Resources and Technology 21 (4): 44-65, 2024
ISSN 2307-0048
http://rt.petrsu.ru

DOI: 10.15393/j2.art.2024.8043
УДК 681.518.5:(631.12+631.3.076)
Статья

Способ и процессная модель предварительной обработки данных автоматизированных систем контроля технического состояния зерноуборочных комбайнов

Помогаев Виталий Михайлович
кандидат экономических наук, доцент, Омский государственный аграрный университет имени П. А. Столыпина (Российская Федерация), vm.pomogaev@omgau.org

Ревякин Павел Игоревич
аспирант, Омский государственный аграрный университет имени П. А. Столыпина (Российская Федерация), pi.revyakin@omgau.org

Басакина Анна Сергеевна
аспирант, Омский государственный аграрный университет имени П. А. Столыпина (Российская Федерация), as.basakina@omgau.org

Получена: 31 июля 2024 / Принята: 15 октября 2024 / Опубликована: 2 ноября 2024

Аннотация: Современные самоходные сельскохозяйственные машины отличаются своей технологичностью, сложностью и высокой стоимостью. Обеспечение надёжности и работоспособности таких машин является ключевой задачей технического сервиса. Развитие технологий технического сервиса позволяет собирать, обрабатывать и строить прогнозы технического состояния машин на основе данных, получаемых в процессе их эксплуатации. Сбор данных осуществляется встроенными системами контроля, с помощью которых происходит диагностирование и оперативное выявление неисправностей в работе узлов и агрегатов машин. Целью исследования являлась разработка и апробация способа предварительной обработки данных, полученных с помощью автоматической системы контроля технического состояния и датчиков, установленных на зерноуборочных комбайнах ACROS, и формализация разработанного алгоритма для последующей автоматизации процесса подготовки данных для технического анализа. Качество данных оценивалось по следующим критериям: объём данных, типы данных, количество атрибутов, наличие и количество пропусков, наличие дубликатов, наличие аномалий, соответствие

категорий, нормализация и согласованность значений, возможная гомогенность, сегментация. В качестве инструментов использованы Python, R, библиотеки Pandas, NumPy, Matplotlib. В результате проведённого исследования установлено, что сырые данные с аналитических систем контроля технического состояния зерноуборочных комбайнов не пригодны для анализа и прогнозирования технического состояния узлов и агрегатов. Прежде всего, это связано с большим количеством пропущенных значений. Построение процессной модели на основе разработанного способа предварительной обработки данных может рассматриваться как концепция информационной системы, позволяющей автоматизировать процесс подготовки данных систем контроля технического состояния зерноуборочных комбайнов для машинной обработки. Представленный способ позволил получить структурированные и информативные данные, корректное заполнение пропусков, устранение выбросов и ошибок. Построенная процессная модель обеспечивает прозрачность, контроль и оптимизацию процессов работы с данными, позволит исключить ошибки и противоречия в их дальнейшем анализе, а также создаст условия для повторяемости действий в дальнейшем при обработке аналогичных датасетов, полученных с систем контроля технического состояния зерноуборочных комбайнов.

Ключевые слова: зерноуборочный комбайн; надёжность; мониторинг; процессная модель

DOI: 10.15393/j2.art.2024.8043
Article

Method and process model for pre-processing data from automated systems for monitoring the technical condition of combine harvesters

Vitaly Pomogaev
Ph. D. in economics, associate professor, Omsk State Agrarian University named after P. A. Stolypin (Russian Federation), vm.pomogaev@omgau.org

Pavel Revyakin
Ph. D. student, Omsk State Agrarian University named after P. A. Stolypin (Russian Federation), pi.revyakin@omgau.org

Anna Basakina
Ph. D. student, Omsk State Agrarian University named after P. A. Stolypin (Russian Federation), as.basakina@omgau.org

Received: 31 July 2024 / Accepted: 15 October 2024 / Published: 2 November 2024

Abstract: Modern self-propelled agricultural machines are characterized by their technological sophistication, complexity and high cost. Ensuring reliability and serviceability of such machines is the key task of technical service. The development of technical service technologies allows collecting, processing and forecasting the technical condition of machines based on the data obtained in the process of machine operation. Data collection is performed by built-in control systems to diagnose and detect malfunctions in the work of machine units and assemblies. The aim of the research was to develop and approbate the method of preliminary data processing obtained with the automatic system of technical condition monitoring and sensors installed on combine harvesters ACROS and to formalize the developed algorithm for further automation of data preparation process for technical analysis. Data quality was assessed according to the following criteria: data volume, data types, number of attributes, presence and number of omissions, presence of duplicates, presence of anomalies, matching categories, normalization and consistency of values, possible homogeneity, and segmentation. The tools used were Python, R, Pandas, NumPy, Matplotlib libraries. The authors established that raw data from analytical systems of technical condition control of combine harvesters were not suitable for analyzing and predicting the technical condition of units and assemblies due to many missing values.

The process model construction based on the developed method of data pre-processing may be considered as a concept of an information system, which allows automating the data preparation process of technical condition control systems of combine harvesters for machine processing. The presented method allowed the authors to obtain structured and informative data and correct filling of omissions and to eliminate outliers and errors. The proposed process model provides transparency, control and optimization of data handling processes, will allow excluding errors and contradictions in their further analysis, and will provide repeatability of actions in the future while processing similar datasets received from the systems of technical condition control of combine harvesters.

Keywords: combine harvester; reliability; monitoring; process model

1. Введение

Проблема обеспечения надёжности сельскохозяйственных машин в последнее время приобретает особую актуальность в связи с возрастанием их сложности и технологичности, последствиями простоя машин в условиях их высокой загруженности, высокой стоимости технического обслуживания и ремонта (ТО и Р). Современные сельскохозяйственные машины в большинстве своём имеют встроенные системы контроля работы узлов и агрегатов. Отдельные ответственные механизмы могут быть дополнительно оснащены системами непрерывного контроля с возможностью накопления и передачи данных [1]. Использование указанных систем контроля позволяет решить важную задачу — обеспечение надёжности машин при минимизации затрат на ТО и Р [2].

В настоящее время встроенные системы контроля на сельскохозяйственных машинах используются как средства диагностирования для оперативного выявления неисправностей в работе узлов и агрегатов машин. Повысить информационную ценность этих данных можно путём их сбора в непрерывном режиме (мониторинг). Собранные системами контроля данные (значения контролируемых параметров) представляют собой многомерные временны*е* ряды. Анализ временны*х* рядов позволяет получить информацию об изменении контролируемых параметров и состоянии машины. Выявление изменений осуществляется различными методами анализа данных и позволяет получить знания о возможных закономерностях в поведении наблюдаемой системы [3].

Однако в научной литературе практически не встречаются исследования качества данных систем контроля сельскохозяйственных машин для целей их последующей обработки. Качество данных существенно влияет на возможность прогнозирования и точность этих прогнозов. В настоящем исследовании поставлена цель разработать и апробировать способ предварительной обработки данных, полученных с помощью автоматической системы контроля технического состояния, датчиков, установленных на зерноуборочных комбайнах ACROS, и формализовать разработанный алгоритм для последующей автоматизации процесса подготовки данных для технического анализа.

2. Материалы и методы

Предварительная обработка данных имеет важнейшее значение для построения качественной аналитики. Необходимость предварительной обработки сырых данных связана с тем, что значения контролируемых параметров могут быть получены в различных единицах, могут быть сбои в работе датчиков, данные могут некорректно выгружаться в хранилище и т. д. Традиционно факторы, влиявшие на качество данных, группируют в три категории: неполнота данных (отсутствуют атрибуты или пропущены значения); зашумлённость данных (ошибочные записи); несогласованность данных (расхождения или конфликт значений).

Предварительная обработка данных — один из самых трудоёмких процессов в техническом анализе, и качество их оценивается по следующим критериям: объём данных, типы данных, количество атрибутов, наличие и количество пропусков, наличие дубликатов, наличие аномалий, соответствие категорий, нормализация и согласованность значений, возможная гомогенность, сегментация.

При неудовлетворительном качестве данных необходимо провести их предварительную обработку:

- Определить исходный объём данных. Источники данных, структуру и другие качественно-количественные характеристики датасета.
- Очистить данные. Выявить наличие дубликатов и произвести их удаление. Определить наличие и выявить природу пропущенных значений с целью определения методов обработки пропусков [4]. Устранить аномалии, зашумления и выбросы. Как правило, применяются такие методы, как удаление, замена пропущенных значений, подстановка среднего значения, подстановка очевидного значения, подстановка регрессионных значений, методы интерполяции, поиск К-соседей [5] и т. д.
- Трансформировать данные. Данные с различных датчиков могут быть получены в разных форматах и единицах измерения, при необходимости выполнить нормализацию и трансформацию данных. Наиболее распространённые методы: метод логарифмического преобразования, дискретизации, кодирования, степенного преобразования [6] и т. д.
- Определить и преобразовать категориальные переменные. Если в датасете присутствуют категориальные данные, необходимо выполнить их преобразование и кодировку. Например, с помощью технологии one-hot encoding.
- Нормализовать данные для уменьшения измерений и шумов, используя методы минимакса, поиска среднего значения стандартного отклонения, сглаживания и т. д.
- Оптимизировать данные для упрощения обработки на основе выявленных закономерностей, примеров из технической документации или известных атрибутов данных. В зависимости от целей обработки, ответственности можно использовать следующие методы: выбор репрезентативного подмножества, установление важных атрибутов, агрегирование данных, сегментация данных [7] и т. д.
- Предварительно визуализировать данные [8]. Для лучшего понимания структуры датасета, распределения данных, идентификации выбросов, корреляции между переменными, сезонными или событийными паттернами можно воспользоваться методами визуализации данных как по определённой выборке, так и по всему датасету в целом. Например, можно построить графики линий тренда, гистограммы, диаграммы рассеяния, линейные графики, тепловые карты.

Предварительная обработка данных и визуализация результатов в настоящем исследовании осуществлялась с помощью следующих программных продуктов:

- В качестве инструментов автоматизации и среды выполнения исходного кода использовались Python и R.

- В качестве инструмента для работы с наборами данных использовались библиотеки: Pandas (https://pandas.pydata.org/), NumPy (https://numpy.org/).
- В качестве инструмента для визуализации использовалась библиотека и Matplotlib (https://matplotlib.org/).

3. Результаты

3.1. Определение качества данных

В рамках данного исследования авторы работали с датасетом, полученным с помощью автоматизированной системы мониторинга технического состояния и датчиков, установленных на зерноуборочных комбайнах серии ACROS.

Датасет охватывает временно́й интервал с 16.07.2021 г. по 10.10.2021 г. В тестовой выборке представлены данные, полученные с одного зерноуборочного комбайна. Записи экспортированы в CSV формат, общее количество записей — 117 245 шт., суммарный объём датасета равен 85 мегабайт. Выгрузка датасета выполнена в табличном виде и содержит множественные временны́е ряды. Структура датасета описана в таблице 1.

Таблица 1. Структура исходного датасета

Table 1. Structure of the original dataset

№ п/п	Наименование столбца	Описание типа данных	Метаданные
1	timestamp_create	Object	Дата и время создания записи
2	params	Object	Служебная диагностическая информация для системы с параметрами переданного пакета данных
3	speed	Float64	Скорость зерноуборочного комбайна, км/ч
4	data	Object	Данные о движении, технологическом режиме, загрузке и скорости зерноуборочного комбайна в формате JSON
5	sensors_data	Object	Данные, полученные с датчиков, установленных на зерноуборочном комбайне, в формате JSON

Таким образом, можно сделать вывод о том, что в текущем датасете собраны данные с различных информационных подсистем зерноуборочного комбайна. В части содержания данные разделены на две категории: представлены в номинальном виде и в строковом формате JSON, в части группировки поделены на пять категорий: временно́й штамп, служебная информация, данные о скорости, данные о загрузке и режимах работы, данные с датчиков, установленных на зерноуборочном комбайне.

На рисунке 1 отображён фрагмент датасета, выгруженного в табличном виде для удобства визуального представления и понимания структуры информации.

timestamp_create	params	speed	data	sensors_data
25.07.2019 12:02	msgid:3:0(	0	{"don": fa	{"ACTIVE": 1564045377000, "LATITUDE": 46.752916666666664, "LONGITUDE": 38.644708333333334, "BUNKER_CAP": 9.0}
25.07.2019 12:05	msgid:3:0(	0	{"don": fa	{"ACTIVE": 1564045527000, "LATITUDE": 46.752916666666664, "LONGITUDE": 38.644708333333334, "BUNKER_CAP": 9.0}
25.07.2019 12:06	msgid:3:0(	0	{"don": fa	{"ACTIVE": 1564045617000, "LATITUDE": 46.752916666666664, "LONGITUDE": 38.644708333333334, "BUNKER_CAP": 9.0}
25.07.2019 12:09	msgid:3:0(	0	{"don": fa	{"ACTIVE": 1564045767000, "LATITUDE": 46.752916666666664, "LONGITUDE": 38.644708333333334, "BUNKER_CAP": 9.0}
25.07.2019 12:07	msgid:3:0(	0	{"don": fa	{"ACTIVE": 1564045662000, "LATITUDE": 46.752916666666664, "LONGITUDE": 38.644708333333334, "BUNKER_CAP": 9.0}
25.07.2019 12:18	msgid:3:0(	0	{"don": fa	{"ON_NK": 0.0, "ACTIVE": 1564046289000, "ON_100": 0.0, "ON_WAY": 0.0, "LATITUDE": 46.752916666666664, "ON_BRAKE": 1.0, "ON_EMPTY": 0.0, "LONGITUDE": 38.644708333333334, "ON_UPLOAD": 0.0,
23.07.2019 8:12	msgid:3:0(	0	{"don": fa	{"ACTIVE": 1563858750000, "LATITUDE": 46.75288333333334, "LONGITUDE": 38.64458833333333, "BUNKER_CAP": 9.0}
23.07.2019 8:10	msgid:3:0(	0	{"don": fa	{"ACTIVE": 1563858645000, "LATITUDE": 46.75288333333334, "LONGITUDE": 38.64458833333333, "BUNKER_CAP": 9.0}
23.07.2019 8:28	msgid:3:0(	0	{"don": fa	{"ACTIVE": 1563859725000, "LATITUDE": 46.75288333333334, "LONGITUDE": 38.64458833333333, "BUNKER_CAP": 9.0}
23.07.2019 8:31	msgid:3:0(	0	{"don": fa	{"ACTIVE": 1563859890000, "LATITUDE": 46.75288333333334, "LONGITUDE": 38.64458833333333, "BUNKER_CAP": 9.0}
23.07.2019 8:22	msgid:3:0(	0	{"don": fa	{"ACTIVE": 1563859335000, "LATITUDE": 46.75288333333334, "LONGITUDE": 38.64458833333333, "BUNKER_CAP": 9.0}
16.07.2019 11:27	msgid:3:0000000000(		{"don": fa	{"ACTIVE": 1563265653000, "BUNKER_CAP": 9.0}
16.07.2019 11:27	msgid:3:0000000000(		{"don": fa	{"ACTIVE": 1563265638000, "BUNKER_CAP": 9.0}
16.07.2019 11:27	msgid:3:0000000000(		{"don": fa	{"ACTIVE": 1563265668000, "BUNKER_CAP": 9.0}
16.07.2019 11:28	msgid:3:0000000000(		{"don": fa	{"ACTIVE": 1563265683000, "BUNKER_CAP": 9.0}
16.07.2019 11:28	msgid:3:0(	0	{"don": fa	{"ACTIVE": 1563265727000, "LATITUDE": 46.75285, "LONGITUDE": 38.64476333333333, "BUNKER_CAP": 9.0}
16.07.2019 11:28	msgid:3:0000000000(		{"don": fa	{"ACTIVE": 1563265698000, "BUNKER_CAP": 9.0}
16.07.2019 11:28	msgid:3:0000000000(		{"don": fa	{"ACTIVE": 1563265713000, "BUNKER_CAP": 9.0}
16.07.2019 11:29	msgid:3:0(	0	{"don": fa	{"ACTIVE": 1563265742000, "LATITUDE": 46.75285, "LONGITUDE": 38.64476333333333, "BUNKER_CAP": 9.0}
16.07.2019 11:29	msgid:3:0(	0	{"don": fa	{"ACTIVE": 1563265757000, "LATITUDE": 46.75285, "LONGITUDE": 38.64476333333333, "BUNKER_CAP": 9.0}
16.07.2019 11:29	msgid:3:0(	0	{"don": fa	{"ACTIVE": 1563265772000, "LATITUDE": 46.75285, "LONGITUDE": 38.64476333333333, "BUNKER_CAP": 9.0}
16.07.2019 11:29	msgid:3:0(	0	{"don": fa	{"ACTIVE": 1563265787000, "LATITUDE": 46.75285, "LONGITUDE": 38.64476333333333, "BUNKER_CAP": 9.0}
16.07.2019 11:30	msgid:3:0(	0	{"don": fa	{"ACTIVE": 1563265802000, "LATITUDE": 46.75285, "LONGITUDE": 38.64476333333333, "BUNKER_CAP": 9.0}
16.07.2019 11:30	msgid:3:0(	0	{"don": fa	{"ACTIVE": 1563265817000, "LATITUDE": 46.75285, "LONGITUDE": 38.64476333333333, "BUNKER_CAP": 9.0}
16.07.2019 11:30	msgid:3:0(	0	{"don": fa	{"ACTIVE": 1563265832000, "LATITUDE": 46.75285, "LONGITUDE": 38.64476333333333, "BUNKER_CAP": 9.0}

Рисунок 1. Фрагмент датасета в табличном представлении [рисунок авторов]

Figure 1. Fragment of a dataset in a tabular representation

С помощью функции библиотеки Matplotlib [9] была построена гистограмма (рисунок 2), где столбцы отражают процент пропущенных записей по параметру от общего числа наблюдений по параметру. Для признака speed значение показателя пропусков составило 0,14 %, для остальных признаков пропусков не обнаружено.

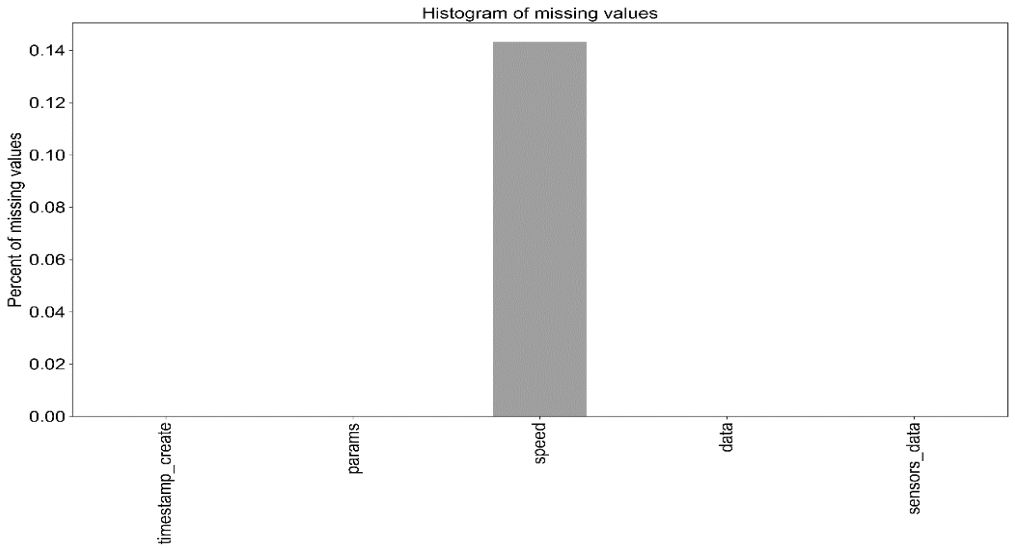

Рисунок 2. Гистограмма пропущенных значений записей в сырых данных [рисунок авторов]

Figure 2. Histogram of missing record values in raw data

Визуальное представление позволяет сделать вывод о том, что общий уровень пропусков в необработанных данных минимален. Важно учесть, что карта отслеживает присутствие или отсутствие записей, не оценивая их количественные характеристики. Для вычисления процентного соотношения пропущенных записей к общему числу записей применялась функция библиотеки Pandas, подсчитывающая пустые значения.

Следующим этапом предварительной обработки данных являлось преобразование строк формата JSON в формат «Признак — Запись». В результате преобразования датасет приведён к структурированному виду (рисунок 3).

timestamp_create	params	speed	data.don	data.mov	data.load	data.mode	data.speed	sensors_data.ACTIVE	sensors_data.LATITUDE	sensors_data.LONGITUDE
16.07.2019 11:27	msgid:3:0000000000000109215		ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265653000		
16.07.2019 11:27	msgid:3:0000000000000109214		ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265638000		
16.07.2019 11:27	msgid:3:0000000000000109216		ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265668000		
16.07.2019 11:28	msgid:3:0000000000000109217,VerSDM:3:2.3.3,VerLinux:3:3.2.0-4.1.15-1.1.1		ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265683000		
16.07.2019 11:28	msgid:3:0000000000000109220	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265727000	46,75285	38,64476333
16.07.2019 11:28	msgid:3:0000000000000109218		ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265698000		
16.07.2019 11:28	msgid:3:0000000000000109219		ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265713000		
16.07.2019 11:29	msgid:3:0000000000000109221	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265742000	46,75285	38,64476333
16.07.2019 11:29	msgid:3:0000000000000109222	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265757000	46,75285	38,64476333
16.07.2019 11:29	msgid:3:0000000000000109223	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265772000	46,75285	38,64476333
16.07.2019 11:29	msgid:3:0000000000000109224	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265787000	46,75285	38,64476333
16.07.2019 11:30	msgid:3:0000000000000109225	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265802000	46,75285	38,64476333
16.07.2019 11:30	msgid:3:0000000000000109226	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265817000	46,75285	38,64476333
16.07.2019 11:30	msgid:3:0000000000000109227	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265832000	46,75285	38,64476333
16.07.2019 11:30	msgid:3:0000000000000109228	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265847000	46,75285	38,64476333
16.07.2019 11:31	msgid:3:0000000000000109229	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265862000	46,75285	38,64476333
16.07.2019 11:31	msgid:3:0000000000000109230	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265877000	46,75285	38,64476333
16.07.2019 11:31	msgid:3:0000000000000109231	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265892000	46,75285	38,64476333
16.07.2019 11:31	msgid:3:0000000000000109232	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265907000	46,75285	38,64476333
16.07.2019 11:32	msgid:3:0000000000000109233	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265922000	46,75285	38,64476333
16.07.2019 11:32	msgid:3:0000000000000109234	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265937000	46,75285	38,64476333
16.07.2019 11:32	msgid:3:0000000000000109235	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265952000	46,75285	38,64476333
16.07.2019 11:32	msgid:3:0000000000000109236	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265967000	46,75285	38,64476333
16.07.2019 11:33	msgid:3:0000000000000109237	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265982000	46,75285	38,64476333
16.07.2019 11:33	msgid:3:0000000000000109238	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563265997000	46,75285	38,64476333
16.07.2019 11:33	msgid:3:0000000000000109239	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563266012000	46,75285	38,64476333
16.07.2019 11:33	msgid:3:0000000000000109240	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563266027000	46,75285	38,64476333
16.07.2019 11:34	msgid:3:0000000000000109241	0	ЛОЖЬ	ЛОЖЬ	0	M2	0	1563266042000	46,75285	38,64476333

Рисунок 3. Фрагмент структурированного датасета [рисунок авторов]

Figure 3. Fragment of a structured dataset

Данные, которые были представлены в номинальном виде, остались неизменными, а строки, содержащие неструктурированные данные, в формате JSON преобразованы в параметры и записи. Алгоритм преобразования построен следующим образом: в датасет добавлялся новый признак с записью, которой соответствует значение переменной в обрабатываемой строке JSON. Имя признака состоит из префикса — имени столбца, в котором находится обрабатываемая строка JSON, и суффикса — имени переменной в обрабатываемой строке JSON. Префикс и суффикс разделяются символом «точка». Для примера рассмотрим столбец data и переменную в строке JSON don. После операции преобразования в датасете появился новый признак data.don с записями, которым соответствовали значения переменной don в строке JSON. Аналогично выполнены преобразования остальных строковых переменных в последовательном порядке во избежание потери данных.

Для приведения записей к единой численной системе измерения была выполнена операция трансформации данных. Логические записи ЛОЖЬ или ИСТИНА преобразованы методом унитарного кодирования в записи с десятичными значениями: 0 и 1 соответственно.

Анализ записей, представленных в десятичной системе, позволил выявить дополнительные категории данных, по которым в дальнейшем можно осуществить

группировку параметров. Категорирование записей параметров, получаемых с датчиков, установленных на зерноуборочном комбайне, можно произвести по следующим признакам:

- Числовые значения, которые изменяются в течение временно́го периода нарастающим итогом с положительной или отрицательной динамикой (уровень топлива в баке, уровень загрузки бункера, остаток пробега до следующего ТО и т. п.).
- Числовые значения, которые характеризуют физические величины (скорость, температура, давление, величина зазора, частота оборотов и т. п.).
- Дискретные значения, которые фиксируют состояние «включено» или «выключено» для элементов, агрегатов или функциональных систем.

В результате преобразований записей и приведения их к единому виду целесообразно повторно рассмотреть датасет в целом на предмет наличия пропусков и аномалий.

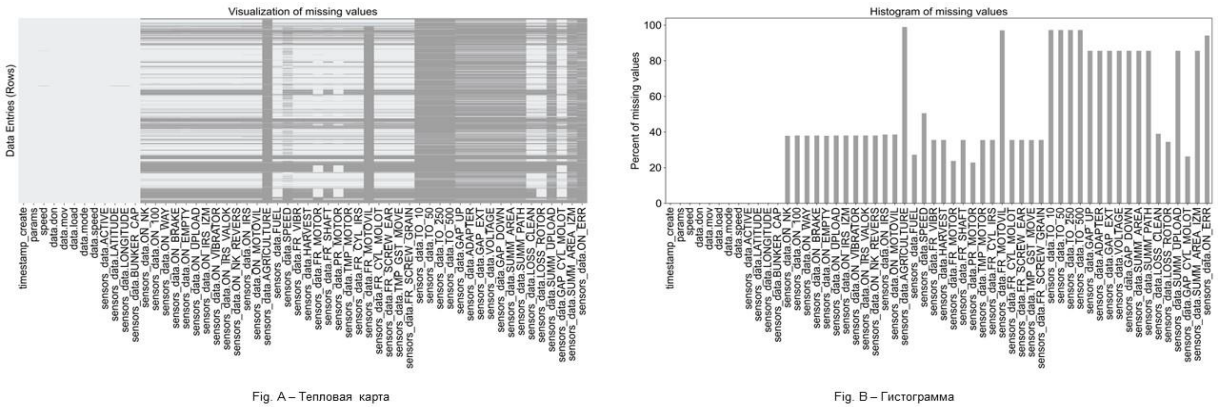

Рисунок 4. Тепловая карта и гистограмма пропущенных значений записей в структурированном датасете без обработки пропусков [рисунок авторов]

Figure 4. Heat map and histogram of missing record values in a structured dataset without gap processing

На рисунке 4 показана тепловая карта и гистограмма пропущенных значений записей структурированного датасета. Карта представлена в двумерной форме, где ось X визуализирует признаки параметров, а ось Y — строки данных наблюдений. Точка пересечения осей посредством различной окраски демонстрирует наличие или отсутствие записи. Светлая палитра обозначает наличие, в то время как более тёмные оттенки указывают на пропуск.

Данные гистограммы указаны в порядке представления их в датасете, для сохранения визуального сопоставления с тепловой картой. Визуальный анализ позволяет сделать вывод, о том, что по сравнению с предыдущей итерацией предварительной обработки данных количество пропущенных значений записей более существенно. Это связано с тем, что в процессе разбора текстовых строк в формате JSON было получено большее число индивидуальных параметров.

В текущей итерации предварительной обработки рассчитанный процент пропущенных значений составил 43,78 % от общего числа записей. Общее число пропущенных значений в датасете определено методом поиска и подсчёта пропущенных значений с помощью функции библиотеки Pandas. Далее определён процент пропусков от общего числа записей датасета.

3.2. Обработка пропусков

Для того чтобы правильно обработать пропуски, необходимо понимать механизм их формирования [10]. Рассмотрим датасет, согласно раннее определённой структуре, опираясь на категории данных (см. таблицу 1). Для упрощения идентификации категорий в дальнейшей работе каждой категории данных присвоены индексы (таблица 2). Произведём расчёт процента пропущенных значений внутри каждой категории. Для определения числа пропущенных значений по категориям определено количество пропусков по каждому наблюдаемому параметру, с помощью функций библиотеки Pandas:

- isnull(Paramn) — определение числа пропусков по параметру.
- sum(Catn) — суммарное значение количества пропусков по параметрам, входящим в каждую категорию: Cat1, Cat2, Cat3.

Таблица 2. Расчёт процента пропущенных значений по категориям

Table 2. Calculation of percentage of missing values by category

Индекс	Префикс и суффикс параметров	Число параметров	Процент пропущенных значений
Cat1	timestamp_create params speed	3	0,1
Cat2	data.*	5	0
Cat3	sensors_data.*	48	43,68

Согласно данным (см. таблицу 2), категории Cat1 и Cat2 практически не имеют пропусков. Опираясь на характер данных и метаданные датасета, можно сделать вывод о том, что на протяжении всего периода наблюдения система сбора и передачи данных работала исправно. В категории Cat3 наблюдается наибольший процент пропущенных значений — 43,68 %, обработка пропусков в данной категории представляет научный интерес, т. к. перечень параметров, входящих в эту категорию, характеризует техническое состояние узлов, механизмов и агрегатов зерноуборочного комбайна, а также позволяет судить о качестве выполнения агротехнологических операций.

Первоочередной задачей данного этапа предварительной обработки является выявление наблюдений, в которых отсутствуют записи по всем параметрам категории Cat3, данные

наблюдения могут быть безопасно удалены из датасета, т. к. не несут в себе информативную часть. Их наличие, возможно, связано с отсутствием сигналов или неисправностью подсистемы сбора и передачи данных с датчиков, установленных на зерноуборочном комбайне.

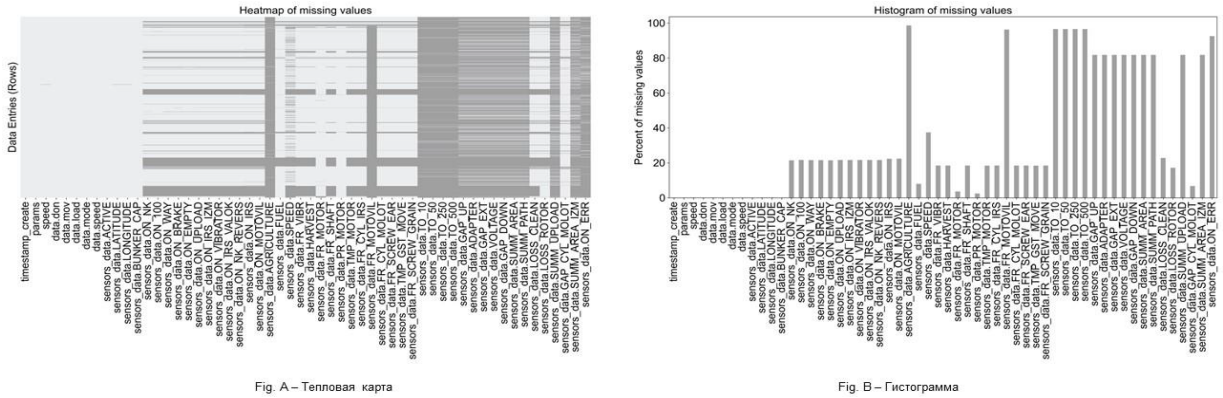

Рисунок 5. Тепловая карта и гистограмма пропущенных значений записей в структурированном датасете после удаления неинформативных наблюдений [рисунок авторов]

Figure 5. Heat map and histogram of missing record values in a structured dataset after removing uninformative observations

Анализируя тепловую карту и гистограмму (рисунок 5), можно сделать вывод о том, что логическая структура датасета после преобразования не изменилась, однако процент пропусков, согласно расчётам, уменьшился и составляет 34,57 % от общего числа записей в датасете. Распределение значений пропусков по периодическим параметрам многомерных временны*х* рядов носит циклический характер. Шаг параметра цикла равен периоду опроса датчиков измеряемых величин. Можно предположить, что данные пропуски формируются по принципам механизма MNAR [21, с. 35], — вероятность появления пропуска зависит в т. ч. от фактора, о котором мы ничего не знаем.

При атрибутировании пропусков MNAR использование методов удаления пропусков неприемлемо, т. к. это приведёт к значительному искажению распределения статистических свойств датасета.

Выбор метода заполнения пропусков зависит от значений записей в рассматриваемой категории. Рассчитав число пропусков по каждому наблюдаемому параметру, процент пропусков от общего числа наблюдений по параметру и период опроса датчика по рассматриваемому параметру (таблица 3), можно определить конкретное значение для заполнения пропуска параметра.

Таблица 3. Расчёт процента пропущенных значений по категориям

Table 3. Calculation of percentage of missing values by category

Параметр	Число пропусков по параметру, шт.	Процент пропусков от общего числа наблюдений по параметру	Период опроса датчика по параметру
timestamp_create	0	0,00	—
params	0	0,00	—
speed	161	0,17	15 с
data.don	0	0,00	15 с
data.mov	0	0,00	15 с
data.load	0	0,00	15 с
data.mode	0	0,00	15 с
data.speed	0	0,00	15 с
sensors_data.ACTIVE	0	0,00	15 с
sensors_data.LATITUDE	150	0,16	15 с
sensors_data.LONGITUDE	150	0,16	15 с
sensors_data.BUNKER_CAP	0	0,00	15 с
sensors_data.ON_NK	22888	23,92	15 с
sensors_data.ON_100	23077	24,11	15 с
sensors_data.ON_WAY	23013	24,05	15 с
sensors_data.ON_BRAKE	23027	24,06	15 с
sensors_data.ON_EMPTY	22900	23,93	15 с
sensors_data.ON_UPLOAD	23053	24,09	15 с
sensors_data.ON_IRS_IZM	23054	24,09	15 с
sensors_data.ON_VIBRATOR	23087	24,12	15 с
sensors_data.ON_IRS_VALOK	23054	24,09	15 с
sensors_data.ON_NK_REVERS	23102	24,14	15 с
sensors_data.ON_IRS	23736	24,80	15 с
sensors_data.ON_MOTOVIL	23752	24,82	15 с
sensors_data.AGRICULTURE	94488	98,73	Не определён
sensors_data.FUEL	10403	10,87	15 с
sensors_data.SPEED	37720	39,42	15 с
sensors_data.FR_VIBR	20113	21,02	15 с
sensors_data.HARVEST	20125	21,03	15 с
sensors_data.FR_MOTOR	6333	6,62	15 с
sensors_data.FR_SHAFT	20134	21,04	15 с
sensors_data.PR_MOTOR	5283	5,52	15 с
sensors_data.TMP_MOTOR	20076	20,98	15 с
sensors_data.FR_CYL_IRS	20130	21,03	15 с
sensors_data.FR_MOTOVIL	92270	96,42	15 с
sensors_data.FR_CYL_MOLOT	20125	21,03	15 с
sensors_data.FR_SCREW_EAR	20100	21,00	15 с
sensors_data.TMP_GST_MOVE	20049	20,95	15 с
sensors_data.FR_SCREW_GRAIN	20110	21,01	15 с
sensors_data.TO_10	92505	96,66	30 мин
sensors_data.TO_50	92505	96,66	30 мин
sensors_data.TO_250	92505	96,66	30 мин

sensors_data.TO_500	92505	96,66	30 мин
sensors_data.GAP_UP	78814	82,36	5 мин
sensors_data.ADAPTER	78845	82,39	5 мин
sensors_data.GAP_EXT	78815	82,36	5 мин
sensors_data.VOLTAGE	78814	82,36	5 мин
sensors_data.GAP_DOWN	78814	82,36	5 мин
sensors_data.SUMM_AREA	78826	82,37	5 мин
sensors_data.SUMM_PATH	78825	82,37	5 мин
sensors_data.LOSS_CLEAN	24173	25,26	15 с
sensors_data.LOSS_ROTOR	18931	19,78	15 с
sensors_data.SUMM_UPLOAD	78827	82,37	5 мин
sensors_data.GAP_CYL_MOLOT	9255	9,67	15 с
sensors_data.SUMM_AREA_IZM	78827	82,37	5 мин
sensors_data.ON_ERR	88825	92,82	Не определён

Периоды опроса датчиков измеряемых величин: 15 с (15″), 5 мин (5′), 30 мин (30′). Описательная статистика процента пропусков от общего числа наблюдений по параметру для каждого значения периода опроса представлена в таблице 4.

Таблица 4. Описательная статистика по процентам пропусков

Table 4. Descriptive Statistics for Missing Rates

Период опроса датчика по параметру	15″	5′	30′
Среднее	17,78	82,37	96,66
Медиана	21,02	82,37	96,66
Мода	0,00	82,36	96,66
Минимум	0,00	82,36	96,66
Максимум	96,42	82,39	96,66
Счёт	39,00	9,00	4,00

Проанализировав сводную таблицу 4 описательной статистики и категории Cat3 параметров, выявили закономерности частоты возникновения пропусков в зависимости от периода опроса датчиков измеряемых величин.

Наибольшее число параметров принадлежит периоду опроса 15 с, обозначим группу этих параметров как Gmissed15″ и построим гистограмму распределения величин по процентам пропусков в каждом параметре (рисунок 6).

Процент пропусков по параметрам в группе Gmissed15″ имеет большой разброс — от 0 до 96,42 %. Учитывая большой разрыв между максимальным и минимальным значениями процентов пропуска в группе Gmissed15″ (см. рисунок 6), оценка тенденции возникновения пропусков данных по среднему значению нецелесообразна. В данном случае в качестве меры оценки тенденции проявления пропусков сто́ит рассматривать значение медианы — Me15″ = 21,02 (см. таблицу 4). В большинстве наблюдений значение процента пропуска близко к значению 21,02 %. Из наблюдаемых параметров группы Gmissed15″

по семи параметрам наблюдается полнота данных (количество пропусков 0), три параметра имеют процент пропуска менее 1 % (0,16 %; 0,17 %), в девяти параметрах процент пропусков составил 20—22% — числовые значения, которые характеризуют физические величины (температура, давление, частота оборотов), 12 параметров (23—25 % пропусков) относятся к дискретным значениям, которые фиксируют состояние «включено» или «выключено» для элементов, агрегатов или функциональных систем.

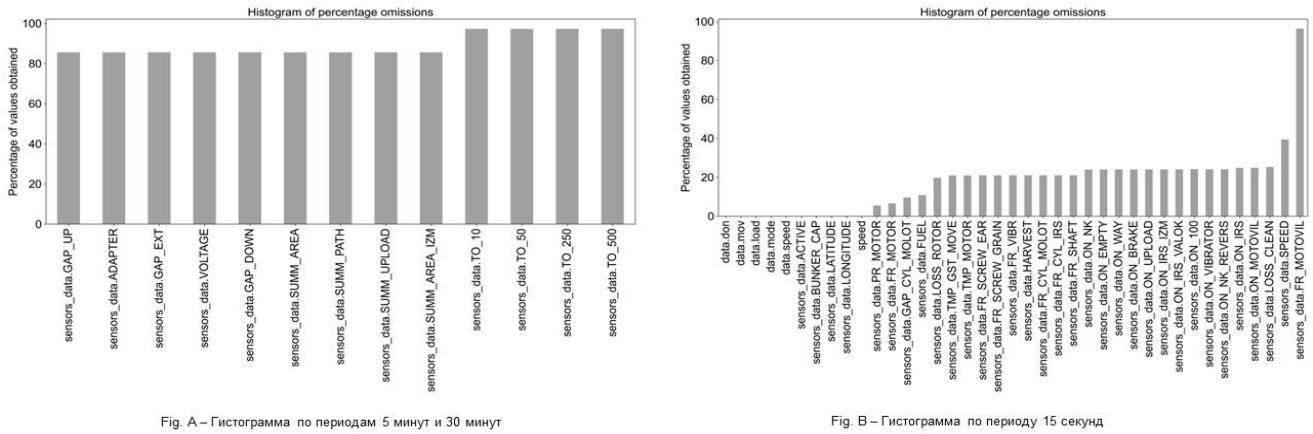

Рисунок 6. Гистограммы процентов пропусков по параметрам и периодам опроса [рисунок авторов]

Figure 6. Histograms of percentage of omissions by parameters and polling periods

Определить единый метод заполнения пропусков для всех параметров группы Gmissed15″ невозможно, выбор метода для каждого параметра будет зависеть от контекста, целей конкретного исследования и ряда иных факторов: к какой категории данных принадлежит параметр, физический смысл параметра, размерность параметра, принадлежность к совокупности генеральных диагностических параметров и т. д.

Группу показателей, принадлежащих к периоду опроса датчиков 5 мин, обозначим Gmissed5′. Количество параметров, принадлежащих к данной группе, 9, значения ключевых показателей описательной статистики равны, в качестве значения процента пропусков по каждому параметру в группе Gmissed5′ принимаем значение 82,37 % (см. таблицу 4). Наблюдается общая тенденция проявления пропусков данных по параметрам в группе Gmissed5′ (см. рисунок 6)

В рассматриваемой группе четыре параметра имеют значения, которые изменяются в течение временно́го периода нарастающим итогом с положительной динамикой. Метод заполнения пропусков в данном случае — по модели арифметической прогрессии, значение разности прогрессии следует определять по каждому периоду наблюдения. Оставшиеся пропуски параметров группы Gmissed5′ заполняются аналогично параметрам группы Gmissed15″.

Группу показателей, принадлежащих периоду опроса датчиков 30 мин, обозначим Gmissed30′ (см. рисунок 6). Параметры группы относятся к категории значений, которые изменяются в течение временно́го периода нарастающим итогом с отрицательной динамикой. Метод заполнения пропусков в данном случае — по модели арифметической прогрессии, значение разности прогрессии следует определять для каждого периода наблюдения.

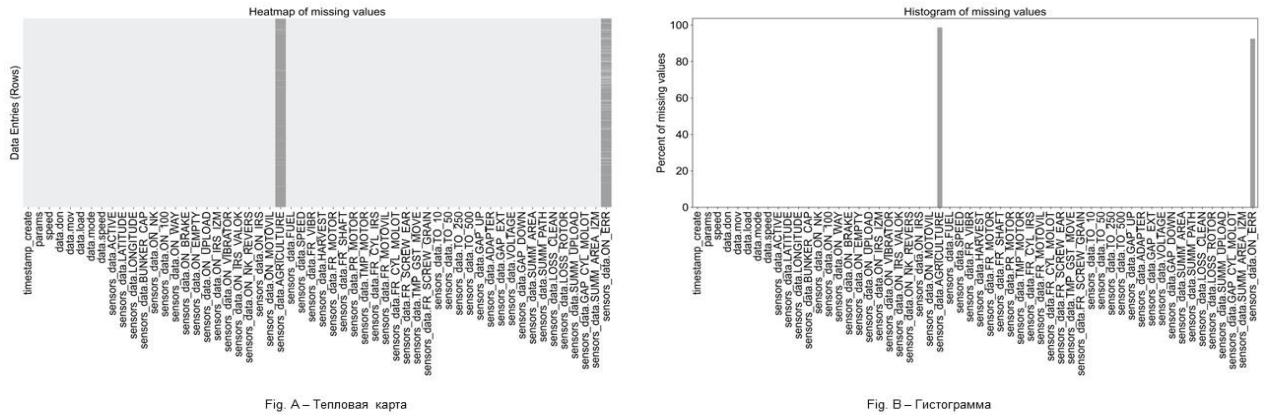

Рисунок 7. Тепловая карта и гистограмма пропущенных значений записей в структурированном датасете после заполнения пропусков, связанных с периодом опроса [рисунок авторов]

Figure 7. Heat map and histogram of missing record values in a structured dataset after filling in the gaps associated with the survey period

Анализируя тепловую карту и гистограмму (рисунок 7), можно сделать вывод о том, что процент пропусков значительно уменьшился и составляет 3,42 % от общего числа записей в датасете, при этом объём записей в датасете остался неизменным.

В датасете встречаются одиночные пропуски значений, не принадлежащие первым двум видам пропусков, определённых ранее:

- Пропуски, связанные с отсутствием сигналов или неисправностью подсистемы сбора и передачи данных с датчиков, установленных на зерноуборочном комбайне.
- Пропуски, связанные с периодами опросов датчиков.

Данные виды пропусков можно отнести к событийным. Алгоритм их обработки определяется исследователями исходя из контекста, характеристики параметра, связи с другими величинами.

Опираясь на метаданные датасета, определены алгоритмы обработки пропусков:

- Параметр sensors_data.AGRICULTURE — заполнение аналогичным значением на период до момента окончания технологической операции.
- Параметр sensors_data.ON_ERR — заполнение информативного значения с преобразованием методом унитарного кодирования в десятичную переменную.

Преобразованный датасет не имеет пропусков, информативен, объём записей в нём остался неизменным.

Для отслеживания динамики изменений основных характеристик датасета по итерациям данные сведены в итоговую таблицу 5.

Таблица 5. Количественные характеристики датасета по итерациям предварительной обработки

Table 5. Quantitative characteristics of the dataset by preprocessing iterations

Итерации предварительной подготовки данных	Исходный датасет	Структурированный датасет	Удалены пропуски	Заполнены пропуски	Обработаны одиночные пропуски
Количество записей, шт.	586225	6565720	5191312	5191312	5191312
Количество наблюдений, шт.	117245	117245	92702	92702	92702
Количество пропусков, шт.	168	2874267	1794402	177318	0
Количество пропусков, %	0,03	43,78	34,57	3,42	0

Анализируя итоговую таблицу 5, можно судить о готовности данных к последующим логическим преобразованиям и анализу.

В ходе детальной работы по предварительной обработке данных было проведено всестороннее исследование исходного датасета и реализован ряд преобразований с целью его оптимизации и улучшения качества данных. Полученный датасет представляет собой структурированный, информативный и консистентный материал, пригодный для дальнейших этапов анализа и моделирования [11].

Ключевые аспекты, достигнутые в результате проведённой обработки данных:

- Структурированность. Данные структурированы, обеспечивая лёгкость их восприятия и обработки в последующих этапах работы.
- Информативность. Благодаря проверке и уточнению данных датасет обрёл высокую степень информативности и точности, что повышает вероятность успешного применения в аналитических моделях.
- Отсутствие пропусков. Все пропущенные значения были идентифицированы, обработаны методами одномерной и многомерной обработки, что обеспечивает целостность и надёжность данных.
- Исключение выбросов и ошибок. Были определены и исключены аномальные значения и выбросы, улучшая тем самым общую надёжность и корректность данных.

- Категоризация данных. Для упрощения анализа и повышения интерпретируемости моделей в будущем данные категоризированы, что также способствует улучшению качества последующих аналитических процессов.

3.3. Процессная модель подготовки данных

Для формализации разработанного и апробированного способа подготовки данных и последующей обработки была построена процессная модель. Использование процессного подхода обусловлено тем, что он позволяет обеспечить целостность описания процесса подготовки данных с отражением функционального и информационного слоя [12]. Модель может рассматриваться как концепция информационной системы, позволяющей автоматизировать процесс подготовки данных систем контроля технического состояния зерноуборочных комбайнов для аналитической обработки.

В качестве инструмента формализации процессной модели была выбрана методология функционального моделирования IDEF0. Методология позволяет создавать модели, отображающие структуру и функции системы, а также потоки информации и материальных объектов, связывающих эти функции. Основу методологии IDEF0 составляет графический язык описания (моделирования) систем.

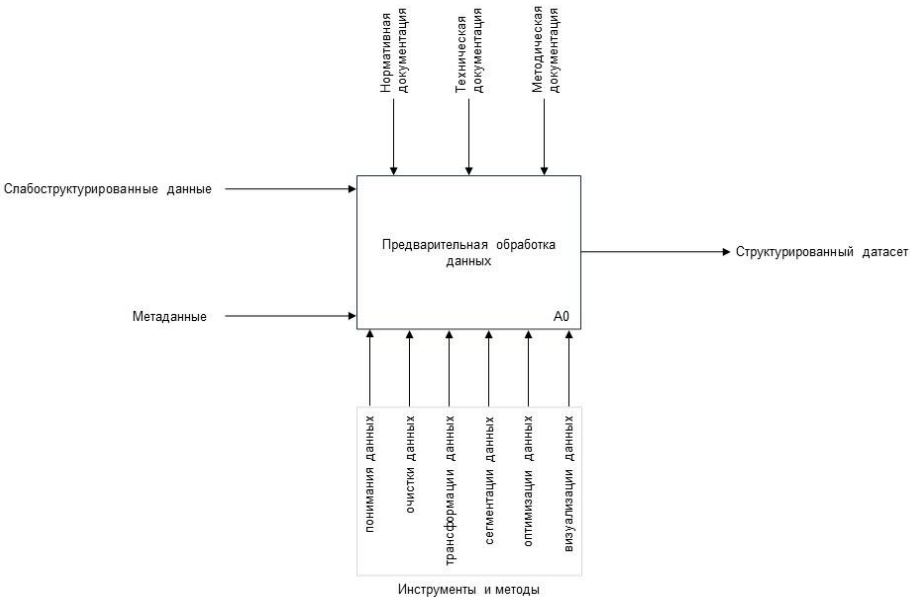

Рисунок 8. Процессная модель предварительной аналитической обработки данных систем контроля технического состояния зерноуборочных комбайнов в нотации IDEF0 [рисунок авторов]

Figure 8. Process model of preliminary analytical processing of data from systems for monitoring the technical condition of combine harvesters in IDEF0 notation

Представленная модель (рисунок 8) состоит из контекстной диаграммы A-0, которая описывает основную цель, с точки зрения исследователя данных, — произвести предварительную обработку данных и подготовить их к дальнейшему машинному анализу. Контекстная диаграмма взаимодействует со следующими компонентами модели:

- Входной сегмент. Модель на входе получает данные, которые в дальнейшем будут преобразованы. В нашем случае на вход модели подаются полуструктурированный датасет и метаданные.
- Сегмент управления. Механизмы модели действуют в соответствии с нормативной, технологической и методологической документацией.
- Механизм. Представляет собой совокупность инструментов, методов и действий по предварительной обработке данных.
- Выходной сегмент. Представляет собой преобразованный моделью объект. В нашем случае это структурированный датасет, полностью подготовленный к дальнейшему машинному анализу.

Для уточнения и алгоритмизации процесса предварительной обработки данных произведена декомпозиция контекстной диаграммы A-0 (рисунок 9).

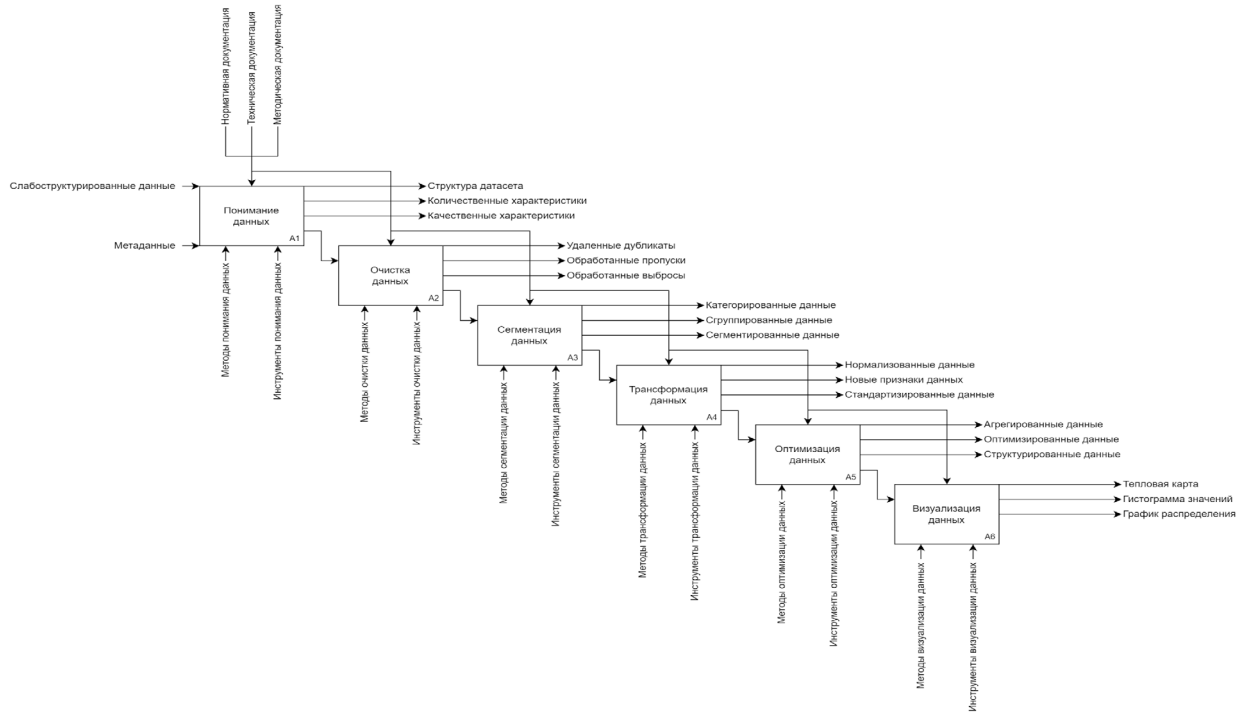

Рисунок 9. Декомпозиция контекстной диаграммы A-0 предварительной обработки данных в нотации IDEF0 [рисунок авторов]

Figure 9. Decomposition of the A-0 context diagram of data preprocessing in IDEF0 notation

Дочерние контекстные диаграммы A1-A6 описывают этапы предварительной обработки датасета. Каждый компонент декомпозированной модели детализирован по основным процессам. Выходная информация текущего этапа является входной для следующего этапа. Стоит отметить, что возможны итерации по этапам в зависимости от полученных результатов и целей этапа предварительной обработки.

Предварительная обработка данных — фундаментальный элемент любого аналитического проекта, успешная реализация этого этапа повышает вероятность достижения точных и надёжных результатов в ходе дальнейших исследований.

4. Обсуждение и заключение

В результате проведённого исследования установлено, что в том виде, в котором данные поступают напрямую с аналитических систем контроля технического состояния зерноуборочного комбайна, они не пригодны для анализа и прогнозирования состояния узлов и агрегатов. Прежде всего, это связано с большим количеством пропущенных значений. Так, для категории Cat3 (данные характеризуют техническое состояние узлов, механизмов и агрегатов зерноуборочного комбайна) в рассматриваемом датасете процент пропущенных значений составляет 43,68 %. С целью сохранения целостности датасета для последующего машинного анализа необходимо подобрать свои инструменты восстановления пропущенных значений для каждой категории данных. Определить единый метод заполнения пропусков для всех параметров невозможно, т. к. они относятся к разным категориям параметров, имеют разный физический смысл, различную размерность параметров и т. д.

Разработанный способ предварительной обработки данных позволил получить структурированность данных; информативность данных; отсутствие пропусков, выбросов и ошибок; категоризацию данных. Достижение таких характеристик позволяет заключить, что результаты исследования достигнуты, а предложенный способ предварительной обработки данных позволяет подготовить датасет для последующей машинной обработки.

Построенная процессная модель предварительной обработки данных обеспечивает прозрачность, контроль и оптимизацию процессов работы с данными, позволит исключить ошибки и противоречия в их дальнейшем анализе, а также обеспечит повторяемость действий в дальнейшем при обработке аналогичных датасетов, полученных с систем контроля технического состояния зерноуборочных комбайнов. Модель может рассматриваться как концепция информационной системы, позволяющей автоматизировать процесс подготовки данных систем контроля технического состояния зерноуборочных комбайнов для машинной обработки.

Список литературы

1. *Помогаев В. М.* Мониторинг технического состояния сельскохозяйственных машин и качества выполнения технологических операций // Вестник Омского ГАУ. 2023. № 2 (50). С. 143—152.
2. *Lüttenberg H., Bartelheimer C., Beverungen D.* Designing Predictive Maintenance for Agricultural Machines // Research papers. 2018. P. 153.
3. Возможности использования данных электронных систем сельскохозяйственных машин для построения предсказательных моделей / В. М. Помогаев, Г. В. Редреев, П. И. Ревякин [и др.] // Вестник Омского ГАУ. 2022. № 2 (46). С. 153—166. DOI: 10.48136/2222–0364_2022_2_153.
4. *Фомина Е. Е.* Сравнительный анализ методов восстановления пропусков в социологических исследованиях // Российский экономический вестник. 2021. Т. 4, № 1. С. 34—40.
5. *Гусев Д. И.* Алгоритм поиска ближайших соседей // Программные продукты и системы. 2012. № 3. С. 231—234.
6. *Богачев И. В.* Эвристический подход к оптимизации структуры кадров телеметрических данных для задачи сжатия // Международный научно-исследовательский журнал. 2022. № 8 (122). С. 1—6.
7. *Клевцов С. И.* Пороговая оценка состояния технического объекта на основе сегментации и идентификации модели контролируемого параметра // Известия ЮФУ. Технические науки. 2023. № 3 (233). С. 201—211.
8. *Вакуленко А. В., Кудрявцев Н. Г.* Визуализация данных с помощью Python и Java Script // Научный вестник Горно-Алтайского государственного университета: сб. ст. / Отв. ред. М. Г. Сухова. Горно-Алтайск: Горно-Алтайский гос. ун-т, 2023. № 17. С. 23—33.
9. *Ефремова А. Н., Полячкова М. А., Васильева Л. В.* Средства визуализации данных в скриптах на языке Python // Труды Братского государственного университета. Серия: Естественные и инженерные науки. 2019. Т. 2. С. 62—67.
10. *Шамрик Д. Л.* Базовые методы восстановления пропусков в массивах данных // Информационные технологии в науке и производстве: Материалы V Всерос. молодёж. научно-техн. конф., Омск, 25—26 апр. 2018 года. Омск: Омский гос. техн. ун-т, 2018. С. 73—83.
11. *Маковейчук Я. Т.* Поиск, анализ и подготовка данных для применения методов машинного обучения // Дистанционные образовательные технологии: сб. тр. VIII Междунар. научно-практич. конф., Ялта, 19—21 сент. 2023 года. Симферополь: ООО «Издательство, Типография «Ариал», 2023. С. 218—222.
12. *Van der Aalst W.* Data Science in Action // In: Process Mining. 2016. P. 3—23. DOI: 10.1007/-978-3-662-49851-4_1.

References

1. Pomogaev V. M. Monitoring the technical condition of agricultural machines and the quality of technological operations. *Bulletin of Omsk State Agrarian University*, 2023, no. 2 (50), pp. 143—152. (In Russ.)
2. Lüttenberg H., Bartelheimer C., Beverungen D. Designing Predictive Maintenance for Agricultural Machines. *Research papers*, 2018, p. 153.
3. Pomogaev V. M., Redreev G. V., Revyakin P. I., Basakina A. S. Possibilities of using data from electronic systems of agricultural machines to build predictive models. *Bulletin of Omsk State*

Agrarian University, 2022, no. 2 (46), pp. 153—166. doi: 10.48136/2222–0364_2022_2_153. (In Russ.)
4. Fomina E. E. Comparative analysis of methods for recovering passes in sociological research. *Russian Economic Bulletin*, 2021, vol. 4, no. 1, pp. 34—40. (In Russ.)
5. Gusev D. I. Algorithm for searching for nearest neighbors. *Software products and systems*, 2012, no. 3, pp. 231—234. (In Russ.)
6. Bogachev I. V. Heuristic approach to optimizing the structure of telemetry data frames for the compression task. *International scientific research journal*, 2022, no. 8 (122), pp. 1—6. (In Russ.)
7. Klevtsov S. I. Threshold assessment of the state of a technical object based on segmentation and identification of the model of the controlled parameter. *Izvestiya SFU. Technical science*, 2023, no. 3 (233), pp. 201—211. (In Russ.)
8. Vakulenko A. V., Kudryavtsev N. G. Data visualization using Python and Java Script. *Scientific Bulletin of the Gorno-Altai State University: collection of articles. executive editor*, *2023*. Gorno-Altaisk, Gorno-Altaisk State University, no. 17, pp. 23—33. (In Russ.)
9. Efremova A. N., Polyachkova M. A., Vasilyeva L. V. Data visualization tools in scripts in Python. *Proceedings of Bratsk State University. Series: Natural and engineering sciences*, 2019, vol. 2, pp. 62—67. (In Russ.)
10. Shamrik D. L. Basic methods for restoring gaps in data arrays. *Information technologies in science and production: materials of the V All-Russian Youth Scientific and Technical Conference, Omsk, April 25—26, 2018*. Omsk, Omsk State Technical University, 2018, pp. 73—83. (In Russ.)
11. Makoveychuk Ya. T. Search, analysis and preparation of data for the application of machine learning methods. *Distance educational technologies: collection of proceedings of the VIII International Scientific and Practical Conference, Yalta, September 19—21, 2023.* Simferopol, Limited Liability Company «Publishing House Typography «Arial», 2023, pp. 218—222. (In Russ.)
12. Van der Aalst W. Data Science in Action. *In: Process Mining*, 2016, pp. 3—23. doi: 10.1007/-978-3-662-49851-4_1.

© Помогаев В. М., Ревякин П. И., Басакина А. С., 2024