UDC 004.942:622.7
DOI https://doi.org/10.32782/2663-5941/2025.2.2/07

Volovetskyi O. O.
Kryvyi Rih National University

DEVELOPMENT AND RESEARCH OF A SYNTHETIC DATA GENERATOR FOR MODELING TECHNOLOGICAL PARAMETERS OF THE IRON ORE BENEFICIATION PROCESS

The paper presents the results of developing a specialized synthetic data generator for modeling technological parameters of the iron ore beneficiation process. The developed generator implements a modular approach that includes parameter configuration, basic signal generation, and controlled noise addition. Methods for generating basic signals are proposed considering the specifics of various technological parameters through the use of normal distribution for parameters with natural grouping tendencies and uniform distribution for parameters with uniform change patterns. The noise addition system provides four intensity modes with individual settings for each parameter. The experimental study included 486 *experiments with different training sample configurations. Analysis of generation quality showed that SNR values are in the range of 25–50 dB, which corresponds to the characteristics of real industrial measurements. A critical dependence of prediction quality on training sample size was established: with sample sizes over* 4096 *records, consistently high prediction quality is achieved* (R^2 = 90–95 %). *The developed generator enables the creation of synthetic datasets with controlled noise characteristics and statistical properties that correspond to real technological parameters. Validation mechanisms for generated data are implemented through comprehensive analysis of statistical characteristics, spectral analysis, and evaluation of correlation dependencies. The results confirmed the effectiveness of the developed generator for creating training datasets in the development and testing of control systems for iron ore beneficiation processes, which significantly reduces time and resources during the control system development phase.*

Key words: *synthetic data generation, iron ore beneficiation, controlled noise, technological parameters, machine learning.*

Problem statement. In modern industrial production, the issue of measurement reliability for technological parameters is extremely relevant as it directly affects product quality and economic efficiency of enterprises. This is especially true for iron ore beneficiation processes, where measurement accuracy determines the quality of the final product. The ore beneficiation process is characterized by complexity, multiple factors, and high requirements for technological parameter measurements, making it relevant to create specialized software tools for modeling these processes.

The development of process control systems for beneficiation requires significant amounts of data for training and testing algorithms. Collecting such data in real production conditions is resource-intensive and often limited, necessitating the use of synthetic data. Existing approaches to synthetic data generation do not take into account the specifics of iron ore beneficiation, which creates a need for developing specialized generators with controlled noise characteristics.

Analysis of recent research and publications. The problems of operational quality control of iron ore and optimization of the enrichment process are discussed in the work of Toporov et al. [1], where methods for improving the efficiency of technological operations through enhancement of measurement systems at the ore preparation input are proposed. These studies emphasize the importance of accurate and reliable measurements of technological parameters to ensure final product quality.

Theoretical aspects of synthetic data generation for modeling technological processes are actively researched in various fields [2, 3]. Specifically, Dankar and Ibrahim [2] proposed comprehensive recommendations for effective synthetic data generation that ensures maximum correspondence to real conditions. In [3], Pei and Zaïane developed a synthetic data generator for clustering analysis and outlier detection, allowing modeling of various distribution types.

Methods of synthetic data generation for the energy sector are presented in the work of Alnumay et al. [4], where an approach to controlled noise addition is proposed, simulating fluctuations of technological parameters based on normal distribution. Murray-Smith and Girard [5] thoroughly examined the application of Gaussian processes with ARMA models to improve prediction accuracy with minimal computational costs, considering the correlation structure of noise.

Mannino and Abouzied [6] proposed methods for generating synthetic data that are visually indistinguishable from real data, which is an important aspect for developing technological process visualization systems. Paper [7] presents an overview of machine learning methods for synthetic data generation with analysis of their advantages and limitations.

Kothare et al. [8] developed the SynGen system for synthetic data generation with parameter adjustment and quality control capabilities.

Analysis of existing research indicates a lack of specialized tools for modeling iron ore enrichment process parameters considering their interconnections and noise characteristics [6–8]. Specifically, methods for controlled noise addition considering normal distribution and different noise levels for various technological parameters of the enrichment process have not been developed.

The aim of the article is to develop a specialized software generator of synthetic data for modeling technological parameters of the iron ore enrichment process. The proposed system aims to create datasets with controlled noise levels that correspond to real production conditions. This will reduce data collection costs and accelerate control system development.

To achieve this goal, the following tasks must be solved:

1. Develop a synthetic data generator architecture with modular structure;
2. Propose methods for generating basic signals for various technological parameters;
3. Develop a controlled noise addition system;
4. Conduct experimental research on generated data quality;
5. Determine the dependency of prediction quality on training sample size.

Presentation of the main research material. The developed data generator is implemented as a NoisedDataGenerator class, which provides a comprehensive approach to creating synthetic technological data with controlled noise levels. The generator's architecture is based on a modular principle and includes four main components: initialization, data handling, signal generation, and noise processing (Fig. 1).

During the initialization phase, detailed configuration of the generator's basic parameters and error level settings for each technological parameter are established separately. This stage is critically important, as the quality of all subsequent operations depends on the correct setup of initial parameters. Data handling includes a comprehensive process of loading the source dataset and careful determination of allowable parameter value ranges, ensuring generated data corresponds to real technological constraints [9]. This stage also includes input data validation and verification of compliance with established criteria.

Signal generation is implemented through creating a sequence of control points and their subsequent interpolation to obtain continuous smooth signals that closely approximate real technological parameters. The final stage is comprehensive noise processing with detailed value verification and correction to ensure data reliability.

Base signal generation is carried out through a carefully selected combination of methods with different types of statistical distributions, considering the specifics of each technological parameter. For parameters that naturally demonstrate a tendency to cluster around the mean value, such as iron content in concentrate and tailings, normal distribution is applied, most accurately reflecting their statistical

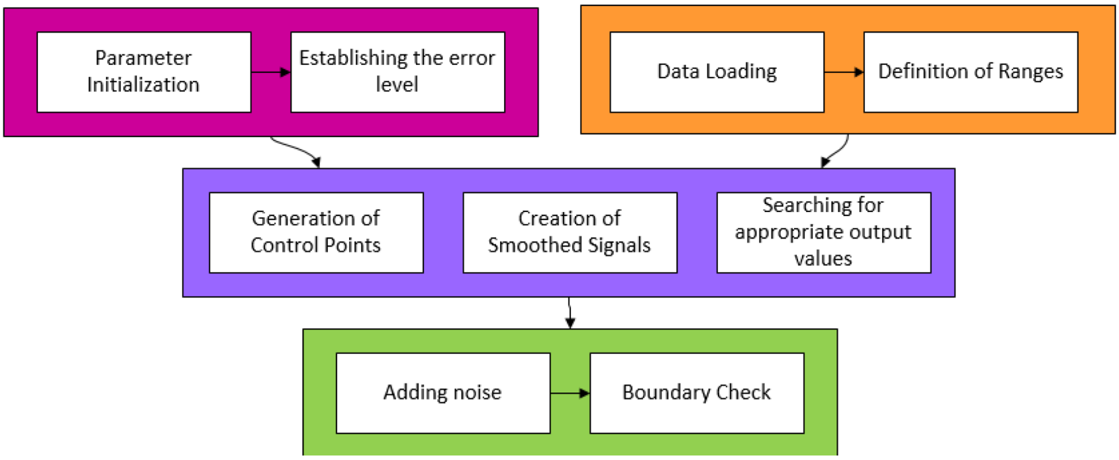
Parameter Initialization
Establishing the error level
Data Loading
Definition of Ranges
Generation of Control Points
Creation of Smoothed Signals
Searching for appropriate output values
Adding noise
Boundary Check

Fig. 1. Structural diagram of a synthetic technological data generator with controlled noise level

nature [10]. Parameters characterized by uniform changes throughout the technological process, such as ore consumption and conveyor speed, are generated using uniform distribution, which best matches their behavior in real conditions. Continuity and smoothness of all signal changes are ensured by applying cubic spline interpolation, allowing for realistic enrichment process dynamics and avoiding sharp transitions between values [11].

The noise addition process is implemented as a multi-stage procedure, considering specific features of each technological parameter and its permissible deviations. The system provides four different noise intensity modes (null, min, aver, max), enabling modeling of various production conditions and equipment operation scenarios. The noise component for each parameter is generated using normal distribution, with its statistical characteristics calculated based on the specified error percentage and specific parameter characteristics. A particularly important aspect is continuous monitoring of allowable value limits and comprehensive validation of results through calculating various statistical characteristics, ensuring generated data corresponds to real technological constraints.

To ensure high quality of generated data, comprehensive analysis of each technological parameter was conducted, with results presented in Fig. 2. Detailed analysis includes multifactor comparison of clean and noisy signals, in-depth study of noise distribution patterns, thorough analysis of signal power spectrum, and comprehensive evaluation of correlations between different signals. The obtained results convincingly demonstrate that generated data fully preserves all main characteristics of the real technological process, while added noise has clearly controlled statistical properties that accurately match real production environment conditions and measurement system characteristics.

For comprehensive evaluation of iron ore enrichment process stability and thorough analysis of experimental data quality, detailed research was conducted using four key metrics, each characterizing an important process aspect. Signal-to-noise ratio (SNR_db) is a fundamental metric used for quantitative assessment of data noise levels, with higher values clearly indicating better signal quality and fewer noise components. Correlation analysis enables deep investigation and evaluation of relationships between different process parameters, critical for understanding overall system dynamics. Relative deviation indicator was chosen as a universal tool enabling correct comparison of various technological parameters' variability regardless of their absolute values and measurement scale. The fourth metric – outliers percentage – is an important indicator clearly showing the frequency of anomalous value occurrence in measurements, allowing comprehensive characterization of overall process stability and measurement system quality.

To evaluate generated data quality, comprehensive analysis was conducted under different noise intensity modes (min, aver, max). Results of detailed analysis of main technological parameters at three data noise levels – minimum (Min exp), average (Average exp), and maximum (Max exp) – are systematized and presented in Table 1. This structured presentation of results allows comparative analysis of different noise levels' impact on data quality and process stability, as well as evaluation of measurement reliability under various experimental conditions.

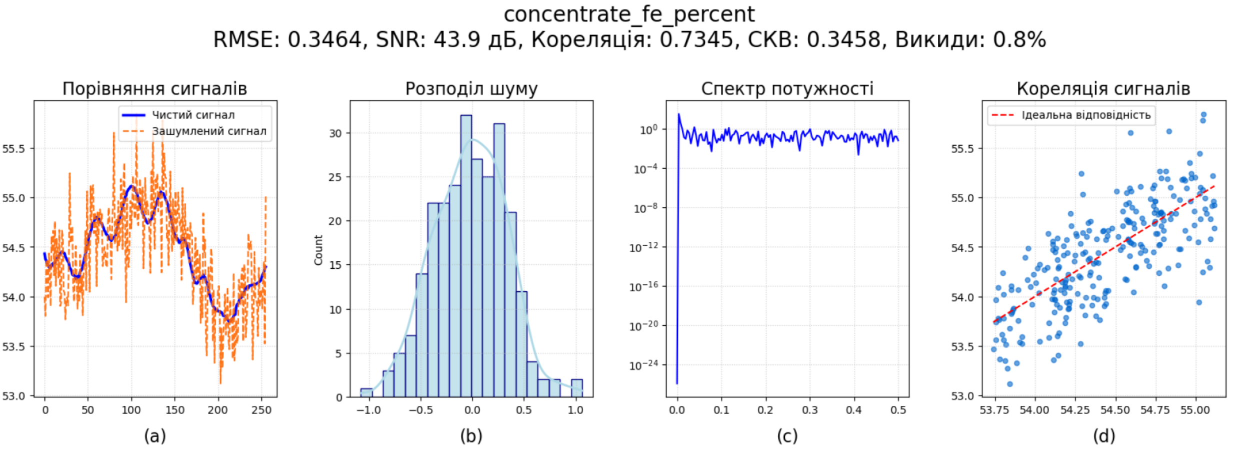

Fig. 2. Complex analysis of generated data quality for iron content parameter in concentrate: (a) comparison of clean and noisy signals; (b) noise distribution histogram; (c) signal power spectrum; (d) correlation relationship between signals

Table 1

Process Stability Metrics at Different Noise Levels

Metric	Min exp	Average exp	Max exp	Difference (Max-Min)
feed_fe_percent				
snr_db	46.0077	42.5046	39.9858	–6.0219
correlation	0.9105	0.8518	0.7650	–0.1456
relative_deviation	0.0009	0.0002	–0.0011	–0.0020
outliers_percent	0.2618	0.2946	0.2584	–0.0034
solid_feed_percent				
snr_db	39.9965	36.4947	34.0018	–5.9948
correlation	0.9894	0.9758	0.9565	–0.0329
relative_deviation	–0.0019	–0.0057	–0.0042	–0.0022
outliers_percent	0.3187	0.3140	0.3022	–0.0164
ore_mass_flow				
snr_db	39.9903	32.0362	27.9772	–12.0131
correlation	0.0031	–0.0003	0.0010	–0.0020
relative_deviation	–0.0066	0.0106	–0.0010	0.0057
outliers_percent	0.2655	0.2579	0.2694	0.0039
concentrate_fe_percent				
snr_db	50.4553	46.0438	43.0792	–7.3761
correlation	0.9252	0.8482	0.7493	–0.1758
relative_deviation	–0.0013	–0.0005	0.0008	0.0021
outliers_percent	0.2801	0.2893	0.2737	–0.0064
concentrate_mass_flow				
snr_db	34.0069	29.1430	26.0346	–7.9723
correlation	0.5692	0.3773	0.2784	–0.2908
relative_deviation	0.0039	–0.0037	–0.0149	–0.0188
outliers_percent	0.2667	0.2453	0.2602	–0.0066
tailings_fe_percent				
snr_db	47.9872	43.7661	40.9070	–7.0803
correlation	0.3099	0.2102	0.1454	–0.1645
relative_deviation	0.0004	–0.0020	–0.0007	–0.0011
outliers_percent	0.2854	0.2632	0.2826	–0.0028
tailings_mass_flow				
snr_db	32.0210	27.9305	25.1859	–6.8351
correlation	0.5815	0.4201	0.3170	–0.2645
relative_deviation	–0.0072	–0.0022	0.0059	0.0131
outliers_percent	0.2798	0.2677	0.2765	–0.0033

Analysis of results presented in Table 1 shows SNR values range from 25–50 dB, corresponding to real industrial measurement characteristics. Iron content parameters (feed_fe_percent, concentrate_fe_percent, tailings_fe_percent) demonstrate best SNR (40–50 dB) with specified errors of 0.3–1.0 %, while mass flow rates are characterized by lower SNR (25–34 dB) with errors of 2.0–5.5 %. Correlation analysis confirms expected dependence of signal quality on specified errors – parameters with small errors have high correlation (> 0.75), while mass flow rates show lower correlation. Relative deviations of all parameters remain close to zero (± 0.015), indicating absence of systematic error.

During experimental research, 486 experiments with different training sample configurations were conducted. Statistical analysis of results showed critical dependence of prediction quality on training sample size. With minimum sample size (256 records), complete degradation of prediction quality is observed with R^2 close to zero. Increasing sample size to 2048–4096 records shows gradual improvement but with significant prediction quality variability (R^2 fluctuates between 65–85 %). When sample size exceeds 4096 records, consistently high prediction quality is achieved with $R^2 = 90–95$ %, with further data volume increase not leading to significant improvement in results.

Experiment results also confirmed high prediction quality for multilayer perceptron models and support vector regression method. Importantly, result stability is maintained regardless of data noise level, confirming reliability of the developed generator.

Conclusions. This paper presents the results of developing and researching a synthetic technological data generator for the iron ore beneficiation process. The developed generator enables creation of datasets with controlled noise levels, implementing a modular approach that includes parameter configuration, base signal generation, and controlled noise addition.

The proposed methods for generating base signals account for the specifics of various technological parameters through the use of normal distribution for parameters with natural grouping tendencies and uniform distribution for parameters with uniform change patterns. The application of cubic spline interpolation ensures the necessary signal smoothness that corresponds to the real process dynamics.

The developed noise addition system provides four intensity modes with individual settings for each parameter. Experimental study of the generated data quality revealed significant differences in the behavior of various technological parameters as noise levels increase. Mass flow parameters proved most sensitive to noise, showing the largest SNR drop (up to 12 dB) and lowest correlation indicators. Iron content parameters demonstrate greater noise resistance, confirmed by lower SNR reduction (6–7 dB) and higher correlation values. The stability of relative deviation indicators and outlier percentages across all parameters confirms the effectiveness of the developed generator in creating realistic datasets with controlled noise levels.

The practical value of the developed generator is confirmed by its potential use in creating training datasets for developing and testing control systems for iron ore beneficiation processes.

Further research should focus on expanding the generator's functionality to model abnormal situations, implement anomaly generation methods, and adapt to other types of technological processes. Another important direction is developing methods for automatic adjustment of generation parameters based on real data analysis.

Bibliography:

1. Toporov, V., Axelrod, V., Tukeyev, U. Operative control of ore quality at the input of ore preparation operations of enrichment fabric. *Journal of Information and Telecommunication*, 2019, vol. 3, no. 4, pp. 465–479. DOI: 10.1080/24751839.2019.1630792

2. Dankar, F.K., Ibrahim, M. Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation. *Applied Sciences*, 2021, vol. 11, no. 5, pp. 1–18. DOI: 10.3390/app11052158

3. Pei, Y., Zaïane, O. A synthetic data generator for clustering and outlier analysis. DOI: 10.7939/R3B23S

4. Alnumay, Y., et al. Synthetic Data Generation for Machine Learning Applications in the Energy Industry. *In Proc. SPE ADIPEC 2022*, Abu Dhabi, UAE, Oct. 2022, pp. 1–8. DOI: 10.2118/211821-MS

5. Murray-Smith, R., Girard, A. Gaussian Process Priors with ARMA Noise Models. *In Proc. Irish Signals and Systems Conference*, Glasgow, UK, 2001, pp. 1–6. URL: https://www.dcs.gla.ac.uk/~rod/publications/MurGir01.pdf

6. Mannino, M., Abouzied, A. Is this real? Generating synthetic data that looks real. *In Proc. 32nd Annual ACM Symposium on User Interface Software and Technology (UIST)*, Oct. 2019, pp. 549–561. DOI: 10.1145/3332165.3347866

7. Lu, Y., et al. Machine Learning for Synthetic Data Generation: A Review. *arXiv preprint arXiv:2302.04062*, Feb. 2023. DOI: 10.48550/arXiv.2302.04062

8. Kothare, A., Chaube, S., Moharir, Y., Bajodia, G., Dongre, S. SynGen: Synthetic Data Generation. *In Proc. Int. Conf. Computational Intelligence and Computing Applications (ICCICA)*, Nagpur, India, Nov. 2021, pp. 53–57. DOI: 10.1109/ICCICA52458.2021.9697232

9. King, R.P. Modeling and Simulation of Mineral Processing Systems. Oxford, UK: Butterworth-Heinemann, 2001. 399 p. DOI: 10.1016/C2009-0-26303-3

10. Maeland, E. On the Comparison of Interpolation Methods. *IEEE Trans. Medical Imaging*, 1988, vol. 7, no. 3, pp. 213–217. DOI: 10.1109/42.7784

Воловецький О.О. РОЗРОБКА ТА ДОСЛІДЖЕННЯ ГЕНЕРАТОРА СИНТЕТИЧНИХ ДАНИХ ДЛЯ МОДЕЛЮВАННЯ ТЕХНОЛОГІЧНИХ ПАРАМЕТРІВ ПРОЦЕСУ ЗБАГАЧЕННЯ ЗАЛІЗНОЇ РУДИ

У роботі представлено результати розробки спеціалізованого програмного генератора синтетичних даних для моделювання технологічних параметрів процесу збагачення залізної руди. Розроблений генератор реалізує модульний підхід, що включає конфігурацію параметрів, генерацію базових сигналів та контрольоване додавання шуму. Запропоновано методи генерації базових сигналів *з урахуванням специфіки різних технологічних параметрів через використання нормального розподілу для параметрів з природною тенденцією до групування та рівномірного розподілу для параметрів з рівномірним характером змін. Система додавання шуму забезпечує чотири режими інтенсивності з індивідуальними налаштуваннями для кожного параметра. Експериментальне дослідження включало 486 експериментів з різними конфігураціями навчальних вибірок. Аналіз якості генерації показав, що значення SNR знаходяться в діапазоні 25–50 дБ, що відповідає характеристикам реальних промислових вимірювань. Встановлено критичну залежність якості прогнозування від розміру навчальної вибірки: при розмірі вибірки понад 4096 записів досягається стабільно висока якість прогнозування (R^2 = = 90–95 %). Розроблений генератор забезпечує можливість створення синтетичних наборів даних з контрольованими характеристиками шуму та статистичними властивостями, що відповідають реальним технологічним параметрам. Реалізовано механізми валідації згенерованих даних через комплексний аналіз статистичних характеристик, спектральний аналіз та оцінку кореляційних залежностей. Результати підтвердили ефективність розробленого генератора для створення навчальних наборів даних при розробці та тестуванні систем керування технологічними процесами збагачення залізної руди, що дозволяє значно скоротити час та ресурси на етапі розробки систем керування.*

Ключові слова: *генерація синтетичних даних, збагачення залізної руди, контрольований шум, технологічні параметри, машинне навчання.*