Софийски университет „Св. Кл. Охридски“

Факултет по математика и информатика

Катедра „Софтуерни технологии“

АВТОРЕФЕРАТ

за придобиване на образователна и научна степен „Доктор“,
професионално направление: 4.6 Информатика и компютърни науки,
докторска програма: Софтуерни Технологии – Софтуерно Инженерство

на тема

„Подпомагане взимането на решения за оптимизиране на обществен транспорт с помощта на големи данни“

Докторант: **Георги Калинов Йосифов**

Научен ръководител:
доц. д-р Милен Йорданов Петров

София, 2022 г.

I. ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

Актуалност на проблема

София е европейски град и най-голямото населено място на територията на Република България. По данни на Националния статистически институт, въпреки намаляващото население на страната (НСИ, 2021), в столицата прогнозите са, че до 2030 година то ще нарасне с още 25 000 души (НСИ, 2018). Прираст се наблюдава и при количеството на автомобилите по пътищата на града. Според данни от Столична община от 2011 г. до 2020 г. броят им е станал почти двоен – от 462 043 на 833 280 (Сантова, 2021). Такава е ситуацията и в много други големи градове. Нарастващото население, нарастващият трафик и нарастващата нужда от качествено и безпрепятствено придвижване в градската среда, в София и градове като нея, създават големи предизвикателства пред администрацията и бизнеса. Съществуват проучвания, които показват потенциалните ползи от провеждане на политики за оптимизиране на транспорта чрез намаляване на нивата на градския трафик, които във финансово изражение достигат и до милиарди левове за по-големите държави (INRIX, 2020). За решаване на този проблем, трябва да се разработят механизми и инструменти за следене на пътната ситуация. С тяхна помощ биха могли да се идентифицират критични пътни отсечки, за които да се вземат конкретни мерки.

Същевременно със засиленото внимание към чувствителността на събирането и управлението на личните данни и регулации като европейската General Data Protection Regulation (GDPR) или калифорнийската California Consumer Privacy Act (CCPA), трябва да се обърне и особено внимание на начините, по които тези механизми и инструменти оперират.

В дисертационния труд се търси решение на проблем касаещ както акуратното определяне на нивата на трафик в градска среда, така и използването на данни, които не се определят като чувствителни за гражданите.

Цел на дисертационния труд, основни задачи и методи за изследване

На база на изложените в този увод разсъждения, можем да дефинираме, че целта на дисертационния труд е да се подпомогне взимането на решения за оптимизация на обществения транспорт чрез определяне, изследване и прогнозиране на нивата на натовареност на движението с помощта на данни, събрани от позиционни координати на периодичния публичен транспорт, използван като проба в трафика.

За реализация на поставената цел се поставят за изпълнение следните задачи:

- Да се състави методология за класификация и анализ на текущото състояние на методите за събиране на данни и определяне нивата на трафика.
- Да се състави обзор на различните методи за анализ и определяне на нивата на трафик в бъдещ момент.
- Да се разработи алгоритъм на база съставената методология, с помощта на който да се определят нивата на трафик в градска среда.
- Да се съставят експериментални сценарии за изследване на качествата и ограниченията на така изготвения алгоритъм.
- Да се разработят инструменти, подпомагащи обработката и управлението на данните от алгоритъма в експерименталните сценарии.
- Да се изследват различни начини за определяне на нивата на трафик в градска среда в бъдещ момент.
- Да се направи сравнителен анализ на представянето на различните способи за прогнозиране и да се определи най-подходящият.

Практическа приложимост и ползи

Данните за натовареността на трафика в градска среда, в текущ момент или бъдещ такъв, могат да бъдат използвани, за да се извлече от тях нужната информация за оптимизация на обществения транспорт. Те могат да бъдат публикувани като отворени данни, за използване от обществото или да са

полезни при създаването на държавни или общински политики, свързани с изграждане на инфраструктура и дистрибуция на обществени ресурси. Чрез корелация с данни от сензори за CO_2 емисии, биха могли да се следят нивата на замърсяване на въздуха и предвиждат бъдещи пикове. Тази информация може да се взима предвид и от маршрутизиращи софтуери, използвани от частни лица, различни компании в спедиторския бранш или такива правещи доставки до домовете, за да могат шофьорите да избират най-оптималните маршрути. Всички тези приложения са част от практическите ползи от резултата на разработения труд.

Публикации

Постиженията на този дисертационен труд са публикувани в общо три публикации. И в трите публикации докторантът е първи автор. Всички те са публикувани в международни научни списания с издатели ACM и Springer, с “impact rank”, като са индексирани в научната база от данни SCOPUS.

Всички публикации са представени на международни научни конференции, като последните две са изнесени в гр. Лондон на Computing Conference 2022 и International Congress on Information and Communication Technology (ICICT) 2022, а първата на CompSysTech 2020, организирана в гр. Русе.

Към 09.2022 г. Публикациите са цитирани общо 4 пъти (проверено в системата Google Scholar).

Списък на публикациите по темата на дисертационния труд

1. Georgi Yosifov, Milen Petrov, Predicting Traffic Indexes on Urban Roads based on Public Transportation Vehicle Data in Experimental Environment, Lecture Notes in Networks and Systems, editor/s:Janusz Kacprzyk , Publisher:Springer, 2022, pages:159-168, ISSN (print):2367-3370, ISSN (online):2367-3389, ISBN:978-3-031-10466-4, doi:10.1007/978-3-031-10467-1_8, Ref, IR ,

SCOPUS, SJR (0.17 - 2020), SCOPUS Quartile: Q4 (2022), др.(INSPEC, WTI Frankfurt eG, zbMATH, SCImago), PhD

2. Georgi Yosifov, Milen Petrov, Review of urban traffic detection approaches with accent of transportation in Sofia, Bulgaria, Proceedings of Seventh International Congress on Information and Communication Technology. Lecture Notes in Networks and Systems, vol 465., editor/s:Yang, XS., Sherratt, S., Dey, N., Joshi, A., Publisher:Springer, 2022, pages:509-517, ISSN (print):978-981-19-2396-8, ISSN (online):978-981-19-2397-5, doi:https://doi.org/10.1007/978-981-19-2397-5_47, Ref, IR , SCOPUS, SJR (0.17 - 2020), SCOPUS Quartile: Q4 (2022), др.(INSPEC, WTI Frankfurt eG, zbMATH, SCImago), PhD

3. Georgi Yosifov, Milen Petrov, Traffic flow city index based on public transportation vehicles data, International Conference on Computer Systems and Technologies - CompSysTech'20 (CompSysTech'2020), editor/s:Vassilev T.,Trifonov R., Publisher:Association for Computing Machinery, 2020, pages:201-207, ISBN:978-145037768-3, doi:10.1145/3407982.3408007, Ref, IR , SCOPUS, SJR (0.182 - 2020), ACM Digital Library, PhD

Доклади на конференции

1. Секционен доклад, Георги Йосифов, Predicting Traffic Indexes on Urban Roads based on Public Transportation Vehicle Data in Experimental Environment

2. Секционен доклад, Георги Йосифов, Review of urban traffic detection approaches with accent of transportation in Sofia, Bulgaria

3. Секционен доклад, Георги Йосифов, Изследване на корелацията между броя на превозните средства и индекса на натовареност на градския трафик, базиран на позиционни данни от периодичен обществен транспорт

4. Секционен доклад, Георги Йосифов, Traffic flow city index based on public transportation vehicles data, International Conference on Computer Systems and Technologies

II. СТРУКТУРА И ОБЕМ НА ДИСЕРТАЦИОННИЯ ТРУД

Дисертационният труд се състои от шест глави. Текстът е написан в 180 страници и съдържа 76 фигури и 17 таблици. Цитирани са 108 литературни източника и интернет страници. Трудът е допълнен с пет приложения. Всяка глава е разделена на тематични секции, спомагащи описанието на решаваните от труда проблеми. В края на дисертационния труд са добавени направените научни публикации по темата, изброени са приносите и е приложена декларация за оригиналност.

В следващите секции представяме конкретното съдържание на всяка една от главите. Шестата и последна глава е заключителна, в която се обобщаваща работата и се предлагат насоки за бъдещо развитие. Номерата на фигурите и таблиците в авторефератa съвпадат с тези от дисертационния труд.

1. Увод

Главата се състои от три тематични части. Първата част разглежда развитието на населението и транспорта в гр. София, България. Показва икономическия ефект, който може да има намаления трафик, като представя данни от COVID-19 пандемията през 2020 г по проучване на INRIX (INRIX, 2020). Обсъжда се важността и различните начини, по които може да се определя нивото на трафик и как би могла да се използва тази информация за подобряване качеството на живот. Обръща се специално внимание на начините на събиране на данни и тяхната чувствителност и регулаторни изисквания (GDPR European Union, 2020).

Във втората част се определят целта и задачите на дисертационния труд, а в третата част е описана структурата на документа.

2. Сравнителен анализ на текущото състояние на изследванията в областта на методи за събиране на данни и определяне на нивата на трафика на обществения транспорт

Втора глава се състои от пет части. В първата част се разглеждат тенденциите на публикациите, индексирани в научната база от данни Scopus по определени ключови думи, свързани с темата на дисертацията.

Във втора част "Методология за класификация и анализ на текущото състояние на методите за събиране на данни и определяне нивата на трафика" се разглеждат различни методи за събиране на информация, която може да бъде използвана за анализ на нивото на трафик и се представя категоризация спрямо избрани критерии на всеки от тях.

В трета част се прави обзор на методите за прогнозиране на нивата на натовареност на трафика с различни примери и цитирани проучвания по темата.

В част четири се разглеждат някои основни типове невронни мрежи, които са използвани в този дисертационен труд. В последната част са представени приносите на главата и къде са представени те.

3. Индекс на натовареност на градския трафик, базиран на позиционни данни от периодичен обществен транспорт

Първата уводна част определя задачата на главата, да се провери дали данните от позиционните сензори на транспортни превозни средства от обществения градски транспорт могат да бъдат използвани, за да се изчисли в реално време нивото на трафика в градска среда.

Втората част се концентрира в това да анализира времеви и позиционни данни от обществения транспорт в два различни града – гр. Единбург, Шотландия и гр. София, България. В следващата част са обособени заключенията от направените изследвания.

В четвърта част, позовавайки се на резултатите от предходната, е предложен алгоритъм за изчисление на индекси на натовареност на трафик, базирани на позиционни данни от превозни средства на публичния транспорт в 30 минутни интервали от време.

В части пет и шест са обръща внимание на възможните начини за събиране на данни от периодичния публичен транспорт, които да бъдат използвани за входни данни на алгоритъма и визуализацията на резултатите му след изпълнение.

В седма част са описани експериментални сценарии, които валидират използването на алгоритъма и определят висока позитивна корелация на Пиърсън между индексите на алгоритъма и реалното ниво на трафик. Главата завършва със заключение и представяне на приносите.

4. Описание на създадените софтуерни инструменти за провеждане на експериментите

Главата съдържа пет части, като последните две са заключение и приноси. В първата част са изложени функционалните и нефункционалните изисквания към софтуерните модули за изчисление на индексите по представения в предишната глава алгоритъм.

Във втора част са описани софтуерните модули, като са описани взаимовръзките им, форматите на входните и изходните файлове, клас диаграмите и са показани екраните за визуализация.

В трета част е описан и показан софтуерът за измерване на времената за преминаване на обществен транспорт от експеримента, направен в гр. София.

5. Предсказване на индекса на натовареност на трафика

Пета глава съдържа общо пет части, като последните две са описание на приносите и заключение. В първата част е описана целта на създадения експеримент и неговите етапи. Той ще бъде използван за определяне на индекса и сравнение на различни методи за прогнозиране на нивото на трафик, изчислени от алгоритъма представен в този труд.

Във втората част е описан сценарият на експеримента и също са дадени различните операции по подготовка на данните, за да бъдат използвани от механизмите за машинно самообучение, избрани за прогнозиране на резултатите.

В третата част са описани резултатите от експеримента, като са показани примери и са дадени сравнения на представянето на различните избрани алгоритми за едностъпкови модели (в един времеви интервал напред в бъдещето) или многостъпкови модели (в няколко времеви интервала напред).

III. СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

1. ГЛАВА 1. Увод

София е европейски град и най-голямото населено място на територията на Република България. По данни на Националния статистически институт, въпреки намаляващото население на страната (НСИ, 2021), в столицата прогнозите са, че до 2030 година то ще нарасне с още 25 000 души (НСИ, 2018). Прираст се наблюдава и при количеството на автомобилите по пътищата на града. Според данни от Столична община от 2011 г. до 2020 г. броят им е станал почти двоен – от 462 043 на 833 280 (Сантова, 2021). Такава е ситуацията и в много други големи градове. Нарастващото население, нарастващият трафик и нарастващата нужда от качествено и безпрепятствено придвижване в градската среда, в София и градове като нея, създават големи предизвикателства пред администрацията и бизнеса. Съществуват проучвания, които показват потенциалните ползи от

провеждане на политики за оптимизиране на транспорта чрез намаляване на нивата на градския трафик, които във финансово изражение достигат и до милиарди левове за по-големите държави (INRIX, 2020). За решаване на този проблем, трябва да се разработят механизми и инструменти за следене на пътната ситуация. С тяхна помощ биха могли да се идентифицират критични пътни отсечки, за които да се вземат конкретни мерки.

Същевременно със засиленото внимание към чувствителността на събирането и управлението на личните данни и регулации като европейската General Data Protection Regulation (GDPR) или калифорнийската California Consumer Privacy Act (CCPA), трябва да се обърне и особено внимание на начините, по които тези механизми и инструменти оперират.

2. ГЛАВА 2. Сравнителен анализ на текущото състояние на изследванията в областта на методи за събиране на данни и определяне на нивата на трафика на обществения транспорт

В тази глава се изследват какви са тенденциите по тематиката в научната литература. След това се създава методология за категоризиране на проучванията в две основни насоки на текущият дисертационен труд:

1. Определяне нивото на трафик в градска среда
2. Определяне на нивото на трафик в бъдещ момент.

При проучване на броя на публикуваните статии в научни издания в базата от данни Scopus по ключови думи свързани с текущия труд – “traffic congestion “ („задръствания в трафика“), "urban traffic" („градски трафик“), “public transport” („градски транспорт“), “public transport big data” („градски транспорт и големи данни“), “traffic prediction” („прогнозиране на трафик“), “congestion prediction” („прогнозиране на задръствания“), за годините 2000 г. до 2020 г., се вижда, че всички следват нарастваща тенденция, което показва интереса на изследователите към тематиката.

2.1. Методология за класификация и анализ на текущото състояние на методите за събиране на данни и определяне нивата на трафика.

Поради важността на тематиката много научни проучвания са направени за засичане и измерване на задръстване на трафика. Има множество начини за събиране на данни и визуализация на текущата динамика на движението в един град. Основните, които се срещат в наши дни и са разгледани тук са:

- Умни устройства / телефони
- GPS в автомобил
- Видеонаблюдение
- Позиционни данни от градския транспорт

2.1.1. Данни от Умни устройства / телефони

Една опция е използването на данни за локацията на мобилни устройства на пътници и водачи (Martín et al., 2019), (Idachaba & Ibhaze, 2016), (Tu et al., 2021). Друга такава е използването на данни от телефонните антени на мобилните устройства, давайки не до толкова прецизна информация, но с обещаващи резултати (S. Li et al., 2020) .

Проблемът с тези данни е, че в повечето случаи са притежание на частни компании или индивиди и не са предоставят на изследователи или правителства, за да бъдат анализирани. Тези данни също биха могли да се счетат за чувствителна информация за хората, които ги предоставят и следователно трябва да преминат през предварителна анонимизация, което би добавило допълнителен слой на сложност, който евентуално може да е предпоставка за компрометиране и риск на сигурността на данните.

2.1.2. GPS в автомобил

Друга опция би била да се използват GPS и интернет свързаността на персоналните моторни превозни средства, за да се постигне подобен резултат (D'Este et al., 1999), като са правени експерименти и с датчиците на автомобилите на таксиметрови компании (Kan et al., 2019). Този подход споделя много от

плюсовете и минусите на мобилните устройства и би могъл да бъде съпътстващ източник на информация.

2.1.3. Видеонаблюдение

Друг възможен подход е използването на данни от камери, чрез анализ на картинен и видео поток (Buch et al., 2011; Nemade, 2016) (Nemade, 2016). Чрез използване на конволуционна невронна мрежа (CNN) и анализ на изображенията изследователи са успели да достигнат 89.5% точност на класификация на нивото на трафик (Kurniawan et al., 2018).

Това обаче идва с цената на използването на комплексни алгоритми, скъпа изчислителна мощ (Esteve et al., 2007) и отчитане на различни фактори като осветления, метеорологични условия и д.р. (Stetsenko & Stelmakh, 2020) (J. Li et al., 2021). В икономически план при липса на внедрена инфраструктура, би се наложила голяма инвестиция за инсталирането на устройства за мониторинг с достатъчно присъствие, за да покрият главните пътни артерии.

2.1.4. Позиционни данни от градския транспорт

Съществуват проучвания, които използват автобусния транспорт като проби в трафика, за да се определя пътната ситуация. Carli et. al. дефинира метрики за измерване на задръстванията в градска среда, използвайки GPS технология в дадена област. Проучването не дискретизира класификация на задръстването на трафика, а сравнява стойностите с най-добрите получени до този момент във времевите серии (Carli et al., 2015). Друго проучване, което използва автобусите като проби е направено от Kumar & Sivanandan, 2019. В него автобусите са снабдени с GPS устройства и се сравняват уникалните характеристики на автобусите спрямо останалите превозни средства на пътя. Проучването дефинира и индекс на задръстване (CI) (Kumar & Sivanandan, 2019).

В Таблица 1 е направено обобщение на посочените по-горе подходи. Те са класифицирани спрямо критериите, които дефинираме, както следва:

- **Наличност** – кореспондира с обема на данните, които се записват, не с лекотата на достъпа до тези данни.
    - Високо – налично покритие на целия град
    - Средно – налично покрити върху главните пътища
    - Ниско – ограничено покритие

- **Поверителност** – данните могат да бъдат налични в изобилие, но спрямо местни регулации (като например европейската GDPR), могат да не са достъпни или ако са достъпни да не е позволена обработката им.
    - Високо – не се използват поверителни лични данни на хора
    - Средно – не във всички случаи се използват лични поверителни данни на хора
    - Ниско – използват се поверителни лични данни на хора

- **Цена** – разходи по изпълнението както на събирането на данни, така и на обработката им.
    - Високо – Цена на единичен модул за отчитане и монтаж над 1000 лв.
    - Средно – Цена на единичен модул за отчитане и монтаж между 100 и 1000 лв.
    - Ниско – Цена на единичен модул за отчитане и монтаж под 100 лв.

От таблицата по-долу става ясно, че подходът от текущото проучване (Позиционни данни от градския транспорт) ни предоставя евтин метод с високо ниво на поверителност на информацията на отделния участник в движението, като същевременно има достатъчно налични данни за осъществяване на практически замервания на пътната обстановка на главните пътища.

Таблица 1 Категоризация на възможни подходи за събиране на данни спрямо дефинирани критерии

№	Подход	Наличност	Поверителност	Цена
1.	GPS в автомобил	Високо	Средно	Средно
2.	Умни устройства / телефони	Високо	Ниско	Средно
3.	Видеонаблюдение	Ниско	Ниско	Високо
4.	Позиционни данни от градския транспорт	Средно	Високо	Ниско

2.2. Обзор на методите за прогнозиране на нивата на натовареност на трафика

Съществуват различни начини за предсказване на нивото на трафик, които са изследвани в литературата. Те биха могли да се разделят на две основни групи – такива, които използват методи за дълбоко машинно самообучение (Deep learning), като невронни мрежи (в техните различни разновидности) и такива, които използват различни изчислителни и статистически модели като Калман филтри (Kalman, 1960), ARIMA (Lan & Miaou, 1999), Експонентни филтри (Ross, 1982), Box–Jenkins метод (Hamed et al., 1995) и др. Съществуват и проучвания, които използват и комбинация от гореизложените модели (Wang et al., 2022). Недостатък на чисто статистическите модели е, че се правят предположения, които не винаги съвпадат с динамиката на данните, събрани от реалния свят. Поради тази причина през последното десетилетие, множеството избирани за целите на предсказването на трафика изследвания са направени с дълбоко машинно обучение чрез невронни мрежи (de Medrano & Aznarte, 2020).

3. ГЛАВА 3. Индекс на натовареност на градския трафик, базиран на позиционни данни от периодичен обществен транспорт

3.1. Увод

В един модерен град повечето главни пътни артерии имат един или друг тип периодичен публичен транспорт. В уводната глава бяха разгледани ползите от използване на данните на обществения транспорт, за измерване на състоянието на пътната обстановка - той има почти постоянно присъствие на пътя и всеки автобус може да се разглежда като измерващ агент – проба в трафика. Също така те по дефиниция са анонимни, тъй като не се използват лични данни на отделни индивиди, а само данните за позицията на конкретен автобус.

Цел на текущата глава е да се изследва хипотезата, дали данните от позиционните сензори на тези транспортни превозни средства могат да бъдат използвани, за да се изчисли в реално време нивото на трафика в градска среда. Нивото на трафика ще бъде дефинирано, чрез индекс на трафик, който да се изчислява на половинчасови диапазони. Индексът трябва да определя нивото в отделен пътен сегмент.

3.2. Анализ на времеви и позиционни данни от обществения транспорт

На база по-горе поставените задачи са анализирани данни за разпределението на времената на преминаване на превозни средства през различни пътни сегменти в две европейски столици – гр. Единбург, Шотландия и гр. София, България. Ще се разглеждат времената в тяхната усреднена стойност в половин часови интервали и ще се изследват хистограмите, наклона и свойствата на разпределението на данните.

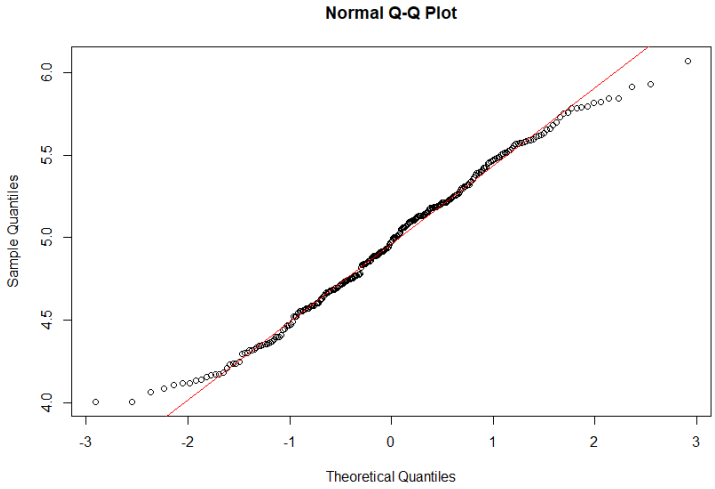
3.2.1. Изследване на данни от гр. Единбург, Шотландия

3.2.1.4. Резултати

В 10 от 10-те или в 100% от изследваните сегменти се наблюдава десен наклон на разпределението на вероятностите. Често използван похват за приближаване

на разпределението на вероятностите на набор от данни до симетрична форма е прилагането на трансформация на данните (Vadali, 2017).

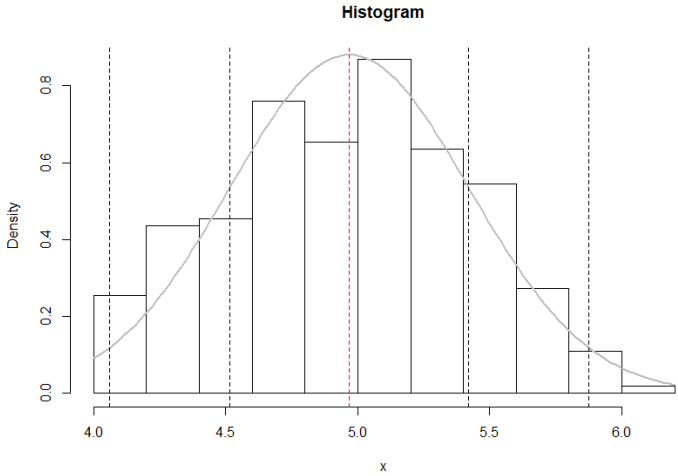
След прилагане на логаритмична трансформация на данните можем да проверим Q-Q графиката на разпределението (Фигура 18), където с изключение на единични стойности в двата ѝ края (екстремните стойности), повечето от точките лежат на Q-Q линията. Това сигнализира, че разпределението на извадката от данни след трансформация е силно приближена с такава от теоретично нормално разпределение.



Фигура 18 Q-Q диаграма на натурален логаритъм от средните времена прекарани в сегмента

Това бива подкрепено и от хистограмата (Фигура 19). Тук се наблюдава една симетрична графика. Чрез използване на функцията skewness от библиотеката moments на програмният език R, може да се сравни наклона на хистограмите преди и след. Изчислявайки го се получават стойностите от 0.905724 за данните преди логаритмичната трансформация и -0.007795457 за тази след. Стойности

близки до нулата означават, че диаграмата е симетрична. Стойности по-големи или по-малки от нула, в зависимост от знака отпред показват, че графиката е наклонена наляво или надясно. В случая стойността на логаритмичните данни показва, че графиката е почти симетрична.



Фигура 19 Хистограма на натурален логаритъм от средните времена прекарани в сегмента

Провеждайки същия експеримент за всички изследвани сегмента от набора от реални данни се наблюдава подобна тенденция. Средно процентът на подобрение на наклона на хистограмите за 10-те сегмента е **74.02%**. Средната стойност на функцията „kurtosis“ на данните преди трансформацията за всички сегменти е 7.4406588, а след нея е 3.0526822, при определена стойност от „3“ за теоретичното нормално разпределение (Pearson, 1905).

3.2.1. Изследване на данни от гр. София, България

В предишната секция беше показан статистическия анализ при обработката на данните от гр. Единбург, Шотландия. Въпреки това, че бяха взети случайни

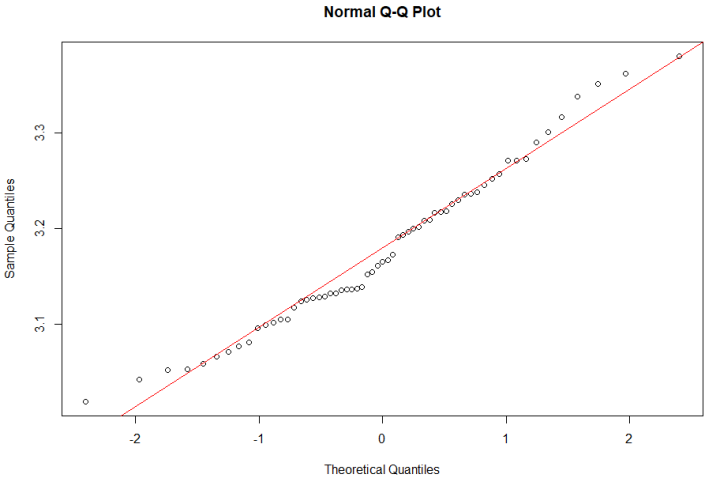
пътни сегменти, които бяха изследвани, съществува възможност този анализ да е валиден само за този определен град и да не бъде приложим в останалите. За да бъде подсигурено, че това не е така е нужно да се намери пример от друг град или държава, който да подкрепя намерените резултати.

В уводната глава беше описано, че град като София би бил подходящ за изследване като текущото, тъй като има значителен брой превозни средства на градския транспорт, покриващи основните му пътни артерии. За съжаление обаче позиционните данни на автобусите и тролеите в столицата на България не са публично достъпни, дори със задна дата. Поради тази причина, за целите на изследването са необходими алтернативи подходи за извличане на данни за гр. София.

Един от тези подходи е използването на камери за следене на метеорологичната обстановка. Текущото изследване, не се нуждае от конкретни координати, по които да се изчисли времето, което отнема на превозното средство да премине през пътен сегмент, а от самото време – то може да бъде засечено директно от направен видео запис на избран пътен сегмент. За целите на експеримента в текущото проучване, видео материали бяха снети от уебсайтът https://weather-webcam.eu (Weather Webcam EU, 2022), като в реално време беше записан екранът с пуснатата уеб страница чрез приложението Zoom (Zoom Video Communications, 2022).

Целта на изследването е да могат да се сглобят материали, заснемащи трафика в продължение на поне два дни, които после да бъдат обработени, като бъдат засечени времената, отнемащи на превозните средства на градския транспорт да преминат през избрания пътен сегмент. Общият брой регистрирани времена след провеждане на експеримента са 320. Първата точка на всеки от двата дни е регистрирана в диапазона 05:30 ч. – 06:00 ч., а последната в 23:30 – 00:00 ч. Следва прилагане на трансформация на данните на база направената хипотеза – взимане на техните логаритмични стойности, усреднени на всеки половин час и изследване на разпределението им.

На Фигура 26 е изобразена Q-Q диаграмата, сравняваща разпределението на стойностите с теоретично нормалното разпределение. Въпреки не големият брои записи се вижда, че двете разпределения се доближават.



Фигура 26 Q-Q диаграмата на не екстремните стойности на логаритъм от средните времена по половин час, измерени от кв. Дружба

3.3. Заключение от изследването на данни от Единбург и София

В предишните две секции беше показано, чрез изследване на реални времена на преминаване през различни и разнородни сегменти, че логаритъм на тяхната средна стойност, взета в диапазони от половин час се приближава до нормално разпределение. В следващите секции е формализиран алгоритъм, който взима предвид това наблюдение и чрез прилагане на статистически методи изчислява 6 стойности на индекс за натовареност на трафик. След това се изследва корелацията между така изчисления индекс и реално ниво на трафик посредством набор от експерименти.

3.4. Предложен алгоритъм

На база изследваните данни от градовете Единбург и София е предложен сленият алгоритъм за изчисление на индекса на трафик. Дефинициите използвани в тази секция са показани в Таблица 4.

Таблица 4 Дефиниции

Символ	Дефиниция
$\mathbf{T_j}$	30 минутен времеви интервал
$\mathbf{CP_i}$	Контролна точка – пътят е разделен от контролни точки
$\mathbf{BS_i}$	Автобусна спирка $\{ BS_0, \dots BS_n \} \subset \{ CP_0, ..., CP_m \}$
$\mathbf{S_i}$	Сегмент. Един сегмент се определя от две последователни контролни точки.
$\mathbf{t_kS_i}$	Времето, за което автобуса е пропътувал през сегмент S_i
$\mathbf{T_jS_i = \{ t_0S_i, .. t_kS_i \}}$	Множество от всички времена записани за автобусите при сегмент S_i в интервал T_j
$\mathbf{MT_jS_i}$	Средната стойност на времената в множество T_jS_i
HDS	Исторически данни
HDS[i]	Исторически данни за сегмент i
$\mathbf{IS_i}$	Трафик индекс за сегмент i
CI	Градски индекс на трафика

Дефинираният индекс на трафика съдържа 6 дискретни стойности от 0 до 5, описани в Таблица 5.

Таблица 5 Индекс на трафика

Индекс стойност	Описание
0	Без трафик
1	Лек трафик
2	Нормален трафик - нисък
3	Нормален трафик - завишен
4	Висок трафик
5	Много висок трафик

„Автобусен маршрут“ се дефинира да бъде предварително избраният маршрут на автобусната линия. Този автобусен маршрут има естествено разделение от автобусните спирки. Представено е и допълнително сегментиране на автобусния

маршрут с помощта на контролни точки (CP). Пътят между две последователни контролни точки бива наречен „сегмент“ (S).

Разделени са 24-те часа на денонощието на 30 минутни интервали, които да бъдат използвани в последващите изчисления. Тези интервали биват обозначени като T_i. Когато автобус преминава през сегмент S_i, се измерва времето за пътуване и се запазва като стойност за този сегмент за този интервал от време. При всяко преминаване на превозното средство се преизчислява средното аритметично време за сегмента. Когато интервалът от време свърши, се взема средно аритметично време за него и се съхранява в историческия набор от данни (HDS). Тези данни се използват, за да се изчисли индексът на трафика за този сегмент. HDS се състои от групи стойности за всеки сегмент. Във всяка група от стойности се съхранява естественият логаритъм на средните аритметични времена за всеки изминал интервал от време.

Забелязва се, че всеки сегмент бива сравняван сам със себе си. Това се прави поради специфичния характер на всеки сегмент - колко пешеходни пътеки има, колко заети са, има ли светофари, пътища без предимство, кръгови кръстовища, ограничения на скоростта и др.

Когато започва нов интервал от време - обозначен с Tj, се създава ново множество за всеки сегмент S_i, а именно T_jS_i. В това множество автоматично се добавя средно аритметичното време за предишния интервал в този сегмент, като първоначално множеството ще изглежда по следния начин $T_jS_i = \{MT_{j-1}\ S_i\}$. Това се прави, тъй като нито един интервал не е изолиран и не се очаква внезапна промяна в трафика между два последователни интервала от време.

Във всеки един момент може да се изчисли индексът на трафик за конкретен сегмент. Ако се предположи, че сегментът е S_i, а интервалът от време, в който се намираме в момента, е T_j. Може да се пресметне средното аритметично на времената, съхранени в множеството T_jS_i, а именно $t = MT_jS_i$. След това е изчислено стандартното отклонение σ, както и средното аритметично μ за историческия набор от данни за сегмента (HDS[i]). После се сравнява

естественият логаритъм на текущо измереното време t със стандартното отклонение от медианата и се намира в кой обхват от данните се намира то. Ако t е между минималното време и μ-2σ - тогава индексът за S_i е 0. По същия начин, ако е между μ-2σ и μ-1σ, тогава е 1; ако е между μ-1σ и μ, тогава стойността ще бъде 2; ако е между μ и μ+1σ, ще бъде 3, след което ако е между μ+1σ и μ+2σ, ще бъде 4 и накрая ако е по-голямо от μ+2σ, то ще бъде 5.

Изчисляването на индекса за града (CI) за времевия интервал Tj се извършва чрез усредняване на индексите за всички сегменти във времевия интервал Tj.

$$CI = \sum_{i=0}^{n} \frac{\text{calc}Index(\log(\text{MTjSi}))}{n}$$

Този подход има някои ограничения. За изчисляване на индекса на задръстванията могат да се използват само пътища, които се подчиняват на определени критерии. Такива критерии са:

- Трябва да има маршрути за превозни средства от обществения транспорт, преминаващи по този път;
- Не трябва да има специална бърза (автобусна) лента за обществения транспорт на този път;

3.7. Експериментална част

3.7.1. Изследване на корелация между брой превозни средства и индекс на трафик изчислен от алгоритъма

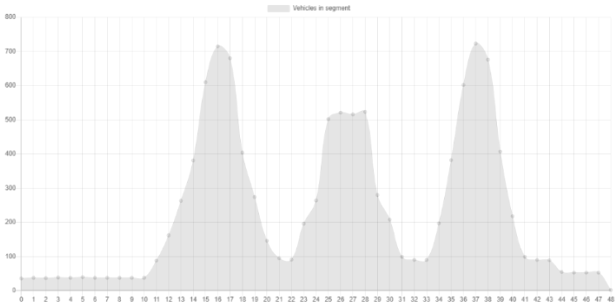
Целта на описания експеримент е да се изследва корелацията между броят на преминали превозни средства през конкретен интервал от време и резултатите изчислени от алгоритъма описан в тази глава. Ако се демонстрира силна положителна корелация, то това би означавало, че алгоритъмът работи правилно и отразява коректно пътната обстановка.

3.7.1.2. Сценарий на експеримента

За съставяне на следния сценарии на експеримента е използван специализираният софтуер "Simulation of Urban MObility" (SUMO). Избрана е отсечка с дължина от 1000 м. в гр. София, съставляваща натоварената част от бул. „Доктор Г. М. Димитров“ през която преминават 3 линии на градския транспорт към момента на писане на дисертационния труд. В участъка няма отредена отделна автобусна лента. Отсечката е разделена на 10 сегмента.

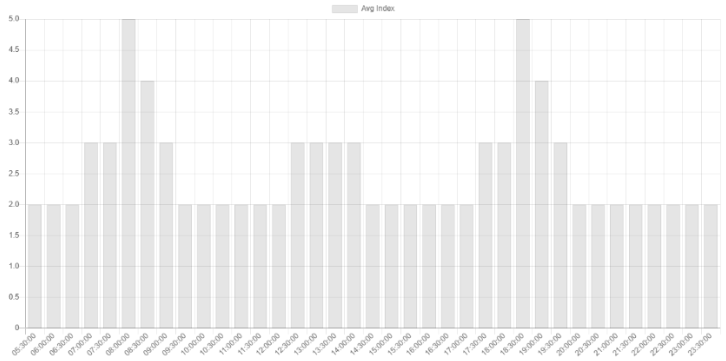
3.7.1.5. Резултати от експеримента

За всеки сегмент от симулацията са представени две графики, описващи характеристиките на входа и изхода на програмата. За пример в тази секция е избран сегмент 2. На Фигура 37 е илюстриран броят на превозните средства преминали през дадения сегмент за конкретен половин час. По абсцисната ос са дадени поредният номер на половинчасието в денонощието – например 0 би съответствал на 00:00 ч. до 00:30 ч., а 18 би бил 09:00 ч. до 09:30 ч. и т.н.. По ординатната ос е обозначен броят на уникалните превозни средства преминали през сегмента. Характеристиките на входните данни са отразени чрез три пика – сутрешен, ранен следобед и вечерен.



Фигура 37 Брой превозни средства преминали през сегмента за конкретен половин час

На Фигура 38 са показани изчислените стойности от алгоритъма за количеството трафик. По абсцисата са отново изброени половинчасовите отрязъци, а по ординатата стойността на индекса в този период. Причината първата стойност да е показана за 05:30 ч. е дефинираното работно време на публичния транспорт. На тази графика отново се отчитат три пика на трафик в сутрешните, следобедните и вечерните часове.



Фигура 38 Изчислен индекс на трафика за сегмент в даден половин час

От сравнението на резултатите на експеримента става ясно, че средният коефициент на корелация на Пиърсън между данните за всичките сегменти от симулацията е **0.8202** или по класификацията – съществува **висока позитивна корелация**. Разглеждайки данните по-подробно виждаме, че приблизително 10% от сегментите имат **много високо ниво на корелация**, а 80% **високо ниво на корелация** между броят на превозните средства в даден пътен сегмент и нивото на трафик, което е изчислено от дефинирания в тази глава алгоритъм.

3.7.2. Изследване на резултатите на алгоритъма във времето

Увеличаването на трафика по пътищата в натоварените часове, би означавало по-голям поток и от пътници, нуждаещи се от градски транспорт. Увеличаването на

пътниците създава нужда, която градският транспорт трябва да запълни, а именно да предостави достатъчно превозни средства, на достатъчно малък период от време, за да задоволи търсенето. За решаване на този въпрос, транспортната компания би могла да бъде гъвкава в избора на това, какъв брой автобуси покриват даден маршрут в даден момент.

Увеличаването на превозните средства означава и увеличаване на броя измерени стойности за сегментите на маршрута. Това би означавало, че в период на половин час на натоварена част от деня ще бъдат събирани повече данни отколкото на не натоварена част. За да реши този казус, алгоритъмът, който е представен в този труд, групира резултатите събрани в даден половинчасов интервал от време и запазва средно аритметичното време, което после се използва за смятане на индекса на трафик.

Целта на представения по-долу експеримент е да се изследва положителното въздействие на това групиране на измерените резултати.

3.7.2.1. Сценарий на експеримента

Създадени са два варианта на алгоритъма:

- Стандартен - който групира измерените стойности в половинчасови диапазони и използва тях за измерване на индекса
- Модифициран – който използва всички измерени стойности за изчисленията

За всеки от диапазоните със завишен трафик са направени няколко експеримента с различен коефициент на увеличение на броят на автобусите 0, 5, 10, 100 пъти и за всеки коефициент е пуснат всеки вариант на алгоритъма – Стандартен и Модифициран, с едни и същи начални данни.

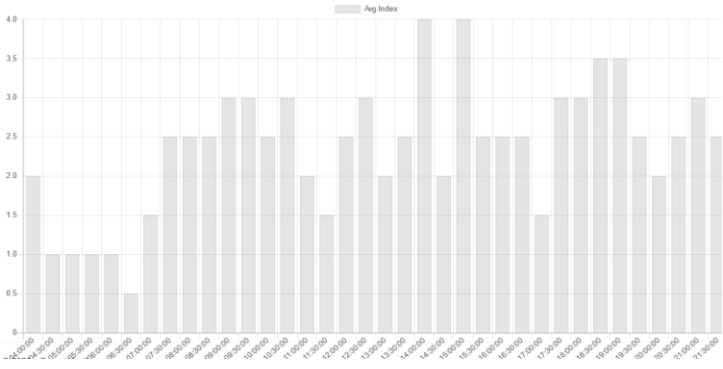
3.7.2.2. Резултати

За разлика от „Стандартния“ алгоритъм, „Модифицираният“ показва деградиране на резултатите с увеличение на коефициента. При „Стандартния“ алгоритъм се наблюдават трите пика на трафик – два по-големи сутрин и вечер и един по-малък в обедните часове при всички коефициенти на учестени измервания. Разглеждайки обаче стойността на коефициента на увеличение на автобусите при „Модифицирания“ алгоритъм, забелязваме, че обедният пик е изчезнал от нея, а вечерният драстично е намалял. В резултатите на коефициент 10, от максимален индекс 5, вече се наблюдава намаление на индекса на 4, докато в екстремния случай при коефициент 100, виждаме, че вечерният пик е достигнал ниво на индекс от само 3.

Горният експеримент демонстрира устойчивостта на алгоритъма от промени в броя на автобусите, обслужващи един маршрут и качеството му да дава консистентни резултати, отговарящи на пътната обстановка.

3.7.3.1. Прилагане на алгоритъма върху сегменти от гр. Единбург, Шотландия

Улица „Princess Street‘ се намира централната част на гр. Единбург. Това е туристическа улица в непосредствена близост до градски забележителности и централната железопътна гара на града. През периода на отчитане минават 569 уникални превозни средства през него. Общият брой на записите, които са били обработени е 8495. На Фигура 41 са показани средно аритметично на няколко дни от индексите на трафика от същия сегмент.



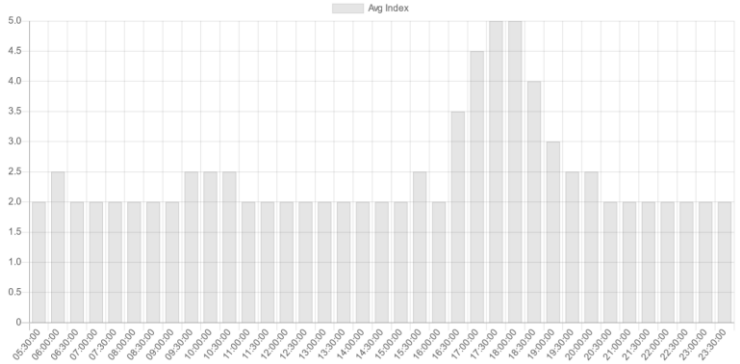
Фигура 41 Разпределение на трафика за втория избран сегмент от гр. Единбург – “Princess Street”

Във фигурата се наблюдават три пика на трафик. Първият е в сутрешните часове в периода 09:00 ч. – 10:30 ч., вторият най-голям пик е през следобедния час 14:00 ч. - 15:00 ч., докато третият е през вечерните часове от 18:30 ч. – 19:00 ч.

3.7.3.2. Прилагане на алгоритъма върху данните от гр. София, кв. Дружба

В тази секция ще се покажат резултатите върнати от алгоритъма за избрания сегмент със записи на времената на градския транспорт от сегмента в кв. Дружба, гр. София. На Фигура 42 и Фигура 43 са показани резултатите от алгоритъма от усреднените индекси в двата разгледани дни. Наблюдават се най-голямо количество трафик в периода 16:30 ч. до 19:30 ч., с пикове в 17:30 ч. и 18:00 ч. Също така може да се отчете и по-висок от нормалния трафик в 9:30 - 10:30 ч, а през останалото време трафикът е нормален за отсечката.

Фигура 42 Визуализация на изчисления индекс върху карта на сегмента в кв. Дружба, гр. София



Фигура 43 Разпределение на трафика за втория избран сегмент от град София, кв. Дружба

3.8. Заключение

В тази глава бяха анализирани два масива от данни, свързани с позицията или времето за преминаване на превозни средства от периодичния обществен

транспорт в градовете Единбург и София и беше приложен статистически анализ върху тях.

На база на този анализ беше направено предположение, което в последствие беше валидирано чрез експерименти, че измереното време има връзка с броя на превозните средства на пътя и беше представен алгоритъм за изчисляване на индекса на трафик както за целия град, така и за конкретен пътен сегмент. Бяха дадени основни дефиниции, приложения и ограничения на алгоритъма. Представен е и механизъм за картна визуализация за състоянието на трафика в града, както исторически така и в реално време.

За валидиране на хипотезата и създадения алгоритъм, бяха изготвени набор от експерименти. Беше изследвана корелацията между броят на превозните средства и резултатите от алгоритъма и беше установена висока положителна корелация между тях. Също така беше изследвана и устойчивостта на алгоритъма при наличие на разнородно количество превозни средства в различните времеви диапазони. С това беше показано, че методът изпълнява заложените му цели.

Така представеният алгоритъм може да се използва от публичната администрация за взимането на решения за промени в инфраструктурата и направата на бъдещи инвестиции. Алтернативно би могъл да се използва за търсене на най-бързи маршрути за града в реално време.

4. ГЛАВА 4. Описание на създадените софтуерни инструменти за провеждане на експериментите

За провеждането на експериментите описани в този дисертационен труд е разработен специализиран софтуер, чиято цел е да извършва дейностите дефинирани от алгоритъма за изчисление на индекса на трафик, както и представяне на резултатите. За целта на създаване на софтуерът е използван .NET Framework и програмният език C#.

4.1. Изисквания към софтуера за провеждане на експериментите

Софтуерът за провеждане на експериментите трябва да покрива следните функционални изисквания:

- Да имплементира алгоритъма за изчисление на индекса за трафик. Главна и основна цел на софтуера е да изчислява индекса, както е описан в този документ.
- Софтуерът трябва да генерира всички нужни изходни файлове за провеждането на експериментите. За целта трябва да се определят стандартни формати, типове и структура на генерираните файлове.
- Софтуерът трябва да предостави модул за графична визуализация на резултатите от алгоритъма, който да използва за вход генерираните изходни файлове от модула за изпълнение на алгоритъма.
- Модулът за визуализация трябва да предоставя възможност за картно представяне на тестваните сегменти. Добавяне на цветовото кодиране над картната визуализация и възможност за проверка на индекса на сегмента в конкретен интервал е също желателно.
- Визуализация на индекса на града е друго нещо, което софтуерът трябва да може да прави. На потребителя на софтуера трябва да се предостави възможност за избор на времевия интервал, в който да се покаже индексът.

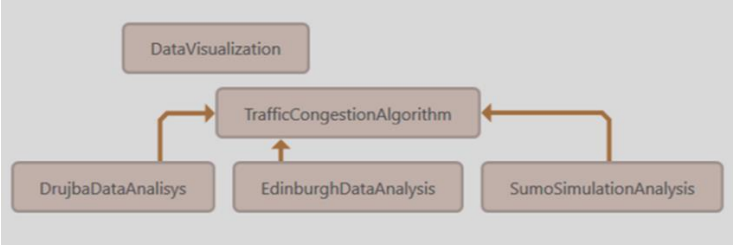
Както и следните нефункционални изисквания:

- Софтуерът трябва да бъде разделен на отделни модули, които да могат да се разработват, обновяват и поддържат независимо един от друг.
- Модулът за алгоритъма да бъде независим от типа на входните данни. Трябва да бъдат създадени нужните абстракции, за да може да се гарантира преизползваемост на надписния код в различните сценарии – генерирани данни, реални данни, данни в различни координатни системи и др.

4.2. Софтуерни модули

На Фигура 45 са показани модулите на софтуера и техните взаимовръзки. Софтуерът има два базови модула:

- TrafficCongestionAlgorithm – модул, който съдържа основните компоненти за изчисление на индекса на трафика. Тук са дефинирани основните модели и обработващи компоненти, както и взаимовръзките между тях. Модулът е направен по начин, който му позволява да бъде агностичен относно формата на входните данни.
- DataVisualisation – това е модулът, който е отговорен за визуализацията на данните. Той предоставя картно представяне, както и спомагателни графики за оценка на крайния резултат.



Фигура 45 Софтуерни модули на системата

Дефинирани са три модула за входни точки на програмата в зависимост от това с какви входни данни ще бъде пуснат експериментът. Те са следните:

- EdinburghDataAnalysis – Това е модулът, който използва за входни данни реални стойности записани от столицата на Шотландия - гр. Единбург и захранва с тях алгоритъма. Резултатите от него са описани в раздел „3.7.3 Прилагане на алгоритъма за изчисление на индекс на трафик върху реални данни.“ от този документ.

- SumoSimulationAnalysis – Това е модулът, който използва за входни данни генерираният файл от SUMO симулацията разгледана в раздел „3.7.1 Изследване на корелация между брой превозни средства и индекс на трафик изчислен от алгоритъм“ от този документ.
- DrujbaDataAnalysis – Това е модулът, който обработва записаните времена от кв. Дружба, гр. София и прилага върху тях алгоритъма, описан в предходната глава.

4.3. Софтуер за измерване на времената за преминаване на обществен транспорт през кв. Дружба, гр. София.

За целите на улесненото измерване на времената на превозните средства в гр. София, кв. Дружба беше разработено и приложение, което да има следните функционални характеристики:

- Да съдържа функциите на хронометър, с възможност за пускане, спиране, рестартиране на времето, както и за спиране със записване на регистрираното време.
- Възможност за показване на досегашните записани времена в обратно хронологичен ред – най-новите да бъдат най-отгоре.
- Възможност за записване на всичките досегашни времена в CSV файл.
- При стартиране, автоматично да изчита записаните времена от последния създаден файл.

Софтуерът е написан на C# с технологията за създаване на интерфейси Windows Forms.

5. ГЛАВА 5. Предсказване на индекса на натовареност на трафика

Изработен е експеримент, отчитащ трафика в 10 пътни сегмента. Експериментът е с данни подготвени от специализиран софтуер за симулация на трафик SUMO в градска среда в продължение на 365 дни, като са приложени и сравнени различни техники за прогнозиране на времеви серии и е направено предложение

за използване на конкретна методология на база резултатите от представянето им.

5.1. Етапи на експеримента

Процесът на подготовка и изпълнение на експеримента се състои в пет основни етапа, като изходните данни на всеки етап захранват следващия. Етапите са както следва:

- Първият етап е подготовка на експеримента – генериране на график на превозните средства, участници в трафика.
- Вторият етап представлява симулиране на трафик на база генерираното разписание от първият.
- Третият етап е изпълнението на алгоритъма за изчисление на индекса на трафик във времевите интервали генерирани от симулацията.
- Четвъртият етап представлява трансформация на изхода от алгоритъма в подходящ формат за консумация от платформата за машинно самообучение.
- Петият и последен етап използва данните от времевите серии с индекси на трафик, за да се направят прогнози за състоянието на трафика в бъдещ момент.

5.2. Сценарий на експеримента

Експериментът се състои в изследване и сравнение на резултатите от различни модели за машинно обучение и прилагането им върху наличните данни, генерирани в първите четири етапа на предишната секция от тази глава.

Фокусът е поставен върху решаването на две основни задачи:

- Предсказване на индекса за натовареност на трафика за следващия времеви интервал в определен пътен сегмент

- Предсказване на индексите на трафик в няколко последователни времеви интервала на база множество от замервания в минал период за определен пътен сегмент

За всяка една от задачите са тренирани различни модели за машинно обучение и се сравняват техните резултати, като накрая на база тези резултати се прави предложение за използване на този, който е дал най-добри показатели.

5.3. Резултати от експеримента

В тази секция се разглеждат в подробности резултатите от различните експерименти и се прави сравнителен анализ на тяхното представяне. Във всички модели за изчисление на функцията за загуба се използва средна квадратична грешка, а за метрики – средна абсолютна грешка.

5.3.1. Едностъпкови модели

Първият тип модели, които ще се разгледат предоставят прогнозиране на една стойност на индекс за натовареност на трафика в бъдещето.

5.3.1.1. Базова линия за моделите

За да се разберат и съпоставят резултатите получени от представянето на моделите за машинно обучение ще се използва прост базов модел като основа. Базовият модел има свойството да вземе стойността в текущия интервал и да я върне като прогноза за следващия, предполагайки, че няма да има резки спадове и покачвания на нивото на трафик в двата съседни интервала.

5.3.1.2. Многостъпков модел, използващ силно свързани слоеве

Първият модел, който ще бъде описан, представлява невронна мрежа с два последователни силно свързани слоя. За да направи едно предсказание, невронната мрежа използва за вход 16-те предишни състояния на индекса.

5.3.1.3. Модел, използващ темпорална конволюция

Традиционно невронните мрежи с темпорална конволюция (CNN) (Tealab, 2018) са били използвани предимно за разпознаване на обекти от изображение, но в последно време излизат проучвания, които показват, че могат успешно да бъдат използвани и за предсказване на времеви серии. В текущия сценарии на експеримента са използвани подобни данни като многостъпковия модел, използващ силно свързани слоеве – входът е 16-те предишни индекса, а изходът е предсказанието за следващия половин час.

5.3.1.4. Рекурентна невронна мрежа с краткосрочна и дългосрочна памет (LSTM)

Рекурентните невронни мрежи (RNN) са невронни мрежи, добре пригодени към работа с времеви серии, тъй като те пазят вътрешно състояние от една времева стъпка към следващата. За следващ модел за предвиждане на индекс на трафика в следващ интервал е избран под-тип на RNN, а именно Long Short Term Memory (LSTM) – краткосрочна и дългосрочна памет.

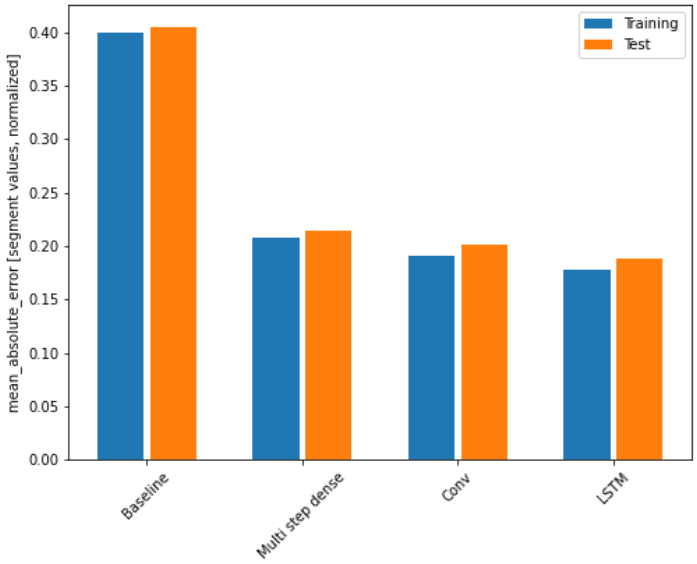
5.3.1.5. Анализ на резултатите от изпълнението на едностъпковите модели

Систематизирано показани могат да бъдат разгледани резултатите получени от изпълнението на моделите, описани в горния раздел. Това, което се наблюдава е, че има близо 50% подобрение на трите модела спрямо базовата линия, като с най-добро представяне е рекурентната невронна мрежа, но също така тя е и по-изискваща откъм време и ресурси за трениране Фигура 69.

В Таблица 16 са показани резултатите на различните едностъпкови модели и техните стойности на средна абсолютна грешка.

Таблица 16 Резултати на едностъпковите модели

Модел	Средна абсолютна грешка
Baseline	0.4052
Multi step dense	0.2146
Conv	0.2013
LSTM	0.1881



Фигура 69 Сравнение на средната абсолютна грешка на едностъпковите модели

5.3.2. Многостъпкови модели

За разлика от едностъпковите модели, многостъпковите такива имат свойството да предсказват няколко стъпки напред във времевите серии. В тази секция от главата ще се разгледат резултатите от различни многостъпкови модели, като накрая ще се направи анализ и сравнение на представянето им.

5.3.2.1. Базова линия на многостъпковите модели

За базова линия е избран модел, който връща последния резултат, като следващи предсказания. Очаквано тук е да се получи по-лош резултат от колкото при линейните модели, тъй като вероятността да се промени трафика в период на 4 часа е по-голяма.

5.3.2.2. Модел, използващ линейна проекция

Линейният модел е най-простият, който ще бъде проверен първи. Той представлява невронна мрежа, която използва за вход последния отчет на индекса на трафик и времевите измерения и на база на тях предсказва какви ще бъдат следващите 8, чрез линейна проекция.

5.3.2.3. Невронна мрежа със силно свързани слоеве

Структурата на тази невронна мрежа се отличава от линейната такава, с това, че между слоевете за вход и изходния такъв има един слой с 512 неврона. Тази конфигурация обаче е подобна на линейната с това, че също като нея приема само последния индекс на трафик и времевите измерения и ги използва, за да определи следващите 8 индекса.

5.3.2.4. Темпорално Конволюционна невронна мрежа

След като са разгледани резултатите от два модела, които използват само последния резултат, за да направят своето предсказание, сега ще бъдат

разгледани такива, които работят с фиксиран брой исторически стъпки назад и използват тази информация, за да изчислят индекса на трафик. Такъв модел е темопорално конволюционната невронна мрежа. За текущия експеримент е конфигурирана да работи с последните 16 записа и отново да дава предсказание за следващите 8. Конволюционният слой на мрежата е съставен от 256 неврона, а използваната активация е ReLU.

5.3.2.5. Рекурентна невронна мрежа

За следващ модел ще се проверят резултатите от рекурентна невронна мрежа и по-специално такава с дългосрочно-краткосрочна памет (LSTM). Мрежата е конфигурирана да акумулира входни данни за период от 48 интервала на време назад и да генерира информация за следващите 4 часа.

5.3.2.6. Ауторегресивна рекурентна невронна мрежа

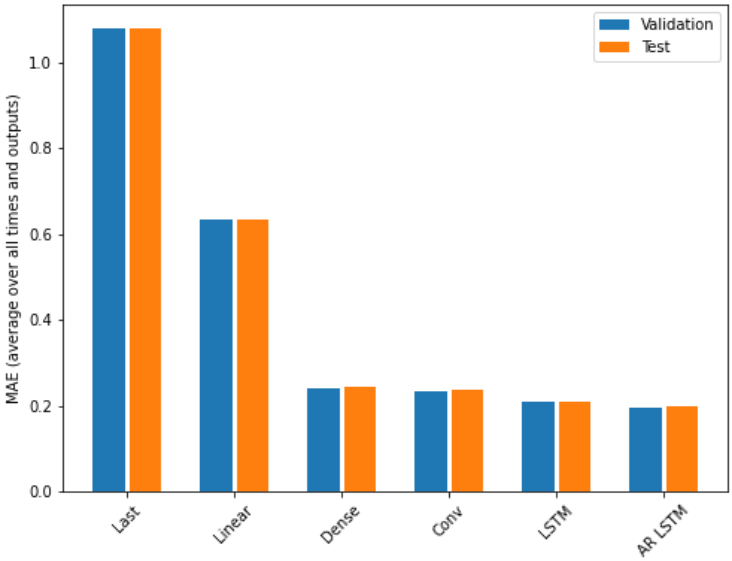
Последният модел, който ще бъде използван за предсказване на времевата серия от индекси на трафик е ауторегресивна рекурентна невронна мрежа. Разликата на този модел с предишните е, че при него резултата се генерира на стъпки и всяка генерирана стъпка се добавя като вход за генерирането на следващата.

5.3.2.7. Анализ на представянето на многостъпковите модели

От резултатите на многостъпковите модели, представени в Фигура 76 и Таблица 17, се вижда, че всички модели освен линейния дават петкратно подобрение на грешката измерена при базовата линия. Моделът с най-добра производителност е ауторегресивната рекурентна невронна мрежа, която е и единственият модел с грешка под 0.2. Този модел обаче е най-ресурсоотнемащ за изчисление, което съответно трябва да се вземе под внимание, когато изчисленията се пуснат върху повече сегменти от реална среда.

Таблица 17 Сравнение на средната абсолютна грешка при многостъпковите модели

Модел	Средна абсолютна грешка
Last	1.0802
Linear	0.6338
Dense	0.2429
Conv	0.2371
LSTM	0.2093
AR LSTM	0.1985



Фигура 76 Сравнение на средната абсолютна грешка при многостъпковите модели

6. ГЛАВА 6. Заключение и бъдещо развитие

6.1. Обобщение на изпълнението на началните цели

В дисертационния труд са разгледани различни механизми за изчисление и предсказване на градски трафик, като специален акцент е отдаден на изследването на трафика с помощта на позиционни данни от обществения транспорт, използван като проби в трафика. Съставена е методология за класификация и анализ на текущото състояние на методите за събиране на данни и определяне нивата на трафика, както и методология за класификация и анализ на текущото състояние на методите за определяне на нивата на трафик в бъдещ момент. Изследвани са разнородни източници на транспортна информация, като чрез приложен статистически анализ са направени изводи за данните и техните разпределения и свойства. На база тази информация е създаден алгоритъм, който категоризира трафика в градска среда, определяйки индекса на трафика. С помощта на редица експериментални сценарии са изследвани и валидирани качествата на изготвения алгоритъм, като се показва, че има силна положителна корелация между стойностите на изчисления индекс и броят на превозните средства на пътното платно. Изследвана е също и устойчивостта на алгоритъма при променлив брой проби в трафика, както и е приложен върху реални данни с цел демонстрация на резултатите. Разгледани са различни методи за събиране на данни, които да могат да бъдат използвани за входен поток на алгоритъма, както и е предложена визуализация на резултатите. Накратко са представени специално разработените софтуерни решения с помощта на които са направени експериментите. След успешно изчисление на индекса на трафик биват изследвани различни начини за определяне на нивата на трафик в градска среда в бъдещ момент, като е сравнено представянето на различните начини за прогнозиране и е определен най-подходящият.

Целта на дисертационния труд за подпомогне взимането на решения за оптимизация на обществения транспорт чрез определяне, изследване и прогнозиране на нивата на натовареност движението, с помощта на данни,

събрани от позиционни координати на периодичния публичен транспорт, използван като проба в трафика е изпълнена, като предложения подход отговаря на всички изисквания, поставени в началото на изследването. Тези резултати биха спомогнали за оптимизиране на транспорта в обществената среда, което би довело до намаляване на разходи и правилно бъдещо планиране.

6.2. Насоки за бъдещо развитие

С оглед на предоставените в този дисертационен труд изследвания и материали могат да се предложат следните насоки за бъдещо развитие.

Изследваният алгоритъм, представен по този начин, не отчита влиянието на съседните пътни сегменти и техния трафик един върху друг, а ги разглежда само в изолация. Изследването на такова влияние не само би могло да подобри определеното на нивото на трафик, но и би могло да се използва за определяне на влиянието от пътя, по който минават превозни средства от периодичния градски транспорт върху съседни пътища, по които не минават, смекчавайки ограниченията на алгоритъма и разширявайки обхвата му на действие.

Преди започване на прилагането на алгоритъма, представен в този дисертационен труд, върху данни от реален град, стои въпросът за оптималното сегментиране на пътищата. Това е не само еднократна стъпка преди евентуалното пускане на мащабната система, но трябва да бъдат поддържани и механизми за изменение на мрежата, при събития на сменени/добавени транспортни линии. Такава задача би било непрактично да бъде решавана ръчно, тъй като за мащабите на град, дори с размерите на гр. София, би означавала сегментация и поддръжка на хиляди километра транспортни маршрути.

Също така наличието на стотици или хиляди транспортни средства, които изпращат всяка секунда позиционните си координати, изисква специална архитектура на система, способна да ги приема, записва и обработва, като в същото време бъде устойчива и постоянно налична. Системата трябва да поддържа и модули за машинно самообучение, които да изпълняват функциите

за прогнозиране на трафика в бъдещ момент. Създаването на подобна архитектура е задача, която би била добро допълнение на текущия труд.

В глава 5 на дисертационния труд бяха разгледани едностъпкови и многостъпкови модели за прогнозиране на трафика. Максималният период на предсказване, който е описан в главата се състои от 8 времеви интервала или продължителност на 4 часа. Съществуват различни приложения обаче, които изискват по-дълги изчисления напред във времето и за които се налагат различен тип статистически анализи и механизми, които не са представени в този труд, но биха били от бъдеща полза.

IV. ПРИНОСИ НА ДИСЕРТАЦИОННИЯ ТРУД

А. Научно-приложни приноси

1. Разгледани и категоризирани са различни видове методи за определяне на натоварването на трафика в градска среда. Направен е анализ на техните характеристики по избрани категории.
2. Направен е статистически анализ на данни за времената на преминаване на превозни средства от периодичния градски транспорт, изчислени от два разнородни източника от гр. Единбург, Шотландия и гр. София, България, през избрани пътни сегменти.
3. Разработен е алгоритъм на база направения статистически анализ, с помощта на който по косвен признак да се определя степента на натовареност на пътен сегмент. Чрез редица експерименти са определени и верифицирани качествата на изложения алгоритъм.
4. Направен сравнителен анализ на резултатите на едностъпкови и многостъпкови модели за машинно обучение, за определяне на нивата на натовареност на трафика в бъдещ момент.

Б. Приложни Приноси

1. Разработен набор от софтуерни решения за събиране, обработване, изчисляване и визуализиране на нивото на натовареност на трафик, посредством представения алгоритъм, предлагащи възможност за модулна интеграция за поддръжка на различен по вид входни данни.

ИЗПОЛЗВАНА ЛИТЕРАТУРА (ИЗВАДКА)

Buch, N., Velastin, S. A., & Orwell, J. (2011). A Review of Computer Vision Techniques for the Analysis of Urban Traffic. *IEEE Transactions on Intelligent Transportation Systems*, *12*(3), 920–939. https://doi.org/10.1109/TITS.2011.2119372

Carli, R., Dotoli, M., Epicoco, N., Angelico, B., & Vinciullo, A. (2015). Automated evaluation of urban traffic congestion using bus as a probe. *IEEE International Conference on Automation Science and Engineering*, *2015-Octob*, 967–972. https://doi.org/10.1109/CoASE.2015.7294224

de Medrano, R., & Aznarte, J. L. (2020). A spatio-temporal attention-based spot-forecasting framework for urban traffic prediction. *Applied Soft Computing*, *96*, 106615. https://doi.org/10.1016/j.asoc.2020.106615

D'Este, G. M., Zito, R., & Taylor, M. A. P. (1999). Using GPS to measure traffic system performance. *Computer-Aided Civil and Infrastructure Engineering*, *14*(4), 255–265. https://doi.org/10.1111/0885-9507.00146

Esteve, M., Palau, C. E., Martínez-Nohales, J., & Molina, B. (2007). A video streaming application for urban traffic management. *Journal of Network and Computer Applications*, *30*(2), 479–498. https://doi.org/10.1016/j.jnca.2006.06.001

GDPR European Union. (2020). *GDPR checklist for data controllers*. Proton Technologies AG. https://gdpr.eu/checklist/

Hamed, M. M., Al-Masaeid, H. R., & Said, Z. M. B. (1995). Short-Term Prediction of Traffic Volume in Urban Arterials. *Journal of Transportation Engineering*, *121*(3), 249–254. https://doi.org/10.1061/(ASCE)0733-947X(1995)121:3(249)

Idachaba, F., & Ibhaze, A. (2016). GSM/GPS Assisted Road and Traffic Congestion Detection System. *International Journal of Applied Engineering Research*, *11*(24), 11610–11613. https://www.researchgate.net/publication/315795939_GSMGPS_Assisted_Road_and_Traffic_Congestion_Detection_System/references

INRIX. (2020). *INRIX 2020. Global Traffic Scoreboard*. https://inrix.com/scorecard/

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, *82*(1), 35–45. https://doi.org/10.1115/1.3662552

Kan, Z., Tang, L., Kwan, M.-P., Ren, C., Liu, D., & Li, Q. (2019). Traffic congestion analysis at the turn level using Taxis' GPS trajectory data. *Computers, Environment and Urban Systems*, *74*, 229–243. https://doi.org/10.1016/j.compenvurbsys.2018.11.007

Kumar, S. V., & Sivanandan, R. (2019). Traffic congestion quantification for urban heterogeneous traffic using public transit buses as probes. In *Periodica Polytechnica Transportation Engineering* (Vol. 47, Issue 4, pp. 257–267). https://doi.org/10.3311/PPtr.9218

Kurniawan, J., Syahra, S. G. S., Dewa, C. K., & Afiahayati. (2018). Traffic Congestion Detection: Learning from CCTV Monitoring Images using Convolutional Neural Network. *Procedia Computer Science*, *144*, 291–297. https://doi.org/10.1016/j.procs.2018.10.530

Lan, C.-J., & Miaou, S.-P. (1999). Real-Time Prediction of Traffic Flows Using Dynamic Generalized Linear Models. *Transportation Research Record: Journal of the Transportation Research Board*, *1678*(1), 168–178. https://doi.org/10.3141/1678-21

Li, J., Xu, Z., Fu, L., Zhou, X., & Yu, H. (2021). Domain adaptation from daytime to nighttime: A situation-sensitive vehicle detection and traffic flow parameter estimation framework. *Transportation Research Part C: Emerging Technologies*, *124*, 102946. https://doi.org/10.1016/j.trc.2020.102946

Li, S., Li, G., Cheng, Y., & Ran, B. (2020). Urban arterial traffic status detection using cellular data without cellphone GPS information. *Transportation Research Part C: Emerging Technologies*, *114*, 446–462. https://doi.org/10.1016/j.trc.2020.02.006

Martín, J., Khatib, E. J., Lázaro, P., & Barco, R. (2019). Traffic monitoring via mobile device location. *Sensors (Switzerland)*, *19*(20), 4505. https://doi.org/10.3390/s19204505

Nemade, B. (2016). Automatic Traffic Surveillance Using Video Tracking. *Procedia Computer Science*, *79*, 402–409. https://doi.org/10.1016/j.procs.2016.03.052

Pearson, K. (1905). "DAS FEHLERGESETZ UND SEINE VERALLGEMEINER-UNGEN DURCH FECHNER UND PEARSON." A REJOINDER. *Biometrika*, *4*(1–2), 169–212.

Ross, P. van. (1982). EXPONENTIAL FILTERING OF TRAFFIC DATA. *Transportation Research Record*.

Stetsenko, I. v., & Stelmakh, O. (2020). Traffic lane congestion ratio evaluation by video data. *Advances in Intelligent Systems and Computing*, *1019*, 172–181. https://doi.org/10.1007/978-3-030-25741-5_18

Tealab, A. (2018). Time series forecasting using artificial neural networks methodologies: A systematic review. *Future Computing and Informatics Journal*, *3*(2), 334–340. https://doi.org/https://doi.org/10.1016/j.fcij.2018.10.003

Tu, W., Xiao, F., Li, L., & Fu, L. (2021). Estimating traffic flow states with smart phone sensor data. *Transportation Research Part C: Emerging Technologies*, *126*, 103062. https://doi.org/10.1016/j.trc.2021.103062

Vadali, S. (2017). *Day 8: Data transformation — Skewness, normalization and much more*. https://medium.com/@TheDataGyan/day-8-data-transformation-skewness-normalization-and-much-more-4c144d370e55

Wang, B., Wang, J., Zhang, Z., & Zhao, D. (2022). *Traffic Flow Prediction Model Based on Deep Learning* (pp. 739–745). https://doi.org/10.1007/978-981-16-5963-8_100

Weather Webcam EU. (2022). *Уеб камера от София от ж.к. Дружба 1 с панорама към Балкана*. https://weather-webcam.eu/sofia-drujba-letishte-stara-planina-live-kamera/

Zoom Video Communications, I. (2022). *Zoom*. https://zoom.us/

НСИ. (2018). *Прогноза за населението по области и пол*. https://www.nsi.bg/bg/content/2996/прогноза-за-населението-по-области-и-пол

НСИ. (2021). *Естествен прираст на 1 000 души от населението по статистически райони, области и местоживеене*. https://www.nsi.bg/bg/content/2989/естествен-прираст-на-1-000-души-от-населението-по-статистически-райони-области-и

Сантова, А. (2021). *Двойно повече са станали колите в София за 10 години*. Капитал. https://www.capital.bg/politika_i_ikonomika/bulgaria/2021/02/08/4171361_dvoino_poveche_sa_stanali_kolite_v_sofiia_za_10_godini/

Декларация за оригиналност на резултатите

Декларирам, че настоящият дисертационен труд съдържа оригинални резултати, получени при проведени от мен научни изследвания, с подкрепата на научния ми ръководител и съавтори. Резултатите, които са получени, описани и/или публикувани от други учени са надлежно и подробно цитирани в библиографията.

Настоящата работа не е прилагана за придобиване на научна степен в друго висше училище, университет или научен институт.

Подпис: _____