International Journal on Advanced Science Engineering Information Technology

Vol.14 (2024) No. 3
ISSN: 2088-5334

Development of a Python Library to Generate Synthetic Datasets for Artificial Intelligence Education

Seul-ki Kim [a], Yong-ju Jeon [b,*]

[a] Department of Computer Education, Korea National Unversity of Education, Cheogju, Chungbuk, 28173, Republic of Korea
[b] Department of Computer Education, Andong National University, Gyeongdongro 1375, Andong, GyeongBuk, 36723, Republic of Korea
Corresponding author: [*]yyongju@anu.ac.kr

***Abstract*—This study aims to improve the quality of AI education for the AI era by developing an educational dataset library and exploring its applicability. Reflecting the needs of teachers engaged in AI educational activities, the dataset library emphasizes the diversity of topics, forms, and sizes of datasets provided. Additionally, it is designed with a feature to generate outliers and missing values suitable for students' accessibility and educational purposes. The library developed in this research is based on Python and employs the random forest modeling method to generate high-quality synthetic datasets. The functionality and suitability of this library for AI education have been evaluated by an expert panel, which has endorsed its applicability in the field. In detailed assessments of the synthetic datasets generated, the library demonstrated its capability to accurately mirror the statistical characteristics of original datasets, achieving high levels of accuracy and cosine similarity in the modeling results. These outcomes confirm the library's efficacy in reconstructing educational datasets specifically for AI education purposes and crafting high-quality synthetic datasets. This approach offers a practical solution to the existing shortage of educational datasets and substantially enhances the overall quality of education. This research proves immensely beneficial for educators and learners, laying a foundation for ongoing and future research focused on creating and utilizing educational datasets in AI. This paves the way for expanding the possibilities and scope of their application in the educational field, potentially transforming AI education practices.**

***Keywords*— Artificial intelligence education; educational datasets; synthetic data; Python library**

Manuscript received 11 Oct. 2023; revised 13 Feb. 2024; accepted 13 Mar. 2024. Date of publication 30 Jun. 2024.

I. INTRODUCTION

The world is undergoing significant changes, often called the 'Artificial Intelligence (AI) era'. Definitions of AI vary among researchers. McCarthy [1] answered some fundamental questions regarding AI in his study "What is Artificial Intelligence?", where he redefined it as the science and engineering of making intelligent machines and brilliant computer programs. He also noted that not all AI researchers agree with this definition. Just as the definition of AI is diverse, so are the definitions of AI education, which are characterized from various perspectives by different researchers. AI education is typically categorized into two perspectives: one that utilizes AI technologies and tools to improve the education and learning process and another that sees AI education as a research topic aimed at providing knowledge and understanding of AI [2]. AI education as a research topic is mainly discussed in terms of enhancing AI literacy, which commonly emphasizes education about understanding AI, using and applying AI, creating and evaluating AI, and AI ethics [3]–[8].

Research on AI education, along with the definition of AI education itself, has focused on effectively enhancing students' AI literacy. The primary methods suggested involving programming exercises and project-based education aimed at developing intelligent systems [9], [10], [11]. Furthermore, a shift in thinking from traditional computer education to a new perspective is deemed necessary, especially highlighting the differences between rule-based programming and the need for low-level educational applications suitable for students' levels. Discussions have also emphasized the distinctions in programming and debugging based on data [12].

The importance of datasets, a key component of AI, has been emphasized in developing high-performance AI technologies and the responsible use of such technologies, leading to dataset-centric research for AI configuration [13]. In AI education, it's essential to present datasets to assist

students in grasping AI principles. Related research has asserted the importance of providing experiences in developing final products by resizing datasets to match students' levels or by sequentially increasing complexity [14]–[16].

Using datasets from real-life data in AI education can provide students with meaningful problem-solving experiences by offering lessons closely related to their life context [17]–[19]. Research focused on developing purpose-driven datasets for AI education based on context has highlighted the importance of considering students' preferences for context, data types, and programming tools, leading to the creation of datasets that reiterate the significance of real-world datasets [20]. Despite these advantages and importance, educators face challenges in utilizing datasets due to difficulties in data exploration, cleansing and restructuring, and evaluating the validity of the data [21]. The AI industry utilizes synthetic dataset creation methods to secure and handle datasets, solving issues related to personal information in datasets and filling in missing data to increase reliability [22].

Synthetic datasets were not actual data but were generated from real data with statistical properties like the original datasets. Synthetic datasets could be categorized into partial synthetic datasets, which partially replace the real data; complete synthetic datasets, which completely replace the real data; and hybrid synthetic datasets, which use both methods [22]. If such methods of synthetic dataset creation used in the industry are applied to reconstruct datasets generated around students' daily lives, including at schools, it could solve the problem of dataset shortage and provide high-quality educational datasets for use in teaching and learning [23]. In this context, this study aims to support AI education for enhancing AI literacy by developing an educational dataset library using synthetic dataset creation methods and assessing its applicability.

II. The Materials and Method

A. Analysis of Needs

It is necessary to analyze teachers' needs to derive the essential features of an educational dataset-generating library for AI education and enhance its suitability for the teaching and learning environment. For this, the current study employed Rossett's Needs Analysis Model, a widely used model in corporate training. Rossett's model is renowned for its focus on the implementation process of needs analysis and for offering an easily applicable procedure [24].

Following Rossett's Needs Analysis Model procedure, this study conducted a step-by-step process to identify problems with AI educational datasets, determine the purpose of the needs analysis to resolve these issues, and make decisions for problem-solving, as shown in Fig. 1.

Fig. 1. Rossett's Needs Analysis Model

Firstly, the purpose of the needs analysis is determined. The needs analysis in this study aims to identify the elements and features the educational dataset library should possess for AI education. Hence, the objective of the needs analysis is set as 'What are the essential elements and features that an educational dataset library must have?' Secondly, the situations causing the issues are identified, and sources with information capable of achieving the purpose of the needs analysis are confirmed. In this study, teachers are designated as the source of needs analysis. They provide information regarding the construction of the teaching and learning environment, methods of dataset provision, and the functions needed when processing datasets for educational purposes from the standpoint of offering datasets to learners and conducting classes within the teaching and learning environment. Thirdly, a tool is selected to collect information for the needs analysis. Due to the need for systematic preliminary studies related to datasets for AI education, the tool for the needs analysis was a questionnaire developed and utilized by the researcher, extracting the necessary elements and features of datasets in the teaching and learning environment. The constructed questionnaire is finalized after being reviewed and modified by one computer education professor and three doctoral-level experts. Fourthly, information is collected using the previously developed tool to achieve the purpose of needs analysis. The completed survey items are distributed to 24 teachers (11 elementary teachers, 13 secondary teachers) active in AI education in the field. Fifthly, the collected information is analyzed, and decisions are made. Based on the survey results, the essential elements and features of the AI educational dataset library should have been derived.

B. Synthetic Dataset Generation and Reconstruction

The method for generating synthetic datasets utilizes the Sequential Regression Multiple Imputation (SRMI) approach, an extension of replacement techniques for filling in missing values using information from the original dataset [25]. This method involves sampling from distributions appropriate to each variable's characteristics. It allows the assumption of various models considering the relationships among variables, making it widely used as a fundamental concept in synthetic dataset generation. Additionally, the SRMI approach can synthesize data through various machine learning methods and learn from the original dataset relatively quickly to generate high-quality synthetic datasets [26].

When generating synthetic data, if modeling the necessary posterior predictive distributions of the original dataset is statistically challenging or the computations are complicated due to many variables, one may use a nonparametric approach or Monte Carlo techniques to approximate the distributions. Since synthetic data use only the generated values—not the statistical inferences of precise models—it is possible to employ machine learning methods to derive the most approximate values [26].

Approximate methods through machine learning have proposed nonparametric approaches using Decision Trees, and more recently, methods employing Random Forest have

also been utilized [27], [28]. Decision Trees are one of the modeling algorithms that can handle both categorical and continuous data, offering the advantage of straightforward interpretation of partitioning results and scalability to accommodate large datasets. While the characteristics of tree algorithms can lead to overfitting or sensitivity in the tree structure as drawbacks, these can be mitigated through techniques such as pruning. Thus, decision trees are extensively used in R packages for generating synthetic datasets and are known to produce stable results in various studies, serving as the base algorithm for the widely used Synthpop package [29], [30].

This study explored various Python packages implementing Synthpop, an R package for generating synthetic datasets. Among these, we identified the synthpop library by Hazy as the most closely aligned with our objectives [31]. The aim of this study was to refine this library to develop a synthetic dataset generation library specifically for AI educational purposes.

C. Selection of Example Datasets

For the experiment of generating synthetic datasets, datasets from the Scikit-learn library, which offers a variety of features related to machine learning and where the relationship between independent and dependent variables is presented, were selected. These datasets are frequently used in AI education and research. Additionally, to carry out experiments with synthetic datasets of various sizes, 30 rows without any missing values across all columns were randomly sampled based on the dependent variable to standardize the number of rows, as shown in Table 1 [32].

TABLE I
INFORMATION OF THE EXAMPLE DATASET FOR THE EXPERIMENT

Example Dataset	Size (Rows*Columns)	Modeling Method	Variable (Dependent Variable)
Iris	30*5	multi-classification	Sepal length, Sepal Width 3 other (Species)
Diabetes	30*8	binary classification	Pregnancies, Glucose 7 other (Outcomes)
Boston Housing	30*13	regression	CRIM, ZN 12 other (MEDV)

The Iris dataset is suitable for multi-classification, where the task is to categorize into one of four species based on various Iris flower characteristics. The Diabetes dataset is apt for binary classification, classifying whether individuals from the Pima Indian heritage have an onset of diabetes within five years based on various diabetes onset factors (Outcomes). The Boston Housing dataset is well-suited for regression analysis, predicting housing prices (MEDV) based on multiple factors influencing prices in the 1970s. Although the distribution of the Boston Housing dataset was officially banned due to ethical concerns, the purpose of this experiment is to generate synthetic datasets, and the meaning of each variable in the dataset is not the subject of research, so ethical issues were not considered [32], [33].

D. Library Suitability Verification

The synthetic dataset generation library developed for AI education in this study has been verified. A group of AI education experts was assembled to review the library's main features and validate its suitability for the AI teaching and learning environment. The questionnaire was designed to evaluate the suitability of the library's main features for teaching and learning activities on a 5-point scale ranging from 'not suitable at all' to 'very suitable.' Additionally, respondents were allowed to provide reasons for their choices if they selected 'not suitable at all' or 'not suitable.' Details of the expert group are presented in Table 2.

TABLE II
EXPERT GROUPS FOR SUITABILITY REVIEWS

Category		Number of Experts (Ratio)	Sum
Target	Elementary School	4(28.6)	14
	Middle School	4(28.6)	
	High School	6(42.8)	
Years of teaching experience	10 ~ 14 years	6(42.8)	14
	15 ~ 19 years	4(28.6)	
	20 years above	4(28.6)	
major	Computer science	2(14.3)	14
	Computer education	12(85.7)	
Final degree	Master	11(78.6)	14
	Doctor	3(21.4)	

To quantitatively evaluate the results of the expert review, responses such as 'not suitable at all' were encoded as 1 and 'very suitable' as 5. The analysis includes the calculation of mean, standard deviation, the number of experts in agreement, and the Content Validity Ratio (CVR) [34]. Secondly, the library developed through this research was used to generate synthetic datasets, and the quality of these datasets was assessed. The evaluation was divided into two experiments simulating a teaching and learning situation: one using a complete synthetic dataset with the same number of rows (30) and another using a complete synthetic dataset with a more significant number of rows (500) created through data augmentation.

The datasets were primarily evaluated based on their 'accuracy' [35]. We generated synthetic datasets using the library developed through this study for a quantitative accuracy comparison, utilizing selected example datasets. To establish a control group for comparison, we similarly generated synthetic datasets using Hazy's Synthpop library and compared their accuracies. To generalize the experiment results, we created 1,000 synthetic datasets and analyzed their accuracies using the method outlined in Table 3.

TABLE III
METHOD FOR COMPARING MODELING RESULTS ACROSS DATASETS

Dataset (Size)	Experiment Size	Modeling Method	Accuracy metric	Model similarity
Iris (30*5)	30*5 500*5	Logistic Regression	Classification accuracy	Cosine similarity of model coefficients
Diabetes (30*8)	30*8 500*8	Logistic Regression	classification accuracy	
Boston Housing (30*13)	30*13 500*13	Linear Regression	mean squared error	

Initially, models were created using synthetic datasets from the library developed in this study and Hazy's synthpop, serving as the control group. To compare the accuracy of these models, all variables except for the dependent variable in each dataset were designated as independent variables, and modeling was conducted utilizing logistic regression and linear regression via Scikit-learn. The example dataset was employed as a test dataset to facilitate the comparison of various accuracy metrics. Furthermore, to compare the similarity with the modeling results of the example dataset, the coefficients of each model were extracted, and the cosine similarity between these coefficients and those of the example dataset model was calculated and analyzed.

III. Results and Discussion

A. Results of Needs Analysis and Library Design

The results of the needs analysis are presented in Table 4. Firstly, it was found that teachers who provide datasets to students and conduct classes in the teaching and learning environment value the diversity of dataset topics, forms, and sizes offered by the dataset library. On the other hand, the demand for additional features such as visualization, programming, and user sharing was relatively low. This is interpreted to mean that data visualization and direct handling of data for data analysis and AI education hold educational significance, and the programming tools preferred by teachers already offer these functionalities.

Regarding the form of datasets for AI education, it was revealed that all teachers prefer structured data in numeric and textual formats, with over 91% of teachers expressing a preference for image data and a minority favoring video datasets. The most preferred method of providing datasets to students was using datasets within programming tools, followed by students downloading them directly and distributing them through class communities or cloud services.

TABLE IV
RESULTS OF NEEDS ANALYSIS

Questionnaire	Response results (Ratio)				
Importance of dataset topic diversity	1 0 (0)	2 0 (0)	3 0 (0)	4 7(29.2)	5 17 (70.8)
Importance of dataset shape diversity	1 0 (0)	2 3 (12.5)	3 2 (8.3)	4 5(20.8)	5 13 (54.2)
Importance of additional functions (visualization, sharing, etc.)	1 2 (8.3)	2 7 (29.2)	3 2 (8.3)	4 5(20.8)	5 8 (33.4)
Choose 2 preferred types of datasets	Data frame 24 (100)	image 22 (91.7)	audio 0 (0)	movie 2 (8.3)	etc. 0 (0)
Choose 2 preferred ways to offering datasets	Student downloads directly Through a programming tool Through the class community Through download links Others (using file-delivering software, etc.)				12(50) 18(75) 9(37.5) 6(25 3(12.5)
Other essential features	Easy student access, address privacy and information identification issues, and generate outlier missing data for educational purposes.				

In the open-ended questions about the essential requirements for an AI education dataset library, there were requests for interoperability with currently used AI educational tools and ensuring student accessibility, creation of outliers and missing values for educational purposes, provision of structured datasets on various topics, and addressing personal information and identification issues for using proprietary datasets. Based on the needs analysis results, we designed the library to provide and reconstruct datasets themselves, focusing on these features, as shown in Fig. 2.

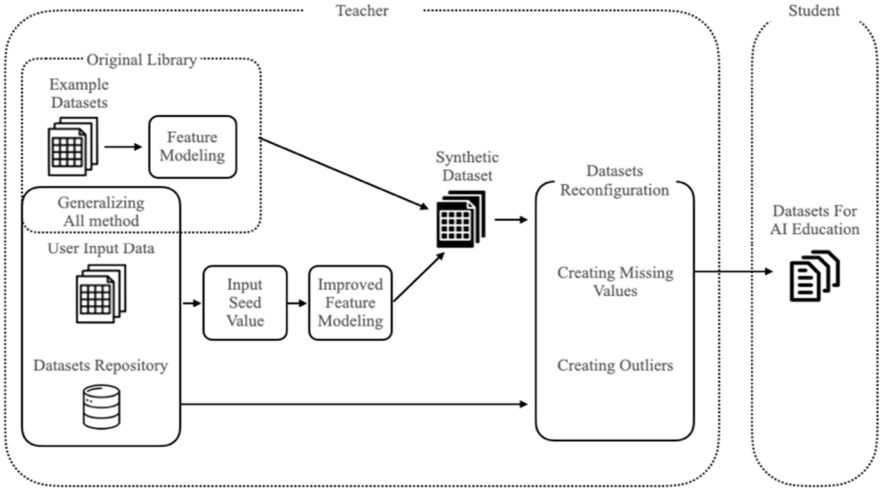

Fig. 2 Designing a Dataset Library for AI Education that is generating synthetic dataset

The programming language was set to Python, widely used in AI education, to enhance accessibility in the teaching and learning environment. The library distribution format utilizes PyPI, enabling direct access from programming tools. The library includes example datasets usable in AI teaching and learning environments. It is designed to generate sophisticated synthetic datasets through improved feature modeling compared to existing libraries. The generated synthetic datasets are limited to structured data and intended to be extracted as DataFrame using the Python Pandas library. Additionally, the functionality to create arbitrary missing

values or outliers has been added to increase educational suitability.

B. Development of Synthetic Dataset Library

The original library is an alpha version library that implements Synthpop, a package for generating synthetic datasets in R, into Python. This library can only create datasets using the example datasets provided within it, and it has a limitation in reproducing the generated datasets due to the randomness in predictions after modeling. To address these limitations, the process was improved to enable the generation of synthetic datasets from general data frames input by users. Furthermore, since the reproducibility of datasets is crucial in educational environments, enhancements were made to reproduce generated synthetic datasets using seed values.

To enhance the original library's modeling method, Random Forest is the primary method for post-prediction distribution. Random Forest, comprising multiple Decision Trees and utilizing ensemble techniques for prediction, generally outperforms most individual Decision Trees. Accordingly, a modeling algorithm utilizing Random Forest was structured as outlined in Table 5 and applied to the library.

TABLE V
RANDOM FOREST ALGORITHM FOR SYNTHETIC DATASET GENERATION

Algorithm Random Forest

Given :
- Df : raw dataframe
- Vs : order of columns to synthesize
- D_p : dataframe the columns referenced have the value 1
- B_s : boolean for column smoothing preprocess status
- B_p: boolean for data shuffling preprocess status
- N_e : number of decision trees, 100
- N_d : maximum depth of decision trees, None
- N_s : seed value for random number generation
- N_k : length of the rows to generate

Modeling Phase :
1: FOR c = Vs
2: col_predictor = all column names with a value of 1 in D_p[c]
3: X_{df} = Df[col_predictor]
4: Y = Df[col]
5: IF data type of Y is categorical type
6: Rf = RandomForestClassifier(N_e, N_d, N_s)
7: IF data type of Y is numeric type
8: Rf = RandomForestRegressor(N_e, N_d, N_s)
9: IF B_s
10: X_{df} , Y = Min Max normalized or One hot encoding(X_{df} , Y)
11: Rf.fit(X_{df} , Y)
12: leaves_y = DataFrame(key : leaf nodes(X_{df}), value : Y)

Generating Phase:
1: synth_df = DataFrame(key : Vs, value : [0]* N_k)
2: synth_df[Vs[0]] = random choice(Df[Vs[0]], N_k, N_s)
3: FOR c = Vs
4: col_predictor = all column names with a value of 1 in D_p[c]
5: X_{df} = synth_df[col_predictor]
6: IF B_s
7: X_{df} = Min Max normalized or One hot encoding(X_{df})
8: ypred = [0] * leaf nodes(X_{df})
9: leaves_pred = DataFrame(key : leaf nodes(X_{df}), value : Y)
10: FOR leaf, indice = leaves_pred.keys, leaves_pred.values
11: IF leaf in leaves_ydict.keys
12: ypred[indices] = random choice(leaves_yt[leaf], indices size, N_s)
13: ELSE
14: ypred[indices] = random choice (leaves_y[near leaf], indices size, N_s)
15: synth_df[c] = ypred

To generate a synthetic dataset, a model that learned the information of the columns was required. According to the order of the columns (V_s) to be generated, they were used as independent variables for modeling. The process involves cumulatively using preceding columns as independent variables for modeling and conducting processes based on the column's data type or synthesis options (B_s ~ N_s). The processed independent variables (X_{df}) were then used as training data, and the column to be generated was modeled as the dependent variable (Y).

Once all columns of the original dataset intended for synthetic dataset generation had been modeled, the SRMI method replaced the data one column at a time. The first column was randomly sampled from the original dataset as many times as needed for generation. All random sampling for creating the synthetic dataset utilizes the Seed value (N_s) entered as an option to ensure reproducibility. Subsequently, the remaining columns were generated and replaced with data using the predicted values from the trained Random Forest, according to the sequence of columns (V_s) to be generated, utilizing previously generated columns as independent variables.

Based on the characteristics of Random Forest, an ensemble-based method, the most frequent outcome by the independent variables was determined, producing a leaf node. If the leaf node corresponds to a value existing in the original dataset, the dataset is generated by randomly sampling the mapped values from that leaf node. If the leaf node does not exist in the original dataset, the dataset is created using the nearest leaf node. Additionally, to add the functionality of creating arbitrary missing or outlier values for educational purposes, a dataset reconstruction algorithm was structured as outlined in Table 6 and applied to the library.

TABLE VI
DATASET RECONSTRUCTION ALGORITHMS FOR AI EDUCATION

Algorithm Dataset Reconfiguration

Given:
- Synth_df: generated synthetic dataset
- N_m: ratio of missing value
- N_o : ratio if outliers
- N_s : seed value for random number generation

Missing value Reconfiguration :
1: count = 0
2: WHILE (whole data * N_m) >= count
3: x = *random(o, len(df), N_s)*
4: y = random(0, len(df.columns) , N_s)
5: synth_df.iloc[x, y] = *NaN*
6: count =+1

Outliers Reconfiguration :
1: count = 0
2: WHILE (whole data * N_o) > count
3: x = *random(o, df.row, N_s)*
4: y = random(0, df.columns, N_s)
5: IF synth_df[y].dtype is numeric type
6: q1 = synth_df[y] .quantile(0.25)
7: q3 = synth_df[y] .quantile(0.75)
8: IF random(N_s) % 2 == 0
9: synth_df[x][y] = q1 – random(1.5, 2 N_s) * (q3-q1)
10: ELSE
11: synth_df[x][y] = q3 + random(1.5, 2 N_s) * (q3-q1)

The addition of missing values and outliers was processed in the final stage after the synthetic dataset had been generated. Given a synthetic dataset, along with the proportions of missing values, outliers, and a seed value (N_s), a specified percentage of the data is replaced with missing values (*NaN*) or outliers. The locations for these replacements

were determined by reproducible random positions generated using the seed value. For outliers, the column's data type at the random position was checked. If it was numerical data, that column's first and third quartiles were derived. Using the IQR (*Q3-Q1*) value, outliers were generated by subtracting the IQR value from the first quartile and adding it or the third quartile to ensure a variety of outliers can be included.

C. Expert Review Results for Library

To validate the suitability of the developed library in the AI teaching and learning environment, a survey focused on the features of 'synthetic dataset generation,' 'dataset size configuration,' 'seed value reproducibility,' 'missing value generation,' and 'outlier generation.' The results are presented in Table 7.

TABLE VII
LIBRARY FEATURE EXPERT REVIEW RESULTS

Feature	Mean	Std	Agree	CVR
Generate synthetic dataset	4.57	0.49	14	1.00
Setting the dataset size	4.64	0.48	14	1.00
Reproduction with seed values	4.86	0.35	14	1.00
Generate missing values	4.50	0.50	14	1.00
Generate outliers	4.79	0.41	14	1.00

Based on the expert panel numbers (*n=14*), it was confirmed that the library's functionalities were all validated (*CVR. ≥ 0.571*). All experts submitted responses that fall under 'suitable' and 'very suitable,' and the average responses, encoded from 1 (not at all) to 5 (very suitable), also indicated that the developed AI educational synthetic dataset generation library is appropriate for the academic field, as the average responses fell between 'suitable' and 'very suitable' (*Mean ≥ 4.50*).

D. Quality Evaluation Results of Synthetic Datasets

To compare the quality of the datasets generated through the example datasets with those generated by Hazy's synthpop, modeling was performed for each, and both accuracy and cosine similarity to the example dataset models were measured. Furthermore, the average and standard deviation of 1,000 measurement results were calculated to generalize the evaluation results. The results of generating a synthetic dataset with the same number of rows (30) as the example datasets are shown in Table 8.

TABLE VIII
RESULTS OF EVALUATING SYNTHETIC DATASET QUALITY THROUGH MODELING (N=30)

Feature	Division	Accuracy		Cosine Similarity	
		Mean	Std	Mean	Std
Multi-classification (Iris)	Experimental	0.96	0.14	0.99	0.00
	Control	0.95	0.14	0.98	0.01
Binary classification (Diabetes)	Experimental	0.77	0.00	0.97	0.01
	Control	0.77	0.00	0.95	0.02
Linear regression (Boston Housing)	Experimental	23.15	0.62	0.97	0.02
	Control	25.21	1.03	0.86	0.06

Regarding the average and standard deviation of accuracy and cosine similarity, there was no significant difference between the library using Random Forest and the existing library that uses Decision Trees. Both libraries demonstrated high accuracy, exceeding 0.95 in multi-class classification, and the cosine similarity was also high, above 0.98, indicating that the datasets well reflected the properties of the original datasets. In binary classification, both libraries showed somewhat lower accuracy, while for linear regression, the library using Random Forest showed relatively higher results in cosine similarity. For comparing each of the 1,000 sessions created, accuracy was visualized in histograms, as shown in Fig. 3.

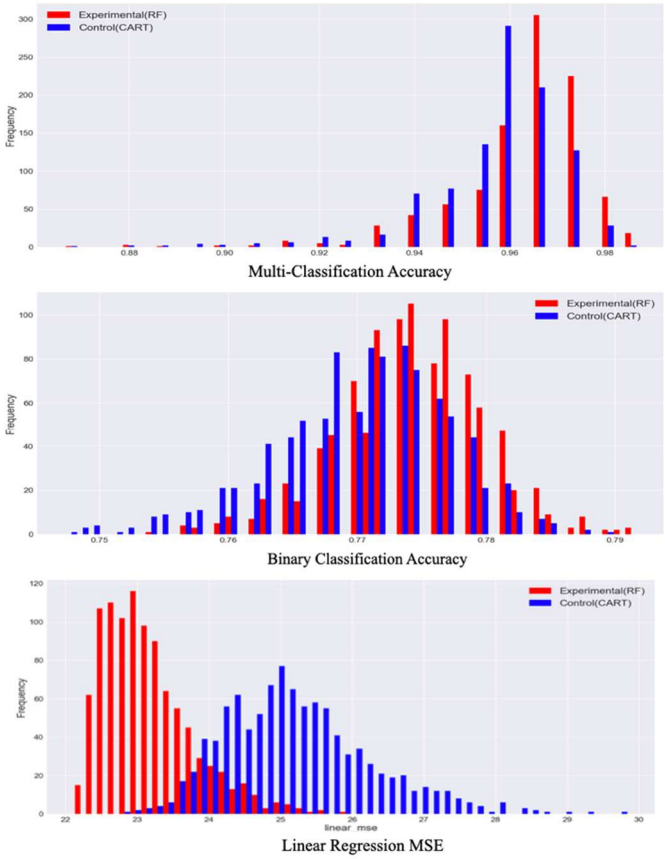

Fig. 3 Synthetic dataset accuracy evaluation results based on example datasets (n=30)

The histogram visualization results revealed differences between the two libraries. While the overall shapes of the distributions for both multi-class and binary classifications were similar across the libraries, the one utilizing Random Forest showed classifications concentrated around higher accuracy levels. In linear regression, the difference between the two libraries was more pronounced. It was observed that the library utilizing Random Forest had Mean Squared Error (MSE) values concentrated around lower values, indicating higher accuracy. The original library using decision tree algorithms exhibited a wider distribution of MSE values, whereas Random Forest's MSE values were centered around lower error values, showing a relatively narrower range.

In accuracy evaluation using the original dataset as test data, the overall accuracy appeared similar for both libraries. However, the study found that the library utilizing Random Forest was able to generate more consistently accurate synthetic datasets. Fig. 4 visualizes the cosine similarity between the model coefficients of the synthetic datasets

generated in each of the 1000 sessions and the model coefficients of the example dataset.

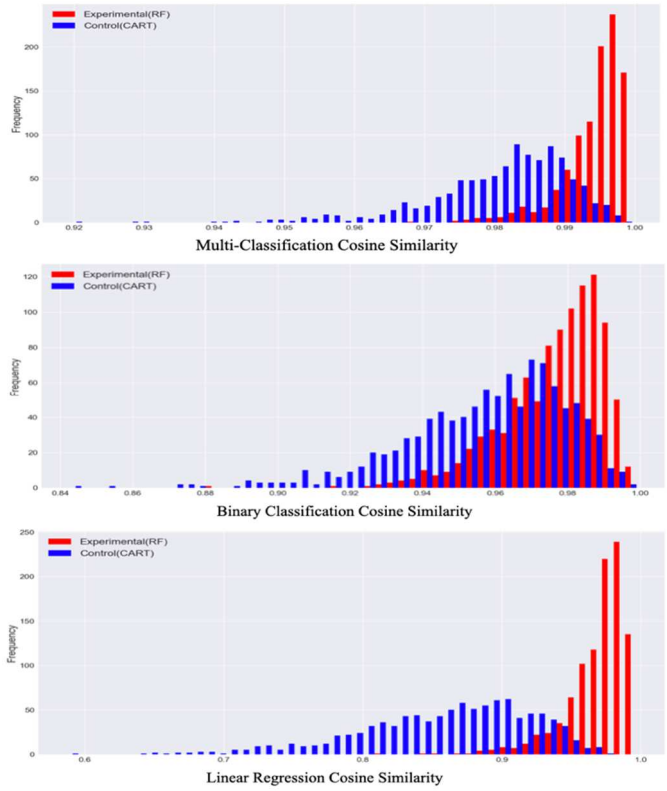

Fig. 4 Synthetic dataset cosine similarity evaluation results based on example datasets (n=30)

The histogram visualization of the cosine similarity also revealed differences between the two libraries. The library utilizing Random Forest showed values close to 1 for most sessions, exhibiting a tighter distribution than the broader distribution seen with the library using Decision Trees, especially in multi-classification and linear regression tasks, where a higher similarity was observed. For multi-classification, the synthetic datasets from Random Forest showed a cosine similarity exceeding 0.97 in most sessions, while Decision Tree results started from 0.92, showing a comparatively wider distribution. The difference was even more pronounced in linear regression. The synthetic results from Random Forest mostly appeared above 0.90, whereas those from Decision Trees were above 0.65, displaying a much more comprehensive range of value distribution.

This indicates that synthetic datasets created using Random Forest could produce properties more similar to the original dataset, and the modeling results using these datasets also showed comparable outcomes. In generating datasets of the same size from a small original dataset, there were no significant differences between the two libraries regarding the mean and standard deviation in the dataset quality comparison, which evaluated accuracy and model coefficient similarity. However, individual analyses of repeated experiments indicated that the library developed in this study, utilizing Random Forest, was more stable and produced datasets that better reflected the characteristics of the original dataset. The results summarizing the average and standard deviation of accuracy and cosine similarity after augmenting an original dataset with 30 rows into a synthetic dataset with 500 rows over 1,000 iterations are shown in Table 9.

TABLE IX
RESULTS OF EVALUATING SYNTHETIC DATASET QUALITY THROUGH MODELING (N=500)

Feature	Division	Accuracy		Cosine Similarity	
		Mean	Std	Mean	Std
Multi-classification (Iris)	Experimental	0.99	0.00	0.96	0.00
	Control	0.99	0.01	0.92	0.00
Binary classification (Diabetes)	Experimental	0.92	0.02	0.91	0.05
	Control	0.78	0.02	0.57	0.10
Linear regression (Boston Housing)	Experimental	3.10	0.38	0.98	0.01
	Control	19.35	2.66	0.43	0.10

Focusing on accuracy, both libraries showed a high accuracy of 0.99 for multiclass classification. However, in binary classification, there was a notable difference with accuracies of 0.92 and 0.78, and in linear regression, the difference was even more significant with accuracies of 3.10 and 19.35. Regarding cosine similarity, except for multiclass classification, there were significant differences in the other two experiments. For multiclass classification, the difference was not substantial, with cosine similarities of 0.96 and 0.92. Still, for binary classification and linear regression, the differences were significant, with 0.91 and 0.57 for binary and 0.98 and 0.43 for linear regression, respectively. Interpreting the results based on the average and standard deviation, it can be observed that the library utilizing Random Forest more precisely generates datasets for binary classification and linear regression when augmenting many datasets. The results of visualizing the accuracy for each of the 1,000 generated sessions as a histogram are displayed in Fig. 5.

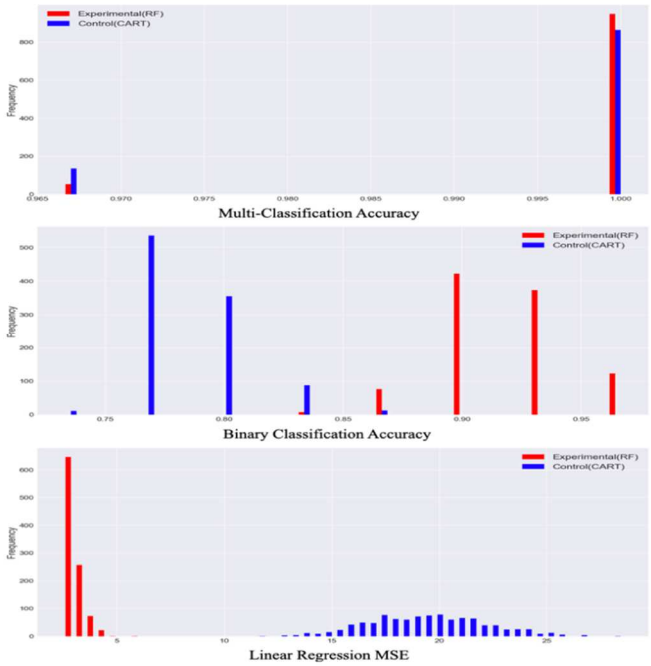

Fig. 5 Synthetic dataset accuracy evaluation results based on example datasets (n=500)

Upon examining the results of the histogram visualization of accuracy, it is observed that, despite applying the same

visualization options (*bins=50*) as in previous sessions of generating synthetic data, the results are more concentrated around specific values overall. Notably, for the accuracy of multi-class classification, both libraries showed over 80% of the values concentrated around 1. Similarly, binary classification results were not dispersed but instead centrally focused on specific values.

An in-depth analysis of each result revealed no significant difference in the outcomes of multi-class classification. However, the library utilizing random forests in binary classification and linear regression demonstrated more precise and concentrated accuracy results. For binary classification, the accuracy of the library employing random forests was mainly distributed above 0.90, centering around this value, whereas the library using decision trees showed most values between 0.75 and 0.80. The difference in results was even more pronounced for linear regression. The random forest outcomes were mainly distributed below 4, whereas the conventional library showed an extensive distribution centered around 20.

Fig. 6 visualizes the cosine similarity between the model coefficients of the synthetic datasets generated in each of the 1000 sessions and the model coefficients of the example dataset.

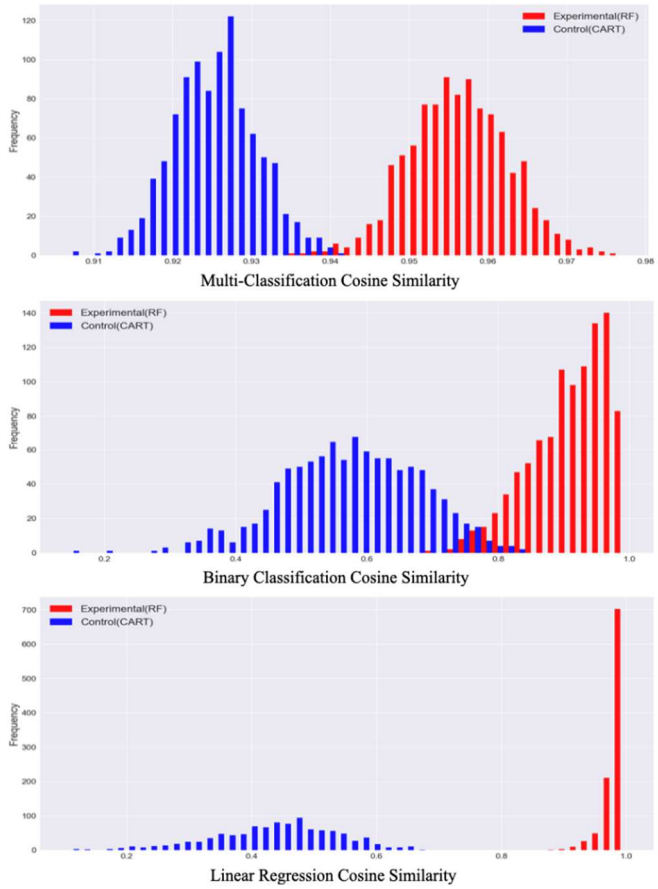

Fig. 6 Synthetic dataset cosine similarity evaluation results based on example datasets (n=500)

Cosine similarity also showed significant differences between the two libraries in all cases. In multi-class classification, although the distribution shapes of the two libraries appeared similar, the library using random forests predominantly showed values distributed between 0.95 and 0.96, whereas the conventional library utilizing decision trees showed distributions between 0.92 and 0.93. The results for binary classification and linear regression indicated distinct distribution shapes between the two libraries. The library employing random forests showed values close to 1 within a narrow range, unlike the decision tree-based library, which displayed lower similarity across a broader range of results, particularly in linear regression, a wide distribution centered around 0.5 was observed, displaying lower similarity compared to results utilizing random forests.

In cases where datasets were augmented to 500, synthetic datasets generated using random forests more precisely reflected the characteristics of the original dataset. They demonstrated more stable outcomes compared to synthetic datasets created through decision trees. When synthesizing the evaluation of accuracy using the original dataset as test data and the assessment of similarity of model coefficients, it became evident that the library utilizing random forests exhibited more stable outcomes in repeated experiments and demonstrated higher accuracy, both when generating a small number of datasets and when augmenting the generation of a more significant number of datasets. Notably, in all experimental outcomes, the cosine similarity of datasets generated using random forests appeared to be over 0.90, indicating a very accurate reflection of the statistical properties of the original dataset. This suggested that such datasets are more suitable for AI education, where modeling practice is crucial.

IV. Conclusion

As AI technology continues to impact society profoundly, the importance of AI education that fosters students' AI literacy is increasingly emphasized. Various studies related to AI education have been conducted, and the need for datasets suitable for students' levels and can be utilized in educational practices, reflecting real-life contexts, has been highlighted. However, there needs to be more research on datasets for AI education from an academic perspective, and academic fields face challenges in securing appropriate educational datasets.

This study identified the essential features of a library for providing datasets in educational environments through an analysis of teachers' needs. Datasets for AI education need to vary in size according to academic objectives, be easily accessible to students, and be free from personal information and data identification issues. Additionally, it was necessary for these datasets to include arbitrary missing values or outliers, depending on the purpose of the education. This aligns with the emphasis on the importance of datasets in AI education, as highlighted in related research [14], [15], [16].

Reflecting the needs of the educational field, an AI education library utilizing synthetic dataset generation techniques was developed. Compared to existing libraries, this library includes various functions tailored to academic purposes and improves the quality of synthetic datasets produced [31]. AI education experts validated the expanded library through a review of its suitability for educational environments and an assessment of the quality of the generated datasets. The datasets created with the developed library, assumed to be used for AI education purposes, were validated based on modeling results. Compared to existing

libraries, it was confirmed that the expanded library could produce more stable synthetic datasets that accurately reflect the statistical properties of the original datasets.

The synthetic dataset generation AI education library, a result of this study, has been generalized for application across all Python DataFrames and offers versatile reconstruction capabilities. Hence, data generated in educational environments is anticipated to augment or reconfigure, thereby supporting the academic field. This addresses the challenges highlighted in related research, where the difficulty of refining data for educational purposes complicates the utilization of real-life data in teaching and learning contexts. Additionally, it is expected to effectively support AI education during computing practices, resolving issues presented in previous studies and enhancing the effectiveness of AI education [9], [10], [11], [19].

The importance of AI education in nurturing the AI literacy of students poised to live in future societies is being increasingly emphasized. However, relative to its significance, more research needs to focus on AI education datasets. This study is meaningful as it centers on datasets for AI education. Moving forward, it aims to evolve this research by meticulously designing the essential requirements for educational datasets and exploring various example datasets to be included in the library. The results of this study will serve as a foundation for diverse research centered on AI educational datasets, thereby aiding in the enhancement of students' AI literacy.

REFERENCES

[1] J. McCarthy, "What is artificial intelligence," 2007, Accessed: Feb. 13, 2024. [Online]. Available: http://cse.unl.edu/~choueiry/S09-476-876/Documents/whatisai.pdf

[2] L. Li, "A comparative study on artificial intelligence curricula," PhD Thesis, Western Ontario Univ., Canada, 2020. Accessed: Feb. 13, 2024.

[3] S. Druga, S. T. Vu, E. Likhith, and T. Qiu, "Inclusive AI literacy for kids around the world," in *Proceedings of FabLearn 2019*, in FL2019. New York, NY, USA: Association for Computing Machinery, Mar. 2019, pp. 104–111. doi: 10.1145/3311890.3311904.

[4] S. G. Han, "Digital Content to Improve Artificial Intelligence Literacy Ability," *Journal of the Korea Society of Computer and Information*, vol. 25, no. 12, pp. 93–100, Dec. 2020, doi:10.9708/jksci.2020.25.12.093.

[5] D. Long and B. Magerko, "What is AI Literacy? Competencies and Design Considerations," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA: ACM, Apr. 2020, pp. 1–16. doi: 10.1145/3313831.3376727.

[6] D. T. K. Ng, J. K. L. Leung, S. K. W. Chu, and M. S. Qiao, "Conceptualizing AI literacy: An exploratory review," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100041, 2021, doi:10.1016/j.caeai.2021.100041.

[7] W. Yang, "Artificial Intelligence education for young children: Why, what, and how in curriculum design and implementation," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100061, 2022, doi:10.1016/j.caeai.2022.100061.

[8] I. T. Sanusi, S. S. Oyelere, H. Vartiainen, J. Suhonen, and M. Tukiainen, "A systematic review of teaching and learning machine learning in K-12 education," *Educ Inf Technol*, vol. 28, no. 5, pp. 5967–5997, May 2023, doi: 10.1007/s10639-022-11416-7.

[9] P. Langley, "An Integrative Framework for Artificial Intelligence Education," *AAAI*, vol. 33, no. 01, pp. 9670–9677, Jul. 2019, doi:10.1609/aaai.v33i01.33019670.

[10] R. M. Martins and C. Gresse Von Wangenheim, "Findings on Teaching Machine Learning in High School: A Ten - Year Systematic Literature Review," *Informatics in Education*, Sep. 2022, doi:10.15388/infedu.2023.18.

[11] W. Chow, "A Pedagogy that Uses a Kaggle Competition for Teaching Machine Learning: an Experience Sharing," in *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*, Yogyakarta, Indonesia: IEEE, Dec. 2019, pp. 1–5. doi:10.1109/tale48000.2019.9226005.

[12] M. Tedre, T. Toivonen, J. Kahila, H. Vartiainen, T. Valtonen, I. Jormanainen, and A. Pears, "Teaching Machine Learning in K–12 Classroom: Pedagogical and Technological Trajectories for Artificial Intelligence Education," *IEEE Access*, vol. 9, pp. 110558–110572, 2021, doi: 10.1109/access.2021.3097962.

[13] B. Hutchinson, A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes and M. Mitchell, "Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event Canada: ACM, Mar. 2021, pp. 560–575. doi: 10.1145/3442188.3445918.

[14] I. Evangelista, G. Blesio, and E. Benatti, "Why Are We Not Teaching Machine Learning at High School? A Proposal," in *2018 World Engineering Education Forum - Global Engineering Deans Council (WEEF-GEDC)*, Albuquerque, NM, USA: IEEE, Nov. 2018, pp. 1–6. doi: 10.1109/weef-gedc.2018.8629750.

[15] R. Biehler and Y. Fleischer, "Introducing students to machine learning with decision trees using CODAP and Jupyter Notebooks," *Teaching Statistics*, vol. 43, no. S1, Jul. 2021, doi: 10.1111/test.12279.

[16] H. Vartiainen, T. Toivonen, I. Jormanainen, J. Kahila, M. Tedre, and T. Valtonen, "Machine learning for middle schoolers: Learning through data-driven design," *International Journal of Child-Computer Interaction*, vol. 29, p. 100281, Sep. 2021, doi:10.1016/j.ijcci.2021.100281.

[17] S. Hooper and L. P. Rieber, "Teaching with technology," *Teaching: Theory into practice*, vol. 2013, pp. 154–170, 1995.

[18] T. K. F. Chiu and C. Chai, "Sustainable Curriculum Planning for Artificial Intelligence Education: A Self-Determination Theory Perspective," *Sustainability*, vol. 12, no. 14, p. 5568, Jul. 2020, doi:10.3390/su12145568.

[19] I. Bosnić, I. Čavrak, and A. Zuiderwijk, "Introducing Open Data Concepts to STEM Students Using Real-World Open Datasets," in *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, Sep. 2021, pp. 1530–1535. doi:10.23919/mipro52101.2021.9596998.

[20] S. Kim, K. Kim, and T. Kim, "Development of PISA Mathematical Context-oriented Dataset for K-12 Artificial Intelligence Education," *Journal of The Korean Association of Information Education*, vol. 27, no. 3, pp. 255–267, Jun. 2023, doi: 10.14352/jkaie.2023.27.3.255.

[21] T. Coughlan, "The use of open data as a material for learning," *Education Tech Research Dev*, vol. 68, no. 1, pp. 383–411, Feb. 2020, doi: 10.1007/s11423-019-09706-y.

[22] K. El Emam, L. Mosquera, and R. Hoptroff, *Practical synthetic data generation: balancing privacy and the broad availability of data*. O'Reilly Media, 2020.

[23] S. Kim, T. Kim, and Y. Jeon, "Research on the Development and Utility Analysis of K-12 Artificial Intelligence Educational Datasets Using Synthetic Datasets Generation Method," *The Journal of Korean Association of Computer Education*, vol. 25, no. 3, pp. 9–21, May 2022, doi: 10.32431/KACE.2022.25.3.002.

[24] A. Rossett, *Training needs assessment*. Educational Technology, 1987. Accessed: Mar. 13, 2024.

[25] T. E. Raghunathan, J. M. Lepkowski, J. Van Hoewyk, and P. Solenberger, "A multivariate technique for multiply imputing missing values using a sequence of regression models," *Survey methodology*, vol. 27, no. 1, pp. 85–96, 2001.

[26] J. Kim and M. Park, "Multiple imputation and synthetic data," *The Korean Journal of Applied Statistics*, vol. 32, no. 1, pp. 83–97, 2019.

[27] J. P. Reiter, "Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study," *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 168, no. 1, pp. 185–205, 2005.

[28] J. Lee, "Review on Statistical Methods for Synthetic Data," M. S. thesis, Dept Statics, UOS, Seoul Univ, Seoul, Korea, 2021.

[29] B. Nowok, G. M. Raab, and C. Dibben, "synthpop: Bespoke Creation of Synthetic Data in R," *J. Stat. Soft.*, vol. 74, no. 11, 2016, doi:10.18637/jss.v074.i11.

[30] S. Yoo and N. Park, "Synthetic Data Generation for Individual Credit Data Using CART," *Journal of the Korean Official Statistics*, vol. 25, no. 1, pp. 1–30, 2020.

[31] Hazy Limeted, "hazy/synthpop.", Dec. 16, 2019. Accessed: Dec. 08, 2022. [Online]. Available: https://github.com/hazy/synthpop

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg,

J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Michel, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[33] M. Carlisle, "racist data destruction?," Medium. Accessed: Jan. 14, 2024. [Online]. Available: https://medium.com/@docintangible/racist-data-destruction-113e3eff54a8

[34] C. H. Lawshe, "A Quantitative Approach to Content Validity," *Personnel Psychology*, vol. 28, no. 4, pp. 563–575, Dec. 1975, doi:10.1111/j.1744-6570.1975.tb01393.x.

[35] M. Bergdahl, M. Ehling, E. Elvers, E. Földesi, T. Körner, A. Kron, P. Lohauß, K. Mag, V. Morais, A. Nimmergut, H. Viggo, K, Szép, U. Timm, and M. J. Zilhão "Handbook on Data Quality Assessment Methods and Tools." Ehling, Manfred Körner, Thomas, 2007.