ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования Национальный исследовательский университет «Высшая школа экономики»

Факультет компьютерных наук

КУРСОВАЯ РАБОТА

БАЙЕСОВСКИЙ ПОДХОД ДЛЯ АНАЛИЗА ДТП
BAYESIAN APPROACH TO CAR ACCIDENTS ANALYSIS

по направлению подготовки 01.04.02 Прикладная математика и информатика

образовательная программа «Науки о Данных»

Студент группы АИД-Б20:

Гармидер Петр Александрович

Научный руководитель:

Борис Демешев

Москва 2021

Содержание

1 Введение **3**

1.1 Актуальность в РФ 3

1.2 Обзор литературы 4

1.3 Методология и цели 6

2 Данные **7**

2.1 Обзор 7

2.2 Графический анализ данных 7

1 Введение

1.1 Актуальность в РФ

Транспорт является важной экономической и социальной составляющей жизни населения. Безопасность и эффективность транспортного передвижения непосредственно влияют на качество жизни. Пусть юридически, дороги могут находиться не только в государственной собственности, но и принадлежать частным, юридическим лицам, но по факту за качество и безопасность большей части дорог отвечает государство. Притом, у государства есть множество вариантов воздействия на дорожное движение: качество и расположение построенных дорог, инфраструктура прилагаемая к дорогам, установление и контроль скоростных режимов на разных участках дорог и т.д.

Правительство Российской Федерации, понимая важность вопроса выше утверждает Стратегию безопасности дорожного движения в Российской Федерации на 2018 — 2024 годы, где как раз подтверждается понимание государства о наличии связи безопастности движения и качестве жизни населения. Принятая стратегия, показывает, что вопрос с безопасностью дорожного движения в России остается открытым.

Основываясь на данных карточек ДТП г. Москва за 2015-2020 года, можно заметить, что кол-во погибших в ДТП снижается из года в год (см. рис. 1), несмотря на отсутствие какого либо тренда в количестве ДТП за указанные года.

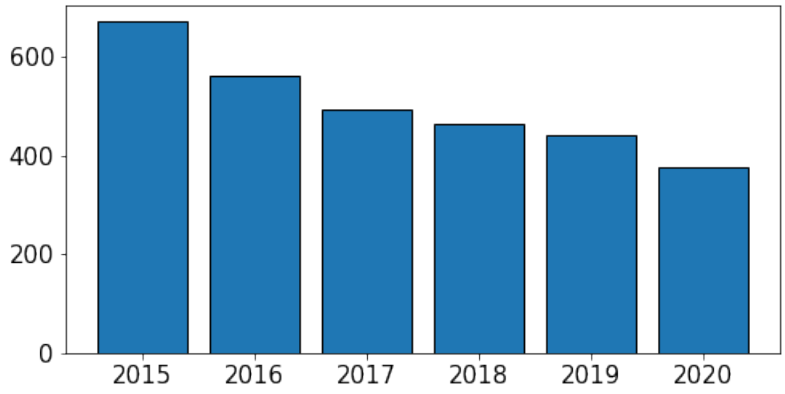

Рис. 1: Количество погибших в ДТП (г. Москва)

Как упоминается в [1] согласно относительно свежим данным, есть несколько основных причин возникновения ДТП в РФ:

- нахождение детей на дороге или недалеко от нее
- загруженность дорог по выходным
- пренебрежение правилами дорожного движения водителем

- опрокидывание транспортного средства
- наезд на пешехода

Каждая вышеперечисленная причина в идеале требует отдельного анализа. Достаточно легко поверить, что не существует единого правила, которое способно ослабить риск каждого ДТП, однако верхнеуровневый анализ позволит понять некоторые самые «проблемные» места, работа над которыми даст максимальный эффект.

Наличие актуальной статистики дорожно-транспортных происшествий, в совокупности с анализом текущей ситуации позволит государству своевременно корректировать меры и действия в рамках принятой стратегии.

1.2 Обзор литературы

Существует не так много статей посвященных анализу ситуации с ДТП в России. Основной причиной для этого считаю доступность и структурированность имеющихся данных. Существующие источники являются достаточно «сырыми», что осложняет задачу для написания больших аналитических работ. Однако, исследования на эту тему все-таки есть.

В [2] авторы рассматривают дорожную ситуацию на верхнем уровне, оперируя федеральными округами и областями. Авторы подмечают важность плотности дорог (площадь на 1000 км территории). Наблюдается рост этого показателя во всей Российской Федерации, однако данный рост тяжело наблюдать оперируя отдельными федеральными округами. Из работы также можно понять, что в недалеком прошлом был небольшой ежегодный рост введения дорог с твердым покрытием. Авторы подмечают, что качество дорог является важный фактором не только для снижения количества, но и для степени тяжести ДТП. На уровень транспортных происшествий также влияет количество автомобилей. По данным на конец 2013 года в Российской Федерации насчитывалось порядка 53 321 тысяч транспортных средств, при росте в 17.5% по отношению к прошлому году. Для того, чтобы оценить влияние количества транспортных средств на количество случае ДТП в статье строится регрессионная модель из которой следует, что это влияние действительно есть. Более того, количество автомобилей объясняют порядка 90% дисперсии случаев ДТП по разным регионам. Статья в целом посвящена итогам Федеральной целевой программы по повышению безопасности дорожного движения в 2006—2012 гг. Авторы отмечают успех данной программы, но заявляют, что ситуация на дорогах все еще требует пристального внимания со стороны правительства.

Другая, достаточно свежая статья [3] посвящена анализу статистики ДТП на уровне всей страны. Авторы также отмечают бурный рост количества автомобилей зарегистрированных в РФ, что конечно ведет к дополнительным транспортным проблемам: снижение скорости сообщения из-за большей дорожной загрузки, увеличение выбросов вредных веществ, а также перерасход топлива. В своей работе, авторы делают разрез пострадавших в ДТП по годам и подмечают, что не смотря на то, что общее количество ДТП и пострадавших в них уменьшается за последние

5 лет, существует рост количества пострадавших/погибших в ДТП детей в возрасте до 16 лет. Также авторы подмечают, что самая частая группа пострадавших в ДТП — это люди с активным образом жизни от 21 до 40 лет. Такая ситуация, по словам авторов негативно сказывается на возрастную пирамиду населения и другие демографические показатели РФ. Авторы в своей работе отмечают снизившееся количество случаев ДТП в первую половину 2020 года, объясняя это снизившимся числом автомобилей на дороге. Однако, совсем недавно в [4] отметили, что этот эффект был временным и в итоге локдаун 2020-го года никак не повлиял на количество фатальных случаев в ДТП. Как отмечают в ГИБДД, несмотря на сокращение случаев в ДТП, смертность в них, как и ситуация с вождением стала «нестабильной»[1]. По мнению службы, водители вновь севшие за руль после локдауна растеряли свои навыки вождения, а некоторые — и чувство опасности, подавшись эйфории после ослабления некоторых ограничений.

В другой работе [5] затрагивался вопрос прогнозирования вероятности ДТП с участием пешеходов. Авторы располагали похожими данным, что использовались и в данном исследовании — карточки ДТП скачанные из сайта stat.gibdd.ru. Авторы описывают методологию сбора и построения витрины данных для последующей подачи на вход моделям машинного обучения. В работе выделяются основные факторы, которые могут иметь предсказательную силу, а также тонкостям сбора подобных данных. В качестве прогнозной модели авторы советуют использовать реализацию градиентного бустинга над деревьями Catboost — библиотека с открытым доступом, разработанный компанией Яндекс. В статье рассказываются преимущества над существующими реализациями, а также даются рекомендации по настройке параметров модели. Авторы подчеркивают «несовершенство» данных в карточках ДТП, которые могут быть результатом человеческих ошибок. В качестве основной мотивации для создания подобной модели, выделяется высокая доля наездов на пешеходов, где наблюдается наивысшее среднее количество пострадавших в ДТП. Остается за кадром построение модели, как и формальная постановка задачи.

В другой работе [6] производится анализ динамики пострадавших в ДТП. На момент исследования данных наблюдалось снижение количества пострадавших и погибших на существенных статистически значимых показателях. Авторы также приводят некоторую сводку информации по ДТП, в частности заявляя, что самыми опасными часами являются вечерние — 30-35 % ДТП случаются именно в этот промежуток времени. Подтверждающий эту статистику график можно увидеть в разделе обзора данных.

В довольно ранней работе [7] рассматривают основные практики 2000-ных годов для выявления особо опасных участков дорог. На тот момент, байесовские методы являлись основным подходом для этой задачи. Строилась некоторая модель, которая каждому участку дороги ставила в соответствие распределение количества аварий, травм, крупных происшествий и др. Полученные оценки распределений сопоставлялись с фактическими данными, после чего ано-

[1] Стоит понимать, что данные могут разнится. В Российской Федерации актуальную информацию о случаях ДТП собирают два ведомства — Госавтоинспеция и Росстат.

мально высокие значения для выведенных распределений трактовались как проблемные места, требующие особого внимания. Ряд стран использовал схожую методологию с поправкой на вид модели, критические пороги для объявление участка дороги проблемным и подобные технические характеристики. Такой подход позволял выявлять подозрительные участки дороги, на которые впоследствии отправлялись специалисты для выявления возможной причины «опасности» участка дороги и её устранения. В работе рассматривают разные спецификации модели, а также особенности практического влияния результатов. Как говорится в статье, на тот момент байесовский подход являлся state-of-the-art методом для подобного рода задач.

1.3 Методология и цели

Основной целью данной работы не является построение особо сложных моделей с использованием байесовской методологии. Автор не является специалистом в столь широкой области, но делает всё для того чтобы поближе познакомиться со столь интересным методом обработки и анализа информации. Данная работа — хороший тренажер чтобы изучить основные концепции байесовского подхода и его отличия от классических статистических методов. Помимо моделирования, ценность работы заключается в обзоре и анализе располагаемых данных: поиск аномалий, выявление закономерностей, распознание общей динамики и тому подобное. Этой части в работе будет посвящена отдельная секция, большую часть которой будут занимать графики.

Работа рассматривается, как учебное упражнение, в связи с чем её результаты не рекомендуются к использованию для принятия каких-либо практических решений.

Дополнительную ценность работе дает тот факт, что не существует популярных статей посвященных анализу дорожно-транспортных происшествий основанный на карточках ДТП с сайта Госавтоинспекции.

Для количественного анализа будут использованы агрегированные данные по дням ДТП для дальнейшего выявления влияния различных погодных или сезонных условий на количество происходящих происшествий в отдельно выбранный день. В качестве модели будет использована классическая байесовская регрессия и представлены вероятностные интервалы для полученных апостериорных распределений коэффициентов на основе которых будут сделаны качественные выводы.

Совершенно очевидно, что не стоит ожидать каких-либо принципиально новых открытий сделанных байесовским подходом в сравнении с классическими частотными методами. Более того, достаточно известный факт, что задача байесовской регрессии, при определенных предпосылках на распределения неизвестных параметров, является одной из интерпретаций стандартных таких техник регуляризаций, как Lasso и Ridge [8]. Поэтому, еще раз подчеркивается, что нет особых явных указателей на то, что в рассматриваемой задачае байесовский подход является явно лучшим вариантом в сравнении с классическим частотным.

2 Данные

2.1 Обзор

Как уже говорилось ранее, не существует актуальных источников по дорожно-транспортным происшествиям в которых данные хранились бы в структурированном виде с возможностью дальнейшей обработки. Открытые для общего пользования ресурсы предоставляют доступ к слабоструктурированным данным за какой-то короткий промежуток на котором тяжело построить дальнейший анализ. Имеющиеся данные были получены благодаря Елене Никитиной, написавшей и выложившей в открытый доступ скрипт позволяющий получить данные по карточкам ДТП за любой доступный период в разрезе по регионам в машиночитаемом формате [9]. Непосредственным поставщиком данных является сайт stat.gibdd.ru — официальный сайт Госавтоинспекции.

При помощи скрипты были собранные данные по дорожно-транспортным происшествиям в г. Москва за промежуток с 2015 по 2020 года. Одно наблюдение — карточка ДТП в которой указана информация заполненная инспектором ГИБДД приехавшим на место происшествия. Основные поля карточки ДТП включают в себя: идентификационный номер, дату, время, место происшествия, количество погибших и пострадавших в ДТП, информация по транспортным средствам, погодные условия, состояние покрытия дороги, данные об освещенности места ДТП, приблизительные координаты и некоторая обезличенная информация о погибших или раненных в инциденте. Стоит понимать, что каждый кусок данных уязвим к человеческим ошибкам и подвержен оценочному суждению, однако несет в себе какую-то информацию доступную для анализа. Почти все описанные поля принимают либо целочисленные значения, либо категориальные.

Также в открытом доступе удалось найти статистику по температуре в Москва[2] в разрезе каждого дня.

2.2 Графический анализ данных

В данной секции будет сделан иллюстративный обзор данных в попытке запечатлеть некие закономерности и паттерны которые удалось выявить и изобразить на данных.

На рис. 2 изображено количество всех случаев ДТП в г. Москва в разрезе по годам. Как видно из графика, начиная с 2016 года, строгого тренда не наблюдается. Явно выделяется результат введения ограничений в связи с COVID-19 в 2020 году, который привел к уменьшению дорожного поток, а вследствие чего к снижению вероятности возникновения ДТП. К сожалению, как уже говорилась ранее, данное наблюдение не согласуется с [4], однако может быть объяснено разными источниками данных.

Также интересно посмотреть на динамику по пострадавшим в ДТП по годам. Можно увидеть,

[2] В аэропорту Шереметьево. Предполагается, что температура различалась не сильно.

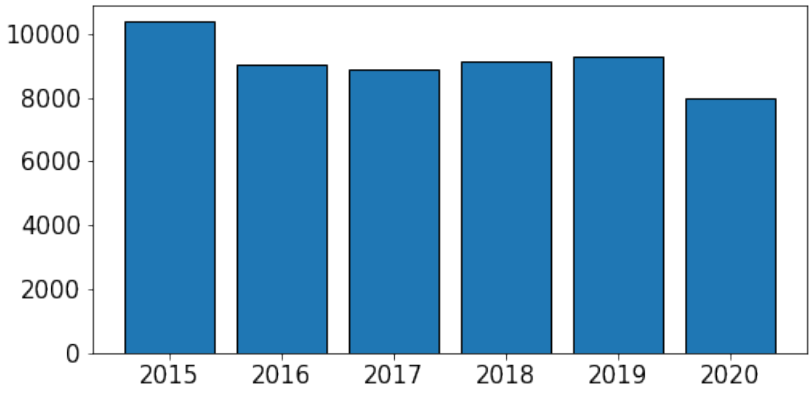

Рис. 2: Общее количество случаев ДТП (г. Москва)

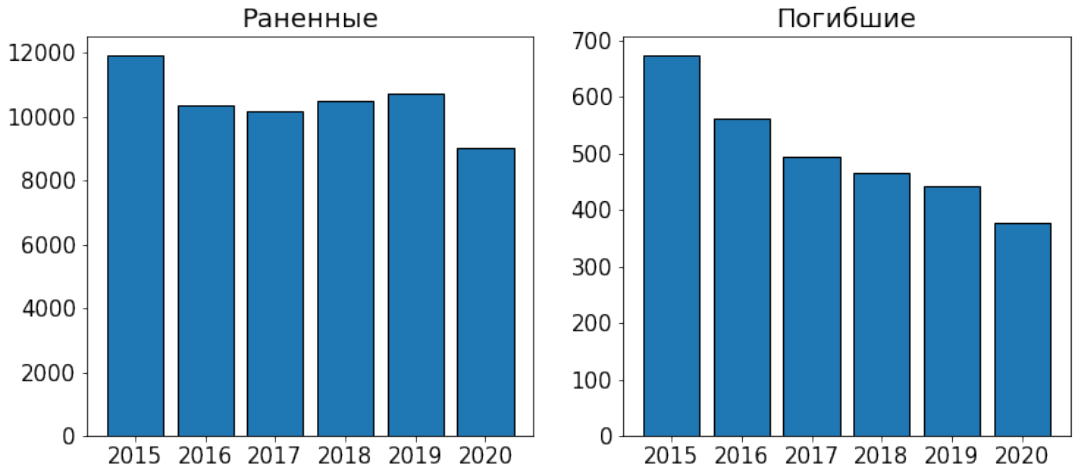

Рис. 3: Количество раненных/погибших (г. Москва)

что количество раненных примерно совпадает с динамикой общего количества случаев ДТП (рис. 2). Однако количество погибших в ДТП падает из года в год. Невооруженным глазом видно, что это падение достигается не за счёт снижения динамики случаев ДТП. Однако, чтобы в этом точно убедиться можно посмотреть на рис. 4. Видим, что отношение количества раненных в ДТП к общему количеству зарегистрированных случаев примерно одинаковое от года к году. Однако, наблюдается снижение отношения погибших к количеству случаев на протяжении всего рассматриваемого периода. Однако спад сходит на нет в 2020 году, что может быть объяснено неосторожностью вождения после отмены локдауна [4].

Для общего контекста полезно понимать распределение погибших/раненных в ДТП за рассматриваемый период. Как и ожидалось, в случившимся ДТП в большинстве своем существует минимум один пострадавший. Любой сдвиг из этой точки большая редкость. С распределением погибших ситуация схожая с разницей в том, что мода количества погибших в ДТП находится в

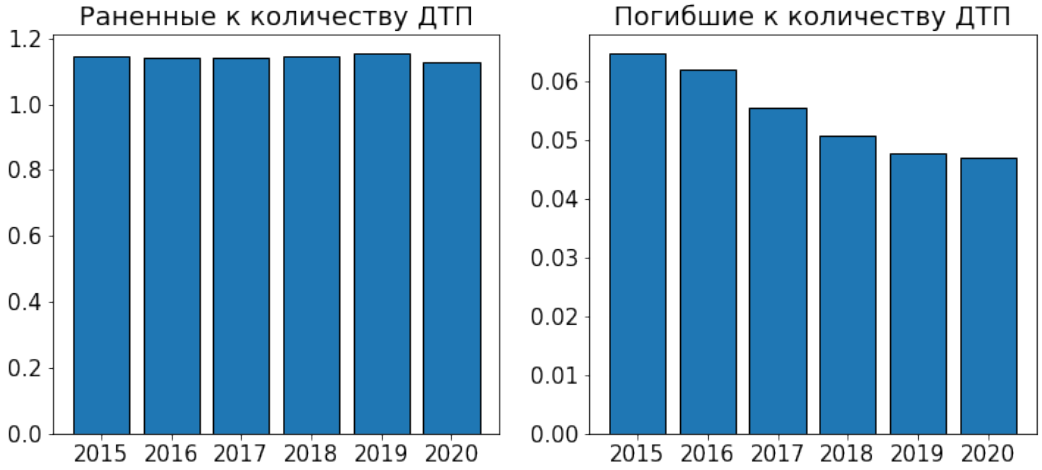

Рис. 4: Отношение раненных/погибших к общему количеству ДТП (г. Москва)

нуле. Тем не менее, данная картина не должна успокаивать ибо это достаточно высокие цифры в перерасчете на количество населения. Есть некоторая статистика, которая показывает что в некотором наборе стран Россия занимает второе место по количеству погибших в ДТП на 100 тысяч населения [10].

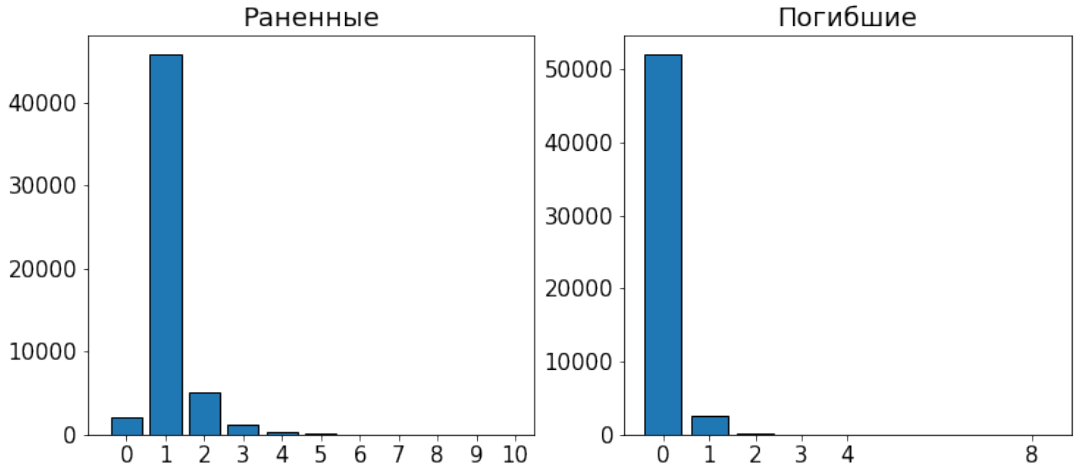

Рис. 5: Гистограммы раненных/погибших (г. Москва)

Для лучшего понимания дорожной ситуации можно также обратиться к похожим данным в разрезе по месяцам. На рис. 6 наблюдается явная закономерность показывающая, что количество ДТП, как и их тяжесть усугубляется к концу года. Причем количество раненных в ДТП резко увеличивается начиная с августа, что может быть объяснено через ухудшающаяся погоду к концу года, что приводит к ухудшению управляемости транспортного средства. Резкое падение в января можно связать с низкой дорожной активностью в этот месяц. Также начало года выпа-

дает на пик снежного сезона. Погодные условия способствуют снижению скоростного режима, что снижает риск и степень тяжести потенциальных ДТП.

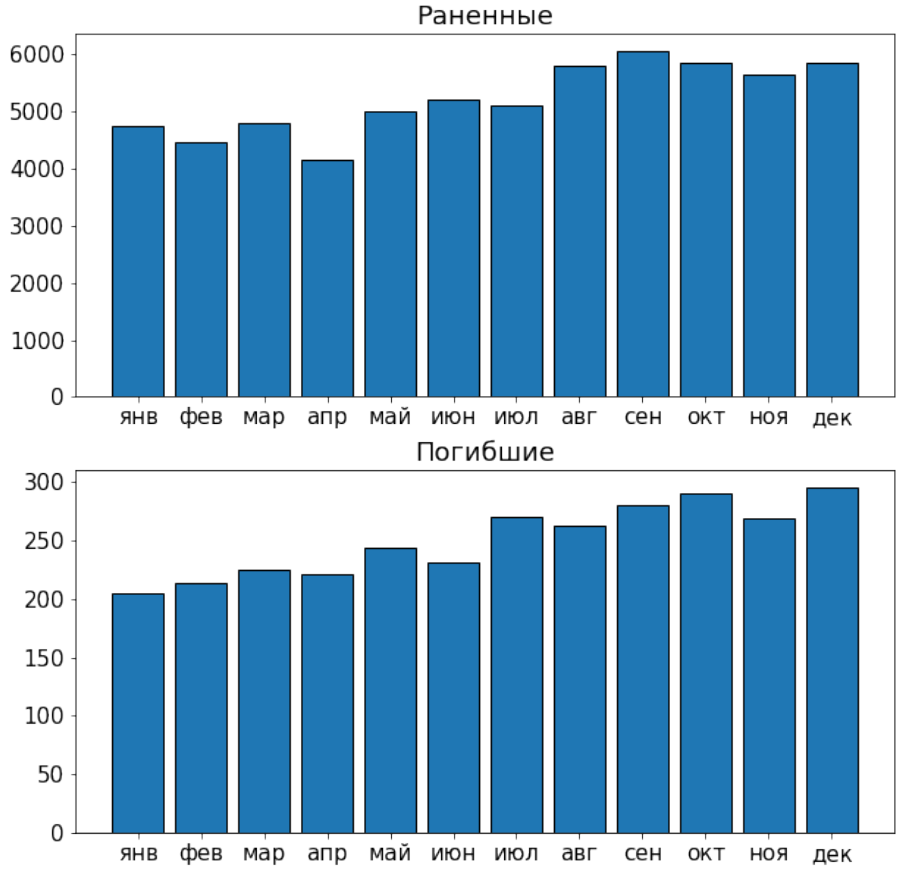

Рис. 6: Динамика раненных/погибших по месяцам (г. Москва)

Для подробного рассмотрения вопроса о связи природных явлений и случаев ДТП полезно исследовать рис. 7. На данном графике отображена динамика количества погибших и раненных при ДТП в контексте средней температуры воздуха в рассматриваемый месяц. К сожалению обнаруживаем для себя, что явных закономерностей в какую-либо сторону обнаружить не удалось. Если всмотреться, то можно заметить, что данный график лишь повторяет динамику «конца-начала» года. Это дает основание считать, что погодные условия не являются важным драйвером к снижению уровня ДТП. Главным источником возможно является повышение активности на участках дорог в связи с наступающими новогодними праздниками в конце года и дальнейшее снижение активности водителей в связи с праздничными каникулами.

От месячного периода можно спустить на одну ступень ниже к разбиению по дням недели (см рис. 8). Можно сразу заметить, что выходные дни являются самыми опасными с точки зрения

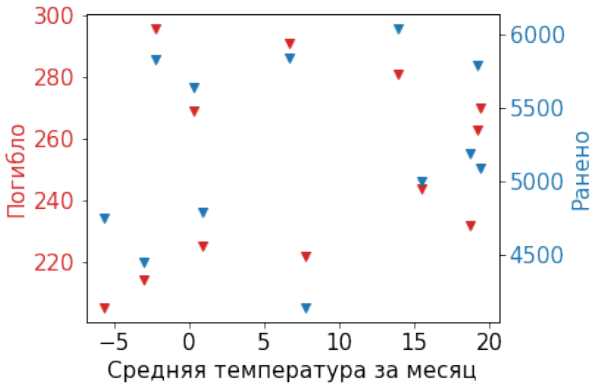

Рис. 7: Динамика раненных/погибших в связке с средней температурой за месяц (г. Москва)

количества потенциальных в таких ДТП. Важно заметить, что пятница является самым популярным днем возникновения ДТП, однако остается «безопасным» в рамках среднего количества погибших. Возникает гипотеза о том, что большая доля жертв ДТП в субботу приходится на ночь с пятницы. Люди отдохнув в последний рабочий день, возможно в нетрезвом состоянии попадают в летальные ДТП. Это также описывает аномально высокую летальность субботы, как и низкую в пятницу. Воскресенье также является неблагоприятным днем, по всей видимости по тем же причинам, что и суббота. С высокой долей вероятности, такую картинку можно наблюдать не во всех регионах, а только в городах с развитым сервисом досуга, которым жители города и желают воспользоваться в свободный от работы день.

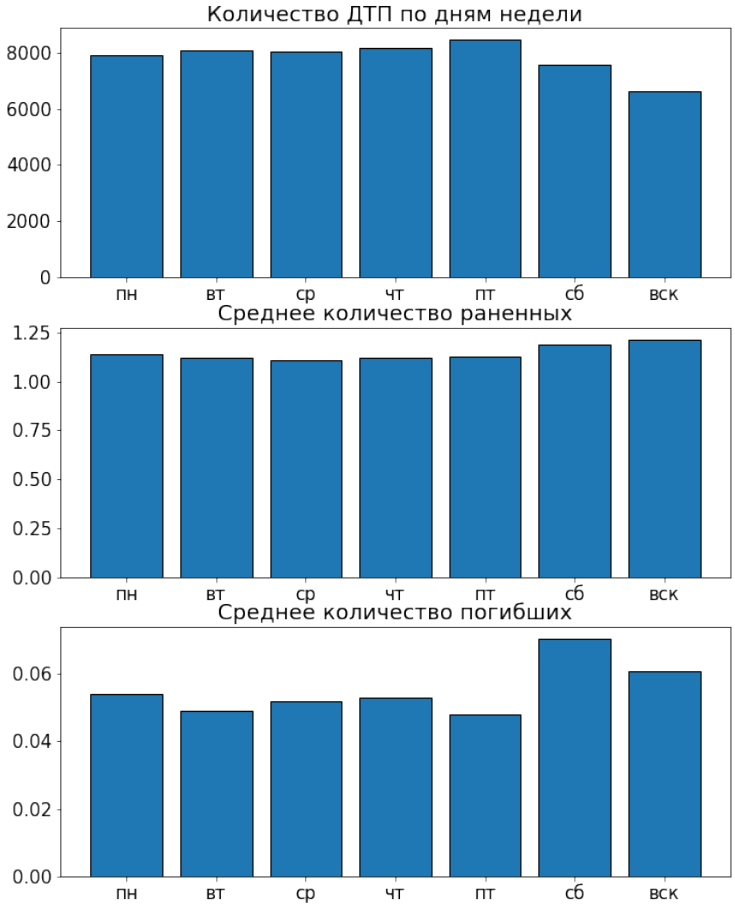

Рис. 8: Среднее количество раненных и погибших по дням недели (г. Москва)

Список литературы

[1] Костычаков ВФ. Дорожно-транспортные происшествия в Российской Федерации: статистика, основные причины возникновения ДТП // Аллея науки. 2018. Т. 1, № 8. С. 538–541.

[2] Ивлиев Михаил Игоревич, Черемисина Наталия Валентиновна. Экономико-статистичекий анализ дорожной ситуации в российской Федерации // Социально-экономические явления и процессы. 2014. Т. 9, № 7.

[3] Клачкова АВ, Семёнова ЕД. АНАЛИЗ СТАТИСТИКИ ДТП В РОССИЙСКОЙ ФЕДЕРАЦИИ // Инновационная наука. 2020. № 12.

[4] Михаил Котляр Иван Ткачёв. Локдаун 2020 года почти не повлиял на смертность в авариях. URL: https://www.rbc.ru/society/13/06/2021/60c4106a9a7947862635eed4.

[5] Кузьменко ЕА, Донченко ДС, Рагозин ВО. Анализ данных для прогнозирования вероятности дорожно-транспортных происшествий с участием пешеходов // Инженерный вестник Дона. 2020. № 6 (66).

[6] Анализ социологических аспектов дорожно-транспортных происшествий на дорогах России методами математической статистики / ВГ Конюхов, АВ Олейник, ГП Конюхова [и др.] // Редакционная коллегия: МЮ Ростовцева, кандидат педагогических наук, профессор; Новикова ЛА, заведующая кафедрой теории и методики гимнастики. 2015. с. 170.

[7] Elvik Rune. State-of-the-art approaches to road accident black spot management and safety analysis of road networks. Transportøkonomisk institutt Oslo, 2007.

[8] Haitovsky Yoel, Wax Yohanan. Generalized ridge regression, least squares with stochastic prior information, and Bayesian estimators // Applied Mathematics and Computation. 1980. Т. 7, № 2. С. 125–154.

[9] Никитина Елена. Парсер статистики ДТП с официального сайта ГИБДД stat.gibdd.ru. URL: https://github.com/Shorstko/GibddStat.

[10] ТАСС. Статистика ДТП в России и мире. Досье. URL: https://tass.ru/info/3233185.