



УДК 629.33

## ФОРМИРОВАНИЕ ПОДХОДА К ОБРАБОТКЕ СТАТИСТИЧЕСКИХ ДАННЫХ ДВИЖЕНИЯ ТРАНСПОРТНЫХ СРЕДСТВ

Поташников М.Д.<sup>1</sup>, Муравьев А.С.<sup>1</sup>, Карташов А.Б.<sup>1</sup>, Пикалов Н.А.<sup>1</sup>, Назаренко С.В.<sup>2</sup>

<sup>1</sup> МГТУ им. Н.Э. Баумана

<sup>2</sup> ПАО «КАМАЗ»



### Информация о статье

Поступила:

5 августа 2023 г.

Рецензирование:

28 августа 2023 г.

Принята к печати:

11 сентября 2023 г.

**Ключевые слова:** анализ данных, машинное обучение, ездовые циклы транспортных средств, навигационное оборудование, статистика

### Аннотация.

В статье описан процесс анализа и обработки статистических данных о движении грузового транспорта, полученных с помощью навигационного оборудования, установленного на рассматриваемых транспортных средствах. Проведен подробный анализ исходных данных, а также все стадии их обработки с использованием специализированных пакетов инструментов языка Python, таких как Pandas, NumPy и Sklearn. Показана эффективность предложенного подхода к обработке данных путем сравнения ездовых циклов транспортных средств, сгенерированных на основе исходных и модифицированных данных. Тема анализа и обработки данных не является новой, однако актуальность статьи состоит в успешном применении распространенных и эффективных методов обработки статистических данных о движении транспортных средств, работа с которыми осложнена тем, что они получены с помощью широко распространенного, но ориентированного на другие задачи оборудования. Таким образом, работа открывает новые возможности к статистическому анализу движения грузового транспорта без необходимости замены навигационного оборудования.

**Для цитирования:** Поташников М.Д., Муравьев А.С., Карташов А.Б., Пикалов Н.А., Назаренко С.В. Формирование подхода к обработке статистических данных движения транспортных средств // Техника и технология горного дела. – 2023. – №3(23). – С. 34-51. – DOI:10.26730/2618-7434-2023-3-34-50, EDN: CJEEHC

### Введение

В современном мире работа с большими данными является неотъемлемой частью многих областей науки и производства, с каждым годом анализ данных становится необходим для решения все новых и новых задач. Машиностроение не является исключением: большинство современных проектных и поверочных расчетов и виртуальных имитационных испытаний проводится на основе анализа статистических данных [1].

В работах [2-5] показано, как можно применять статистические данные для генерации ездовых циклов движения, соответствующих условиям движения в разных регионах, с целью их дальнейшего применения при определении нагрузочных режимов для трансмиссионных узлов и агрегатов [6-12]. В перспективе такой подход может значительно улучшить приспособленность автотранспорта к конкретным условиям эксплуатации и существенно увеличить его ресурс.

Для мониторинга транспортных средств и сбора статистических данных применяют различное телематическое оборудование, представляющее собой навигационные терминалы, подключенные к бортовой CAN-шине автомобиля и оборудованные ГЛОНАСС/GPS приемниками. Основной проблемой при использовании статистической информации,



полученной с помощью данного типа оборудования, является ее низкое качество: наборы полученных данных имеют большое количество пропущенных значений, где-то встречается достаточно низкое разрешение измерений или относительно большой интервал между измерениями.

Стоит отметить, что получаемые данные удовлетворяют требованиям, предъявляемым к результатам, получаемым с помощью данного типа оборудования, и в полной мере обеспечивают решение задач повышения эффективности эксплуатации автопарков и сокращения издержек предприятий в различных отраслях, таких как грузоперевозки, пассажирский транспорт, добыча полезных ископаемых, строительство и во многих других [13-16]. Но для решения задач анализа данных качества получаемой информации недостаточно. Указанную проблему возможно решить двумя способами – за счет использования дорогостоящего специализированного оборудования или за счет применения различных алгоритмов обработки данных. Применение специализированного оборудования для большого количества транспортных средств затруднительно как с финансовой точки зрения, так и по причине необходимости периодического сопровождения измерений техническим специалистом. Поэтому наиболее предпочтительным вариантом является применение различных алгоритмов обработки, позволяющих преобразовать статистические данные движения транспортных средств к виду, удовлетворительному для дальнейшего использования при генерации ездовых циклов и проведении виртуальных испытаний транспортных средств.

#### **Анализ исходных данных**

В данной работе используются статистические данные о движении грузовых автомобилей с максимальной массой, приходящейся на ось, равной 11,5 тонн, в частности, магистральных тягачей КАМАЗ-54901 и грузовых автомобилей КАМАЗ-5325. Данные собирались с 29 машин, временной период записей варьируется от 1,5 месяцев до 1 года. Набор данных содержит в себе следующую информацию (столбцы):

- «time» – время в формате ГГ-ММ-ДД ЧЧ: ММ:СС;
- «latitude» и «longitude» – координаты транспортного средства в текущий момент времени;
- «altitude» – высота транспортного средства над уровнем моря;
- «speed» – скорость транспортного средства, вычисленная на основе данных с GPS-приемника, выраженная в км/ч;
- «axle\_weight» – вес, приходящийся на ведущую ось транспортного средства;
- «high\_resolution\_total\_vehicle\_distance» – пробег транспортного средства, выраженный в километрах с точностью до метра;
- «engine\_speed» – скорость вращения ДВС;
- «high\_resolution\_engine\_total\_fuel\_used», «engine\_total\_fuel\_used» – расход топлива;
- wheel\_based\_vehicle\_speed – скорость транспортного средства, рассчитанная на основе сигналов с датчиков частоты вращения колес.

В наборе данных достаточно большое количество записей, более 100 млн, однако далеко не все они являются информативными из-за большого количества пропущенных значений измеряемых параметров. В Таблице 1 указан процент пропущенных значений от общего количества измерений соответствующих параметров.

В обыкновенной практике анализа данных и машинного обучения [17] принято считать, что если в столбце более 5-10% пропущенных значений, то такие данные не подлежат восстановлению и являются неинформативными [18]. Но в некоторых случаях наборы данных с большим количеством пропущенных значений также могут быть полезны и использование их при анализе может оказать существенное влияние на конечный результат.



Таблица 1. Количество пропущенных значений в наборе данных  
Table 1. Number of missing values in the dataset

Название столбца	Процент пропущенных значений, %
time	0 %
latitude	0 %
longitude	0 %
altitude	11 %
speed	11 %
axle_weight	86.4 %
engine_speed	32.9 %
high_resolution_engine_total_fuel_used	52.8 %
high_resolution_total_vehicle_distance	41.3 %
engine_total_fuel_used	99.3 %
wheel_based_vehicle_speed	78.4 %

При обработке в первую очередь необходимо избавиться от данных, которые не представляют интереса при анализе или являются неинформативными и не имеют возможности восстановления. При формировании нагрузочного режима для агрегатов электромеханических трансмиссий неинформативными являются значения частоты вращения ДВС и расхода топлива («engine\_speed», «high\_resolution\_engine\_total\_fuel\_used», «engine\_total\_fuel\_used»). Также в полученных данных значения скорости транспортного средства, измеренной на основании датчиков частоты вращения колес («wheel\_based\_vehicle\_speed»), имеют большое количество пропусков, восстановить которые не представляется возможным. Набор данных, с которым будут производиться дальнейшие манипуляции, представлен на Рис. 1.

	time	latitude	longitude	altitude	speed	axle_weight	high_resolution_total_vehicle_distance
1010	2021-05-31 05:11:50	52.576963	39.453584	178.0	63.0	NaN	NaN
1011	2021-05-31 05:11:50	52.576963	39.453584	178.0	63.0	NaN	53736.925
1012	2021-05-31 05:11:52	52.576656	39.453482	177.0	63.0	9.5	53736.940
1013	2021-05-31 05:11:52	52.576656	39.453482	177.0	63.0	NaN	53736.955
1014	2021-05-31 05:11:54	52.576350	39.453384	177.0	63.0	NaN	NaN
1015	2021-05-31 05:11:54	52.576350	39.453384	177.0	63.0	9.6	53736.975
1016	2021-05-31 05:11:54	52.576350	39.453384	177.0	63.0	9.5	53736.990
1017	2021-05-31 05:11:56	52.576042	39.453282	177.0	63.0	NaN	53737.010
1018	2021-05-31 05:11:56	52.576042	39.453282	177.0	63.0	NaN	53737.025
1019	2021-05-31 05:11:58	52.575733	39.453172	177.0	64.0	NaN	NaN
1020	2021-05-31 05:11:58	52.575733	39.453172	177.0	64.0	9.6	53737.045
1021	2021-05-31 05:11:58	52.575733	39.453172	177.0	64.0	9.5	53737.060
1022	2021-05-31 05:12:00	52.575426	39.453050	177.0	63.0	NaN	NaN
1023	2021-05-31 05:12:00	52.575426	39.453050	177.0	63.0	9.6	53737.080

Рис. 1. Фрагмент набора данных  
Fig. 1. Data set fragment



### Обработка данных скорости, высоты и нагрузки на ось

Устранить пропущенные значения без какой-либо корректировки данных возможно в столбцах, содержащих значения скорости и нагрузки на ось. При удалении строк набора данных важно следить за тем, что при этом не теряется никакая полезная информация.

Анализируя исходный набор данных (Рис. 1), можно отметить, что пропуски значений в столбце «speed» появляются в момент отключения навигационного контроллера. В других ситуациях пропущенных значений в данном столбце не возникает. Также оказалось, что в столбцах «altitude» и «speed» одинаковое количество пропущенных значений, из чего можно сделать вывод, что они появляются в одно и то же время – когда транспортное средство остановлено и происходит процесс выключения питания. Участки данных, в которых транспортное средство стоит на месте, не представляют интереса для анализа, поэтому строки с пропущенными значениями в столбцах скорости и высоты были удалены из набора данных.

Анализ столбца с информацией о нагрузке на ось показал, что в нем содержится 86,4% пропущенных значений (Таблица 1), однако полностью убирать ее из рассмотрения нежелательно, так как она полезна при большом количестве расчетов, к примеру, в расчете тормозного момента транспортного средства. Посмотрев на данные в столбце нагрузки на ось (Рис. 1), можно заметить, что датчик, измеряющий нагрузку на ось, присылает новое значение, только когда нагрузка на ось изменяется. Соответственно, пропущенные значения сигнализируют о том, что изменений в нагрузке на ось не происходит, то есть пропущенные значения в этом столбце несут в себе информацию о том, что значение нагрузки на ось не изменилось, поэтому удалять их нельзя. Для избавления от пропусков достаточно каждое пропущенное значение в столбце заменить на ближайшее верхнее непропущенное значение, в таком случае потери информации не будет.

В результате выполнения вышеперечисленных действий в наборе данных были удалены все пропущенные значения, за исключением столбца с пробегом.

Еще одной важной характеристикой при расчетах и виртуальных испытаниях является текущее ускорение транспортного средства. Для примерного расчета моментального ускорения достаточно знать разницу скоростей и разницу во времени между двумя соседними записями в наборе данных, то есть в расчете используется следующая формула:

$$a = \frac{\Delta V}{\Delta t}, \quad (1)$$

где  $a$  – текущее ускорение транспортного средства,  $\Delta V$  – разница скоростей между двумя записями набора данных,  $\Delta t$  – разница во времени между двумя записями набора данных.

### Обработка данных пробега и времени

Анализ временных данных стоит производить исходя из трех столбцов: времени, текущей скорости и пробега. Рассмотрим небольшой фрагмент данных (Рис. 1).

В фрагменте данных на рисунке видно сразу две проблемы: пропущенные данные в столбце пробега и повторяющиеся данные в столбце времени. Предположительно, такая ситуация вызвана плохой синхронизацией измерительных устройств, а также возможной потерей или пренебрежением десятой долей секунд при записи текущего времени.

Для решения указанной проблемы можно удалить все повторяющиеся строки в столбце времени, при этом сохраняя строки, в которых не пропущено значение пробега, а затем заполнить небольшое количество оставшихся пропусков в столбце пробега, добавляя к предыдущему значению пробега расстояние, полученное исходя из текущей скорости и разницы во времени в соответствующих строках. Однако такой вариант нежелателен, так как будет потеряно большое количество полезной информации. В таком случае будет удалено примерно 60% всех строк в наборе данных, а также будет потеряно значительное количество информации о пробеге, так как в строках с повторяющимися временными данными зачастую содержатся разные значения пробега (1017 и 1018 строки на Рис. 1). Более того, будет потеряно много временной информации, так как после вышеописанной обработки будут часто появляться



строки, значение времени в которых различаются на 2-3 секунды, что, очевидно, не соответствует действительности.

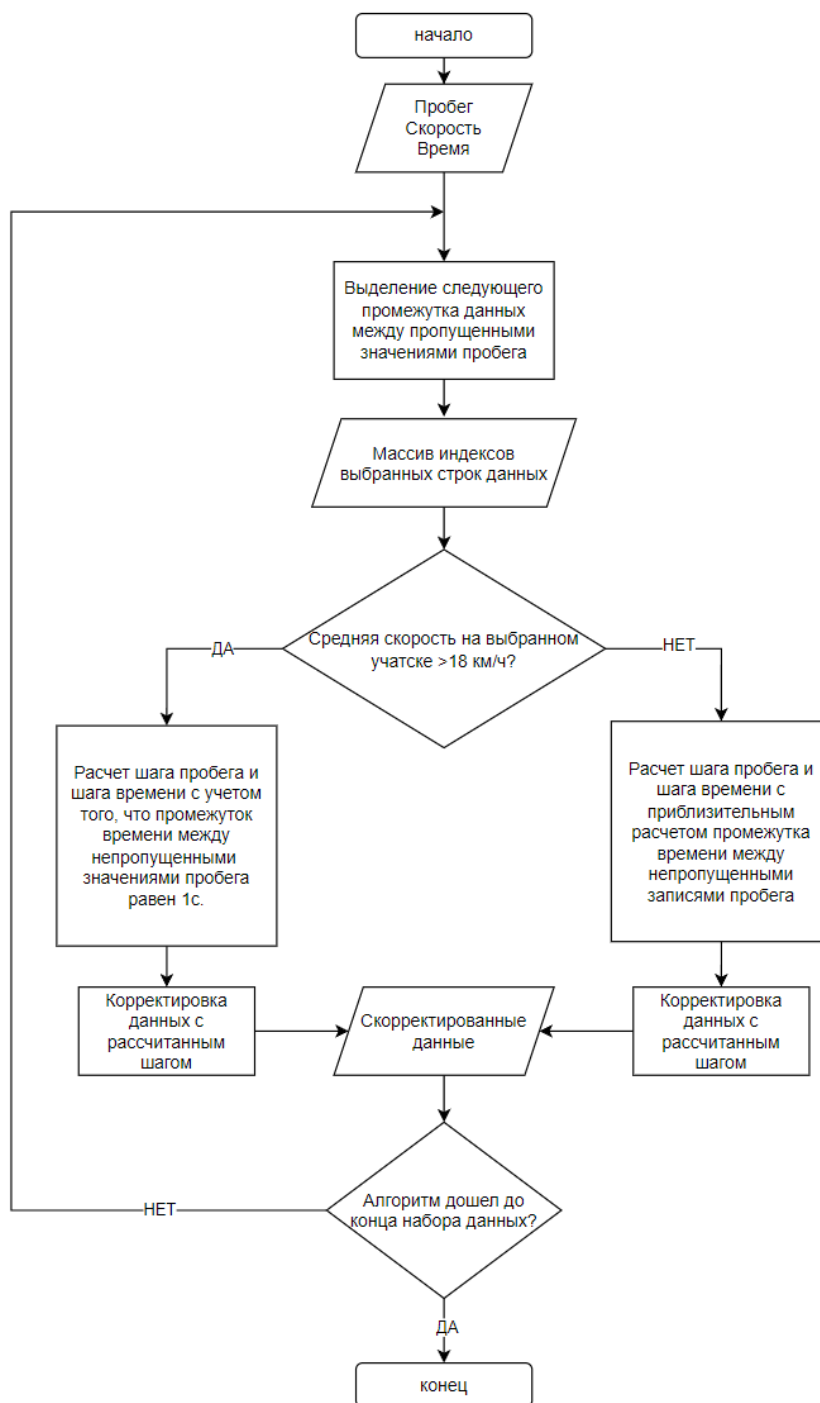


Рис. 2. Блок-схема алгоритма обработки столбцов пробега и времени  
Fig. 2. Flow diagram of the algorithm for processing mileage and time columns

Эффективнее будет рассматривать и обрабатывать участки с повторяющимися временными данными по отдельности, не удаляя строки из набора данных, а корректируя их. При детальном анализе данных был получен вывод о том, что устройство измерения пробега присылает значения раз в секунду при условии, что скорость транспортного средства больше 18 км/ч или 5 м/с. Если транспортное средство за секунду проезжает менее 5 метров, то счетчик пробега не



обновляется и в таком случае не присылает значение. В связи с этим алгоритм обработки должен разделяться на две части: для участков данных, где скорость больше 18 км/ч и для участков, где скорость меньше 18 км/ч. На основании этого решения был разработан соответствующий алгоритм, блок-схема которого представлена на Рис. 2.

Опорным значением в таком случае будет значение пробега. Если посмотреть на динамику изменения пробега, можно заметить, что он изменяется с точностью до 5 метров, а также то, что в фрагменте на Рис. 1 его значения изменяются на 15-20 метров каждую итерацию при постоянной скорости 62-63 км/ч. Предполагая, что устройство измерения пробега присылает данные с фиксированным интервалом, можно рассчитать этот интервал путем вычислений по формуле 2.

$$\Delta t = \frac{\Delta S}{V_{\text{тек}}} = \frac{17.5}{62.5/3.6} \approx 1 \text{ с.}, \quad (2)$$

где  $\Delta t$  – интервал времени между записями в столбце пробега,  $\Delta S$  – путь, пройденный за этот интервал (разница между соседними непропущенными значениями в столбце пробега),  $V_{\text{тек}}$  – соответствующая текущая скорость транспортного средства.

В результате работы алгоритма, были устранены все пропущенные значения пробега и повторяющиеся временные данные без существенных потерь и искажений данных.

### Обработка данных высоты

Обращаясь к столбцу «altitude» на Рис. 1, можно выделить, что высота изменяется «ступенчато», и для последующей обработки графика это ступенчатое изменение необходимо сгладить. При разработке алгоритма сглаживания значений высоты было принято допущение, что на участках, где значения высоты повторяются, она на самом деле изменяется линейно. То есть работа алгоритма состоит в поиске участков повторяющихся данных и замене их на линейно-возрастающие или линейно-убывающие, в зависимости от первого неповторяющегося значения после каждого участка. Результат работы алгоритма сглаживания продемонстрирован на Рис. 3 а) и 3 б).

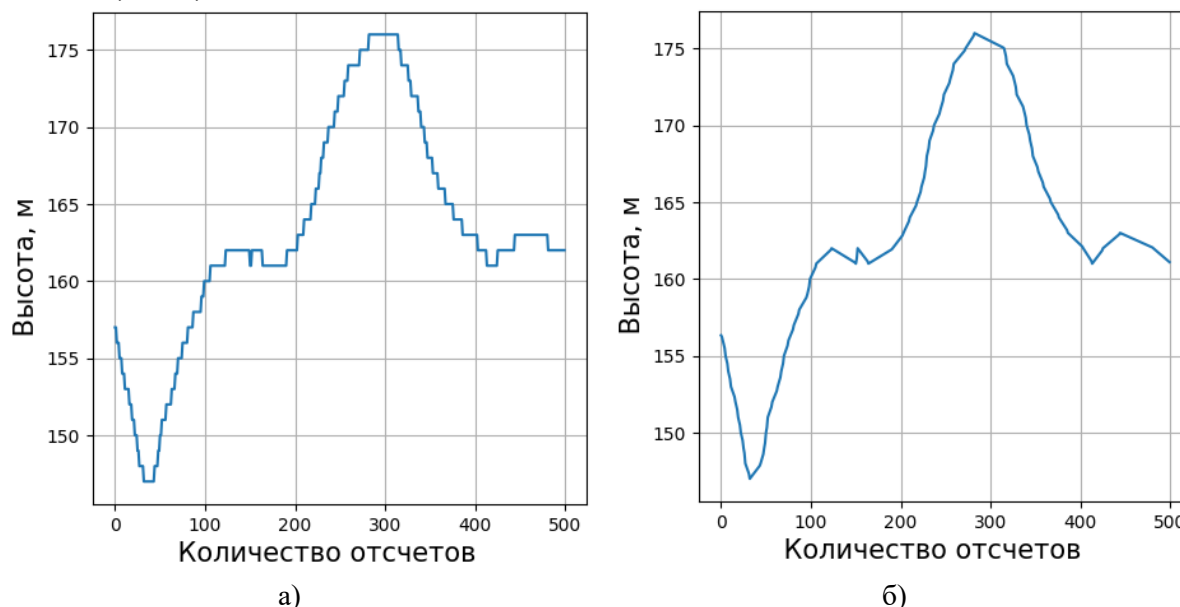


Рис. 3. Графики зависимости высоты ТС от времени: а – до обработки высотных данных; б – после обработки высотных данных

Fig. 3. Graphs of the dependence of the vehicle height on time: a – before processing altitude data; b – after processing altitude data

### Расчет угла уклона дороги

Исходя из имеющихся данных, есть возможность рассчитать уклон дороги в текущий момент времени. Значение уклона является важным параметром, к примеру, при расчете нагрузок на трансмиссионные узлы и агрегаты [7]. Благодаря имеющимся качественным данным пробега транспортного средства и высоты (обработанных в предыдущих пунктах) уклон дороги можно вычислить по формуле 3, аналогично тому, как подобные расчеты проводились в работе [19].

$$\alpha = \arcsin\left(\frac{\Delta h}{\Delta S}\right), \quad (3)$$

где  $\alpha$  – уклон дороги в текущий момент,  $\Delta h$  – разница высоты между двумя записями в наборе данных,  $\Delta S$  – разница пробега между двумя записями в наборе данных.

Существует альтернативный вариант расчета уклона дороги с использованием значений долготы и широты транспортного средства – используя формулу Хаверсина (формула 4). В случае использования этой формулы нет необходимости проводить обработку данных пробега транспортного средства, так как в формуле он не используется.

$$\sin(\alpha) = \frac{\Delta h}{2R \arcsin\left(\sqrt{\sin^2\frac{\Delta\varphi}{2} + \cos(\varphi_1)\cos(\varphi_2)\sin^2\frac{\Delta\lambda}{2}}\right)}, \quad (4)$$

где  $\Delta\varphi = \varphi_2 - \varphi_1$  – разница широт,  $\Delta\lambda = \lambda_2 - \lambda_1$  – разница долгот,  $R$  – радиус Земли.

Проведем сравнение двух расчетных уклонов дороги. На Рис. 4а представлен уклон, посчитанный по формуле (2) с использованием столбца реального пробега транспортного средства, а на Рис. 4б – уклон, посчитанный с использованием формулы (3).

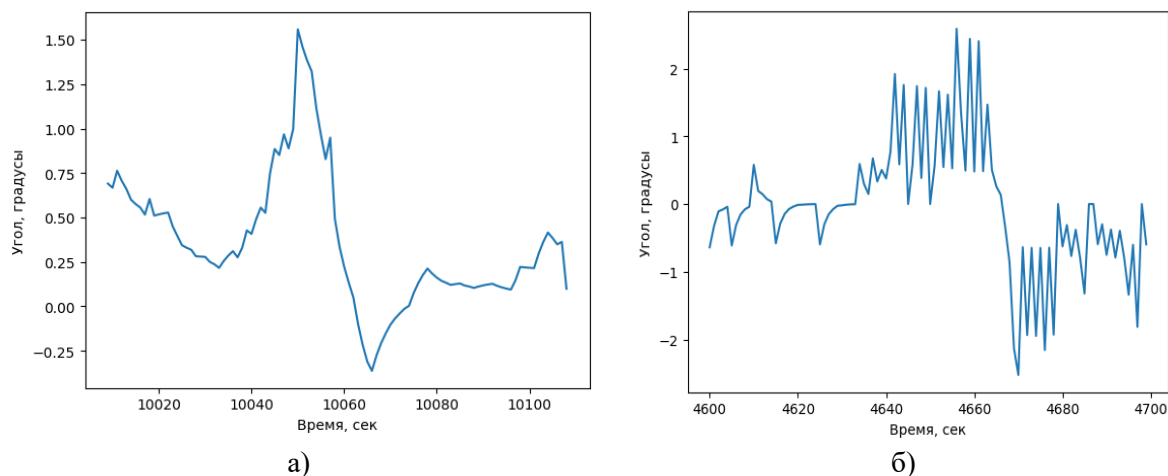


Рис. 4. График зависимости уклона дороги от времени: а – полученный на основе реальных записей пробега; б – полученные на основе расчетов по формуле Хаверсина  
Fig. 4. Graph of the dependence of the road slope on time: a – obtained on the basis of real mileage records; b – obtained on the basis of calculations using the Haversin formula

По графикам видно, что угол, рассчитанный на основе обработанных записей пробега, изменяется более плавно, имеет значительно меньше колебаний, что является существенным преимуществом по сравнению с данными об уклоне, рассчитанными с использованием формулы Хаверсина. Формула 4 дает такой результат вследствие того, что данные широты и долготы в наборе данных являются также недостаточно точными и их значения имеют колебательный характер, который отражается на расчете уклона. С этим можно бороться путем фильтрации значений уклона с помощью, к примеру, фильтра Калмана или скользящего среднего. Однако



отфильтрованные данные все равно будут уступать расчету по формуле 3, а также будут смещены по времени из-за особенностей используемых фильтров.

### Преобразование данных для использования в методе генерации ездового цикла.

В работе [6] показано, что на основе статистических данных о движении можно генерировать качественные ездовые циклы транспортных средств, однако в ней не раскрывается, каким образом данные предварительно обрабатываются для того, чтобы ездовой цикл удовлетворял требованиям сходимости по ряду параметров с полной реальной записью движения транспортного средства. В текущем разделе показано, что правильная подготовка данных для таких алгоритмов зачастую является определяющим фактором успеха.

Для генерации циклов будет использован метод «микропоездок» с применением кластеризации [6]. Соответственно, из набора данных необходимо выделить «микропоездки» и далее присвоить им определенные критерии, аналогично работе [6], по которым они будут выбираться для участия в итоговом ездовом цикле. Для выделения отдельной микропоездки достаточно найти в общем наборе данных индексы ее начала и конца. Блок схема разработанного алгоритма выделения «микропоездок» представлена на Рис. 5.

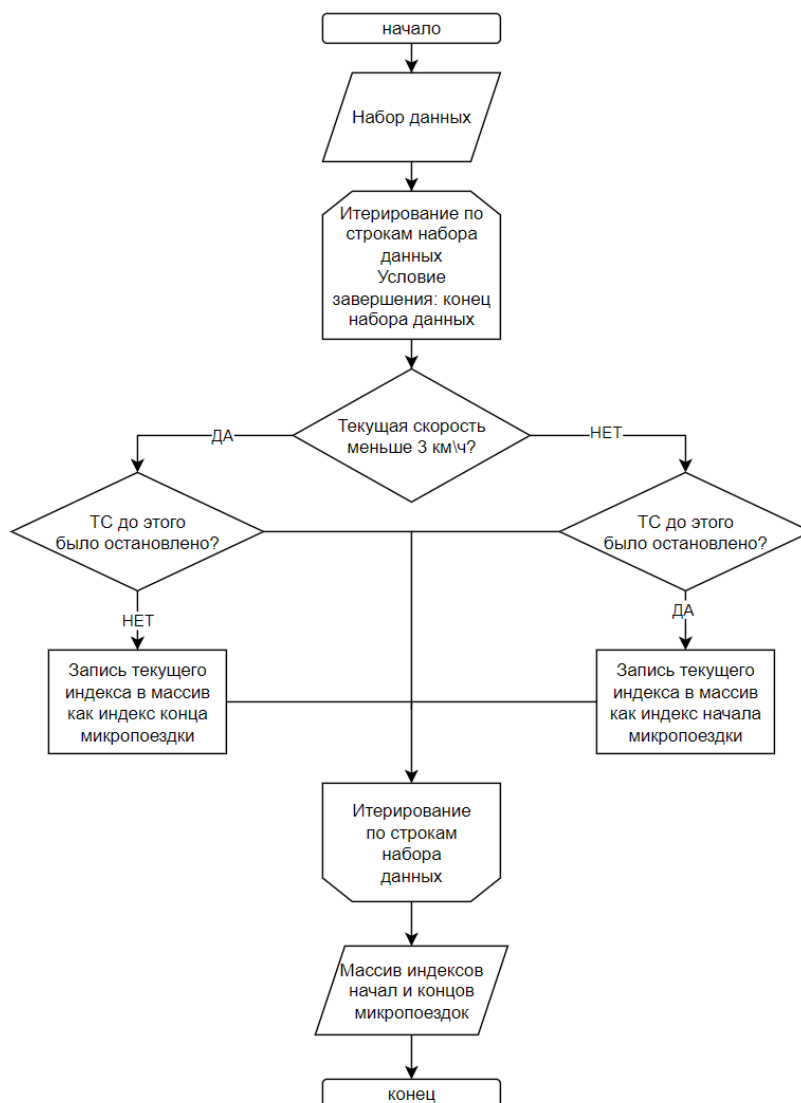


Рис. 5. Блок-схема алгоритма выделения «микропоездок» из набора данных  
Fig. 5. Block diagram of the algorithm for selecting «micro-trips» from the data set





После разбиения набора данных на «микрорезды» их необходимо охарактеризовать несколькими параметрами, для того чтобы выбирать «микрорезды», входящие в ездовой цикл, не случайно, а на основе анализа этих самых параметров. Вычислив соответствующие параметры, получаем набор данных, каждая строка в котором соответствует одной «микрорезде» и содержащий значения параметров для каждой «микрорезды».

При взгляде на корреляционную матрицу (Рис. 6) для параметров видно, что среднее замедление сильно коррелирует со средним ускорением, а длина «микрорезды» сильно коррелирует с ее продолжительностью. Так как далее будет использована кластеризация для отбора «микрорездок», от корреляции следует избавиться, так как практически все методы машинного обучения, в том числе методы кластеризации, очень плохо работают с данными, имеющими сильно коррелирующие признаки. Поэтому из рассмотрения необходимо исключить столбцы среднего замедления и времени «микрорезды».

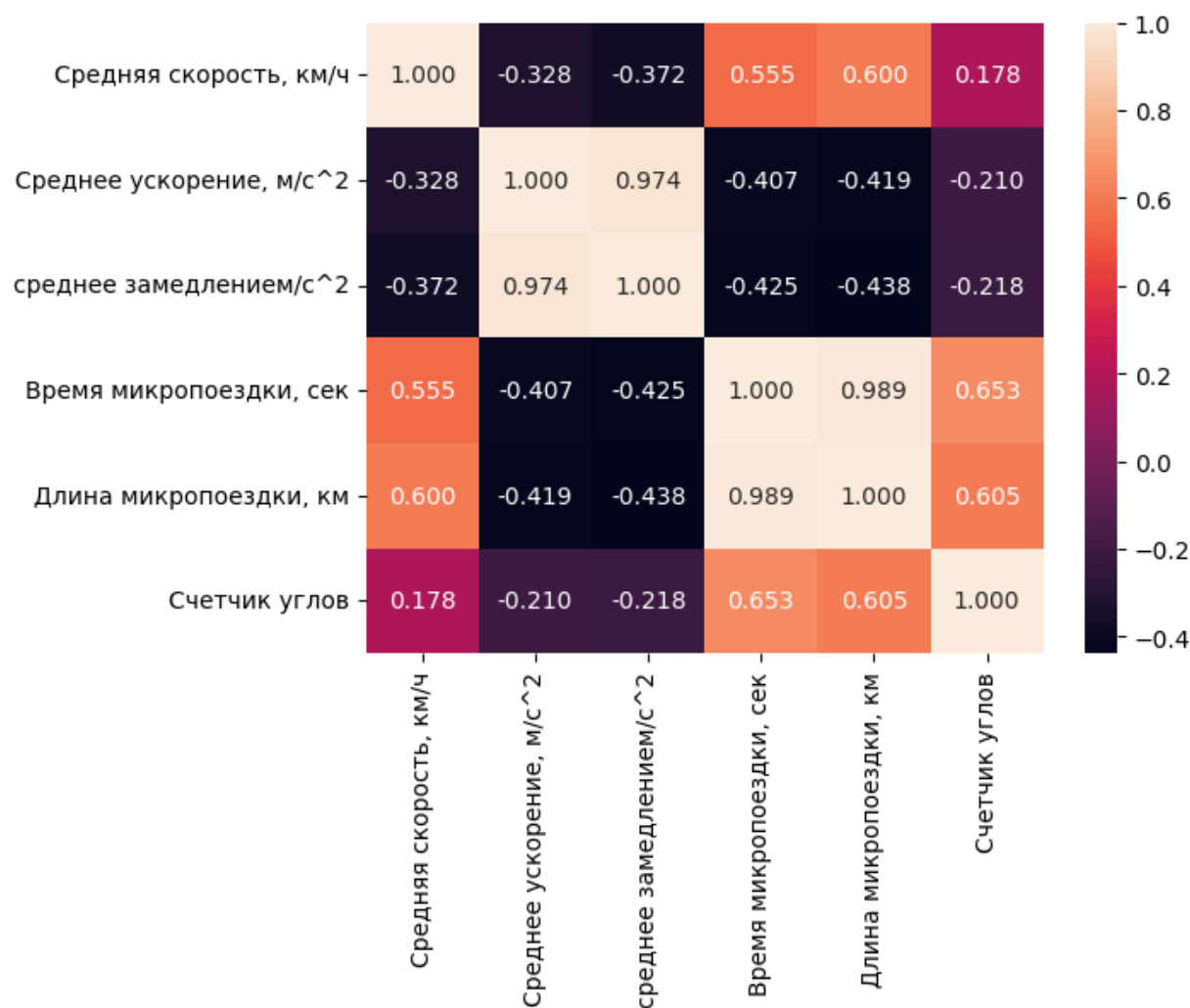


Рис.6. Корреляционная матрица для параметров, характеризующих «микрорезды»  
Fig. 6. Correlation matrix for parameters characterizing «micro-trips»

Также при работе с данными желательно, чтобы признаки имели нормальное распределение. Это облегчает работу методам машинного обучения, в том числе методам кластеризации, так как в основном такие методы могут «замечать» лишь линейные зависимости в данных. Одним из наиболее эффективных методов нормализации данных является метод Йео-Джонсона [20]. Разницу в распределениях данных до и после нормализации можно увидеть на Рис. 7 а и б.

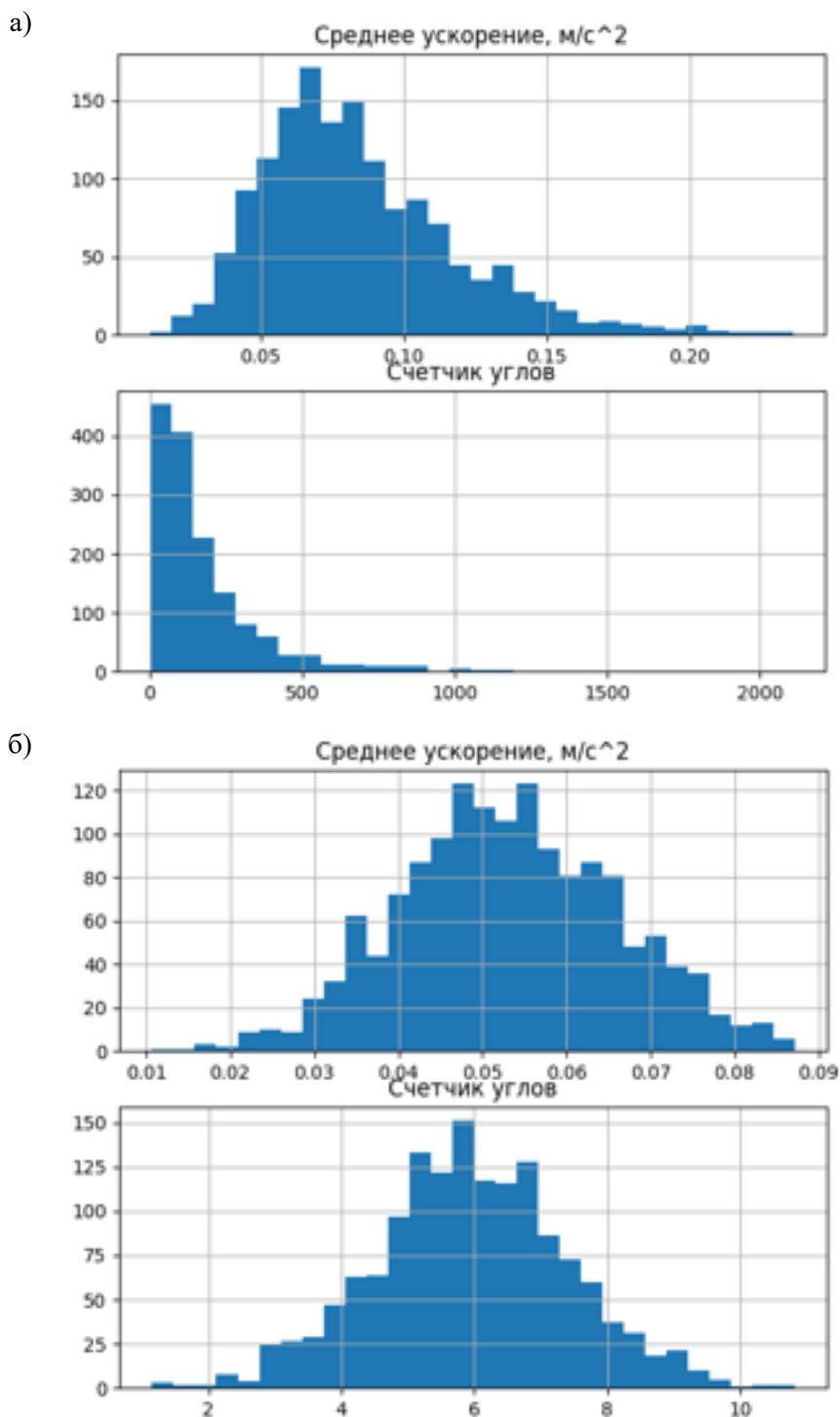


Рис. 7. Распределение значений некоторых параметров: а – до применения нормализации; б – после применения нормализации  
Fig.7. Distribution of values of some parameters: a – before applying normalization; b – after applying normalization

Следующая процедура, без которой методы кластеризации не будут адекватно работать – это масштабирование данных. Очень важно, чтобы все признаки принимали значения только в пределах какого-либо одинакового числового интервала, иначе при работе алгоритмов кластеризации признак, имеющий больший масштаб, будет оказывать наибольшее влияние по сравнению с остальными признаками.



Также в данных часто встречаются так называемые «выбросы», то есть данные, резко выделяющиеся на фоне остальных. Такие данные чаще всего являются ошибками в измерениях или расчетах и от них следует избавиться, чтобы они не исказили результаты.

Для масштабирования будет использован алгоритм масштабирования по медиане, который выполняет сразу две функции: масштабирование данных и устранение выбросов (путем «отрезания» хвостов распределения). Он предусматривает следующую формулу преобразования:

$$x' = \frac{x - \text{median}(x)}{IQR}, IQR = Q3(x) - Q1(x), \tag{5}$$

где  $x$  – исходные данные,  $x'$  – масштабированные данные,  $IQR$  – разница между 1 и 3 квартилями распределения признака.

Результаты применения масштабирования видны на Рис. 8.

### Генерация ездового цикла

После подготовки данных их можно передать в алгоритм кластеризации и на основании полученных результатов осуществить генерацию ездового цикла. В данной работе применяется метод кластеризации «k-means» [6], с его помощью «микрорездки» разделяются на кластеры, из которых они отбираются для формирования конечного ездового цикла.

Так как целью генерации ездовых циклов является статистическое совпадение полученного цикла с полной записью движения транспортного средства, для оценки качества данных до и после обработки было проведено сравнение ездовых циклов, сгенерированных на обработанных и необработанных данных, с исходными статистическими данными с использованием следующих критериев [6]:

- средняя скорость (км/ч);
- среднее ускорение (м/с<sup>2</sup>);
- среднее замедление (м/с<sup>2</sup>);
- время поездки (сек);
- длина поездки (м);
- количество уклонов свыше 5%.

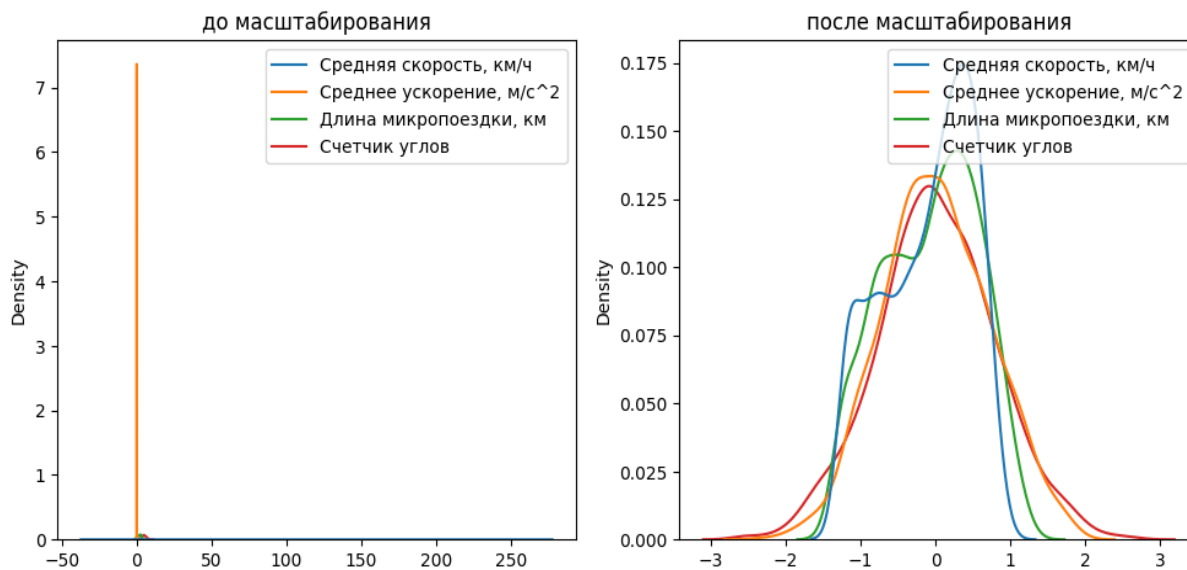


Рис. 8. Иллюстрация масштаба значений параметров до и после масштабирования

Fig. 8. Illustration of the scale of parameter values before and after scaling



Исходный выбор набора критериев был основан на рекомендациях [21] и включал в себя больше пунктов, но в ходе проведенного анализа были выявлены критерии, вносящие наиболее существенный вклад в работу алгоритма кластеризации, а остальные были исключены из рассмотрения.

Для эксперимента были сгенерированы несколько ездовых циклов разной продолжительности из обработанных и необработанных данных, и затем посчитано отклонение каждого из них от реальной записи. Такой метод оценки наиболее объективен, так как показывает и процент отклонения от реальных записей, и зависимость отклонения от продолжительности цикла. На Рис. 9 представлен результат сравнения.

При анализе графиков на Рис. 9 видно, что циклы, полученные из необработанных данных, значительно хуже сходятся с реальной записью. К примеру, циклы, полученные из обработанных данных, имеют отклонение менее 5% при продолжительности 10 часов, в то время как циклы из необработанных данных достигают 5% только при продолжительности в 20 часов.

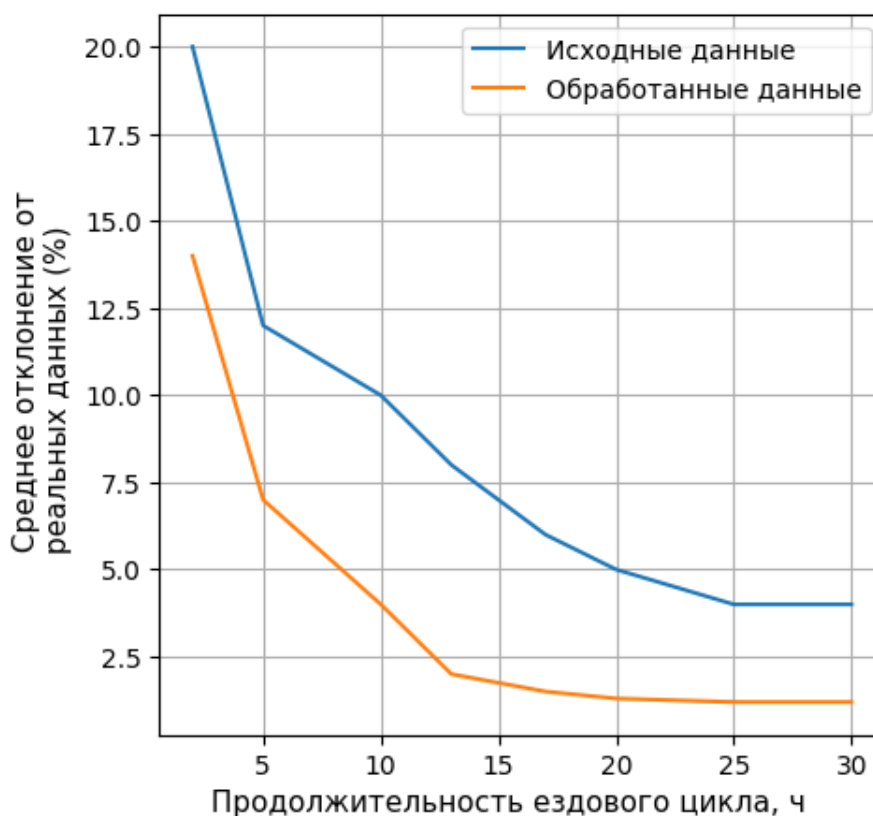


Рис. 9. Зависимость сходимости цикла с полным набором данных по статистическим параметрам

Fig. 9. Dependence of the convergence of a cycle with a complete set of data on statistical parameters

### Заключение

По результатам работы можно сделать вывод, что обработка статистических данных движения ТС является неотъемлемой и крайне важной частью работы со статистическими данными, так как она позволяет избавиться от недостатков в данных, вызванных различными физическими аспектами, такими как большой период отправки сообщения датчиковой аппаратуры, неточности в измерениях, какие-либо особенности измерительного оборудования и так далее. В работе также представлены приемы работы с данными для обеспечения качественной работы алгоритмов машинного обучения применительно к статистическим данным о движении транспортных средств и показана их эффективность при генерации ездовых циклов методом «микрорезидов» с применением кластеризации. В качестве меры эффективности были использованы статистические параметры, по которым сравнивались



ездовые циклы. При сравнении качества ездовых циклов, сгенерированных на основе обработанных и необработанных данных, оказалось, что циклы на основе обработанных данных обеспечивают лучшую сходимость с исходным набором данных. Ездовой цикл, полученный на основе обработанных данных, обеспечивает отклонение от записей реального движения менее 5% при вдвое меньшей продолжительности по сравнению с необработанными данными. Таким образом, в случае задачи генерации ездовых циклов предложенное решение проблемы качества данных позволяет сформировать более короткий ездовой цикл и сократить время, затрачиваемое на проведение виртуальных испытаний транспортных средств.

#### **Благодарности**

Работа выполнена при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках соглашения № 075-11-2022-015 от 07.04.2022 г. с ПАО «КАМАЗ» по комплексному проекту «Создание высокотехнологичного производства мехатронных трансмиссий перспективных грузовых автомобилей и автобусов КАМАЗ с электрическими накопителями энергии и водородными топливными элементами» в части выполнения научно-исследовательских, опытно-конструкторских и технологических работ.

#### **Конфликт интересов**

Авторы заявляют об отсутствии конфликта интересов.

© 2023 Авторы. Издательство Кузбасского государственного технического университета имени Т.Ф. Горбачева. Эта статья доступна по лицензии Creative Commons «Attribution» («Атрибуция») 4.0 Всемирная (<https://creativecommons.org/licenses/by/4.0/>)

#### **Информация об авторах**

**Поташиников Максим Дмитриевич**, программист  
e-mail: potashnikov@bmstu.ru

**Муравьев Александр Сергеевич**, ведущий программист  
e-mail: aleksandr.muravev@bmstu.ru

**Карташов Александр Борисович**, к.т.н., доцент  
e-mail: kartashov@bmstu.ru

**Пикалов Никита Андреевич**, заместитель директора  
e-mail: nikitapikalov@bmstu.ru

МГТУ им. Н.Э. Баумана, 105005, Российская Федерация, г. Москва, ул. 2-я Бауманская, 5

**Назаренко Сергей Владимирович**, главный конструктор по инновационным автомобилям  
e-mail: Sergey.Nazarenko@kamaz.ru

ПАО «КАМАЗ», 420000, Российская Федерация, Республика Татарстан, г. Набережные Челны, Транспортный проезд, 70

#### **Список литературы**

1. Козлова Т.А. Методика поиска рациональных конструктивных параметров тягового привода электромобиля // Интернет-журнал «НАУКОВЕДЕНИЕ» Том 8, №5 (2016) <http://naukovedenie.ru/PDF/86TVN516.pdf> (доступ свободный). Загл. с экрана. Яз. рус., англ.
2. Антипов С. И., Дементьев Ю. В. Современные Испытательные ездовые циклы и их актуальность при создании алгоритма работы системы управления автомобиля с КЭУ, Известия волгоградского технического университета. Серия: наземные транспортные системы. – 2013. – №10(113). – С. 8-11.
3. Гусаков С.В., Марков В.А., Михрячев Д.В. Расчетно-экспериментальная методика корректировки ездового цикла для фазы движения транспортного средства в городских условиях // Известия высших учебных заведений. Машиностроение. – 2012. – № 5. – С. 23-30.



4. Акунов Б.У., Ездовые циклы для оценки топливной экономичности легковых автомобилей // Вестник Таджикского Технического Университета. – 2014. – №1(25). – С. 92-95.
5. Маняшин С.А. Моделирование расхода топлива автомобилями на базе ездового цикла в низкотемпературных условиях эксплуатации: дис. ... канд. техн. наук: 05.22.10, Оренбург, 2013. 172 с.
6. Maksim Potashnikov, Victoria Shishkina, Alexandr Muravev and Alexandr Kartashov. Development of vehicle driving cycles based on the real traffic dataset // E3S Web of Conferences. – 2023. – Vol. 402. – Article 04001.
7. Valeriy Grinin, Evgeniy Shkarupelov, Aleksandr Muravev, Aleksandr Kartashov, Sergey Nazarenko and Aleksandr Klimov, Method for applying vehicle driving cycles to assess the durability of electromechanical transmissions of trucks // E3S Web of Conferences. – 2023. – Vol. 402. – Article 10008.
8. Dubinkin D., Kartashov A., Muraviev A., Buzunov N., Khlobystov I. Study of the control algorithm of the braking system of an autonomous haul truck braking system with the use of imitational models // E3S Web of Conferences. – 2021. – Vol. 315. – Article 03021.
9. Dubinkin D., Kulpin A., Stenin D. Justification of the Number and Type of Tire Size for a Dump Truck with a Lifting Capacity from 90 to 130 Tons // E3S Web of Conferences. – 2020. – Vol. 174. – Article 03015. – DOI 10.1051/e3sconf/202017403015.
10. Muravyev A.S., Shishkina V.A., Buzunov N.V., Kartashov A.B., Dubinkin D.M., Nozirzoda S. Research of control algorithm of traction drive of a mining dump truck using simulation models of motion // Journal of Physics: Conference Series. – 2021. – Vol. 2052. – Article 012028. – DOI 10.1088/1742-6596/2052/1/012028.
11. Buzunov N.V., Pirozhkov R.D., Kartashov A.B., Dubinkin D.M., Efremenkov A.B. Simulation of operation of a sequential hybrid drive of a haul truck with a traction battery and a bilateral DC-to-DC converter // IOP conference series: materials science and engineering. – 2020. – Vol. 939. – Article 012017. – DOI 10.1088/1757-899X/939/1/012017.
12. Muraviev A.S., Pirozhkov R.D., Buzunov N.V., Kartashov A.B., Dubinkin D.M., Efremenkov A.B. Studying the algorithm for controlling the generator set of a mining dump truck using simulation models // AIP Conference Proceedings. – 2022. – Vol. 2486. – Article 020031. – DOI 10.1063/5.0106475.
13. Филиппова Н.А., Мушта Б.М., Сидоренко А.В., Анализ развития навигационной системы диспетчерского управления грузовым транспортом, Синергия Наук, eISSN: 2500-0950, 2019.
14. Моисейкин Д.А., Кожевникова С.А. Инновационные процессы в транспортной индустрии // Концепт. – 2014. – Спецвыпуск № 04. ART 14538.
15. Filippova N.A., Vlasov V.M., Belyaev V.M. Navigation Control of Cargo Transportation in the North of Russia. World of Transport and Transportation. – 2019. – №17(4). – С. 218-231. DOI:10.30932/1992-3252-2019-17-4-218-231.
16. Lisa Park, Katie Fender, Assessing the Use of Navigation Systems in the Trucking Industry. – 2014. – Transportation Research Record Journal of the Transportation Research Board, DOI: 10.3141/2411-13
17. Вьюгин В. Математические основы машинного обучения и прогнозирования. М., 2013. – 387 с.
18. Soledad Galli. Python Feature Engineering Cookbook. Packt Publishing, 2020. ISBN 9781789806311
19. Xinyi Jia, Hewu Wang, Liangfei Xu, Qing Wang, Hang Li, Zunyan Hu, Jianqiu Li, Minggao Ouyang, Constructing representative driving cycle for heavy duty vehicle based on Markov chain method considering road slope, Energy and AI 6 (2021) 100115
20. I.K. Yeo and R.A. Johnson. A new family of power transformations to improve normality or symmetry // Biometrika, 87(4), pp.954-959, (2000)
21. Ramya Madhuri Desineedi, Srinath Mahesh, Gitakrishnan Ramadurai, Developing driving cycles using k-means clustering and determining their optimal duration, WCTR 2019 Mumbai 26-31 May 2019, DOI: 10.1016/j.trpro.2020.08.268



## FORMATION OF AN APPROACH TO THE PROCESSING OF STATISTICAL DATA ON VEHICLE MOVEMENTS

Maksim D. Potashnikov<sup>1</sup>, Alexandr S. Muravev<sup>1</sup>, Aleksandr B. Kartashov<sup>1</sup>,  
Nikita A. Pikalov<sup>1</sup>, Sergey V. Nazarenko<sup>2</sup>

<sup>1</sup> Bauman Moscow State Technical University

<sup>2</sup> PJSC "KAMAZ"



### Article info

Received:

05 August 2023

Revised:

28 August 2023

Accepted:

11 September 2023

**Keywords:** data analysis,  
machine learning, driving cycles  
of vehicles, navigation  
equipment, statistics

### Abstract.

The article describes the process of analysis and processing of statistical data on truck traffic obtained with the help of navigation equipment installed on the vehicles under consideration. The detailed analysis of initial data, as well as all stages of their processing, using specialized packages of Python language tools, such as Pandas, Numpy and Sklearn, is carried out. The effectiveness of the proposed data processing approach is demonstrated by comparing the driving cycles of vehicles generated from the original and modified data. The topic of data analysis and processing is not a new one, but the relevance of the paper lies in the successful application of common and efficient processing methods to statistical vehicle motion data, the handling of which is complicated by the fact that they are obtained with widely used but task-oriented equipment. Thus, the work opens new opportunities for statistical analysis of truck traffic without the need to replace navigation equipment.

---

**For citation** Potashnikov M.D., Muravev A.S., Kartashov A.B., Pikalov N.A., Nazarenko S.V. (2023) Formation of an approach to the processing of statistical data on vehicle movements, *Journal of mining and geotechnical engineering*, 3(22):34. DOI: 10.26730/2618-7434-2023-3-34-50, EDN: CJEEEXC

---

*This research work was financially supported by the Ministry of Science and Higher Education of the Russian Federation under agreement No. 075-11-2022-015 dated 07.04.2022 with PJSC "KAMAZ" under the complex project "Creation of high-tech production of mechatronic transmissions for promising KAMAZ trucks and buses with electrical energy storage devices and hydrogen fuel cells" in terms of R&D and technological works*

### References

1. Kozlova T.A. Metodika poiska ratsional'nykh konstruktivnykh parametrov tyagovogo privoda elektromobilya // Internet-zhurnal «NAUKOVEDENIE» Tom 8, №5 (2016) <http://naukovedenie.ru/PDF/86TVN516.pdf> (dostup svobodnyy). Zagl. s ekrana. Yaz. rus., angl.
2. Antipov S. I., Dement'ev Yu. V. Sovremennye Ispytatel'nye ezdovye tsikly i ikh aktual'nost' pri sozdanii algoritma raboty sistemy upravleniya avtomobilya s KEU, *Izvestiya volgogradskogo tekhnicheskogo universiteta*. Seriya: nazemnye transportnye sistemy. – 2013. – №10(113). – S. 8-11.
3. Gusakov S.V., Markov V.A., Mikhryachev D.V. Raschetno-eksperimental'naya metodika korrekcirovki ezdovogo tsikla dlya fazy dvizheniya transportnogo sredstva v gorodskikh usloviyakh // *Izvestiya vysshikh uchebnykh zavedeniy. Mashinostroenie*. – 2012. – № 5. – S. 23-30.
4. Akunov B.U., Ezdovye tsikly dlya otsenki toplivnoy ekonomichnosti legkovykh avtomobiley // *Vestnik Tadzhikskogo Tekhnicheskogo Universiteta*. – 2014. – №1(25). – S. 92-95.
5. Manyashin S.A. Modelirovanie raskhoda topliva avtomobilyami na baze ezdovogo tsikla v nizkotemperaturnykh usloviyakh ekspluatatsii: dis. ... kand. tekhn. nauk: 05.22.10, Orenburg, 2013. 172 s.
6. Maksim Potashnikov, Victoria Shishkina, Alexandr Muravev and Alexandr Kartashov. Development of vehicle driving cycles based on the real traffic dataset // *E3S Web of Conferences*. – 2023. – Vol. 402. – Article 04001.



7. Valeriy Grinin, Evgeniy Shkarupelov, Aleksandr Muravev, Aleksandr Kartashov, Sergey Nazarenko and Aleksandr Klimov, Method for applying vehicle driving cycles to assess the durability of electromechanical transmissions of trucks // E3S Web of Conferences. – 2023. – Vol. 402. – Article 10008.
8. Dubinkin D., Kartashov A., Muraviev A., Buzunov N., Khlobystov I. Study of the control algorithm of the braking system of an autonomous haul truck braking system with the use of imitational models // E3S Web of Conferences. – 2021. – Vol. 315. – Article 03021.
9. Dubinkin D., Kulpin A., Stenin D. Justification of the Number and Type of Tire Size for a Dump Truck with a Lifting Capacity from 90 to 130 Tons // E3S Web of Conferences. – 2020. – Vol. 174. – Article 03015. – DOI 10.1051/e3sconf/202017403015.
10. Muravyev A.S., Shishkina V.A., Buzunov N.V., Kartashov A.B., Dubinkin D.M., Nozirezoda S. Research of control algorithm of traction drive of a mining dump truck using simulation models of motion // Journal of Physics: Conference Series. – 2021. – Vol. 2052. – Article 012028. – DOI 10.1088/1742-6596/2052/1/012028.
11. Buzunov N.V., Pirozhkov R.D., Kartashov A.B., Dubinkin D.M., Efremkov A.B. Simulation of operation of a sequential hybrid drive of a haul truck with a traction battery and a bilateral DC-to-DC converter // IOP conference series: materials science and engineering. – 2020. – Vol. 939. – Article 012017. – DOI 10.1088/1757-899X/939/1/012017.
12. Muraviev A.S., Pirozhkov R.D., Buzunov N.V., Kartashov A.B., Dubinkin D.M., Efremkov A.B. Studying the algorithm for controlling the generator set of a mining dump truck using simulation models // AIP Conference Proceedings. – 2022. – Vol. 2486. – Article 020031. – DOI 10.1063/5.0106475.
13. Filippova N.A., Mushta B.M., Sidorenko A.V., Analiz razvitiya navigatsionnoy sistemy dispetcherskogo upravleniya gruzovym transportom, Sinergiya Nauk, eISSN: 2500-0950, 2019.
14. Moiseykin D.A., Kozhevnikova S.A. Innovatsionnye protsessy v transportnoy industrii // Kontsept. – 2014. – Spetsvyпуск № 04. ART 14538.
15. Filippova N.A., Vlasov V.M., Belyaev V.M. Navigation Control of Cargo Transportation in the North of Russia. World of Transport and Transportation. – 2019. – №17(4). – S. 218-231. DOI:10.30932/1992-3252-2019-17-4-218-231.
16. Lisa Park, Katie Fender, Assessing the Use of Navigation Systems in the Trucking Industry. – 2014. – Transportation Research Record Journal of the Transportation Research Board, DOI: 10.3141/2411-13
17. V'yugin V. Matematicheskie osnovy mashinnogo obucheniya i prognozirovaniya. M., 2013. – 387 s.
18. Soledad Galli. Python Feature Engineering Cookbook. Packt Publishing, 2020. ISBN 9781789806311
19. Xinyi Jia, Hewu Wang, Liangfei Xu, Qing Wang, Hang Li, Zunyan Hu, Jianqiu Li, Minggao Ouyang, Constructing representative driving cycle for heavy duty vehicle based on Markov chain method considering road slope, Energy and AI 6 (2021) 100115
20. I.K. Ye and R.A. Johnson. A new family of power transformations to improve normality or symmetry // Biometrika, 87(4), pp.954-959, (2000)
21. Ramya Madhuri Desineedi, Srinath Mahesh, Gitakrishnan Ramadurai, Developing driving cycles using k-means clustering and determining their optimal duration, WCTR 2019 Mumbai 26-31 May 2019, DOI: 10.1016/j.tpro.2020.08.268

### **Conflicts of Interest**

The authors declare no conflict of interest.

© 2023 The Authors. Published by T. F. Gorbachev Kuzbass State Technical University. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

### **Information about the authors**

**Maksim A. Potashnikov**, programmer  
e-mail: potashnikov@bmstu.ru

**Aleksandr S. Muravev**, lead programmer  
e-mail: aleksandr.muravev@bmstu.ru

**Aleksandr B. Kartashov**, PhD (Tech.), Associate Professor  
e-mail: kartashov@bmstu.ru





**Nikita A. Pikalov**, deputy director  
e-mail: nikitapikalov@bmstu.ru

Bauman Moscow State Technical University,  
5 2nd Baumanskaya Street, Moscow, 105005, Russian Federation

**Sergey V. Nazarenko**, chief designer of innovative vehicles  
e-mail: [Sergey.Nazarenko@kamaz.ru](mailto:Sergey.Nazarenko@kamaz.ru)

PJSC "KAMAZ"  
70 Traspportniy proezd, 420000, Naberezhnye Chelny, Republic of Tatarstan, Russian Federation

