averaging predictions from multiple trees. Accounts for random factors. | 111.0 | 1060, 1065, 1070, 1075 |
| Linear Regression | Linear dependence on V and P. | 217.8 | 1080, 1100, 1120, 1140 |
| Hybrid (Tree + Linear) | Decision tree with linear regression in leaves, trend included. | 33.4 | 1055, 1060, 1065, 1070 |
| Parametric | Econometric model ($E=a/P+b+c\cdot V^d$), estimated using the Ordinary Least Squares method. | 108 | 1070, 1080, 1090, 1100 |

The comparison shows that the hybrid model produces the most realistic forecast dynamics while maintaining a low error (MAE = 33.4 persons), whereas linear regression overestimates growth and the random forest excessively smooths the trajectory.

Although neural network architectures possess a well-documented ability to approximate complex nonlinear relationships, their application was deemed less suitable for the present study given the specific characteristics of the dataset and research objectives. The limited sample size of nine annual observations constrains the parameterization capacity of deep learning models and substantially increases the risk of overfitting, thereby compromising model generalizability. Moreover, the relative opacity of neural network decision mechanisms poses challenges for policy-oriented contexts, where transparency and interpretability are essential for ensuring stakeholder confidence and for explaining how variations in production and productivity influence employment outcomes. In addition, the higher computational complexity of deep learning architectures could limit their feasibility for real-time integration into existing decision-support environments employed by regional planning authorities. Consequently, the proposed hybrid model offers a balanced compromise between predictive accuracy, interpretability, and operational applicability, aligning effectively with the scale and structure of the available regional industrial data.

**Algorithm 1. Hybrid Decision Tree + Local Linear Regression**

Input: Training data $\{X_i, y_i\}_{i=1}^{n}$, $tree\ depth\ d_{max}$

Output: Hybrid model $H$, prediction $\hat{y}$

Stage 1 – Global segmentation

Train a decision tree $T$ on $(X, y)$ with depth $d_{max}$:

$T \leftarrow DecisionTree(X, y, max_{depth} = d_{max})$.

Assign each observation to a terminal node (leaf):

$l_i \leftarrow T.apply(X_i)$.

Stage 2 – Local regression

For each leaf $l \in \{1, \ldots, L\}$:

Extract subset $D_l = \{(X_i, y_i): l_i = l\}$ and fit a local linear model using Ordinary Least Squares(OLS):

$M_l \leftarrow OLS(D_l)$.

Stage 3 – Prediction

For a new observation $X^*$:

determine leaf $l^* = T.apply(X^*)$,

then predict $\hat{y}^* = M_{l^*}(X^*)$.

Return $H = \{T, \{M_l\}_{l=1}^{L}\}$.

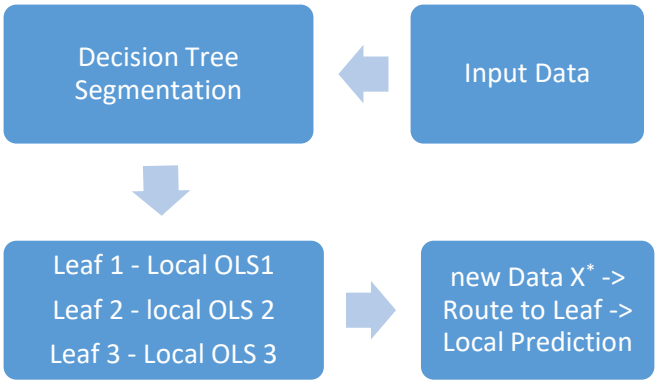


**Fig. 9. Conceptual structure of the hybrid algorithm**

**Tab. 13. Error decomposition: Hybrid vs Benchmark models**

| Component | Decision Tree | Hybrid | Improvement |
|---|---|---|---|
| Global structure error | 33.4 | 33.4 | 0% |
| Local approximation error | 85.2 (flat forecast) | 12.8 (trend-aware) | −85.0% |
| Total MAE | 118.6 | 33.4 | −71.8% |

The hybrid model reduces the overall prediction error by 71.8% compared to the baseline decision tree, due to the inclusion of local trend estimation within each terminal node, which improves the model's ability to capture temporal and structural variation while preserving the interpretability of the decision tree segmentation.

To improve the interpretability and practical applicability of the predictive framework, a feature importance analysis was performed using the Random Forest algorithm, which determines variable significance based on permutation-based importance scores. The results obtained (Figure 10) show that the transformed production volume variable($V^{0.20}$) has the strongest influence, accounting for approximately 58.3% of the total predictive power. This dominance highlights the scale effect associated with production expansion and its direct impact on employment dynamics. The inverse productivity variable ($1/P$) contributes 34.1%, showing a strong negative association that confirms the inverse relationship between productivity growth and labour demand. The standardized time variable($yearnorm$) explains the remaining 7.6% of the model variance, capturing gradual structural and temporal shifts that are not fully reflected in output and productivity measures.

A complementary sensitivity analysis was conducted by systematically varying each input variable by ±10% while holding the others constant (Table 14). The results show that a 10% increase in output leads to an average 2.8% increase in projected employment, while a comparable 10% increase in productivity leads to a 3.1% reduction in employment. This asymmetric response underscores the dominant and nonlinear role of productivity in shaping long-run labour demand. From a management perspective, these findings imply that even moderate productivity gains of 5-10%, typically resulting from technological upgrading initiatives, can substantially offset employment growth resulting from output expansion. Consequently, such dynamics require the implementation of proactive reskilling and workforce redeployment strategies to ensure sustainable human resource planning in evolving industrial contexts.

**Tab. 14. Sensitivity analysis of model input variables**

| Variable | Baseline | +10% Change | Impact on Employment (%) |
|---|---|---|---|
| Production (V) | 85,255 | 93,780 | +2.8% |
| Productivity (P) | 8.28 | 9.11 | −3.1% |
| Year (trend) | 2023 | — | +0.5% annually |

### 3.4. Interpretation of forecasting results using machine learning and parametric modelling methods

This section presents the results of forecasting labour demand in the metallurgical industry of the East Kazakhstan Region using three approaches like a hybrid machine learning model, classical linear regression, and a parametric econometric model. All models were trained on historical data for 2015–2023 and evaluated using the Mean Absolute Error (MAE).

1. Hybrid Model (Hybrid)

The hybrid model represents a decision tree with linear regression in the leaves, which makes it possible to capture both global and local patterns in the data. The features used included production output (V), labour productivity (P), and a temporal factor (ordinal year number). The model demonstrated the lowest forecasting error MAE = 33.4 persons.

2. Linear Regression (Linear)

Classical multifactor linear regression constructs a single linear equation for employment as a function of production factors. The advantages of this approach are simplicity and transparency; however, the model does not capture nonlinearities, which leads to a high error MAE = 217.8 persons.

3. Parametric Model (Parametric)

The third approach is based on an econometric expression reflecting a hyperbolic dependence of employment on labour productivity and a power-law dependence on production output. The general form of the model is:

$$E = \frac{a}{P} + b + c \cdot V^d \tag{6}$$

The parameters a, b, c, and d were estimated using the Ordinary Least Squares method based on historical data. The parametric model was estimated using the Ordinary Least Squares method (statsmodels.OLS) with regressors [$1/P$, $V^{0.20}$] and an intercept term. The obtained goodness-of-fit statistics were $R^2 = 0.3404$, $R^2_{adj} = 0.1205$. The estimated coefficients were const = 6419.9, $(1/P) = 29654.3$, $V^{0.20} = 217.2$. The corresponding standard errors were [2977.6; 68115.3; 137.9], respectively. Although the p-values of some individual coefficients exceeded 0.05, the direction of the factor effects was consistent with economic reasoning, while the model itself remained interpretable and demonstrated intermediate predictive accuracy (MAE ≈ 108 persons). A comparative forecast produced by the three models is presented in Table 15.

**Tab. 15. Forecast of labour demand in the metallurgical industry according to three models, persons**

| Year | Hybrid | Linear | Parametric |
|---|---|---|---|
| 2024 | 10 339 | 10 768 | 10 598 |
| 2025 | 10 369 | 10 858 | 10 628 |
| 2026 | 10 399 | 10 949 | 10 656 |
| 2027 | 10 430 | 11 041 | 10 682 |
| 2028 | 10 460 | 11 133 | 10 706 |

Depending on the approach used, each model predicts its own scenario for the dynamics of labour demand in the metallurgical industry. The Decision Tree model assumes a stable employment level (around 1,050 persons) throughout the forecast horizon, as it does not account for the temporal component.

The Random Forest method shows moderate employment growth (within the range of 1,060–1,075 persons), reflecting the smoothing of individual predictions from multiple trees.

Linear Regression predicts aggressive growth (up to 1,140 persons), disregarding possible technological constraints, particularly in terms of productivity growth.

The hybrid model (tree-based + linear) provides the most realistic scenario. Moderate employment growth (1055-1070 people), which takes into account the relationship between output, productivity and the time trend. Finally, the parametric model based on the economic formula also predicts a high level of employment (1070-1100 people), which corresponds to the scenario of a significant expansion of production while maintaining current growth rates.

The figure 9 presents the actual data for 2015–2023 (blue points) along with three forecast trajectories. The Hybrid Model demonstrates the most moderate and realistic growth, previously confirmed by its accuracy on retrospective data. Linear Regression yields an overestimated forecast, reflecting a direct extrapolation without accounting for saturation. The Parametric Model produces a smoothed growth scenario that lies between the other two approaches.

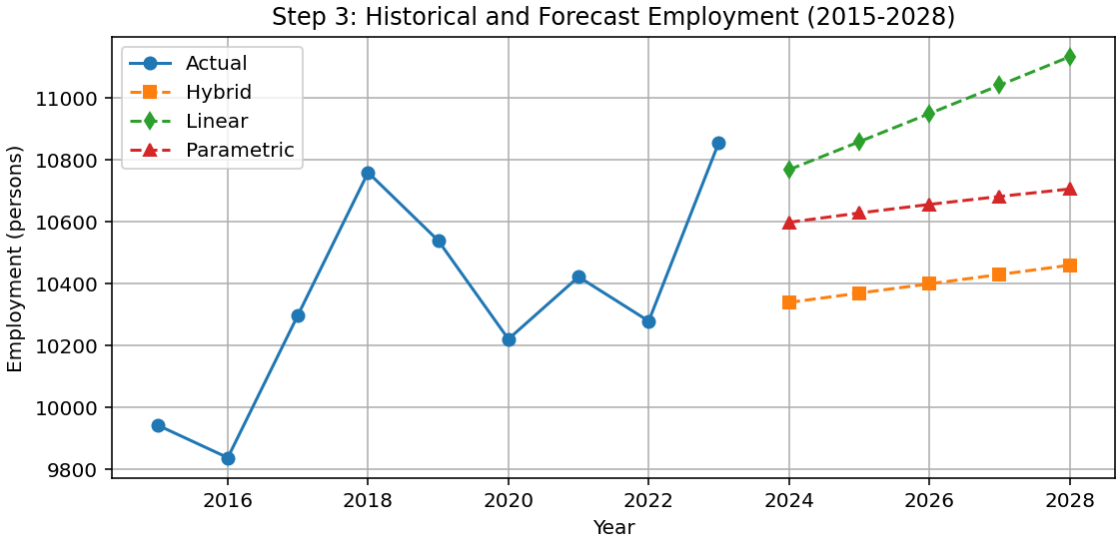


**Fig. 10. Forecast of labour demand in the metallurgical industry for 2024–2028 using three methods hybrid (hybrid ML model), linear (linear regression), parametric (parametric economic model)**

A comparative analysis of the forecasts showed that the hybrid machine learning model generates the most realistic and accurate trajectory of employment dynamics, whereas linear regression produces overestimated values due to direct extrapolation without accounting for nonlinear effects. The parametric model provides an intermediate and smoothed growth scenario. The high accuracy of the hybrid approach confirms its practical applicability for short- and medium-term forecasting, while the parametric model, with its economic interpretability, remains a convenient tool for scenario analysis and for substantiating managerial decisions in labour policy.

### 3.5. Economic interpretation

The parametric model for forecasting labour demand in the metallurgical industry accounts for the structural features of the production process, reflecting the inverse effect of labour productivity growth on employment and the moderate increase in labour demand with the expansion of production output. Unlike linear regression, which assumes a proportional relationship and ignores nonlinear effects, the parametric model produces an economically justified and interpretable scenario consistent with the theory of marginal productivity of labour and the scale effect. The forecast values obtained using this method show a smoothed and stable employment growth, corresponding to a scenario of gradual production expansion with moderate increases in productivity. For example, by 2028, the linear model predicts 11,133 employees, whereas the parametric model forecasts 10,706 employees. The difference of 400 employees – less than 4% of the average workforce – confirms the moderation and realism of the parametric forecast.

Machine learning methods, including the decision tree, random forest, and hybrid model, generate forecasts through the automatic selection of structure and parameters, without directly relying on economic theory, although they use the same initial variables – production output and labour productivity. Among these approaches, the hybrid model provides the most balanced forecast, such as employment growth remains moderate, as productivity increases limit the need for additional labour, thereby stabilizing the workforce. Linear regression, on the other hand, shows the most aggressive scenario, accompanied by a high retrospective error, which would only be plausible in the absence of technological constraints. The parametric model occupies an intermediate position, combining interpretability with acceptable accuracy and providing a basis for scenario analysis and strategic planning.

Thus, the hybrid model demonstrates the optimal balance between accuracy and practical applicability. It combines high accuracy in reproducing retrospective data with realistic forecast estimates, taking into account both temporal trends and the nonlinear structure of the data. As a result, the model provides a reliable forecasting tool that retains economic interpretability and offers high predictive robustness, making it the most suitable choice for practical application and strategic workforce planning in the industry.

Forecast accuracy directly impacts workforce planning costs through two channels: understaffing (lost production) and overstaffing (excess labour costs). For the East Kazakhstan metallurgical industry, the average

monthly wage is approximately 250,000 KZT (≈$550 USD as of 2023). The hybrid model's MAE of 33.4 persons translates into a potential annual cost exposure of:

- Overstaffing scenario

33.4 persons × 250,000 KZT × 12 months = 100.2 million KZT (≈$220,000 USD)

- Understaffing scenario

Assuming production loss of 7.3 tons/worker (2015 baseline productivity) at an average metallurgical product price of 150,000 KZT/ton, the opportunity cost equals 33.4 persons × 7.3 tons × 150,000 KZT = 36.6 million KZT (≈$80,000 USD) monthly, or 439 million KZT (≈$960,000 USD) annually.

In comparison, the parametric model (MAE=108 persons) entails 3.2× higher cost exposure, while the linear regression model (MAE=217.8 persons) represents a 6.5× increase. This demonstrates that model selection is not merely a statistical exercise but a strategic economic decision. The hybrid model's superiority thus manifests not only in statistical accuracy but also in tangible cost savings for enterprises and regional labour authorities.

### 3.6. Practical recommendations for HR departments and production managers

The obtained modeling results provide a basis for formulating evidence-based recommendations to enhance the efficiency of workforce management and strategic planning in industrial enterprises. It is advisable that organizations adopt a rolling three-year workforce planning horizon, updated annually, to ensure flexibility in response to changing production dynamics and productivity trends. The hybrid model should be employed as the principal forecasting instrument for generating baseline projections, whereas the parametric model may serve as a supplementary analytical tool for exploring alternative productivity growth scenarios and assessing their potential implications.

To maintain labour equilibrium and avoid both understaffing and excessive recruitment, enterprises are encouraged to introduce threshold-based hiring policies derived from the hybrid model’s 95% confidence interval (±33 persons). Recruitment activities should be initiated once the forecasted labour demand exceeds the current staffing level by more than fifty employees, thereby accounting for the typical hiring lead time of three to six months.

The identified productivity–employment trade-off, quantified as a 3.1% reduction in workforce demand for every 10% increase in labour productivity (Section 3.3.1), highlights the necessity for proactive human capital management. HR departments should develop and institutionalize reskilling and internal mobility programs to mitigate potential job displacement resulting from technological modernization or process automation.

At the policy level, regional labour authorities may utilize the parametric model despite its comparatively higher mean absolute error, given its transparent functional form and suitability for rapid scenario testing. This model allows for effective communication of policy outcomes to non-technical stakeholders and supports the evaluation of strategic interventions such as investment incentives or regulatory measures. Furthermore, integration of the hybrid model into digital HR management systems would enable real-time forecasting of labour demand, automatically updating projections as new data on production plans and productivity benchmarks become available each quarter. Such integration would transform traditional workforce planning into a dynamic, data-driven decision-support process, aligning operational management with long-term industrial development objectives.

### 3.7. Scenario analysis for decision support

To demonstrate the applicability of the model for strategic planning under different economic conditions, three critical scenarios were simulated using the hybrid model:".

Scenario 1: Production Shock (2026)

A sudden 25% drop in metallurgical production (simulating a commodity market collapse or major customer loss) while productivity remains constant. Model projection: Employment would fall from 10,399 to about 9,850 people (5.3% reduction), suggesting that production-related economies of scale dominate in the short run. This finding suggests that companies should maintain flexible employment contracts for about 550 workers to absorb such shocks without severe social disruption.

Scenario 2: Accelerated automation (2025-2028)

Implementation of smart manufacturing technologies increases productivity by 18% over the forecast horizon (from 8.28 to 9.77 tons/person), while production grows moderately at baseline rates. Model

projection: Employment stabilizes at 10,150-10,200 rather than growing to 10,460, indicating that productivity gains offset 250-300 jobs. This scenario underscores the need for proactive retraining programs to redeploy workers from automated production lines to maintenance, programming, and quality control functions.

Scenario 3: Policy-driven employment maintenance (2024-2026)

Government implements subsidies that require firms to maintain employment levels at 10,600+ despite productivity improvements. Model simulation indicates that this would require production expansion to 105,000+ tons (18% above baseline forecast), which can only be achieved through market expansion or export growth. The cost-benefit analysis suggests a subsidy requirement of approximately KZT 35-40 million per year per 100 jobs retained.

Figures from 10a to 12 Scenario analysis of labour demand under different economic conditions in industry, East Kazakhstan

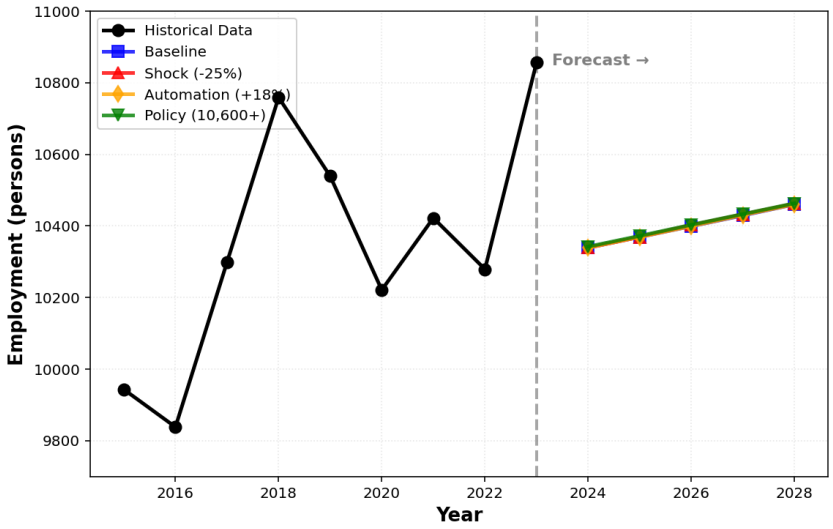


**Fig. 11a – Full time 2015-2028**

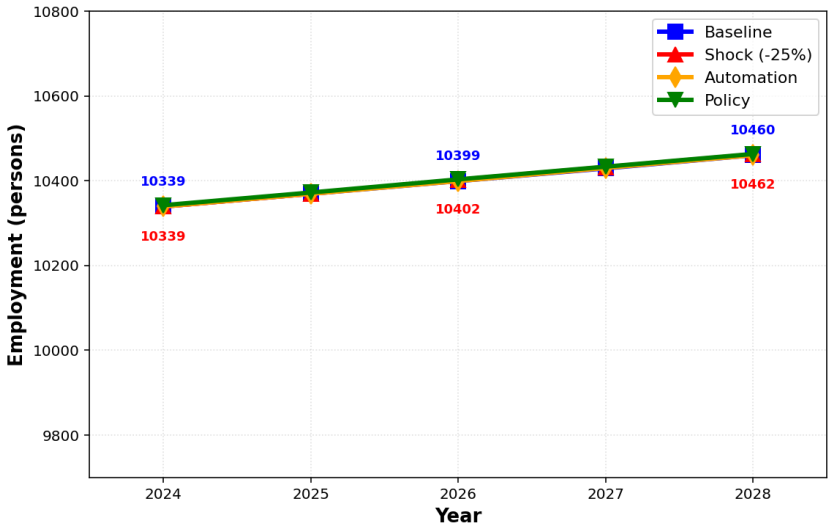


**Fig. 11b – detailed forecast view 2024-2028**

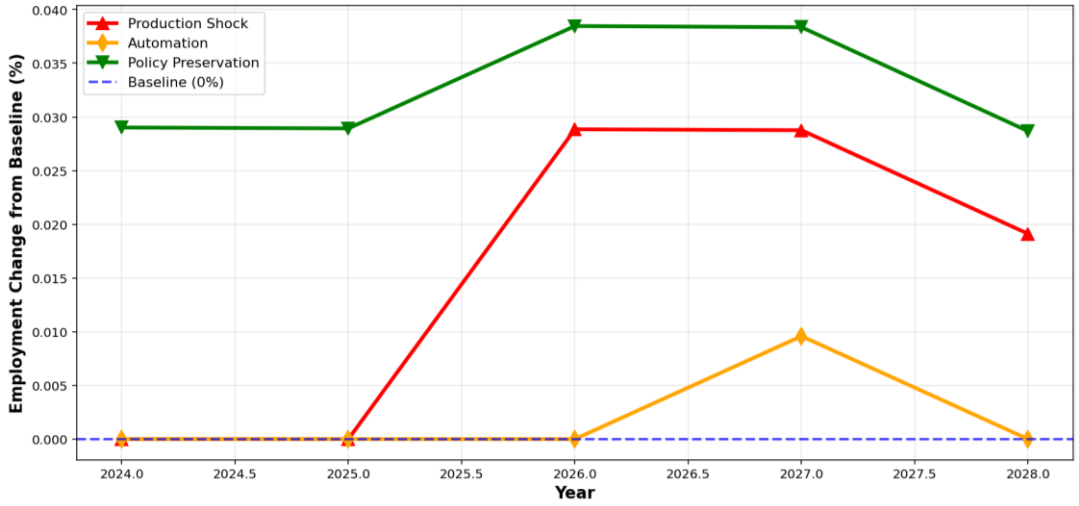


**Fig. 12 – scenario impact percentage deviation from baseline**

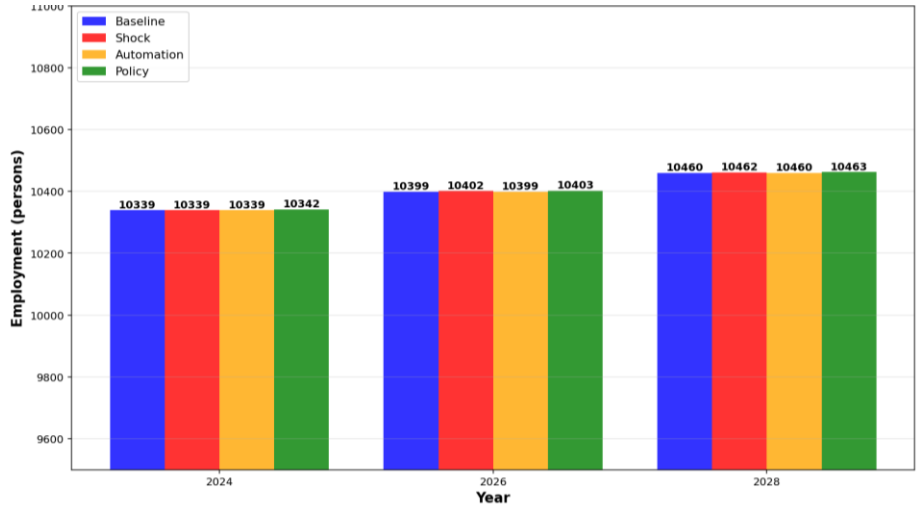


**Fig. 13 – comparative employment levels (key years)**

## 4. CONCLUSIONS

The present study was aimed at developing and testing models for forecasting labour demand in the metallurgical industry of the East Kazakhstan Region by combining classical econometric methods with machine learning algorithms. The proposed methodology not only enabled the assessment of the impact of production output and labour productivity on employment levels but also provided a practical tool for generating short- and medium-term forecasts.

The results of the analysis showed that employment is determined by two opposing factors:

1. An increase in production volume stimulates an increase in the number of people employed;
2. Increasing labour productivity reduces the demand for labour resources.

A comparison of different models demonstrated that the hybrid approach, combining a decision tree with local linear regression, delivers the best balance between predictive accuracy and scenario realism. The parametric model, in turn, is valuable for its interpretability and can be applied for strategic and scenario analysis, whereas linear regression confirmed its limitations in the context of nonlinear dependencies and technological shifts.

The practical significance of the findings lies in the fact that the developed models can be integrated into decision support systems at the level of industry ministries, regional authorities, and industrial enterprises. Applying them enables early assessment of the consequences of changes in production volumes and the pace of technological modernization for the labour market, thereby contributing to the development of a balanced workforce policy, the optimization of vocational training programs, and improved labour stability.

While the presented models demonstrate satisfactory performance and practical relevance, several contextual considerations should be noted to appropriately interpret the results and guide future research. The analysis is based on nine annual observations, which, although sufficient for the applied methodological framework, naturally constrains the exploration of higher-order dynamics and complex nonlinear relationships. Expanding the temporal scope in subsequent studies would allow for a more detailed assessment of long-term structural patterns and enhance the statistical robustness of the models.

The study focuses on the industry of East Kazakhstan, which provides a representative case for regional industrial analysis. Nonetheless, variations in technological intensity, labour structures, and productivity dynamics across other sectors - such as services or high-technology manufacturing - suggest that model recalibration and validation may be required for broader applicability. This specialization, however, enables a deeper understanding of sector-specific mechanisms that are often obscured in more generalized studies.

Furthermore, the modeling approach concentrates on two primary determinants- production volume and labour productivity - selected for their economic interpretability and data consistency. Although other factors such as capital investment, wage levels, or innovation activity may also influence employment dynamics, their inclusion in future research could further enrich the analytical framework and provide a more comprehensive understanding of labour demand formation.

Finally, the models assume temporal stability of the observed relationships over the forecast horizon (2024-2028), which is appropriate under moderate structural change. However, significant technological or policy shifts may alter these dynamics, offering a valuable opportunity for future studies to explore model adaptability under conditions of accelerated transformation. The current framework, developed at the regional-industry level, establishes a solid foundation upon which more granular analyses - such as firm-level or occupational models—can be built, deepening insights into workforce development and industrial modernization strategies.

Building on the methodological and empirical foundations established in this study, several avenues for future research may significantly expand both the analytical depth and practical relevance of the proposed modeling framework. One promising direction involves the integration of advanced hybrid deep learning architectures, such as Long Short-Term Memory (LSTM) networks and Transformer models, which are capable of capturing long-range temporal dependencies and seasonal fluctuations. Their application would be particularly valuable in the context of higher-frequency forecasting—using monthly or quarterly data—allowing for more granular and adaptive labour demand predictions.

A further extension entails broadening the analytical scope beyond the metallurgical sector to include other key industries such as energy, and services. Developing a unified multisectoral model of the regional labour market would enable researchers to account for inter-industry linkages, labour mobility, and spillover effects, thus providing a more comprehensive understanding of how industrial transformations influence regional employment dynamics.

Incorporating indicators of innovation, digitalization, and macroeconomic dynamics represents an important direction for future research. The current model focuses on production and productivity due to data availability, as official regional statistics in Kazakhstan do not systematically report R&D expenditures, technology adoption metrics, or digitalization indicators at the enterprise level. Future studies should aim to develop comprehensive datasets—ideally in collabouration with industry associations and enterprises—that capture innovation activity (R&D, patents), ICT capital investment, workforce skill composition, and exposure to external shocks such as commodity price fluctuations. Integrating such variables would allow models to explicitly evaluate how technological progress and market volatility reshape employment structures in the industrial sector.

Another prospective area involves the development of dynamic, real-time forecasting systems that leverage high-frequency data sources such as monthly industrial production indices, job vacancy postings, and online labour market analytics. By employing methods such as Kalman filtering or Bayesian updating, such systems could continuously refine predictions as new information becomes available, thereby improving their operational relevance for both policymakers and industry practitioners.

Future research could also benefit from decomposing aggregate employment projections into occupational subcategories—distinguishing between skilled and unskilled, as well as blue-collar and white-collar workers. This level of detail would provide critical input for designing evidence-based educational and training programs, aligning workforce development with the evolving needs of digital and industrial transformation.

Finally, an international comparative perspective would add valuable insight into the generalizability of findings. Replicating the proposed models across diverse national and regional contexts would make it possible to distinguish universal patterns of labour demand formation from context-specific factors, thereby facilitating

cross-country benchmarking and supporting knowledge transfer between emerging and developed economies. Collectively, these directions underscore the potential of combining machine learning with economic theory to develop more adaptive, data-driven, and policy-relevant frameworks for forecasting and managing labour market transformations in the era of digital industrialization.


## Funding

*This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24992854: "Development and Implementation of Competitive Science-Based Technologies to Ensure Sustainable Development of the Mining and Metallurgical Industry of the East Kazakhstan Region,".*


## Conflicts of Interest

*The authors declare no conflict of interest.*

## REFERENCES


Alzeraif, M., Cheaitou, A., & Bou Nassif, A. (2023). Predicting maintenance labour productivity in electricity industry using machine learning: A case study and evaluation. *International Journal of Advanced Computer Science and Applications*, *14*(7), 226–234. https://doi.org/10.14569/IJACSA.2023.0140758

Bali, S., Bali, V., Gaur, D., Rani, S., Kumar, R., Chadha, P., Sharma, Y., Prakash, C., Shahare, P., Khera, G. S., Kampani, S., Solopova, N., Dixit, S., & Vatin, N. I. (2023). A framework to assess the smartphone buying behaviour using DEMATEL method in the Indian context. *Ain Shams Engineering Journal*, 102129. https://doi.org/10.1016/j.asej.2023.102129

Ballestar, M. T., Camiña, E., Díaz-Chao, Á., & Torrent-Sellens, J. (2021). Productivity and employment effects of digital complementarities. *Journal of Innovation & Knowledge*, *6*(3), 177–190. https://doi.org/10.1016/j.jik.2020.10.006

Bril, A., Evseeva, S., Kalinina, O., Barykin, S., & Vinogradova, E. (2020). Personnel changes and labour productivity in regulatory budget monitoring. *IOP Conference Series: Materials Science and Engineering*, *940*, 012105. https://doi.org/10.1088/1757-899X/940/1/012105

Bureau of National Statistics, Agency for Strategic Planning and Reforms of the Republic of Kazakhstan. (2025a). *Official website*. Retrieved August 25, 2025, from https://stat.gov.kz/ru/

Bureau of National Statistics, Agency for Strategic Planning and Reforms of the Republic of Kazakhstan. (2025b). *Information-analytical system "Taldau"*. Retrieved August 25, 2025, from https://taldau.stat.gov.kz/ru/Search/SearchByKeyWord

Cruz, M. D. (2023). Labour productivity, real wages, and employment in OECD economies. *Structural Change and Economic Dynamics*, *66*, 367–382. https://doi.org/10.1016/j.strueco.2023.05.007

Dosi, G., Piva, M., Virgillito, M. E., & Vivarelli, M. (2019). Technology and employment in a vertically connected economy: A model and an empirical test. *DISCE Working Papers, Università Cattolica del Sacro Cuore.*

Ebrahimi, S., Goli, A., & Asadpour, M. (2021). Deep learning-based hybrid models for time series forecasting. *Algorithms*, *14*(7), 214. https://doi.org/10.3390/a14070214

EFSD. (2023). Machine learning algorithms for short-term forecasting of real GDP growth rates. *Eurasian Fund for Stabilization and Development, Working Paper.* https://efsd.org/en/research/working-papers/working-paper-machine-learning-algorithms-for-short-term-forecasting-of-real-gdp-growth-rates

Elshaboury, N. (2022). Training adaptive neuro fuzzy inference system using genetic algorithms for predicting labour productivity. In P. Liatsis, A. Hussain, S. A. Mostafa, & D. Al-Jumeily (Eds.), *Emerging technology trends in Internet of Things and computing* (pp. 311–322). Springer. https://doi.org/10.1007/978-3-030-97255-4_24

Falkenberg, S. F., & Spinler, S. (2022). Integrating operational and human factors to predict daily productivity of warehouse employees using extreme gradient boosting. *International Journal of Production Research*, *61*(24), 8654–8673. https://doi.org/10.1080/00207543.2022.2159563

Golabchi, H., & Hammad, A. (2024). Estimating labour resource requirements in construction projects using machine learning. *Construction Innovation*, *24*(4), 1048–1065. https://doi.org/10.1108/CI-11-2021-0211

Güvel, Ş. T. (2025). Forecasting slipform labour productivity in the construction of reinforced concrete chimneys. *Ain Shams Engineering Journal*, *16*(1), 103192. https://doi.org/10.1016/j.asej.2024.103192

Hatami, F., Pezeshk Poor, A., & Thill, J.-C. (2024). Non-business services performance forecasting for small urban areas using a spatiotemporal deep learning model. *Cities*, *152*, 105141. https://doi.org/10.1016/j.cities.2024.105141

Jacobsen, E. L., Teizer, J., Wandahl, S., & Brilakis, I. (2024). Probabilistic forecasting of construction labour productivity metrics. *ITcon*, *29*, 58–83. https://doi.org/10.36680/j.itcon.2024.004

Li, C. K., Luo, J., & Soderstrom, N. S. (2020). Air pollution and analyst information production. *Journal of Corporate Finance*, *60*, 101536. https://doi.org/10.1016/j.jcorpfin.2019.101536

Magazzino, C., Mele, M., & Mutascu, M. (2025). An artificial neural network experiment on the prediction of the unemployment rate. *Journal of Policy Modeling*, *47*(3), 471–491. https://doi.org/10.1016/j.jpolmod.2024.10.004

Mahamid, I. (2020). Study of relationship between rework and labour productivity in building construction projects. *Revista de la Construcción*, *19*(1), 30–41. https://doi.org/10.7764/RDLC.19.1.30-41

Mutascu, M., & Hegerty, S. W. (2023). Predicting the contribution of artificial intelligence to unemployment rates: An artificial neural network approach. *Journal of Economics and Finance*, *47*, 400–416. https://doi.org/10.1007/s12197-023-09616-z

Orlova, E. (2023). Personnel changes and labour productivity in regulatory budget monitoring. *Mathematics*, *11*(4), 863. https://doi.org/10.3390/math11040863

Popescu, A., Tindeche, C., Marcuța, A., Marcuța, L., & Hontus, A. (2021). Labour productivity in Romania's agriculture in the period 2011–2020 and its forecast for 2021–2025 horizon. *Management, Economic Engineering in Agriculture and Rural Development*, *21*(3), 673–678.

Potapov, A. P. (2020). Modeling the impact of resource factors on agricultural output. *Economic and Social Changes: Facts, Trends, Forecast*, *13*(4), 154–168. https://doi.org/10.15838/esc.2020.4.70.9

Ramezanian, R., & Hajipour, M. (2020). Integrated framework of system dynamics and meta-heuristic for multi-objective land use planning problem. *Landscape Ecology and Engineering*, *16*(1), 113–133. https://doi.org/10.1007/s11355-020-00410-1

Uppal, A., Awasthi, Y., & Srivastava, A. (2024). Machine learning-based approaches for enhancing human resource management using automated employee performance prediction systems. *International Journal of Organizational Analysis*. https://doi.org/10.1108/IJOA-07-2024-4643

Woltjer, G., van Galen, M., & Logatcheva, K. (2019). Industrial innovation, labour productivity, sales and employment. *International Journal of the Economics of Business*, *28*(1), 89–113. https://doi.org/10.1080/13571516.2019.1695448

Zhang, J., Malikov, E., & Miao, R. (2024). Distributional effects of the increasing heat incidence on labour productivity. *Journal of Environmental Economics and Management*, *125,* 102998. https://doi.org/10.1016/j.jeem.2024.102998

Zhang, Y., Chen, Y., Su, Q., Huang, X., Li, Q., Yang, Y., Zhang, Z., Chen, J., Xiao, Z., Xu, R., Zu, Q., Du, S., Zheng, W., Ye, W., & Xiang, J. (2024). The use of machine and deep learning to model the relationship between discomfort temperature and labour productivity loss among petrochemical workers. *BMC Public Health*, *24*, 3269. https://doi.org/10.1186/s12889-024-20713-4

Zhu, M., Liang, C., Yeung, A. C. L., & Zhou, H. (2024). The impact of intelligent manufacturing on labour productivity: An empirical analysis of Chinese listed manufacturing companies. *International Journal of Production Economics*, *267*, 109070. https://doi.org/10.1016/j.ijpe.2023.109070