УДК 519.246.8

СТАТИСТИЧЕСКИЙ АНАЛИЗ И МОДЕЛИРОВАНИЕ ДИНАМИКИ ЧИСЛА ДОРОЖНО-ТРАНСПОРТНЫХ ПРОИСШЕСТВИЙ В КЕМЕРОВСКОЙ ОБЛАСТИ В РАЗРЕЗЕ ПОМЕСЯЧНОЙ СТАТИСТИКИ

Касьяненко Д.А., студент гр. ПМИ-221, III курс
Научный руководитель: Мешечкин В.В., к.ф.-м.н., доцент
Кемеровский государственный университет
г. Кемерово

Проблема дорожно-транспортных происшествий (ДТП) в Кемеровской области – Кузбассе является актуальной научно-практической задачей в контексте обеспечения безопасности дорожного движения. Высокий уровень аварийности в регионе представляет серьёзную угрозу для жизни и здоровья населения, что подтверждается статистическими данными. Помимо этого, ДТП влекут за собой значительный экономический ущерб для региона, снижая общий уровень благосостояния и требуя серьёзных затрат на ликвидацию последствий и восстановление инфраструктуры. В условиях роста автомобилизации и интенсивности дорожного движения актуальность разработки эффективных моделей прогнозирования и принятия мер по предотвращению ДТП возрастает, что обусловливает необходимость комплексного подхода к изучению данной проблемы.

В то же время, существующий математический аппарат позволяет использовать доступные статистические данные для построения эффективных моделей прогнозирования статистических показателей, связанных с безопасностью дорожного движения, что, в свою очередь, позволяет оценивать текущую и предполагаемую эффективность реализации мер, направленных на обеспечение безопасности дорожного движения, выявлять тренды в существующих данных с целью критического оценивания текущей динамики дорожно-аварийной обстановки, определять степень взаимосвязи статистических показателей аварийности для поиска потенциальных путей воздействия на текущую дорожную сеть с целью снижения общей дорожно-транспортной аварийности, смертности и ранения в ДТП, уменьшения возможного социально-экономического ущерба их последствий и пр.

В данной статье представлены результаты исследования динамики числа дорожно-транспортных происшествий в Кемеровской области с помощью метода математического моделирования.

Для построения модели использовались статистические данные, размещённые в свободном доступе Министерством внутренних дел Российской Федерации и формируемые на основании приказа Федеральной службы государственной статистики (Росстат) от 21 мая 2014 года N 402 «Об утвержде-

нии статистического инструментария для организации Министерством внутренних дел Российской Федерации федерального статистического наблюдения за состоянием безопасности дорожного движения» (до 2017 года включительно) и приказа Росстата от 07.12.2017 N 810 «Об утверждении статистического инструментария для организации Министерством внутренних дел Российской Федерации федерального статистического наблюдения за состоянием безопасности дорожного движения» (начиная с 2018 года). Данные доступны всем гражданам без ограничений.

В целях моделирования из исходных данных был сформирован временной ряд, содержащий сведения о количестве ДТП за каждый месяц с января 2015 года по декабрь 2024 года включительно (10 полных лет). На рисунке 1 представлен график этого ряда.

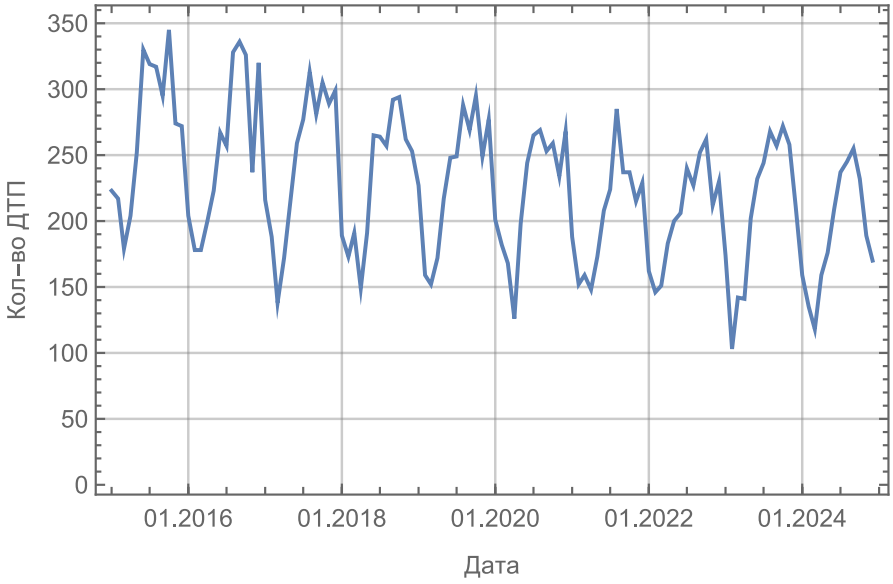

Рисунок 1. График количества ДТП по месяцам

В целях графического анализа данных были построены гистограмма частот ДТП (рисунок 2), вероятностный график точек ряда по отношению к нормальному распределению (рисунок 3), коробчатая диаграмма (рисунок 4).

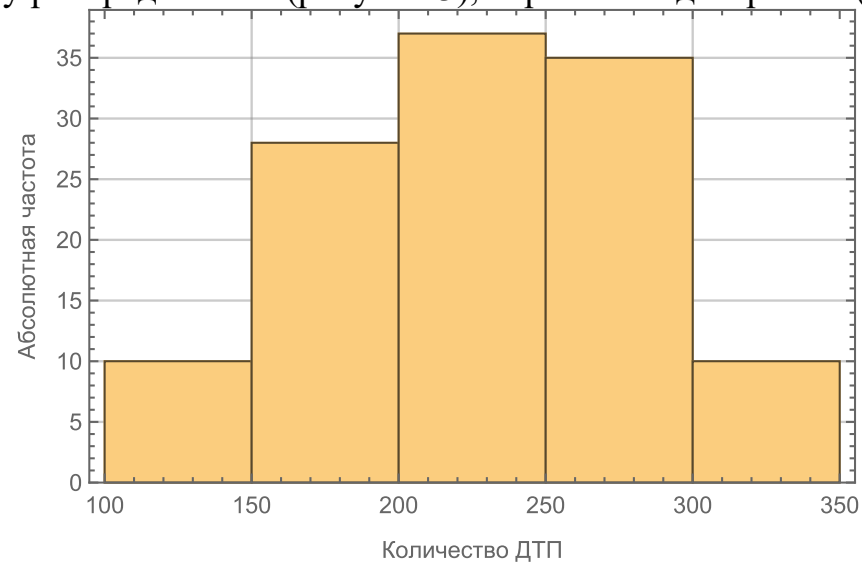

Рисунок 2. Гистограмма частот ДТП

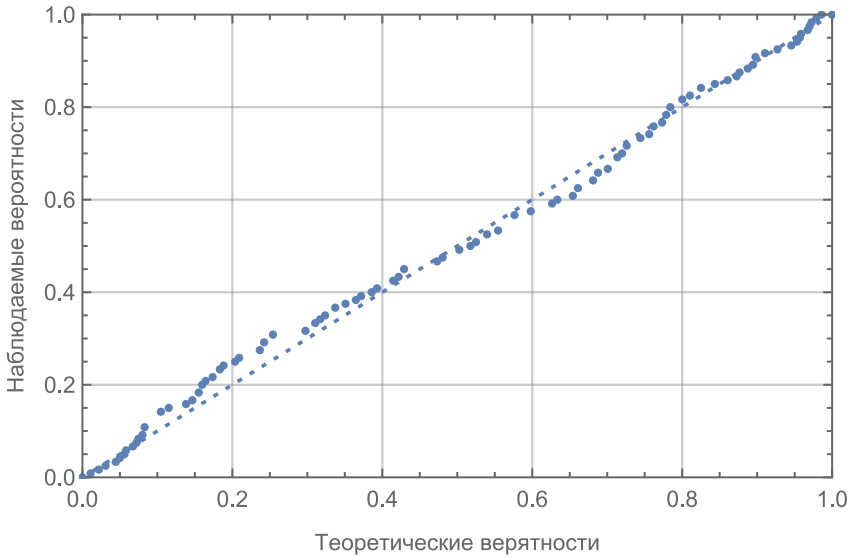

Рисунок 3. Вероятностный график ряда по отношению к нормальному распределению

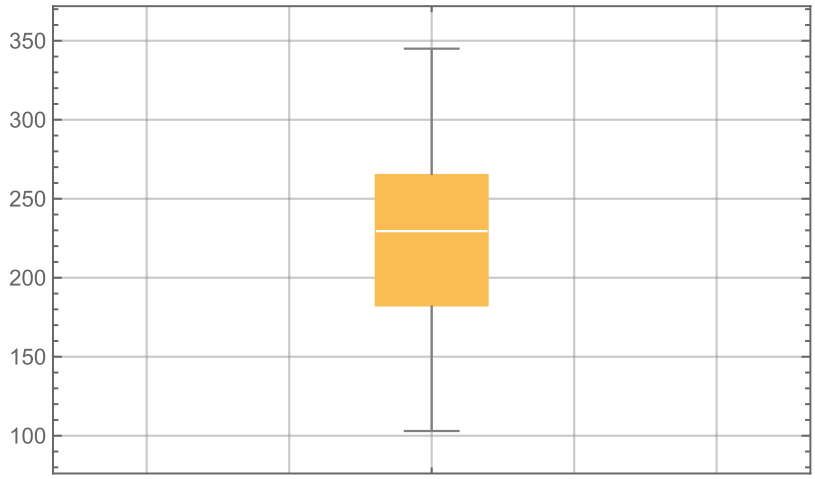

Рисунок 4. Коробчатая диаграмма данных ряда

Приведённые графики позволили сделать следующие выводы о структуре данных:

1. Временной ряд, вероятно, имеет тенденцию на убывание.
2. В значениях ряда, вероятно, присутствует периодичность, то есть ряд имеет сезонную составляющую.
3. Данные о количестве ДТП, вероятно, распределены нормально и без выбросов.

Далее была проведена дескриптивная аналитика данных временного ряда, основные результаты которой представлены в Таблице 1.

Таблица 1. Описательные статистики ряда

Показатель	Значение
Количество точек	120
Среднее значение	226,63
Медиана	229,5
Дисперсия	2926,47
Стандартное отклонение	54,097

Минимум	103
Максимум	345
Коэффициент вариации (*CV*)	0,239
Коэффициент асимметрии (*As*)	0,0043
Коэффициент эксцесса (*Es*)	2,296

Выводы из описательного анализа данных:

1. Сравнение показателей среднего (226,63) и медианы (229,5) свидетельствует о левосторонней асимметрии данных.

2. Значение коэффициента вариации $CV = 0{,}239$ свидетельствует об однородности исходных данных ($CV < 0{,}33$) и значительной степени их рассеяния ($CV > 0{,}2$).

3. Значение показателя асимметрии $As = 0{,}0043$ свидетельствует о крайне незначительной степени асимметрии ($|As| < 0{,}25$).

4. Значение коэффициента эксцесса $Es = 2{,}296$ свидетельствует о плосковершинном распределении ($Es < 3$), близком к нормальному ($Es \sim 3$).

Из графического и дескриптивного анализа данных было выдвинуто предположение о нормальности распределения, которое было проверено на статистических тестах Андерсона-Дарлинга, Барингхауса-Хензе, Крамера-Мизеса-Смирнова, Харке-Бера, трёх тестах Мардиа, критериях Пирсона и Шапиро-Уилка. Все 9 статистических тестов позволили принять гипотезу о нормальности распределения данных на уровне значимости в 0,05.

Далее для проверки предположения о наличии тенденции в ряду динамики ряд был разделён на две половины по 60 точек в каждой. На основании тестов Бартлетта, Брауна-Форсайта, Коновера, Фишера и Левена была принята гипотеза о равенстве дисперсий половин ряда, что позволило применить критерий разностей средних уровней. Ряд статистических тестов, включая параметрический критерий Стьюдента, подтвердили статистическую значимость различий средних значений половин ряда. Таким образом, на основании метода проверки разностей средних уровней было подтверждено наличие тенденции в ряду динамики.

Для регрессионного моделирования трендовой составляющей ряда было выбрано уравнение прямой. Так как данные были распределены нормально, не содержали выбросов и аномальных значений, для оценки параметров уравнения прямой был использован обычный (неробастный) метод наименьших квадратов. Полученное уравнение линии тренда имеет вид:

$$y = 260{,}294 - 0{,}556372t.$$

В этом уравнении y – кол-во ДТП в месяц, t – номер месяца, отсчитываемый от января 2015 года (январь 2015 года принимается за $t = 1$).

Для оценки значимости коэффициентов полученной модели были рассчитаны t-статистики, на основании которых был определён p-уровень значи-

мости, а также были рассчитаны доверительные интервалы параметров регрессии. Данные приведены в таблице 2.

Таблица 2. Данные о значимости коэффициентов регрессии

	Значение коэффициента	*t*-статистика	*p*-уровень	Доверительный интервал
1	260,294	27,9279	$7{,}829 \cdot 10^{-54}$	(241,837; 278,75)
t	– 0,556372	– 4,16165	0,0000603	(– 0,82112; – 0,29163)

Оба коэффициента уравнения регрессии принимаются статистически значимыми, так как их *p*-уровни не превосходят принятого значения 0,05, а точка 0 не включена в доверительные интервалы коэффициентов.

На рисунке 5 изображён исходный временной ряд с нанесённой на график линией тренда. Поскольку текущая модель описывает лишь трендовую составляющую ряда, оценка качества модели на данный момент не имеет смысла, однако текущее уравнение тренда позволяет выделить сезонную составляющую ряда (рисунок 6).

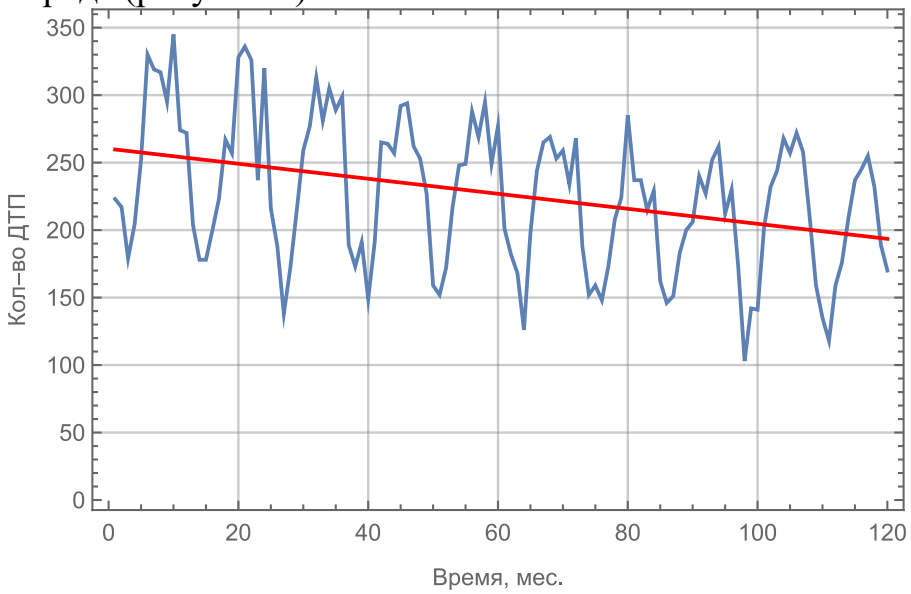

Рисунок 5. График временного ряда (синий) с нанесённой линией тренда (красный)

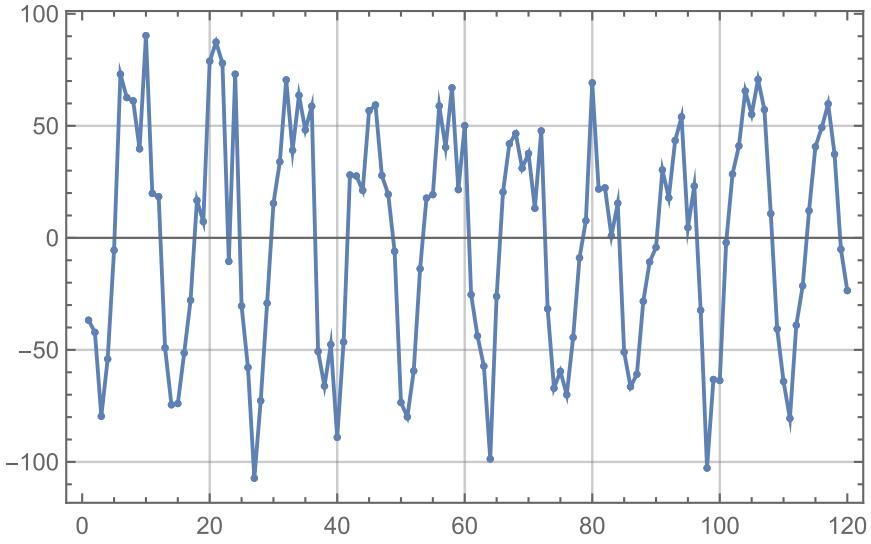

Рисунок 6. График сезонной составляющей ряда

Для выделенной сезонной составляющей временного ряда можно определить периодичность колебаний, например, графически изобразив коэффициенты серийной корреляции (см. рисунок 7)

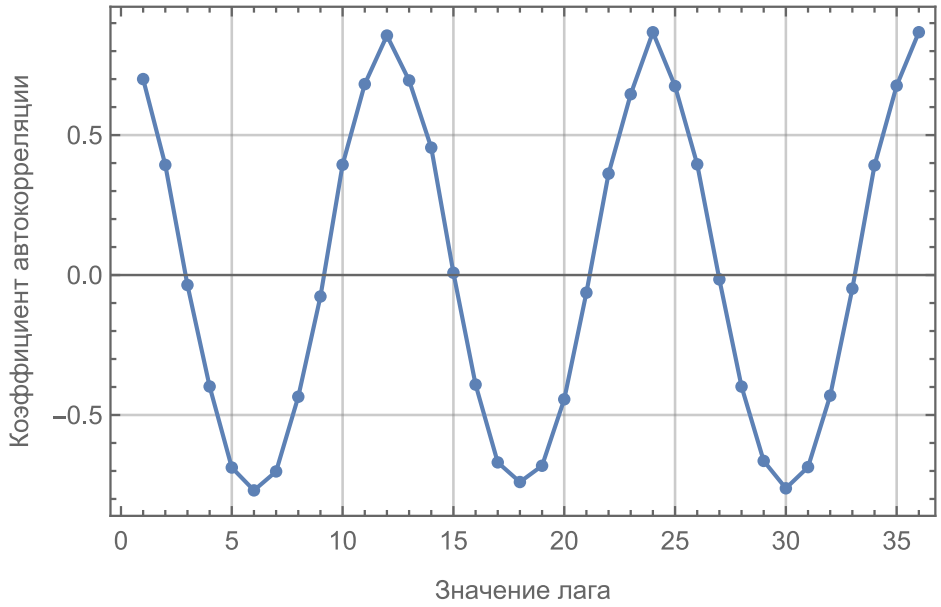

Рисунок 7. График серийной корреляции ряда

По графику коэффициентов автокорреляции видно, что период колебаний сезонной составляющей ряда равен 12 (так как пиковые значения коэффициентов корреляции приходятся на значения лага, кратные 12). Для моделирования сезонной составляющей ряда был выбран ряд Фурье из 6 гармоник с частотами вида $\pi k/6$. Для аппроксимации значений был также выбран обычный метод наименьших квадратов из соображений, описанных ранее. Коэффициенты полученной модели приведены в таблице 3, а хорошие аппроксимирующие свойства полученного уравнения были подтверждены расчётами метрик, включающих относительную ошибку аппроксимации (*MAPE* = 6,817%) и коэффициент детерминации модели (R^2 = 0,86), статистическая значимость которого была подтверждена критерием Фишера.

Таблица 3. Таблица коэффициентов сезонной составляющей модели

Гармоника	Значение коэффициента	Доверительный интервал коэффициента
$\mathrm{Cos}(\pi x/6)$	–2,11917	(–7,18433; 2,94598)
$\mathrm{Sin}(\pi x/6)$	–63,3664	(–68,4316; –58,3013)
$\mathrm{Cos}(2\pi x/6)$	12,7397	(7,67455; 17,8049)
$\mathrm{Sin}(2\pi x/6)$	–0,934797	(–5,99995; 4,13036)
$\mathrm{Cos}(3\pi x/6)$	1,42304	(–3,64212; 6,48819)
$\mathrm{Sin}(3\pi x/6)$	1,87696	(–3,18819; 6,94212)
$\mathrm{Cos}(4\pi x/6)$	6,42304	(1,35788; 11,4882)
$\mathrm{Sin}(4\pi x/6)$	1,17989	(–3,88527; 6,24504)

$\mathrm{Cos}(5\pi x/6)$	5,41525	(0,350095; 10,4804)
$\mathrm{Sin}(5\pi x/6)$	6,02428	(0,959122; 11,0894)
$\mathrm{Cos}(6\pi x/6)$	5,44485	(1,86325; 9,02646)

Для остатков регрессионной модели были проверены условия выполнения теоремы Гаусса-Маркова: остатки модели распределены нормально (на основании тех же тестов, что проводились для исходных данных), математическое ожидание остатков равно нулю (6 тестов, включая *t*-критерий Стьюдента), остатки гомоскедастичны (тесты Уайта, Голдфелда-Квандта, Бриша-Пэгана) и в них отсутствуют автокорреляции (критерий Дарбина-Уотсона, тесты Бокса-Пирса и Льюнг-Бокса). Выполнение условий теоремы подтверждает эффективность и несмещённость оценок коэффициентов регрессии, полученных методом наименьших квадратов.

В результате проведённого исследования была разработана математическая модель, описывающая динамику числа дорожно-транспортных происшествий в Кемеровской области в разрезе помесячной статистики. Модель позволяет выявить закономерности и тенденции в изменении количества ДТП как на протяжении календарного года, так и между годами.

Полученные результаты могут быть использованы для прогнозирования числа дорожно-транспортных происшествий в регионе, что, в свою очередь, может помочь в планировании мероприятий по обеспечению безопасности дорожного движения, распределении ресурсов служб экстренного реагирования и оптимизации работы правоохранительных органов.

Результаты моделирования представлены в виде графика на рисунке 8, где наряду с расчётными уровнями ряда (красная линия) для сравнения показаны фактические уровни (синяя линия).

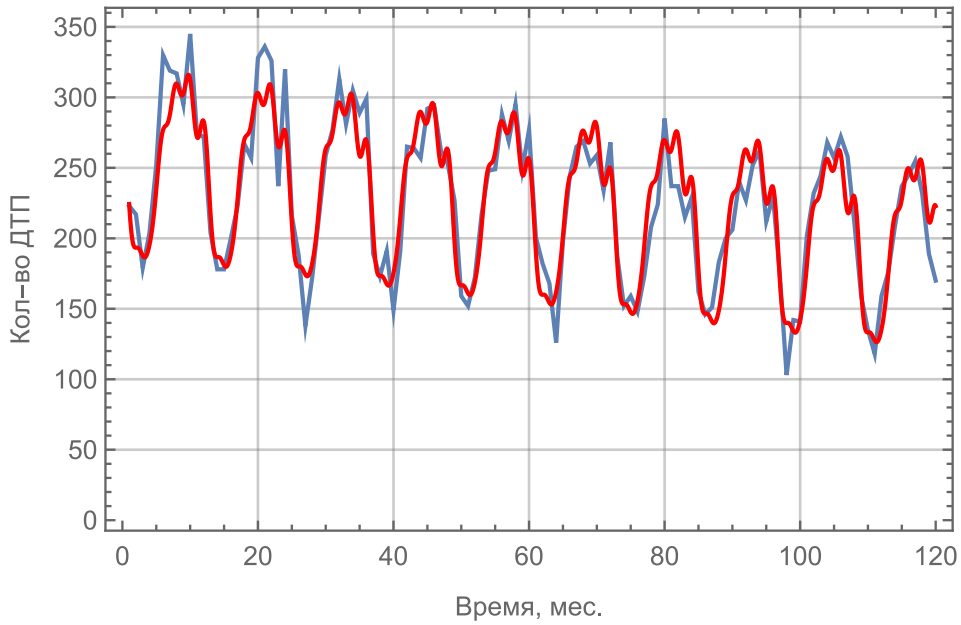

Рисунок 8. График временного ряда (синий) с наложенной кривой аппроксимирующего уравнения (красной)

Список литературы:

1. Кобзарь, А. И. Прикладная математическая статистика: для инженеров и научных работников / А. И. Кобзарь. – М.: ФИЗМАТЛИТ, 2006. – 816 с.

2. Фёрстер, Э. Методы корреляционного и регрессионного анализа / Э. Фёрстер, Б. Рёнц: пер. с нем. – М.: Финансы и статистика, 1983. – 302 с.

3. Брюс, П. Практическая статистика для специалистов Data Science / П. Брюс: пер. с англ. – СПб.: БХВ-Петербург, 2018. – 304 с.

4. Уатт, Дж. [и др.] Машинное обучение: основы, алгоритмы и практика применения / пер. с англ. – СПб.: БХВ-Петербург, 2022. – 640 с.

5. Сведения о показателях состояния безопасности дорожного движения [Электронный ресурс] / ГИБДД МВД России. – URL: http://stat.gibdd.ru (дата обращения: 18.02.2025).