Факультет экономических наук

Экономика

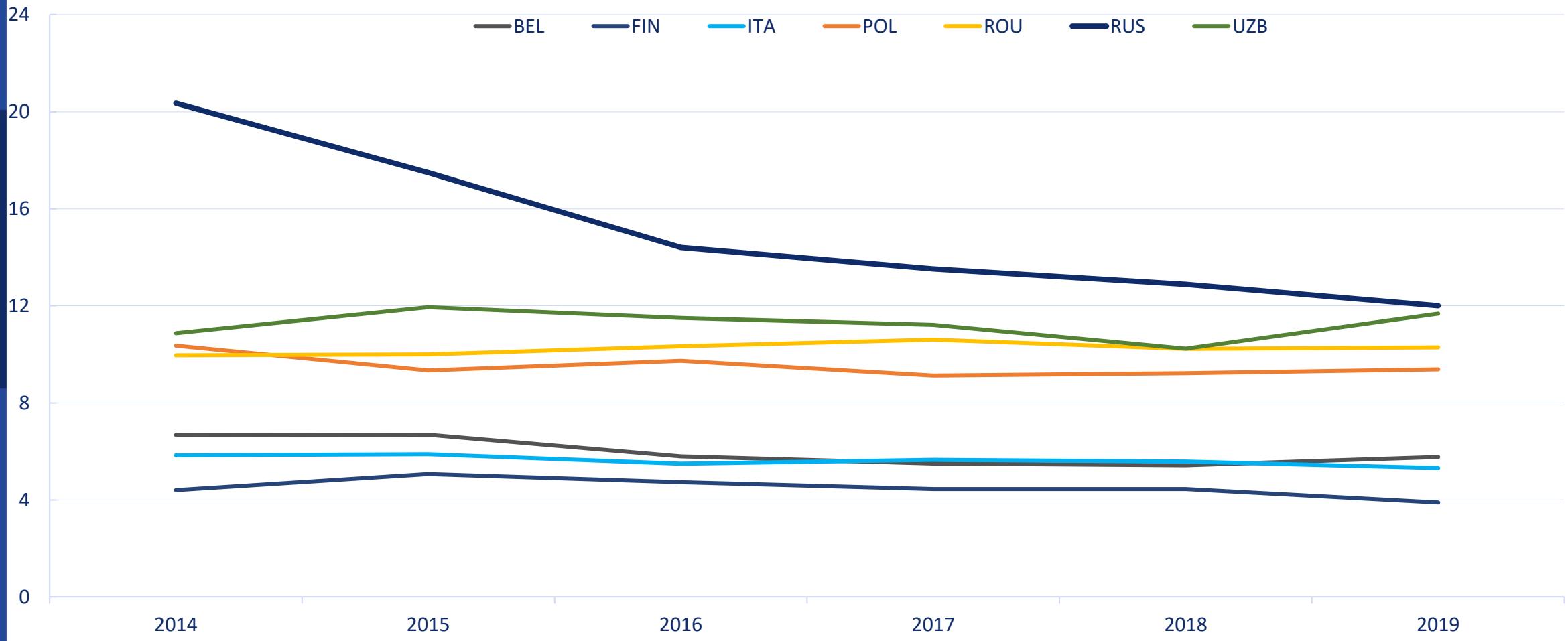
Москва, 2022

Estimation of Influence of Road Construction and Maintenance on Accident Severity and Frequency

Оценка влияния ремонтных дорожных работ на частоту и тяжесть ДТП

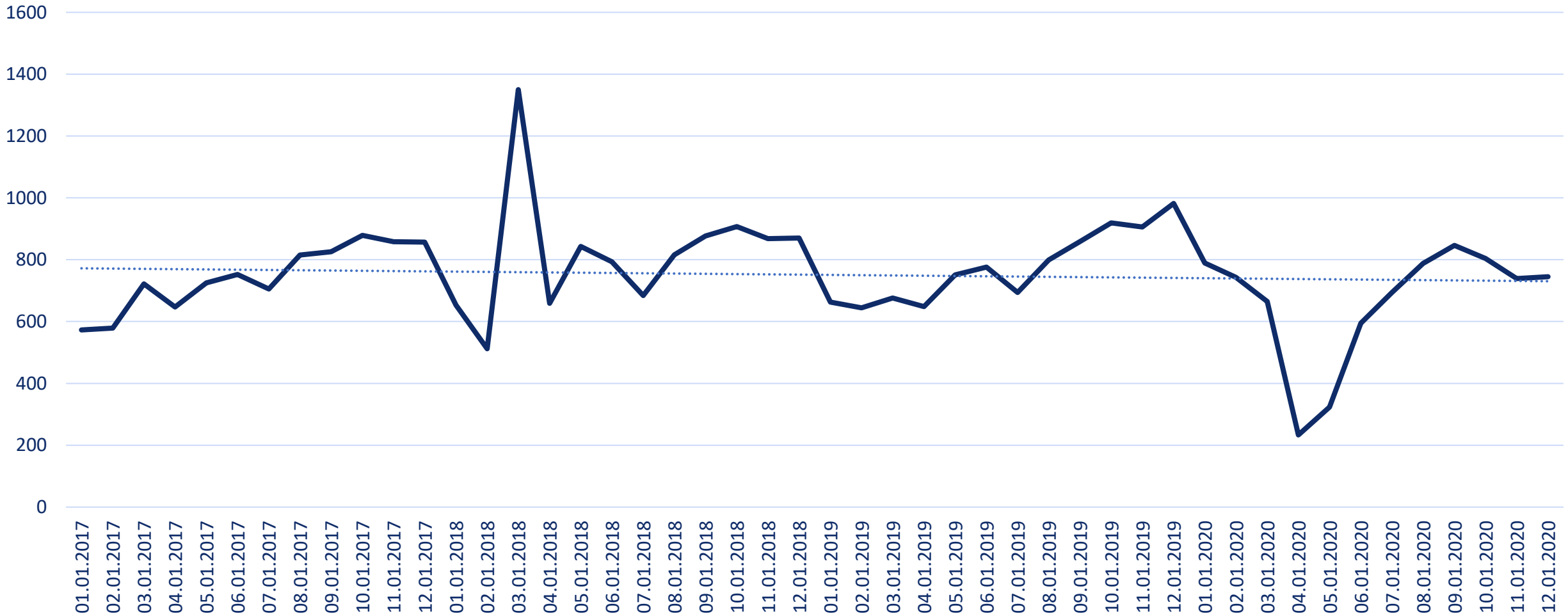
Босколо Фьоре Стефано, БЭК189
Научный руководитель: Станкевич Иван Павлович

Смертей из-за ДТП на 100 000 населения

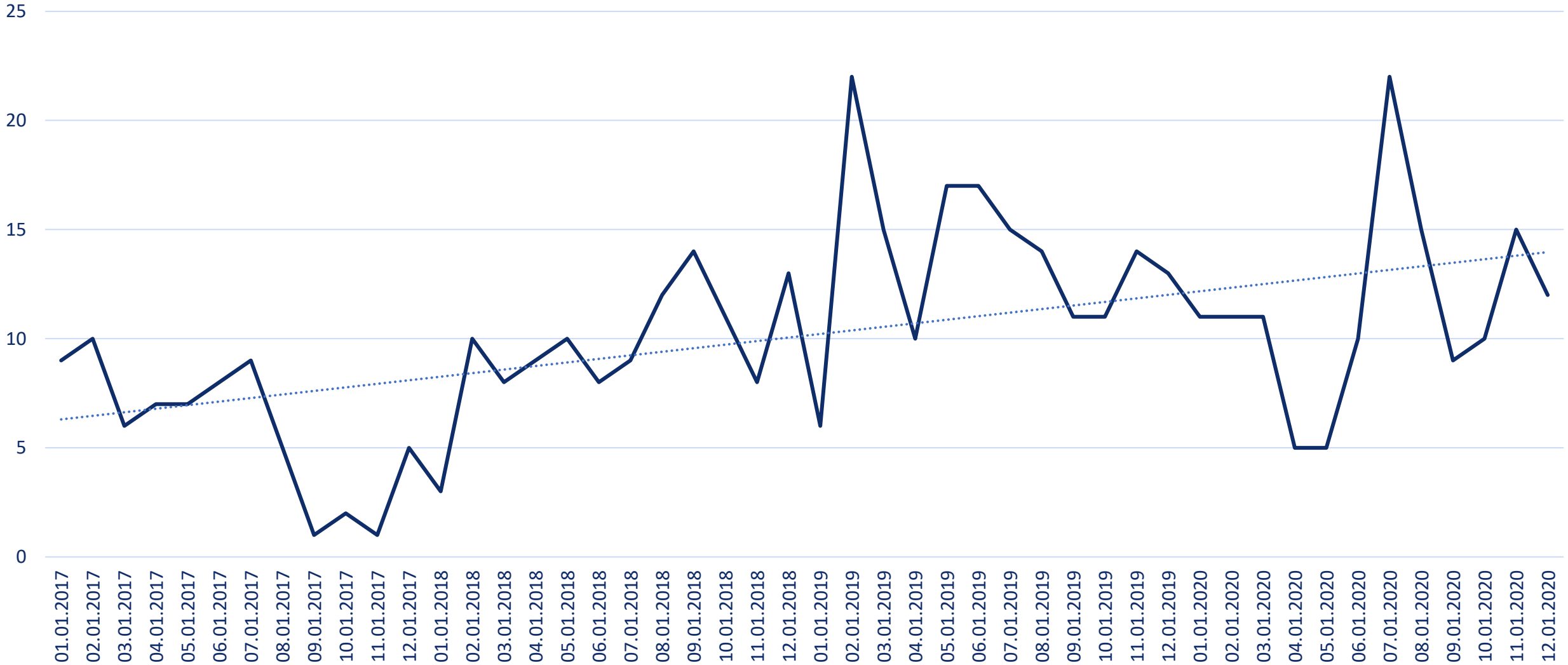



Смертей из-за ДТП на 100 000 населения в России, Бразилии и Швеции
35
30
25
20
15
10
5
0
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
RUS
BRA
SWE
Линейная (RUS)

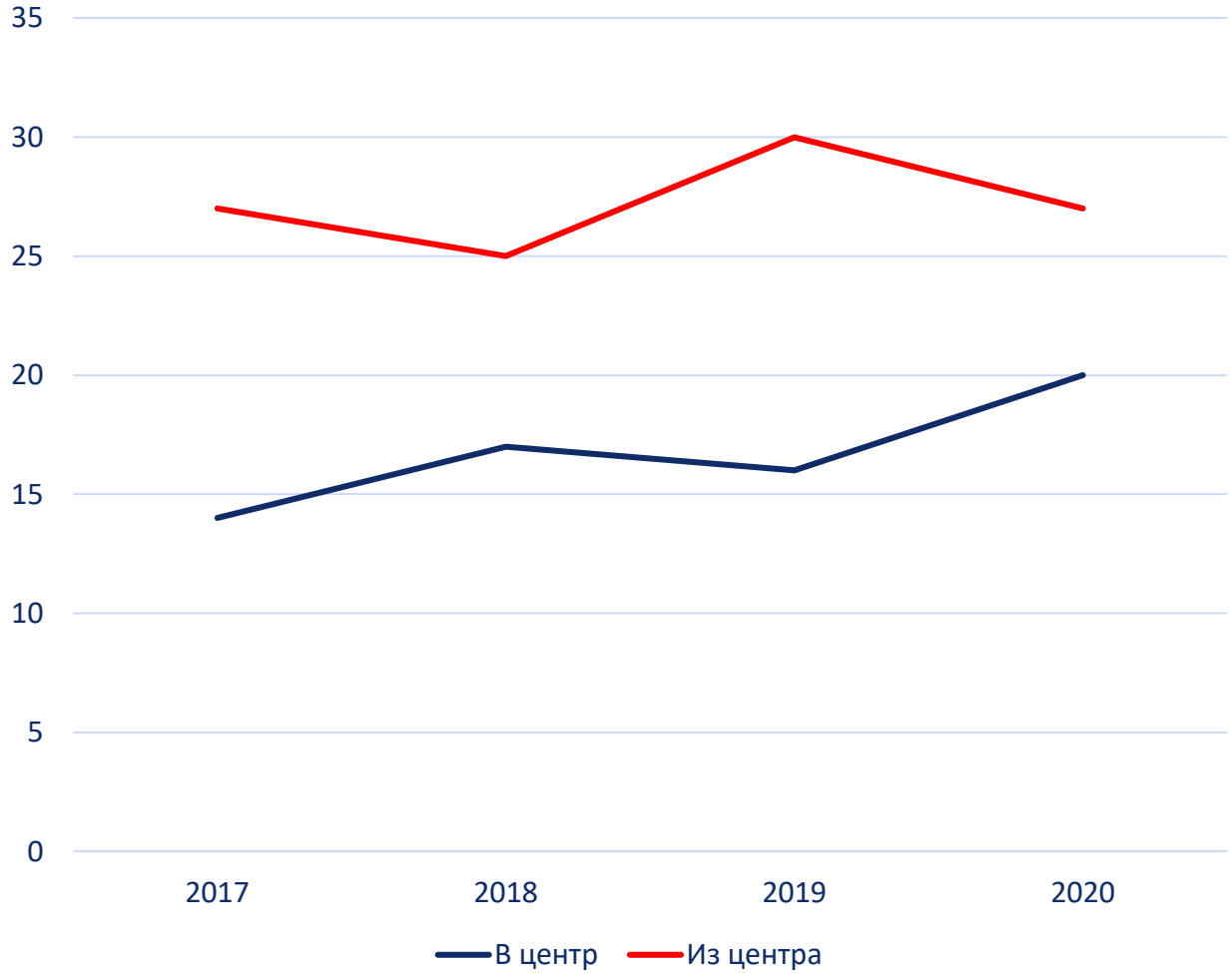
Количество ДТП в Москве



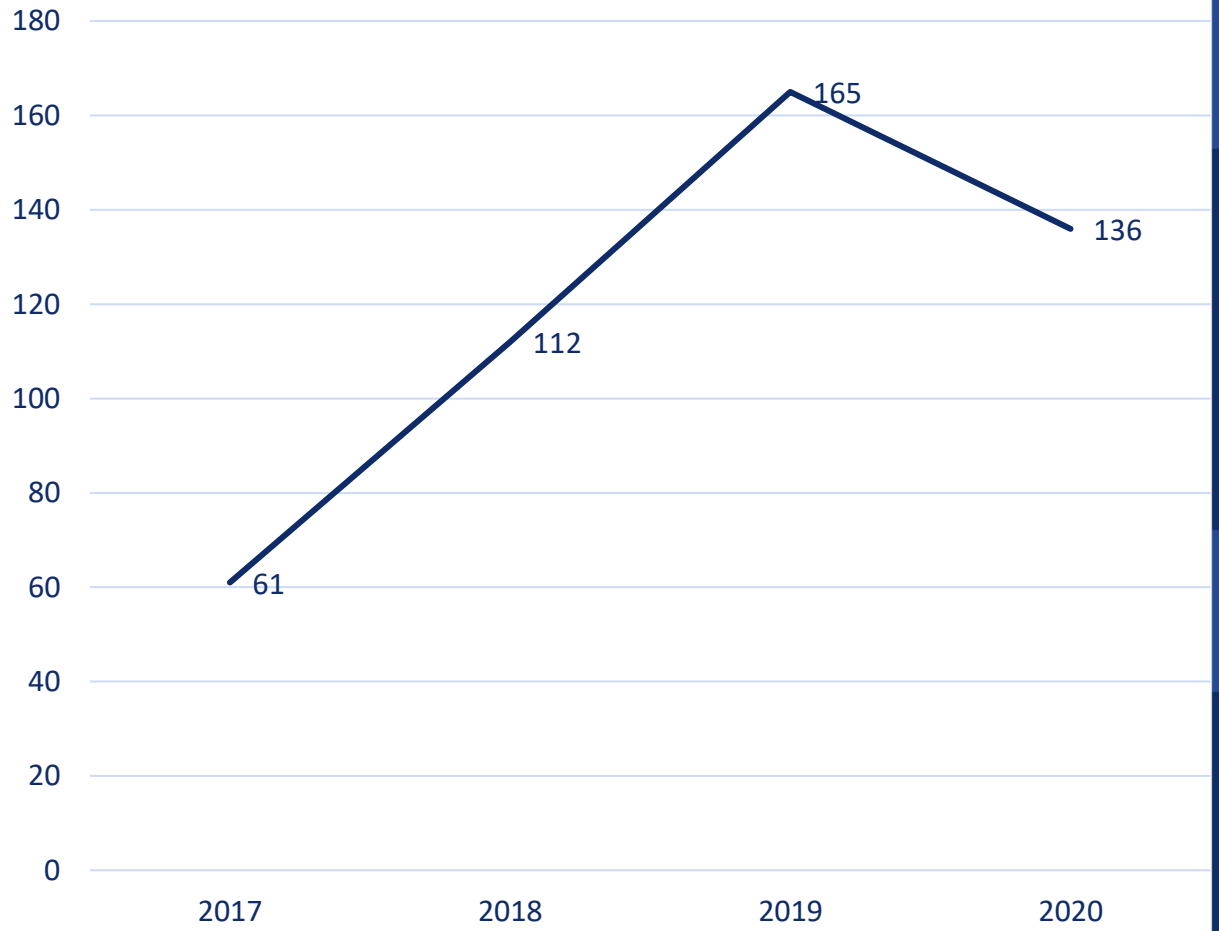
Количество ДТП на Ленинградском шоссе



Средняя скорость в час пик



Количество ДТП на Ленинградском шоссе



Процесс создании базы данных 2017 – 2020

- Вручную загружены, соединены и отформатированы 150 XML файлов с карточками ГИБДД.
- С помощью координат отфильтровано более 40000 наблюдений, в результате осталось 500.
- К каждому наблюдению добавлены данные погоды из GSOD.
- Добавлены данные более чем 20 ремонтных работ в этом периоде.

RStudio

Разделение на участки

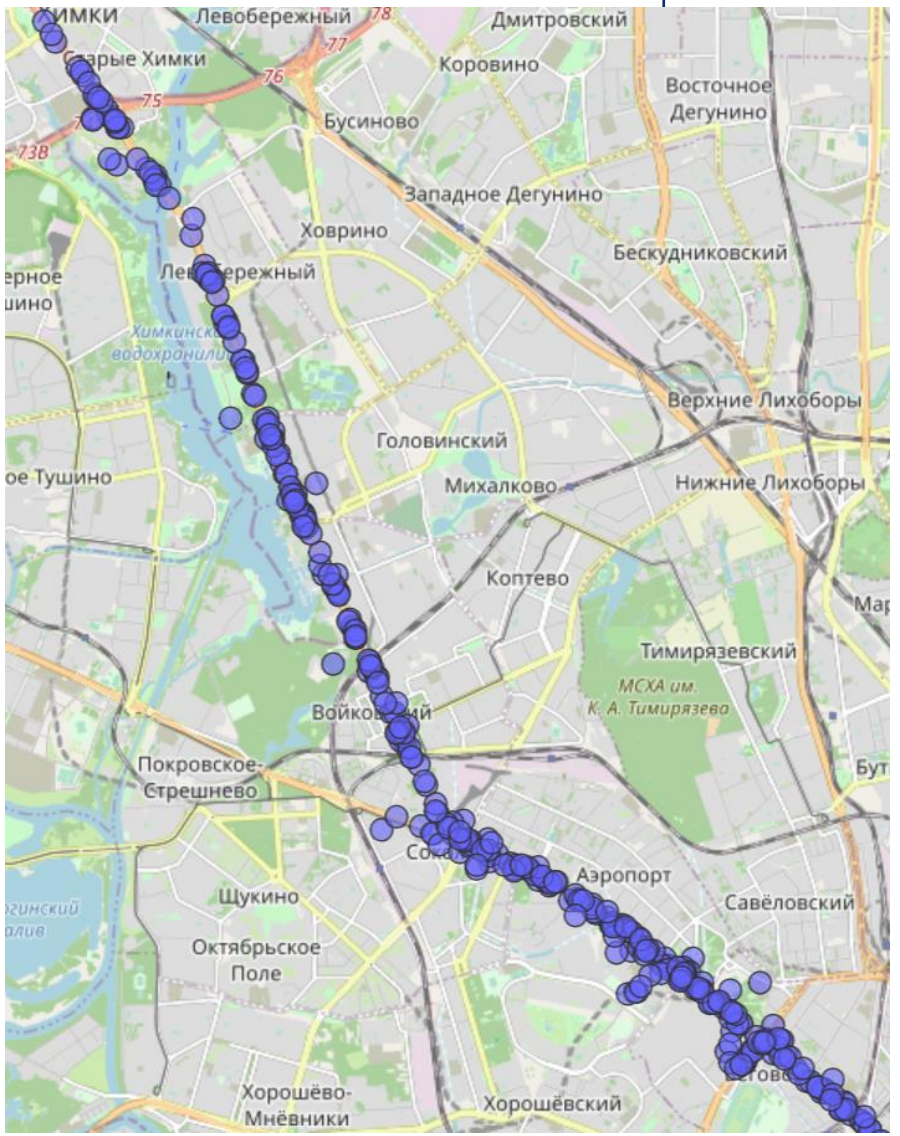
Начало	Конец
Тверская улица	Метро Белорусская
Мост у метро Белорусская	пересечение с ТТК у метро Динамо
пересечение с ТТК у метро Динамо	Институт Гидропроект у метро Сокол
Институт Гидропроект у метро Сокол	Метро Водный стадион
Метро Водный стадион	Пересечение с МКАД
Пересечение с МКАД	Уровень аэропорта Шереметьево

Таблица 2, выделенные участки Ленинградского шоссе

Факторы, влияющие на ДТП на маленьких улицах и на больших дорогах отличаются, поэтому было решено исследовать только часть Ленинградского шоссе.

Наивно предполагать, что факторы ДТП на всей протяжённости шоссе одинаковые.

Разделение на участки по пересечению с другой крупной транспортной артерией и развязкой увеличивает однородность данных.



Картинка 1, точки ДТП на Ленинградском шоссе 2017-2020

Методология: модели степени тяжести ДТП

Логистическая регрессия

- Зависимая переменная: «0» – нет пострадавших, «1» – хотя бы один участник получил травмы.
- Распределение по классам: $N_0 = 260$, $N_1 = 213$

Мультиномиальная логистическая регрессия

- Показывает переход от базового класса
- Зависимая переменная: «0» – нет пострадавших, «1» – хотя бы один участник находился на амбулаторном лечении или в условиях дневного стационара, «2» – хотя бы один участник находился на стационарном лечении, был тяжело ранен или погиб в результате ДТП.
- Распределение по классам: $N_0 = 260$, $N_1 = 167$, $N_2 = 46$

Алгоритм случайного леса

- Случай бинарной и многоклассовой классификации

Алгоритм градиентного бустинга

- Случай бинарной и многоклассовой классификации

Методология: модели частотности ДТП

Панельная Пуассон регрессия

- Классический метод анализа счетных данных и частотности ДТП.
- Накладывает ограничение: $\mathrm{Var} = mean$

Панельная отрицательная биномиальная регрессия

- Более современный метод анализа частотности ДТП
- **Не** накладывает ограничение: $\mathrm{Var} = mean$
- Плохо работает с overdispersed данными

Панельная Hurdle модель

- Аналог zero-inflated моделей
- Отличается структура 0 в наблюдениях

Алгоритм случайного леса на панельных данных

Алгоритм градиентного бустинга на панельных данных

Логистическая регрессия

Зависимая переменная бинарная степень тяжести: «0» – нет пострадавших в ДТП, «1» – есть хотя бы 1 пострадавший.

Были включены группы переменных: погодные условия, характеристика покрытия, условия местности, нарушения участников ДТП, объекты рядом с ДТП, тип ДТП, условия освещения.

Результаты оценки бинарной логистической модели:

- Ремонт не влияет на тяжесть ДТП.
- Наезд на ТС увеличивает, в то время как наезд на пешехода уменьшает степень тяжести ДТП.
- Объекты рядом, влияющие на разную скорость транспортного потока увеличивают тяжесть ДТП.

Переменная	Оценка
Близость к ремонту	0.044
Тип ДТП: наезд на ТС	1.355*
Тип ДТП:Наезд на пешехода	-3.827***
Условия местности: Остановка	1.578*
Объект рядом: Остановка	1.502**
Объект рядом: АЗС	1.643*
Нарушение: перестроение	-2.283***
ROC-AUC	0.95
Уровни значимости: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’	

Таблица 1, значимые оценки параметров логистической модели

Мультиномиальная логистическая регрессия

Зависимая переменная степень тяжести: «0» – нет пострадавших, «1» – хотя бы 1 участник находился на амбулаторном лечении или в условиях дневного стационара, «2» – хотя бы один участник находился на стационарном лечении, был тяжело ранен или погиб в результате ДТП. Были включены группы переменных: погодные условия, характеристика покрытия, условия местности, нарушения участников ДТП, объекты рядом с ДТП, тип ДТП, условия освещения.

Результаты оценки мультиномиальной логистической модели:

- Ремонт не влияет на тяжесть ДТП.
- Уменьшают тяжесть только факторы наезд на пешехода, подход к мосту и нарушение правил перестроения
- Факторы, изменяющие однородность и предсказуемость потока делают случающиеся ДТП более тяжкими
- Практически все параметры влияют как на «1» так и на «2» класс.
- Одновременная значимость переменных в переходах между классами говорит о монотонном влиянии на обе категории.

	Переход от 0 к 1	Переход от 0 к 2
Близость к ремонту	0.033	-0.056
Наезд на ТС	0.484	2.191**
Наезд на препятствие	119.398***	120.241***
Наезд на пешехода	-3.574***	-52.874***
Падение пассажира	35.716***	36.604***
Опрокидывание	83.375***	86.891***
Дополнительный фактор: мешающее ТС	6.392***	5.709***
Условия местности: подход к мосту	-21.284***	-21.056***
Объект рядом: Остановка	1.320**	1.052
Объект рядом: отделенная парковка	64.293***	66.033***
Мужчина и женщина	89.283***	90.024***
Нарушение: несоблюдение бокового интервала	25.278***	27.197***
Нарушение: перестроение	-1.616**	-32.081***

Уровни значимости: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

Таблица 2, значимые оценки параметров мультиномиальной логистической модели

Алгоритмы машинного обучения

5 важнейших переменных: бинарный случай	
Случайный лес	**Градиентный бустинг**
Наезд на пешехода	Наезд на пешехода
Мужчина и женщина	Мужчина и женщина
Наезд на препятствие	Температура
Нарушение: перестроение	Наезд на препятствие
Опрокидывание	Температура^2
ROC-AUC = 0.8	ROC-AUC = 0.85

Таблица 3, важность переменных алгоритмов машинного обучения : бинарный случай

5 важнейших переменных: многоклассовый случай	
Случайный лес	**Градиентный бустинг**
Мужчина и женщина	Наезд на пешехода
Наезд на пешехода	Мужчина и женщина
Наезд на препятствие	Наезд на препятствие
Падение пассажира	Температура
Нарушение: превышение скорости	Температура^2

Таблица 4, важность переменных алгоритмов машинного обучения : многоклассовый случай

- Похожий набор важных переменных.
- Ремонт не влияет на предсказания в алгоритме.
- Фактор участия и мужчины и женщины в ДТП на втором месте.
- У бинарных моделей ROC-AUC ниже, чем у простой логистической регрессии.

Сравнение точности алгоритмов: бинарный случай

Матрица ошибок логистической регрессии			
N = 136		Actual	
		0	1
Predicted	0	74	6
	1	9	47
accuracy = 0.89			

Таблица 5, матрица ошибок линейной регрессии

Матрица ошибок алгоритма случайного леса			
N = 136		Actual	
		0	1
Predicted	0	65	12
	1	15	44
accuracy = 0.80			

Таблица 6, матрица ошибок случайного леса

Матрица ошибок алгоритма градиентного бустинга			
N = 136		Actual	
		0	1
Predicted	0	72	12
	1	8	44
accuracy = 0.85			

Таблица 7, матрица ошибок градиентного бустинга

- Тренировочная подвыборка – 75%, 2017-2019 год, 337 наблюдений; тестовая подвыборка – 25%, 2020 год, 136 наблюдений.
- Логистическая регрессия имеет наибольшую точность предсказаний.
- Все алгоритмы точны.
- Эти методы подходят для данной выборки.

Сравнение точности алгоритмов: многоклассовый случай

Матрица ошибок мультиномиальной логистической регрессии

N = 136		Actual		
		0	1	2
predicted	0	73	7	0
	1	10	35	3
	2	1	3	4
accuracy = 0.82				

Таблица 8, матрица ошибок мультиномиальной линейной регрессии

Матрица ошибок алгоритма случайного леса

N = 136		Actual		
		0	1	2
predicted	0	68	20	2
	1	12	28	6
	2	0	0	0
accuracy = 0.71				

Таблица 9, матрица ошибок многоклассового случайного леса

Матрица ошибок алгоритма градиентного бустинга

N = 136		Actual		
		0	1	2
predicted	0	74	18	3
	1	6	26	4
	2	0	4	1
accuracy = 0.74				

Таблица 10, матрица ошибок многоклассового градиентного бустинга

- Тренировочная подвыборка – 75%, 2017-2019 год, 337 наблюдений; тестовая подвыборка – 25%, 2020 год, 136 наблюдений.
- Мультиномиальная логистическая регрессия имеет точность предсказаний значительно более высокую.
- У алгоритмов машинного обучения возникают трудности при предсказании тяжелых ДТП (2 класс).

Панельная Пуассон и отрицательная биномиальная регрессии

Все переменные указаны в месячной частоте.

Зависимая переменная количество ДТП в месяц на определенном участке.

- Наличие активного ремонта является значимым и увеличивает частоту ДТП.
- Остальные переменные связаны с неблагоприятными погодными условиями, усложняющие езду

Переменная	Оценка	
	Пуассон регрессия	Отрицательная биномиальная регрессия
Константа	-	9.636**
Наличие активного ремонта	0.653***	0.083*
Среднемесячная температура	0.084*	0.538***
Кол-во дождливых дней	-	-0.349*
Кол-во дней с моросящим дождем	0.057**	-
Кол-во дней с покрытием лед	0.403*	-
Уровни значимости: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’		

Таблица 11, значимые параметры панельной Пуассон и отрицательной биномиальной регрессий

Hurdle-модели

Все переменные указаны в месячной частоте.

Зависимая переменная количество ДТП в месяц.

- Наличие активного ремонта не является значимым фактором частоты ДТП
- Результаты оценки Hurdle-моделей подтверждают гипотезу о том, что нашем случае наблюдения «0» ДТП не могут быть структурными, а сам процесс не может быть разделим единственным образом на нулевую и счетную части.

Переменная	Оценка	
	Счетная часть Пуассон hurdle-модели	Счетная часть Отрицательной биномиальной hurdle-модели
Температура	0.091*	0.093
Ремонт	0.230	0.242
	Нулевая часть Пуассон hurdle-модели	Нулевая часть Отрицательной биномиальной hurdle-модели
Ремонт	4.372	334.944
Моросящий дождь	0.140**	5.474
Снег	-1.049*	-1.058
Покрытие: лед	1.270**	24.852
Дождь	-0.826*	13.099
Уровни значимости: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’		

Таблица 12, hurdle-модели

Алгоритмы машинного обучения: частотный случай

5 важнейших переменных	
Случайный лес	**Градиентный бустинг**
Скорость ветра	Покрытие: мокрое
Покрытие: мокрое	Снег
Ремонт	Ремонт
Температура^2	Туман
Туман	Температура^2
RMSE = 2.04	RMSE = 1.51

Таблица 16, важность переменных алгоритмов машинного обучения: частотный случай

- Практически идентичный набор важных переменных.
- Наличие активного ремонта оказывает влияние на частоту ДТП.
- Невысокая точность предсказаний, вероятно, из-за небольшой выборки.
- У наивной модели, предсказывающей самое частое количество ДТП – «0», RMSE = 2.52
- На небольших выборках алгоритмы склонны переобучаться

$$\mathrm{RMSE} = \frac{\sqrt{\Sigma(\hat{y}_i - y_i)^2}}{n}$$

Выводы исследования

- Выбранные подходы применимы для исследования частоты и тяжести ДТП на Ленинградском шоссе.
- Установлено значимое влияние ремонтных работ на частоту, но не тяжесть ДТП.
- На тяжесть ДТП наибольшее влияние оказывают переменные, которые не предполагают от водителя снижение скорости заранее, являющиеся неожиданностью.
- На частоту ДТП влияют погодные условия, усложняющие езду, увеличивающие плотность потока, а также ремонтные работы, изменяющие геометрию дороги.

Это исследование представляет новые для России методы моделирования аварийности, а применение панельных данных для отдельных частей шоссе стало одним из первых в мире развитием подхода по анализу отдельных перекрестков. Эта работа показывает применимость различных методов анализа и моделирования на российских данных, а также способы создания и объединения массивов.

Данная тема является актуальной для дорожной безопасности в РФ, ее изучение поможет оптимизировать бюджетные расходы и снизить частоту и смертность вследствие ДТП.

Дальнейшие исследования тяжести и частоты ДТП в России могут стать стимулом для более точного и комплексного сбора данных.

Спасибо за внимание