УДК 004.413

ПАРЕТО-АНАЛИЗ КАЧЕСТВА РАБОТЫ СЕРВИСНЫХ ЦЕНТРОВ АВТОПРОИЗВОДИТЕЛЕЙ

© 2025 В.Г. Мосин[1], К.А. Брагина[2], В.Н. Козловский[1], А.В. Гусев[1]

[1] Самарский государственный технический университет, г. Самара, Россия
[2] Московский государственный технический университет имени Н.Э. Баумана, г. Москва, Россия

Статья поступила в редакцию 27.05.2025

Применение метода Парето анализа для решения задач в области управления качеством остается востребованным и крайне актуальным. В настоящее время мы наблюдается резкий рост электронных данных отражающих протекание различных процессов в системе менеджмента организаций, а также данных отражающих этапы жизненного цикла продукции. Соответственно росту данных, требуется решение научно-технических задач направленных на развитие методов и инструментов мониторинга и анализа данных. Требуется их совершенствование и соответствующее развитие в контексте решений задач в области цифровизации и информатизации процессов менеджмента качества. Представляется, что применение классических инструментов аналитической деятельности корпоративных служб качества автосборочных производств, с учетом решения перспективных задач в области их развития создает предпосылки для улучшений процессов управления и повышения их эффективности. В данной статье рассмотрен метод Парето, который реализует задачу выделения наиболее значимых фрагментов данных, что становится основой для создания высокоэффективных моделей прогнозирования качества продукции в процессе эксплуатации. Метод реализован на реальных данных о поступивших заявках на сервисное обслуживание автомобилей одного из ведущих отечественных автопроизводителей.

Ключевые слова: управление качеством, анализ данных, машинное обучение, прогнозирующие модели, детекция аномалий.

DOI: 10.37313/1990-5378-2025-27-3-92-98
EDN: OZPLPU

1. ВВЕДЕНИЕ

В современном мире, насыщенном данными, эффективное использование информации становится критически важным для достижения высоких результатов в области машинного обучения. Одним из ключевых этапов в процессе разработки предсказательных моделей является выделение значимых признаков, оказывающих наибольшее влияние на результаты прогнозирования, и исключение малозначимых, которые не вносят существенного вклада в обучение. Этот процесс не только способствует повышению точности моделей, но и минимизирует риск переобучения, улучшая обобщающую способность алгоритмов.

Кроме того, оптимизация данных позволяет сократить время, необходимое для обучения модели, что является важным фактором в условиях ограниченных вычислительных ресурсов и необходимости быстрой обработки информации [4].

Таким образом, применение методов редукции данных становится неотъемлемой частью разработки эффективных предсказательных моделей.

1.1. Теоретическая часть

Для оценки качества обслуживания сервисных центров необходимо разработать модель с высокой прогнозирующей способностью. Это возможно только при обучении модели на примерах, содержащих значимую информацию и отражающих реальные закономерности в данных. Однако на этапе построения модели может возникнуть проблема с наличием шумных данных и аномалий, которые способны существенно исказить результаты прогнозирования [2].

В связи с этим возникает необходимость выделить наиболее значимые фрагменты данных, способные улучшить качество прогноза. Мы используем принцип Парето для выявления тех данных, которые имеют наибольшее влияние на результаты прогнозирования, что не только оптимизирует вычислительные ресурсы, но и обеспечит высокую прогностическую способность модели [3].

Мосин Владимир Геннадьевич, кандидат физико-математических наук, доцент. E-mail: yanbacha@yandex.ru
Брагина Катарина Александровна, аспирант. E-mail: katarsis.996@gmail.com
Козловский Владимир Николаевич, доктор технических наук, профессор, заведующий кафедрой. E-mail: Kozlovskiy-76@mail.ru
Гусев Алексей Викторович, аспирант. E-mail: gusevav@aviacor.ru

В качестве наблюдаемой характеристики выберем общее число заявок, полученных каждым сервисным центром за день. Это позволит нам оценить эффективность работы центров и выявить закономерности в их обслуживании.

1.2. Постановка задачи

Выделить наиболее значимые фрагменты данных, которые позволяют обнаружить максимальное количество закономерностей и оказывают наибольшее влияние на результаты прогнозирования.

1.2.1. Предмет исследования

Предметом нашего исследования является метод, позволяющий выделить наиболее значимые примеры для обучения модели в данных.

1.2.2. Методика исследования

В данной методике применяется метод Парето для выделения наиболее значимых примеров в данных, используемых для обучения модели. Этот подход основывается на принципе 80/20, согласно которому примерно 80% результатов достигается за счет 20% причин. В контексте анализа данных это означает, что относительно небольшое количество объектов может оказывать значительное влияние на общие результаты модели. Для определения ключевых примеров мы оцениваем вклад каждого объекта в общую вариацию данных и отбираем те, которые находятся в верхних процентах по значимости. Объекты, не соответствующие установленным критериям значимости, классифицируются как менее важные и могут быть исключены из дальнейшего анализа.

1.2.3. Цель исследования

Цель — разработать алгоритм, который будет классифицировать объекты выборки как значимые, основываясь на принципе метода Парето.

1.3. Технологии

Для обработки и анализа данных мы используем язык программирования Python, среду разработки Jupyter Notebook и одни из основных библиотек: Pandas и Matplotlib. Библиотека Pandas является одной из наиболее популярных и мощных инструментов для анализа и обработки данных. Pandas предоставляет высокоуровневые структуры данных и функции, которые позволяют эффективно управлять табличными данными, что делает ее незаменимой для работы с большими объемами информации [1].

2. ОПИСАНИЕ ДАННЫХ

Используются данные о заявках одного из ведущих отечественных автопроизводителей в период с 01.01.2023 по 30.12.2024. Данные содержат информацию о заявках в 336 сервисных центрах и представлены в виде датафрейма (табл. 1).

Таблица 1 - Датафрейм с данными о заявках в сервисный центр

service_id	service_name	model	…	defect_id	defect_name	causer
44976	САМ	11183	…	2915004024000	ТЕЧЬ АМОРТИЗА	51000
44976	САМ	11183	…	6106082083000	УСТАНОВКА ШАЙ	30000
44976	САМ	11183	…	5402334000000	ДЕФЕКТ ОКАНТО	0
…	…	…	…	…	…	…
38555	ТОЛ	21703	…	8118022183000	ОБРЫВ ЦЕПИ ДО	51000
38555	ТОЛ	21703	…	8118020000000	ДЕФЕКТ ЭЛЕКТР	51000
38555	ТОЛ	21703	…	1703055004000	ДЕФОРМАЦИЯ 06	30000

Таблица имеет 20 атрибутов. Разведочный анализ показал, что в приведенном датафрейме отсутствуют пропуски. Атрибуты имеют названия, смысловое значение и типы, которые представлены в таблице 2.

Таблица 2 - Атрибуты, описывающие заявки на сервисное обслуживание

№	Название	Тип	Описание
1	service_id	int64	уникальный идентификатор заявки
2	service_name	object	город расположения сервиса
3	model	int64	модель автомобиля
4	component	int64	компонент автомобиля
5	product_id	int64	уникальный идентификатор услуги
6	product_part1_id	int64	идентификатор первой запчасти
7	product_part2_id	int64	идентификатор второй запчасти
8	date_in	object	дата поступления заявки
9	date_out	object	дата закрытия заявки
10	mileage	int64	пробег автомобиля
11	warranty	int64	гарантийный срок (в месяцах)
12	labor_cost	float64	стоимость рабочей силы
13	operating_time	float64	затраченное время
14	service_price	float64	стоимость услуги
15	part_price	float64	стоимость использованных запчастей
16	materials_price	int64	стоимость расходных материалов
17	general_price	float64	общая стоимость
18	defect_id	int64	уникальный идентификатор дефекта
19	defect_name	object	название выявленного дефекта
20	causer	int64	причина обращения

3. ПРИВЕДЕНИЕ ДАННЫХ К ФОРМАТУ ML

3.1. Определение формата ML и обоснование его применения

ML (Moment, Location) форматом данных является набор значений наблюдаемой характеристики, организованный в виде таблицы 3.

Таблица 3 – Формат ML

	location$_1$	**location$_2$**	**location$_3$**	**…**	**location$_m$**
moment$_1$	c_{11}	c_{12}	c_{13}	…	c_{14}
moment$_2$	c_{21}	c_{22}	c_{23}	…	c_{2m}
moment$_3$	c_{31}	c_{32}	c_{33}	…	c_{3m}
…	…	…	…	…	…
moment$_n$	c_{n1}	c_{n2}	c_{n3}	…	c_{nm}

Данная таблица имеет n строк, промаркированных моментами наблюдений, и m столбцов, промаркированных локациями.

Поскольку будем наблюдать значение такой характеристики, как количество заявок в сервисных центрах в разные дни, приведем необходимые нам данные из основного датафрейма (табл. 1) к формату ML (Moment, Location) [6].

3.2. Приведение данных к формату ML

3.2.1. Чтение данных

С помощью функции read_csv() из библиотеки Pandas произведем чтение данных о заявках на сервисное обслуживание автомобиля в формате датафрейма (табл. 1).

3.2.1. Группировка данных при помощи агрегирующей функции

Чтобы организовать данные в формате ML, применим к датафрейму функцию для агрегации данных pivot_table() из библиотеки Pandas. При настройке параметров функции в качестве индекса возьмем атрибут ‘date_in’, в качестве столбца выберем ‘service_id’ и установим агрегирующую функ-

цию size() для подсчета количества заявок на каждую дату для каждого сервисного центра:

ML = df.pivot_table(index='date_in', columns='service_id', aggfunc='size')

Получим сводную таблицу 4, содержащую 729 строк и 335 столбцов, на пересечении которых располагается значение, характеризующее количество заявок в указанную дату (Moment) в конкретном сервисном центре (Location).

Таблица 4 – Данные в формате ML

	10363	**10461**	**10901**	...	**68163**	**69063**	**69363**
2023-01-01	0	0	0	...	0	0	0
2023-01-02	0	28	0	...	0	0	0
2023-01-03	0	19	0	...	0	0	0
...	...	...	...	...	...	...	...
2024-12-28	2	12	7	...	0	6	0
2024-12-29	2	4	0	...	0	4	0
2024-12-30	0	0	0	...	0	0	0

4. ПАРЕТО-АНАЛИЗ ЛОКАЦИЙ

При визуализации портретов локаций[1] из таблицы 4 мы можем увидеть ситуацию, изображённую на рисунке 1. На втором графике (рис. 1) почти все значения наблюдаемой характеристики находятся на нулевом уровне, за исключением нескольких пиков в начале. Это указывает на малое количество заявок в сервисном центре.

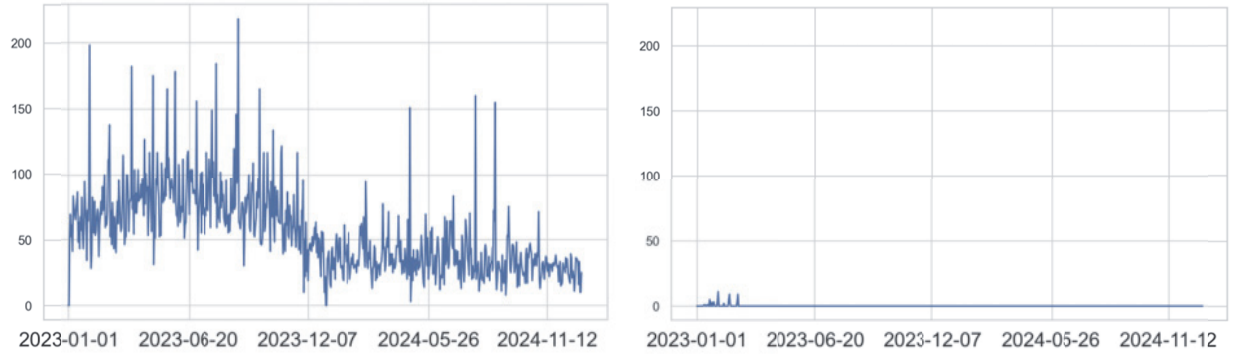

Рис. 1. Портрет локаций

Таким образом, портрет локации на втором графике (рис. 1) является примером, который не вносит существенного вклада в обучение модели. Он искажает данные и лишает возможности выявить реальные закономерности. В результате это затрудняет построение высокоточного прогноза.

Совсем иначе обстоит ситуация на первом графике (рис. 1), где значений наблюдаемой характеристики достаточно для обнаружения реальных закономерностей, и именно такие локации будем считать значимыми для дальнейшего построения прогнозирующей модели. Построим алгоритм, позволяющий произвести редукцию данных из таблицы 3 при помощи метода Парето, чтобы сконцентрироваться только на значимых локациях.

4.1. Расчет общего количества заявок

Общим количеством заявок является количество записей в основном датафрейме (табл. 1). Для определения количества записей применим к датафрейму df функцию len() из стандартной библиотеки Python:

total_profit = len(df)

4.2. Расчет процентов

Определим количество заявок относительно общего числа заявок при помощи умножения общего числа заявок на 0.8:

target_profit = total_profit * 0.8

[1] Портрет локации — это некая кривая, характеризующая изменение характеристики с течением времени в данной локации

4.3. Расчет количества заявок в каждом сервисном центре и сортировка данных

Сгруппируем данные по сервисным центрам, применив к датафрейму df функцию groupby() и агрегирующую функцию size() из библиотеки Pandas для подсчета количества заявок в каждом сервисном центре:

service_requests = df.groupby(‹service_id›).agg(‹size›)

Далее отсортируем полученную таблицу по количеству заявок в порядке убывания с помощью функции sort_values() из библиотеки Pandas и ее параметра ascending=False:

service_requests = service_requests.sort_values(ascending=False)

Переименуем столбец с количеством заявок в полученной таблице с помощью функции rename() из библиотеки Pandas:

service_requests = service_requests.rename(columns={0: ‘count’})

Теперь сервисные центры с наибольшим количеством заявок располагаются в начале таблицы. Получим сводную таблицу 5.

Таблица 5 - Рейтинг сервисных центров

№	service_id	count
1	44360	76282
2	11785	39951
3	44560	37695
...	...	...
333	42013	1
334	44525	1
335	39402	1

Таблица 6 – Кумулятивная сумма количества заявок

№	service_id	cumsum
1	44360	76282
2	11785	116233
3	44560	153928
...	...	...
333	42013	1227901
334	44525	1227902
335	39402	1227903

Результаты распределения заявок по сервисным центрам представлены на рисунке 2, где на графике по оси абсцисс отображаются номера рейтингов сервисных центров из таблицы 5, по оси ординат — количество заявок.

4.4. Расчет кумулятивного количества заявок

Вычислим кумулятивное количество заявок для каждого сервисного центра. Для вычисления применим функцию cumsum() из библиотеки Pandas к столбцу ‘count’ из таблицы 5:

accumulated_profit = service_requests[[‘count’]].cumsum()

Получим таблицу 6, в каждой строке которой содержится кумулятивное количество заявок. Это значение вычисляется путем суммирования количества заявок из таблицы 5. Таким образом, каждая строка таблицы 6 показывает общее количество заявок, поступивших до текущего сервисного центра включительно. Это позволяет нам увидеть, как растет общее количество заявок по мере добавления новых данных.

4.5. Определение ключевых сервисных центров

На этом шаге мы находим сервисные центры, которые в сумме приносят 80% от общего количества заявок (рис. 2). Для этого произведем фильтрацию таблицы, полученной на предыдущем шаге, по столбцу ‘cumsum’:

accumulated_profit[accumulated_profit[‘cumsum’] <= target_profit]

График на рисунке 2. демонстрирует закон Парето, показывая, что большинство заявок сосредоточено в небольшом количестве сервисных центров, а большинство остальных имеют значительно меньшее количество заявок.

4.6. Редуцированные данные в формате ML

В результате применения Парето анализа получено 68 ключевых сервисных центров, используем их для редукции данных в таблице 4. Выделим ключевые сервисные центры в массив с именем top и выберем столбцы с этими сервисными центрами из таблицы ML (табл. 4):

ML_new = ML[top]

Получим таблицу 7, которая имеет 729 строк и 68 столбцов.

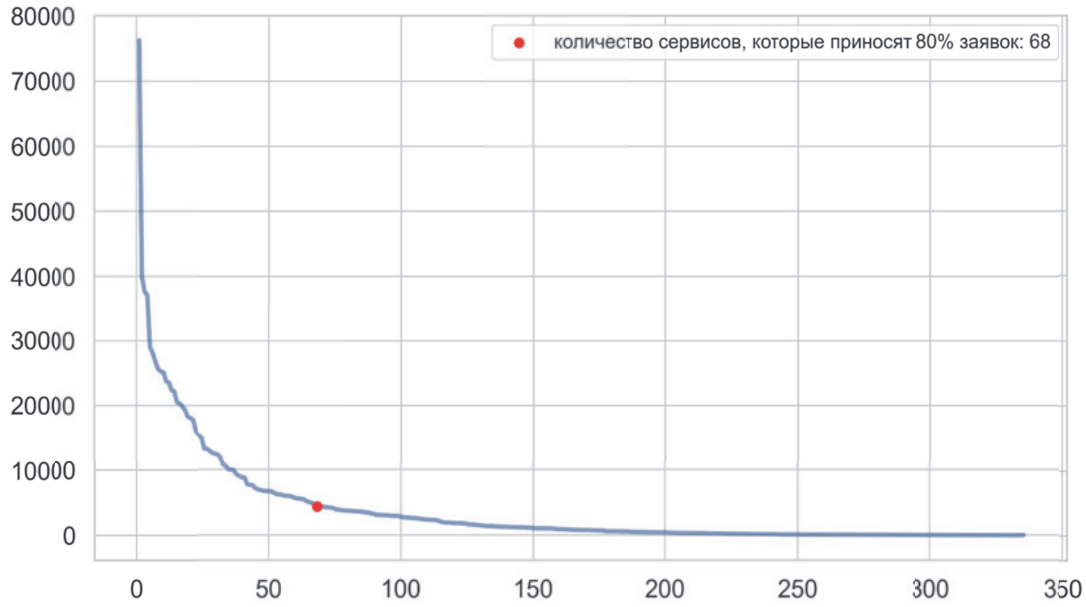

Рис. 2. Распределение заявок по сервисным центрам

Таблица 7 – Редуцированные данные в формате ML

	44360	**11785**	**44560**	...	**14060**	**16463**	**12963**
2023-01-01	0	0	0	...	0	0	0
2023-01-02	25	0	0	...	0	0	0
2023-01-03	88	50	0	...	0	0	0
...	...	...	...	...	...	...	...
2024-12-28	121	10	21	...	15	0	1
2024-12-29	78	10	13	...	4	0	0
2024-12-30	57	25	0	...	0	0	0

5. РЕЗУЛЬТАТЫ И ВЫВОДЫ

В данной работе мы сосредоточились на применении метода Парето для редукции данных, что является ключевым этапом в подготовке информации для анализа и построения моделей. Мы исследовали, как принцип 80/20 может быть использован для выявления наиболее значимых фрагментов данных, что способствует более эффективному и точному прогнозированию. В результате применения метода Парето было выделено 68 ключевых сервисных центров из общего числа 335.

Обучение модели на качественных данных, собранных из этих сервисных центров, существенно снижает влияние белого шума, который может искажать результаты анализа. Это позволяет нам более эффективно выявлять значимые тенденции и паттерны, открывая возможность для создания высокоточной прогнозирующей модели в автосервисной отрасли. Кроме того, редукция данных уменьшает необходимые вычислительные мощности для построения модели, что делает процесс более экономичным и быстрым.

Сосредоточив внимание на выявленных ключевых сервисных центрах, мы сможем глубже анализировать тренды и разрабатывать стандарты обслуживания. Это позволит быстро выявлять аномалии и проблемные сервисы — те, где обнаружены отклонения от норм, установленных прогнозирующей моделью. В результате качество аудита значительно возрастет, так как мы сможем концентрироваться на конкретных участках, требующих особого внимания. Это, в свою очередь, позволит существенно сократить время и ресурсы, затрачиваемые на аудит, а также повысить точность принимаемых решений. В конечном итоге, применение метода Парето в нашем анализе создаст более устойчивую и эффективную систему управления качеством услуг в автосервисной отрасли, что, безусловно, будет способствовать улучшению общего уровня обслуживания клиентов и повышению их удовлетворенности [5, 6].

СПИСОК ЛИТЕРАТУРЫ

1. *Хейдт, М.* Изучаем Pandas / М. Хейдт. – М.: ДМК Пресс, 2018. – 438 с. – ISBN 978-5-97060-625-4.
1. *Бурков, А.* Машинное обучение без лишних слов / А. Бурков. – СПб: Питер, 2020. – 192 с. – ISBN 978-5-4461-1560-0.
3. *Вьюгин, В.В.* Математические основы теории машинного обучения и прогнозирования / В. В. Вьюгин. – М.: МЦИМО. – 2013. – 387 с. ISBN: 978-5-4439-2014-6.
4. *Бринк, Х.* Машинное обучение / Х. Бринк, Дж. Ричардс, М. Феверолф. – СПб.: Питер, 2017. – 336 с. – ISBN 978-5-496-02989-6.
5. *Мосин, В.Г.* Предиктивная детекция аномалий в процессе гарантийного обслуживания автомобилей / В.Г. Мосин // Известия Тульского государственного университета. Технические науки. – 2024. – № 12. – С. 506–513.
6. *Мосин, В.Г.* Методика МССР (modeling, calibration, challenge, production) в сравнительном анализе пунктов гарантийного обслуживания автомобилей / В.Г. Мосин, В.Н. Козловский, А.С. Клентак, О.В. Пантюхин // Известия Тульского государственного университета. Технические науки. – 2025. – № 1. – С. 272–289.

PARETO ANALYSIS OF THE PERFORMANCE QUALITY OF AUTOMAKERS' SERVICE CENTERS

© 2025 V.G. Mosin[1], K.A. Bragina[2], V.N. Kozlovsky[1], A.V. Gusev[1]

[1] Samara State Technical University, Samara, Russia
[2] Bauman Moscow State Technical University, Moscow, Russia

The application of the Pareto analysis method to solve problems in the field of quality management remains in demand and is extremely relevant. Currently, we are seeing a sharp increase in electronic data reflecting the flow of various processes in the management system of organizations, as well as data reflecting the stages of the product life cycle. Accordingly, the growth of data requires solving scientific and technical problems aimed at developing methods and tools for monitoring and analyzing data. Their improvement and corresponding development is required in the context of solving problems in the field of digitalization and informatization of quality management processes. It seems that the use of classical tools for analytical activities of corporate quality services of car assembly plants, taking into account the solution of promising problems in the field of their development, creates the prerequisites for improving management processes and increasing their efficiency. This article discusses the Pareto method, which implements the task of identifying the most significant fragments of data, which becomes the basis for creating highly effective models for predicting product quality during operation. The method is implemented on real data on incoming requests for service of cars of one of the leading domestic automakers.

Keywords: quality management, data analysis, machine learning, predictive models, anomaly detection.

DOI: 10.37313/1990-5378-2025-27-3-92-98
EDN: OZPLPU

REFERENCES

1. *Hejdt, M.* Izuchaem Pandas / M. Hejdt. – M.: DMK Press, 2018. – 438 s. – ISBN 978-5-97060-625-4.
2. *Burkov, A.* Mashinnoe obuchenie bez lishnih slov / A. Burkov. – SPb: Piter, 2020. – 192 s. – ISBN 978-5-4461-1560-0.
3. *V'yugin, V.V.* Matematicheskie osnovy teorii mashinnogo obucheniya i prognozirovaniya / V. V. V'yugin. – M.: MCIMO. – 2013. – 387 s. ISBN: 978-5-4439-2014-6.
4. *Brink, H.* Mashinnoe obuchenie / H. Brink, Dzh. Richards, M. Feverolf. – SPb.: Piter, 2017. – 336 s. – ISBN 978-5-496-02989-6.
5. *Mosin, V.G.* Prediktivnaya detekciya anomalij v processe garantijnogo obsluzhivaniya avtomobilej / V.G. Mosin // Izvestiya Tul'skogo gosudarstvennogo universiteta. Tekhnicheskie nauki. – 2024. – № 12. – S. 506–513.
6. *Mosin, V.G.* Metodika MCCP (modeling, calibration, challenge, production) v sravnitel'nom analize punktov garantijnogo obsluzhivaniya avtomobilej / V.G. Mosin, V.N. Kozlovskij, A.S. Klentak, O.V. Pantyuhin // Izvestiya Tul'skogo gosudarstvennogo universiteta. Tekhnicheskie nauki. – 2025. – № 1. – S. 272–289.

Vladimir Mosin, Ph.D. in Physics and Mathematics, Associate Professor. E-mail: yanbacha@yandex.ru
Katarina Bragina, Postgraduate Student. E-mail: katarsis.996@gmail.com
Vladimir Kozlovsky, D.Sc. (Eng.), Professor, Head of Department. E-mail: Kozlovskiy-76@mail.ru
Aleksey Gusev, Postgraduate Student. E-mail: gusevav@aviacor.ru