|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ \_\_\_\_\_ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_\_СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

***Обработка и анализ навигационных данных движения грузовых транспортных средств\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

Студент \_\_ИУ5-34М\_\_\_\_\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_М.Д. Поташников\_\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** **\_\_\_\_\_****Ю.Е. Гапанюк\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2023 г.***Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технический университет имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

УТВЕРЖДАЮ

Заведующий кафедрой \_\_\_ИУ5\_\_\_\_

(Индекс)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_В.И. Терехов\_\_

(И.О.Фамилия)

«\_04\_» \_\_\_\_сентября\_\_\_\_\_\_\_ 2023 г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме Обработка и анализ навигационных данных движения грузовых транспортных средств \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент группы \_ИУ5-34М\_\_

Поташников Максим Дмитриевич\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ИССЛЕДОВАТЕЛЬСКАЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Источник тематики (кафедра, предприятие, НИР) \_\_\_\_\_КАФЕДРА\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

График выполнения НИР: 25% к \_\_\_\_ нед., 50% к \_\_\_\_ нед., 75% к \_\_\_ нед., 100% к \_\_\_\_ нед.

***Техническое задание*** провести обработку и анализ навигационных данных движения грузовых транспортных средств, провести качественную оценку обработки данных\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Оформление научно-исследовательской работы:***

Расчетно-пояснительная записка на \_18\_ листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата выдачи задания «\_04\_» \_\_\_\_сентября\_\_\_\_\_\_\_ 2023 г.

**Руководитель НИР**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_Ю.Е. Гапанюк\_\_\_

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_**М.Д. Поташников**\_**

(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

**Содержание**

[Введение 3](#_Toc154660582)

[Анализ исходных данных 4](#_Toc154660583)

[Обработка данных скорости и нагрузки на ось 7](#_Toc154660584)

[Обработка данных пробега и времени 8](#_Toc154660585)

[Обработка данных высоты 10](#_Toc154660586)

[Расчет угла уклона дороги 11](#_Toc154660587)

[Преобразование данных для использования в методе генерации ездового цикла 12](#_Toc154660588)

[Генерация ездового цикла 16](#_Toc154660589)

[Заключение 18](#_Toc154660590)

# **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 проводятся на основе анализа статистических данных [1].

В работах [2-3] показано, как можно эффективно применять статистические данные движения автомобиля при расчетах нагрузок на трансмиссионные узлы и агрегаты, а также как на основе этих данных генерировать ездовые циклы [4-7], соответствующие условиям движения в разных регионах. 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 данному типу 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 [8-11]. 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.

В данной работе используются сбора статистические данных о движении грузовых автомобилей с весом, приходящимся на ось, равным 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.

# **Анализ исходных данных**

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.

Таблица 1**–** Количество пропущенных значений в наборе данных

|  |  |
| --- | --- |
| **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 % |

В обыкновенной практике анализа данных и машинного обучения [12] принято считать, что если в столбце более 5-10% пропущенных значений, то такие данные не подлежат восстановлению и являются неинформативными [13]. 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.

При обработке в первую очередь необходимо избавиться от данных, которые не представляют интереса при анализе или являются неинформативными и не имеют возможности восстановления. Такими данными являются измеренные значения частоты вращения ДВС и расхода топлива («engine\_speed», «high\_resolution\_engine\_total\_fuel\_used», «engine\_total\_fuel\_used»), а также скорость транспортного средства, измеренная на основании датчиков частоты вращения колес («wheel\_based\_vehicle\_speed»). Набор данных, с которым будут производиться дальнейшие манипуляции, представлен на рисунке 2.

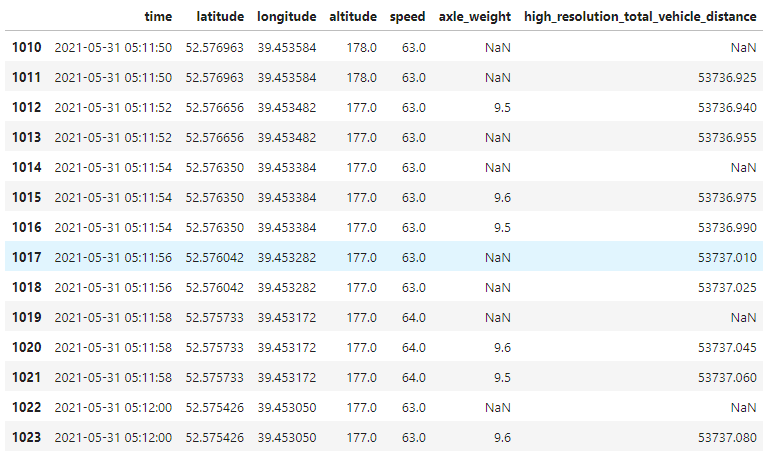


Рисунок 2 –Фрагмент набора данных, иллюстрирующий характер пропусков данных в столбце скорости

При дальнейшем анализе необходимо восстановить пропущенные значения измеренной скорости на основе данных с GPS приемника («speed»). Анализируя фрагмент набора данных, 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, то есть когда транспортное средство остановлено и происходит процесс выключения питания.

Проанализировав информацию в столбце «time» видно, что в нем имеются повторяющиеся записи, к примеру, в строках 1010 и 1011 или 1019 и 1020. В большом количестве операций с данными может быть использована разность между отсчетами времени в наборе данных, поэтому недопустимо, чтобы она была равна нулю. К примеру, это может привести к получению бесконечного ускорения в какой-либо момент времени. Соответственно, столбцу с данными времени также требуется дополнительная обработка.

На графиках на рисунках 3(а) и 3(б) показаны фрагменты зависимости скорости и высоты от времени. Видно, что графики имеют ступенчатый характер. Для данных скорости движения ТС это не является критичным из-за высокого разрешения измеренного сигнала. (подразумевается, что сигнал измеряет с точностью до 0,1 км/ч, а не ступенчато с точностью 1 км/ч). Но для значений высоты шаг измерения в 1 метр является неприемлемым, так как в таком случае получение адекватных значений уклона с помощью высотных данных невозможно.

|  |  |
| --- | --- |
|  |  |
| а) | б) |

Рисунок 3 – Фрагменты данных в столбцах скорости и высоты

При дальнейшем анализе данных пробега транспортного средства (рисунок 2), можно заметить, что пропуски в данных носят системный характер. Устранять такие пропущенные значения можно двумя способами, либо удалить из набора данных все строки с пропущенными значениями, либо постараться аппроксимировать показания пробега исходя из значений столбцов времени и скорости. Первый вариант является неприемлемым, так как в таком случае нужно будет удалить 41% всех строк в наборе данных, поэтому столбец со значениями пробега также необходимо дополнительно обработать.

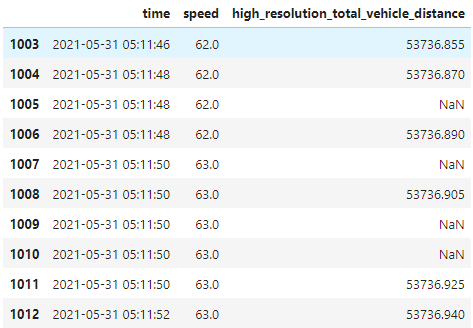


Рисунок 4. – Фрагмент набора данных, иллюстрирующий характер пропущенных значений в столбце пробега

Анализ столбца с информацией о нагрузке на ось показал, что в нём содержится 86,4 % пропущенных значений, однако полностью убирать её 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 (рисунок 5) 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.

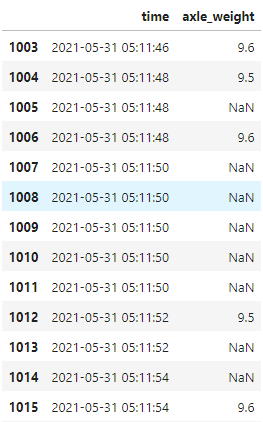


Рисунок 5 –Фрагмент набора данных, иллюстрирующий характер пропущенных значений в столбце нагрузки на ведущую ось

# **Обработка данных скорости 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.

Как было сказано выше, в столбцах скорости и высоты пропущенные значения встречаются в одних и тех же записях, при этом в это время в транспортном средстве отключается питание и его скорость равна нулю. Участки данных, в которых транспортное средство стоит на месте не представляют интереса для анализа, поэтому пропущенные значения в столбцах скорости и высоты были удалены, так как при этом полезная информация не теряется.

В предыдущем разделе было показано, что пропущенные значения в столбце нагрузки на ось возникают из-за особенностей датчика, с помощью которого измеряются значения нагрузки на ведущую ось. Поэтому пропущенные значения в этом столбце также 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 (таблица 2).

Таблица 2 –Количество пропущенных значений в наборе данных

|  |  |
| --- | --- |
| **Название столбца** | **Количество пустых записей, шт** |
| time | 0 |
| latitude | 0 |
| longitude | 0 |
| altitude | 0 |
| speed | 0 |
| axle\_weight | 0 |
| high\_resolution\_total\_vehicle\_distance | 25454143 |

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:

|  |  |
| --- | --- |
|  | (4) |

где – текущее ускорение транспортного средства, – разница скоростей между двумя записями набора данных, – разница во времени между двумя записями набора данных.

# **Обработка данных пробега и времени**

Анализ временных данных стоит производить исходя из трёх столбцов: времени, текущей скорости и пробега. Рассмотрим небольшой фрагмент данных (рисунок 6).

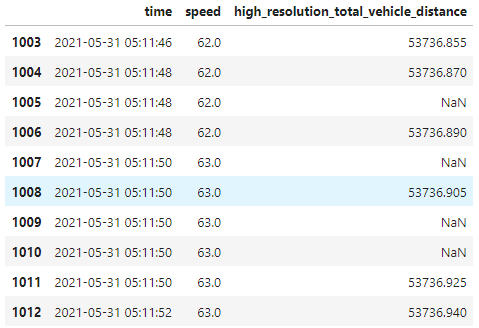


Рисунок 6– Фрагмент набора данных, иллюстрирующий проблемы в столбцах пробега и времени

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 (1004 и 1006 строки на рисунке 4). 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.

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 со скоростью большей 18 км/ч. На основании этого решения был разработан соответствующий алгоритм, блок схема которого представлена на рисунке 7. 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 что в фрагменте на рисунке 6 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 1.

|  |  |
| --- | --- |
|  | (1) |

где – интервал времени между записями в столбце пробега, – путь, пройденный за этот интервал (разница между соседними непропущенными значениями в столбце пробега), – соответствующая текущая скорость транспортного средства.

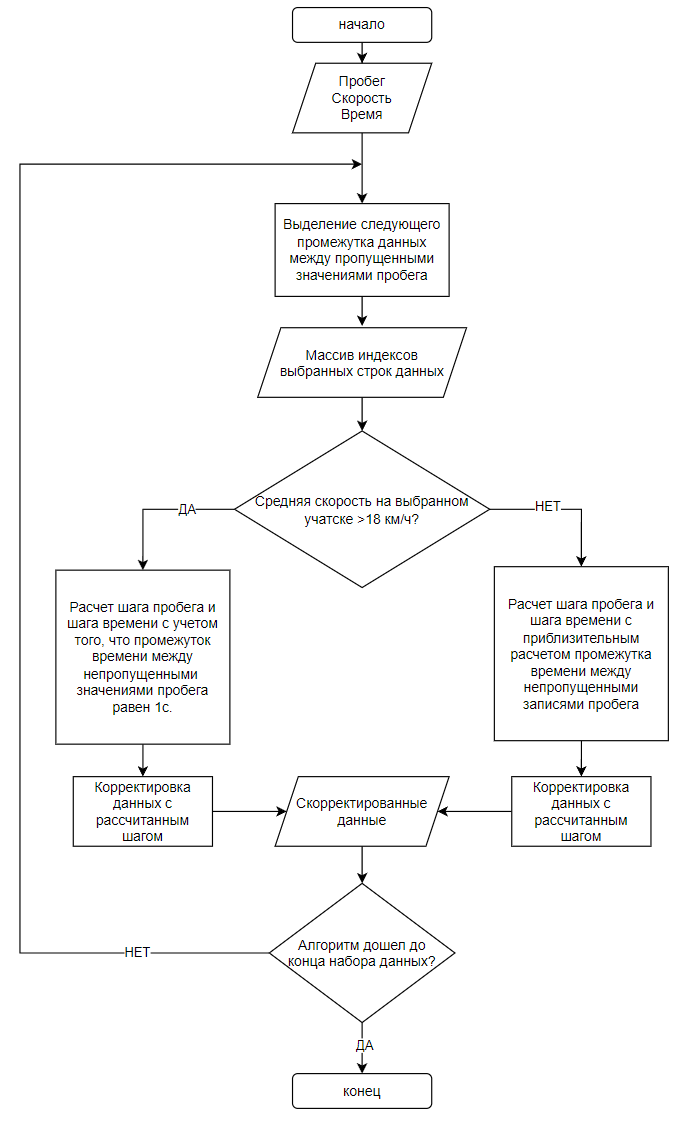


Рисунок 7 **–** Блок-схема алгоритма обработки столбцов пробега и времени

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 графику на рисунке 3(б), можно выделить, что высота изменяется «ступенчато», и для последующей обработки график это ступенчатое изменение необходимо сгладить. Для выяснения структуры данных высоты, обратимся к рисунку 2, на котором видно, что в столбце высоты значения являются целыми и часто повторяются. Для того чтобы сгладить данные, необходимо добавить к значениям высоты несколько знаков после запятой, предполагая, что высота изменяется линейно в промежутках, где её значения повторяются.

Алгоритм для корректировки столбца высоты будет подобен алгоритму для столбцов пробега и времени, указанном на рисунке 7, так как в нем также будут выделяться отдельные участки данных в которых будут корректироваться значения.

На рисунке 8 представлены графики одного и того же участка высотных данных до (рисунок 8а) и после (рисунок 8б) обработки столбца.

|  |  |
| --- | --- |
|  |  |
| а) | б) |

**Рисунок 8 –** Графики зависимости высоты ТС от времени: а – до обработки высотных данных; б –после обработки высотных данных

# **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 [3]. Благодаря имеющимся качественным данным пробега транспортного средства и высоты (обработанных в предыдущих пунктах), уклон дороги можно вычислить по формуле 2, аналогично тому, как подобные расчеты проводились в работе [14].

|  |  |
| --- | --- |
|  | (2) |

где – уклон дороги в текущий момент, - разница высоты между двумя записями в наборе данных, – 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 3). В случае использования этой формулы, нет необходимости проводить обработку данных пробега транспортного средства, так как в формуле он не используется.

|  |  |
| --- | --- |
|  | (3) |

где разница широт, разница долгот, R – радиус Земли.

Проведем сравнение двух расчетных уклонов дороги. На рисунке 9а представлен уклон, посчитанный по формуле (2) с использованием столбца реального пробега транспортного средства, а на рисунке 9б уклон, посчитанный с использованием формулы (3).

|  |  |
| --- | --- |
|  |  |
| а) | б) |

Рисунок 9 – График зависимости уклона дороги от времени: а – полученный на основе реальных записей пробега; б – полученные на основе расчетов по формуле Хаверсина

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 3 дает такой результат в следствие 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 2, а также будут смещены по времени из-за особенностей используемых фильтров.

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

В работе [2] 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 [2]. Соответственно, из набора данных необходимо выделить «микропоездки» и далее присвоить им определенные критерии, по которым они будут выбираться для участия в итоговом ездовом цикле. Для выделения отдельной микропоездки достаточно найти в общем наборе данных индексы её начала и конца. Блок-схема алгоритма определения микропоездок представлен на рисунке 10.

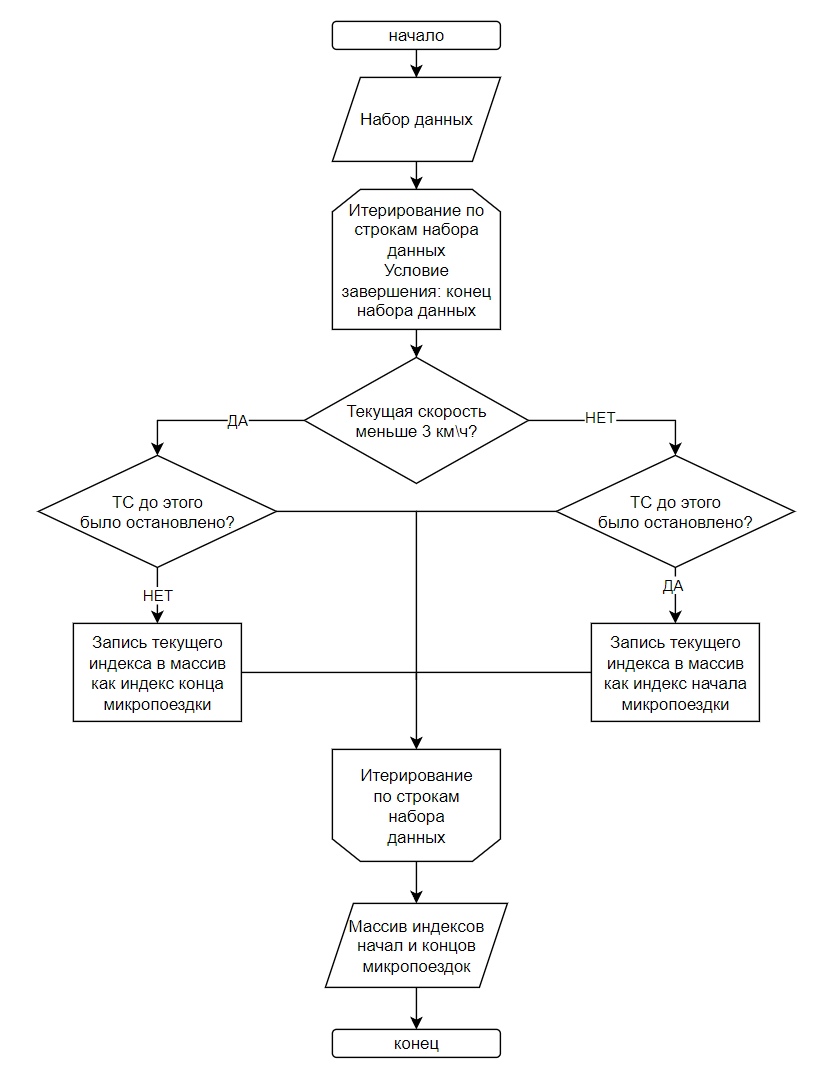


Рисунок 10 **–** Блок-схема алгоритма выделения "микропоездок" из набора данных

После разбиения набора данных на «микропоездки», их необходимо охарактеризовать несколькими параметрами, для того чтобы выбирать «микропоездки», 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 «микропоездке» и содержащий значения параметров для каждой «микропоездки» (рисунок 11).

Взглянув на корреляционную матрицу (рисунок 12) для параметров, видно, что среднее замедление сильно коррелирует со средним ускорением, а длина «микропоездки» сильно коррелирует с её продолжительностью. Так как далее будет использована кластеризация для отбора «микропоездок», 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 «микропоездки».

Рисунок 11 – Набор параметров, характеризующих отдельную "микропоездку"

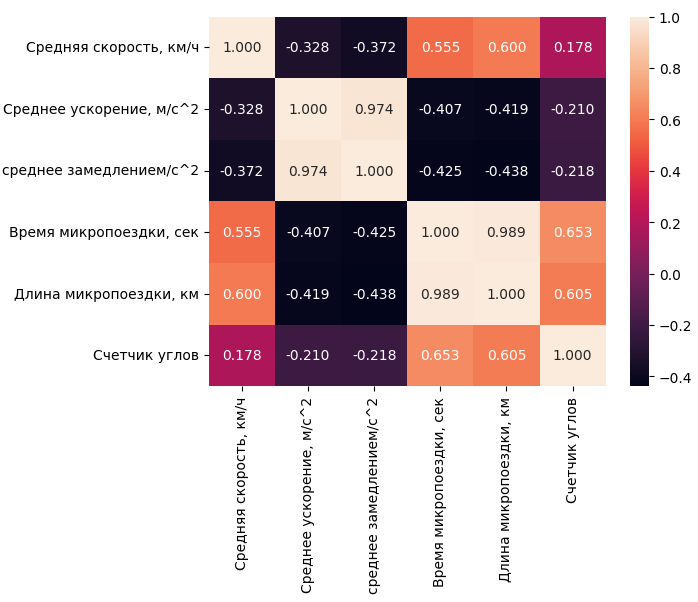


Рисунок 12 – Корреляционная матрица для параметров, характеризующих "микропоездки"

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 [15]. Разницу в распределениях данных до и после нормализации можно увидеть на рисунках 13 а и б.

|  |  |
| --- | --- |
|  |  |
| а) | б) |

Рисунок 13 **–** Распределение значений некоторых параметров: а – до применения нормализации ; б – после применения нормализации

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 искажали 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:

где – исходные данные, - масштабированные данные, – разница между 1 и 3 квартилями распределения признака.

Результаты применения масштабирования видны на рисунке 14.

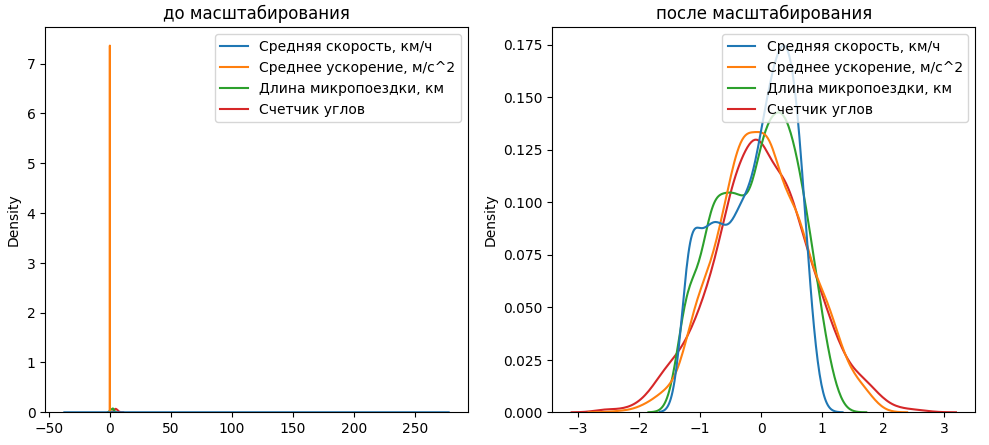


Рисунок 14 **–** Иллюстрация масштаба значений параметров до и после масштабирования

# **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», с его помощью «микропоездки» разделяются на кластеры, из которых они отбираются для формирования конечного ездового цикла. Результаты генерации ездового цикла представлены на рисунке 15.

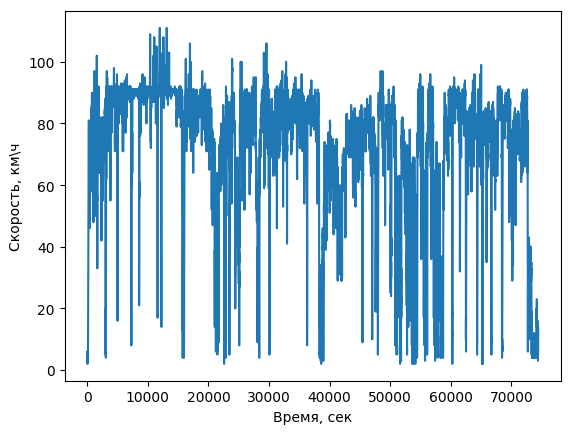


Рисунок 15 – График зависимости скорости от времени в ездовом цикле

По графику зависимости скорости от времени сложно рассуждать о качестве ездового цикла, поэтому для оценки качества данных до и после обработки было проведено сравнение ездовых циклов, сгенерированных на обработанных и необработанных данных, с используемыми статистическими данными с использованием следующих критериев [2]:

* средняя скорость (км/ч);
* среднее ускорение (м/);
* среднее замедление (м/);
* время поездки (сек);
* длина поездки (м);
* количество уклонов свыше 5%.

Исходный выбор набора критериев был основан на рекомендациях [16] и включал в себя больше пунктов, но в ходе проведенного анализа были выявлены критерии, вносившие 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 рисунке 16 представлен результат сравнения.

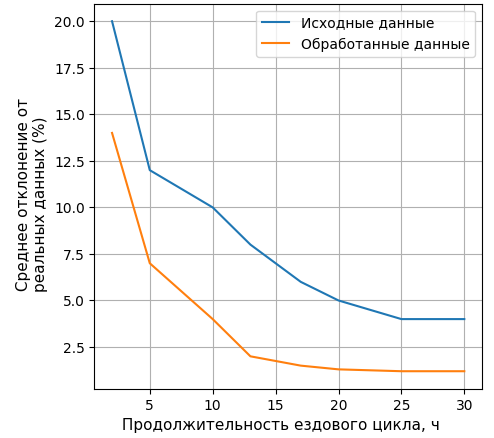


Рисунок 16 – Зависимость сходимости цикла с полным набором данных по статистическим параметрам

При анализе графиков на рисунке 16 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.

# **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 «микропоездок» 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.

**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. Maksim Potashnikov, Victoria Shishkina, Alexandr Muravev and Alexandr Kartashov, Development of vehicle driving cycles based on the real traffic dataset, E3S Web of Conf., Volume 402, 2023, International Scientific Siberian Transport Forum - TransSiberia 2023
3. 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 Conf., Volume 402, 2023, International Scientific Siberian Transport Forum - TransSiberia 2023
4. С. И. Антипов, Ю. В. Дементьев 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.
5. С. В. Гусаков, В. А. Марков, Д. В. Михрячев, Расчетно-экспериментальная методика корректировки ездового цикла для фазы движения транспортного средства в городских условиях, Известия высших учебных заведений. - М.: Машиностроение. - 2012. - № 5. - С. 23-30. EDN: OXDJPP
6. Акунов Б.У., Ездовые циклы для оценки топливной экономичности легковых автомобилей, Вестник Таджикского Технического Университета, 2014, №1(25), с.92-95.
7. Маняшин, Сергей Александрович. Моделирование расхода топлива автомобилями на базе ездового цикла в низкотемпературных условиях эксплуатации: диссертация кандидата технических наук: 05.22.10, Оренбург, 2013. 172 с.: ил. РГБ ОД, 61 13-5/1777
8. Филиппова Н.А., Мушта Б.М., Сидоренко А.В., Анализ развития навигационной системы диспетчерского управления грузовым транспортом, Синергия Наук, eISSN: 2500-0950, 2019.
9. Моисейкин Д.А., Кожевникова С.\_А. Инновационные процессы в транспортной индустрии // Концепт. — 2034. - Спецвыпуск № 04. АRТ 14538.
10. 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):218-231. <https://doi.org/10.30932/1992-3252-2019-17-4-218-231>
11. Lisa Park, Katie Fender, Assessing the Use of Navigation Systems in the Trucking Industry, December 2014, Transportation Research Record Journal of the Transportation Research Board, DOI: 10.3141/2411-13
12. Владимир Вьюгин, Математические основы машинного обучения и прогнозирования
13. Soledad Galli, Python Feature Engineering Cookbook
14. 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
15. 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)
16. 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](http://dx.doi.org/10.1016/j.trpro.2020.08.268)