



**МИННО-ГЕОЛОЖКИ УНИВЕРСИТЕТ „СВ. ИВАН РИЛСКИ”,
СОФИЯ**

**МИННО-ЕЛЕКТРОМЕХАНИЧЕН ФАКУЛТЕТ
КАТЕДРА „МАТЕМАТИКА И ИНФОРМАТИКА“**

ас. маг. инж. Николина Йорданова Иванова

**ИЗСЛЕДВАНЕ НА ПОДХОДИ ЗА УНИФИЦИРАНЕ НА ГОЛЕМИ ДАННИ ОТ
РАЗЛИЧНИ ЕКОЛОГИЧНИ ИЗТОЧНИЦИ С ЦЕЛ УСЪВЪРШЕНСТВАНЕ НА
ТЕХНИЯ АНАЛИЗ**

А В Т О Р Е Ф Е Р А Т

на дисертационен труд за присъждане на
образователна и научна степен **"ДОКТОР"**

Научна област: 5. Технически науки

Професионално направление: 5.13. Общо инженерство

Докторска програма: Компютърни технологии в инженерната дейност

Научен ръководител:

доц. д-р Веселин Иванов Христов

СОФИЯ, 2025 г.

Дисертационният труд е обсъден и насочен за защита от Разширен катедрен съвет на катедра „Математика и информатика“ към Минно-електромеханичен факултет на МГУ „Св. Иван Рилски“, София, на 11.06.2025 г., съгласно Ректорска заповед № РД-13-28 от 28.05.2025 г.

Публичната защита на дисертационния труд ще се състои пред Научно жури, утвърдено със заповед № РД-13-32 от 19.06.2025 г. на Ректора на МГУ „Св. Иван Рилски“ и ще се проведе на 17.09.2025 г. от 13:30 часа в зала 320 на Минно-електромеханичен факултет, МГУ „Св. Иван Рилски“, София.

Материалите по защитата са на разположение на интересуващите се в канцеларията на Сектор „Студентска и преподавателска мобилност“ на МГУ „Св. Иван Рилски“, Ректорат, ет. 3, стая № 79, тел. 02/ 80 60 209.

Утвърденото Научно жури е в състав:

1. доц. д-р Мариана Трифонова Драганова, МГУ „Св. Иван Рилски“;
2. доц. д-р инж. Йорданка Найденова Анастасова, МГУ „Св. Иван Рилски“;
3. проф. д-р инж. Даниела Асенова Гоцева, ТУ - София;
4. доц. д-р инж. Борислав Иванов Николов, ТУ - София;
5. доц. д-р инж. Ива Нацкова Николова, ТУ - София;

Резервни членове:

1. доц. д-р инж. Георги Станчев Запрянов, ТУ - София;
2. доц. д-р инж. Весела Емилова Петрова, МГУ „Св. Иван Рилски“;

Дисертантът е редовен докторант към катедра „Математика и информатика“ към Минно-електромеханичен факултет.

Изследванията по дисертационната разработка са направени от автора, като някои от тях са подкрепени от проект „Прогнозни модели в минната промишленост, използвайки техники на Data mining и Big Data“, Договор МЕМФ-176/2023 г.

Автор: маг. инж. Николина Иванова

Заглавие: Изследване на подходи за унифициране на големи данни от различни екологични източници с цел усъвършенстване на техния анализ

СЪДЪРЖАНИЕ

I. ОБЩИ ХАРАКТЕРИСТИКИ НА ДИСЕРТАЦИОННИЯ ТРУД	4
I.1. Актуалност на проблема	4
I.2. Обект и предмет на дисертационния труд	5
I.3. Цел на дисертационния труд, основни задачи и методи за изследване	6
I.4. Научна новост	7
I.5. Практическа приложимост	7
I.6. Аprobация на резултатите	8
I.7. Структура и обем на дисертационния труд	8
II. СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД	10
II.1. Теоретични основи и актуално състояние на изследваната област (ГЛАВА I) ...	10
II.2. Системи за пречистване на отпадни води и метеорологични влияния като източници на данни за анализ (ГЛАВА II)	12
II.3. Интегриран подход за анализ и запълване на липсващи данни в системите за мониторинг на отпадни води (ГЛАВА III)	13
II.4. Аналитичен подход за моделиране на зависимости между атмосферни фактори и екологични показатели (ГЛАВА IV)	17
II.4.1. Корелационен анализ	17
II.4.2. Крос-корелационен анализ	17
II.4.3. Факторен анализ	18
II.4.4. Формулиране и проверка на статистически хипотези на база резултатите от факторния анализ	21
II.4.4.1. Хипотеза 1: Влияние на Температурно-химичния фактор (Ф1) върху процесите в ПСОВ	21
II.4.4.2. Хипотеза 2: Анализ на влиянието на Атмосферния фактор (Ф2) върху ПСОВ през преходните сезони	22
II.4.4.3. Хипотеза 3: Анализ на въздействието на Хидравличния фактор (Ф3) върху концентрациите на азот и фосфор	23
II.4.4.4. Хипотеза 4: Влияние на редукцията на броя на наблюдаваните променливи върху достоверността на анализа и информационната цялост	26
III. ОБОБЩЕНИ ИЗВОДИ И ЗАКЛЮЧЕНИЕ	27
IV. ОГРАНИЧЕНИЯ НА ИЗСЛЕДВАНЕТО	28
V. ЛИТЕРАТУРА, ПОСОЧЕНА В АВТОРЕФЕРАТА	29
VI. ПРИНОСИ НА ДИСЕРТАЦИОННИЯ ТРУД	29
VII. СПИСЪК НА ПУБЛИКАЦИИТЕ ПО ДИСЕРТАЦИОННИЯ ТРУД	31

I. ОБЩИ ХАРАКТЕРИСТИКИ НА ДИСЕРТАЦИОННИЯ ТРУД

I.1. Актуалност на проблема

В условията на нарастваща дигитализация обемът от данни, генерирани от разнообразни източници, се увеличава с висока интензивност. Тези данни произлизат от технически, бизнес и екологични системи и често се различават по структура, честота на измерване и качество. Едновременно с това нараства необходимостта от тяхното интегриране и използване за аналитични цели, което поражда редица методологични и технологични предизвикателства.

Сред множеството източници на екологични и инженерни данни особено място заемат информационните масиви, генерирани в рамките на функционирането на пречиствателните станции за отпадни води (ПСОВ). Те съдържат оперативни и технологични параметри, които отразяват състоянието на водния поток и ефективността на прилаганите пречиствателни процеси. В ПСОВ - гр. Кърджали до този момент генерираните данни се записват на хартиен носител и се използват единствено за текущ мониторинг, въпреки че притежават значим потенциал за по-задълбочени аналитични изследвания. Паралелно с това, метеорологичните станции натрупват големи масиви от данни, които позволяват изследване на възможни връзки между климатични условия и технологични процеси. Обединяването на данните от тези два източника създава условия за по-пълното разбиране на взаимодействията между атмосферните фактори и процесите в пречиствателната система, както и за изграждане на модели за мониторинг и прогнозиране.

Работата с подобни хетерогенни масиви от данни изисква решаване на редица ключови проблеми, свързани с предварителна обработка - наличие на липсващи стойности, различна структура и честота на измерване, както и несъгласуваност във форматите и измервателните интервали. За да бъдат данните подложени на съвместен количествен анализ, е необходимо предварително унифициране, включително почистване, синхронизация и импутация (Fadlallah et al., 2023).

Данните, използвани в настоящото изследване, произлизат от два независими и несъвместими по структура източника - ПСОВ и метеорологичната станция (МС) на Националния институт по метеорология и хидрология (НИМХ) в гр. Кърджали. Въпреки съществените различия в начина на събиране и представяне, тяхното обединяване създава основа за изграждане на модел за анализ на връзките между екологични, климатични и технологични фактори.

Значимостта на подобен тип изследвания се засилва и от нарастващата роля на концепцията за Big Data в природните и инженерните науки. Характеристиките на „големите“ данни - обем, разнообразие, скорост на натрупване, достоверност и стойност (5V модел) - са напълно приложими към разглежданите масиви. Това изисква прилагането на подходи, които съчетават класически статистически методи с алгоритми за структурна редукция и обработка на реални оперативни и екологични данни.

Настоящото изследване се фокусира върху изграждането на методологичен подход за обработка и анализ на оперативни и метеорологични данни с цел разкриване

на значими зависимости, които да подпомогнат управлението и прогнозирането на процесите в ПСОВ. За постигането ѝ са използвани подходи като унифициране, импутация на липсващи стойности, корелационен и крос-корелационен анализ, факторен анализ и редукция на размерността. Акцент е поставен върху цялостната количествена оценка, включително проверка на устойчивостта на аналитичните резултати - както при използване на различни импутационни техники, така и при сезонно разделение на данните и интерпретация на латентната структура.

Част от използваните в изследването данни са резултат от преки наблюдения и съвместна работа с екипа на ПСОВ - гр. Кърджали, в рамките на които са извършени проверки на място, верификация на технологични стойности и уточняване на оперативни практики. Данните от МС на НИМХ - гр. Кърджали са предоставени под формата на ръчно въвеждани дневници и съдържат по осем измервания дневно. Анализираният период обхваща 2021-2022 г., като данните и от двата източника са подложени на цялостна предварителна обработка в съответствие с целите на изследването.

Авторката на дисертационния труд, маг. инж. Николина Йорданова Иванова, е асистент към катедра „Математика и информатика“ във Филиала на Минно-геоложки университет „Св. Иван Рилски“ - гр. Кърджали.

I.2. Обект и предмет на дисертационния труд

Обект на изследването в дисертационния труд са **реални, хетерогенни и несинхронизирани оперативни данни**, постъпващи от две независими системи за мониторинг - ПСОВ и МС на НИМХ, функциониращи в рамките на един и същ географски регион. Данните представляват типичен пример за сложни екологични и атмосферни времеви редове с непълна и нееднородна структура, каквито често се срещат в инженерната и екологичната практика, но рядко са обект на систематичен съвместен количествен анализ.

Предмет на изследването е разработването и приложението на интегриран аналитичен подход, чрез който хетерогенни екологични и метеорологични данни да бъдат уеднаквени, допълнени при наличие на липсващи стойности и анализирани чрез статистически и факторни методи. Основната цел е да се разкрият скритите зависимости между технологичните параметри на пречиствателния процес и въздействието на атмосферните фактори. Предложената методология има за цел не само да опише наблюдаваната система, но и да се създаде основа за оптимизация, прогнозиране и подобро управление на пречиствателната дейност чрез количествени модели.

На базата на формулираните обект и предмет се извеждат конкретни цели и задачи, които структурират етапите на изследването и осигуряват логическа последователност в анализа.

I.3. Цел на дисертационния труд, основни задачи и методи за изследване

Основната цел на дисертационния труд е насочена към **разработването и прилагането на методологичен подход за предварителна обработка, унифициране и анализ на големи обеми хетерогенни данни (Big Data) от различни източници**, с акцент върху оперативни данни от ПСОВ и метеорологични наблюдения. Изследването спомага да се разкрият статистически зависимости между технологични и климатични фактори, които биха могли да подпомогнат управлението на пречиствателните процеси.

За постигане на поставената основна цел са формулирани следните изследователски задачи:

Задача 1. Изследване на характеристиките и проблемите при обработка и анализ на хетерогенни и непълни масиви, включващи екологични и метеорологични данни, с различна резолюция, структура и качество.

Задача 2. Разработване на методология за унифициране на данни с цел постигане на съвместимост между различните източници. Проучване и сравнение на съществуващи методи и подходи за почистване, стандартизиране и интеграция на данни с различни формати и структури.

Задача 3. Изследване на методи за запълване на липсващи стойности. Прилагане и сравнение на различни импутационни техники, оценка на точността на възстановените данни и тяхното влияние върху аналитичните резултати.

Задача 4. Избор на подходящи аналитични методи за обработка на унифицирани и запълнени данни - корелационен, крос-корелационен и факторен анализ, както и оценка на устойчивостта на изведените модели при различни условия.

Задача 5. Приложение на разработената методология върху реалните данни от ПСОВ и МС - Кърджали, с цел извличане на значими зависимости между климатичните и технологични показатели. Анализ на резултатите и оценка на приложимостта на методите в реална инженерна среда.

За решаването на тези задачи, данните от ПСОВ са подложени на запълване на липсващите стойности чрез три подхода - кубична интерполация на Ермит (PCHIP), сплайн-интерполация (SPLINE) и авторегресионен модел (AR). За оценка на точността на методите, за всяка променлива са избрани по два интервала с известни реални стойности, които са занулени и впоследствие възстановени чрез трите метода, като резултатите са сравнени с оригиналните стойности. Количествена оценка на резултатите от възстановяването е направена чрез метриците MAE, MAPE и RMSE.

Данните от МС са осреднени до дневна база и са синхронизирани по дати с оперативните данни от ПСОВ. След уеднаквяване на структурата, променливите от двата източника са обединени в общ масив. Чрез корелационен и крос-корелационен анализ са премахнати силно корелирани променливи с цел ограничаване на излишната свързаност в модела.

Върху така получения масив е приложен факторен анализ с Varimax-ротация с цел извличане на латентни зависимости. Сравнени са резултатите при факторен анализ с 2,

3, 4 и 5 фактора, като най-устойчив се оказва моделът с 3 фактора - температурно-химичен, атмосферно-метеорологичен и хидравличен.

Въз основа на резултатите са формулирани и проверени четири изследователски хипотези, свързани с устойчивостта на отделните фактори и ефективността на извлечената латентна структура, както и с изследване на загубата на информация при редуциране на броя на променливите въз основа на резултатите от факторния модел.

I.4. Научна новост

Научната новост на дисертационния труд се изразява във въвеждането на системен подход за предварителна обработка, уеднаквяване и анализ на хетерогенни оперативни и метеорологични данни, генерирани в реални експлоатационни условия. Въпреки че в България подобни данни от ПСОВ се събират ежедневно, тяхната употреба извън рамките на текущия мониторинг е силно ограничена. Данните от ПСОВ – гр. Кърджали не са били подлагани на цифрова трансформация и аналитична обработка до настоящото изследване, а интеграцията им с външни метеорологични източници се реализира за първи път.

Предложена е методология, която включва дигитализация на оперативни масиви, осредняване на данни с различна честота на измерване, оценка и запълване на липсващи стойности чрез сравнителен анализ на импутационни техники, както и синхронизация по време и обединение в унифицирана структура. Включена е и количествена оценка на точността на възстановените стойности чрез зануляване на реални интервали и сравнение с оригиналните данни.

Новостта не се изчерпва с отделните използвани методи, а се изразява в цялостната аналитична рамка, в която те са подредени и приложени. Последователно са използвани корелационен и крос-корелационен анализ за редукция на силно свързани показатели, чрез които първоначалният набор от 59 променливи е сведен до 32; факторен анализ за извличане на латентни структури; както и подбор на представителен поднабор от 15 променливи на базата на факторните товари за всеки фактор. По този начин се постига намаляване на размерността без значителна загуба на информация и с възможност за по-добра интерпретация. Предложената последователност от стъпки представлява приложим аналитичен пакет, който може да бъде използван и при други системи с непълни, хетерогенни и оперативни данни.

I.5. Практическа приложимост

Резултатите от дисертационния труд притежават висока практическа приложимост в контекста на управлението и анализа на данни в ПСОВ. Предложените аналитични подходи - включително факторен, корелационен и крос-корелационен анализ - създават основа за разпознаване на закономерности, които могат да подпомогнат вземането на оперативни и стратегически решения. Разработената методология представлява основа за разработване на автоматизирани системи за обработка, анализ и визуализация на данни, включително в реално време.

В контекста на ПСОВ - гр. Кърджали, изследването демонстрира как рутинно събираните технологични параметри могат да бъдат използвани не само за текущ мониторинг, но и за изграждане на модели за диагностика, оценка на ефективността и прогнозиране на поведението на системата при различни атмосферни условия. Обобщените факторни структури и извършената оценка на сезонната устойчивост на връзките между променливите могат да подпомогнат процесите по оптимизация, както и да бъдат използвани за изготвяне на конкретни препоръки към експлоатационния персонал.

I.6. Аprobация на резултатите

Основни резултати от дисертационния труд са апробирани чрез публикации в рецензирани научни издания и представяне на международна конференция. По темата на дисертацията са публикувани четири научни статии, от които три в съавторство и една самостоятелна. Част от резултатите са представени на 5th International Conference on Communications, Information, Electronic and Energy Systems (CIEES), IEEE, което е потвърждение за тяхната актуалност и научна значимост. Допълнително, изводи от изследването са обсъдени и в рамките на работни срещи със специалисти от ПСОВ в гр. Кърджали, с което се потвърждава тяхната практическа приложимост.

I.7. Структура и обем на дисертационния труд

Дисертационният труд е в обем от **145** страници, като включва увод, **четири** основни глави за решаване на формулираните основни задачи, списък на основните приноси, списък на публикациите по дисертацията и използвана литература. Цитирани са общо **98** литературни източници, като **91** са на латиница и **1** на кирилица, а останалите са интернет адреси. В основния текст са включени общо **26** фигури и **28** таблици. Номерата на фигурите и таблиците в автореферата съответстват на тези в дисертационния труд.

Първа глава разглежда теоретичните основи на концепцията за Big Data, нейните основни характеристики и типовете данни, които обхваща. Описани са предизвикателствата, свързани с анализа на хетерогенни и непълни масиви, както и необходимостта от тяхното унифициране с цел последващ анализ. Представен е преглед на съвременни аналитични методи, приложими в контекста на обработката на големи данни.

Във **Втора глава** са описани използваните източници на данни - ПСОВ и МС на НИМХ в гр. Кърджали. Анализирани са структурата на данните, честотата на измерване и спецификата на показателите, както и проблемите, свързани с дигитализацията, преобразуването и предварителната им обработка. Описани са предприетите стъпки за осигуряване на съвместимост между двата набора от данни.

Трета глава е посветена на обработката на липсващите стойности в данните от ПСОВ. Разгледани са механизмите на възникване на липси, както и съответните методи за тяхното идентифициране и възстановяване. Разработена е методология за запълване на липсващите стойности чрез сравнение между различни импутационни

методи, приложени в среда MATLAB. Ефективността на методите е оценена чрез метрики за точност и грешка на възстановяване.

Четвърта глава съдържа основните аналитични изследвания, извършени върху предварително обработените и запълнени данни. Проведени са корелационен и крос-корелационен анализ за откриване на зависимости между екологични и климатични показатели, както и факторен анализ за идентифициране на латентни фактори и обединяване на наблюдаваните променливи в логически обосновани групи (Shrestha, 2021). Обоснован е изборът на оптимален брой фактори. Формулирани и проверени са изследователски хипотези, свързани със стабилността на факторната структура и влиянието на климатичните и технологичните фактори върху показателите на входа и изхода на ПСОВ.

В заключението са обобщени основните изводи от проведеното изследване. Представени са приносите на дисертационния труд, ограниченията на изследването, списък с публикации и използваните литературни източници.

Дисертационният труд е допълнен с три приложения, които разширяват представените в основния текст резултати. Приложения 1 и 2 съдържат допълнителни графики, илюстриращи резултатите от запълването на липсващи стойности и проверката на ефективността на използваните методи. Приложение 3 съдържа част от MATLAB скрипта, използван при прилагането на факторния анализ върху обединения масив от данни.

II. СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

II.1. Теоретични основи и актуално състояние на изследваната област (ГЛАВА I)

В началото на Първа глава се очертава концептуалната рамка, в която Big Data се разглеждат като съвкупност от подходи, методи и технологични решения за събиране, съхранение и обработка на обемни, разнообразни и динамични информационни потоци. Подчертава се, че Big Data не се дефинират единствено чрез обем, а и чрез сложност, честота на генериране и необходимост от нови аналитични и инфраструктурни решения. Разгледани са основните източници на такива данни, включително сензори, автоматизирани системи, индустриални процеси и онлайн платформи, които генерират комплексна, многомерна и често полуструктурирана информация, която не подлежи на обработка чрез традиционни инструменти. Анализът на Big Data предполага наличие на специализирана инфраструктура, поддържаща паралелна обработка, разпределени изчисления, алгоритми за откриване на зависимости и визуализация на резултатите (Ivanova, 2023).

Специфичният контекст на дисертационния труд поставя Big Data в сферата на екологични приложения, с фокус върху използването на полуструктурирани данни от ПСОВ и МС. Тяхната хетерогенност, различна времева резолюция и наличие на липсващи стойности изискват специфични стъпки по предварителна обработка и унифициране, които формират основата на разработената аналитична рамка в дисертационния труд.

Особено внимание се отделя на основните характеристики, чрез които се дефинират големите данни, известни като петте „V” измерения - обем (volume), скорост (velocity), разнообразие (variety), достоверност (veracity) и стойност (value), които определят качествените и аналитични предизвикателства при работата с Big Data (Dimitrov et al., 2021). В контекста на екологичните системи тези характеристики се проявяват чрез различна честота на измерване, разнородни формати на представяне, несъгласуваности и пропуски, което изисква допълнителна обработка за извличане на значима информация. Адекватният отговор на тези предизвикателства предполага съвместно отчитане на техническите и аналитичните аспекти на петте измерения на Big Data.

В рамките на глава първа е разгледана и класификация на видовете данни според тяхната структура - структурирани, неструктурирани и полуструктурирани. Именно към последната категория спадат и използваните в настоящото изследване информационни масиви, които се характеризират с нееднородна структура, различна честота, наличие на липсващи стойности и вариращи формати на представяне. Това е типично за екологични и технологични системи, в които данните постъпват от различни източници, често несинхронизирани и без пълно покритие. Тези особености поставят необходимостта към предварителна обработка, синхронизация и стандартизация, за да се гарантира съвместимост и надеждност при последващия анализ.

Като част от теоретичния преглед са представени основните измерения на качеството на данните: пълнота, точност, съвместимост, валидност, актуалност, уникалност и интегритет. Тези характеристики са от съществено значение за

надеждността и приложимостта на аналитичните резултати, особено при работа с хетерогенни, оперативни и полуструктурирани масиви (Anastasova, 2019).

В рамките на литературния обзор са разгледани аналитични подходи, доказали своята ефективност при анализа на екологични Big Data. Обърнато е внимание на приложимостта на корелационния и крос-корелационния анализ за оценка на взаимовръзки между технологични и климатични показатели, както и на факторния анализ като метод за редукция на размерността и извличане на латентни структури. Именно тези методи са избрани за основа на аналитичната рамка на дисертационния труд.

Значителна част от главата е посветена на преглед на съвременни приложения на Big Data в различни индустриални и екологични сектори, включително интелигентно управление на водни ресурси, мониторинг на замърсяване, оптимизация на водоснабдителни и пречиствателни процеси, IoT-базирани системи за въздух и отпадъци, прогнозиране на климатични рискове и „умни“ градски инфраструктури. Въпреки широкото приложение на Big Data в екологичния контекст, литературният преглед показва, че оперативните данни от ПСОВ рядко са обект на систематичен анализ, особено в комбинация с метеорологични наблюдения. Тази празнина в литературата очертава необходимостта от разработване на нови методологични решения и е пряко свързана с целта на настоящото изследване.

В обобщение, първа глава утвърждава концепцията за Big Data като ключов инструмент при вземането на решения в сложни и динамични системи. Акцентът върху петте измерения, качеството на данните и спецификата на екологичните приложения създава теоретичната основа за разработване на методология, съобразена с особеностите на изследваните данни от ПСОВ и МС.

Анализът на литературните източници показва, че въпреки натрупания опит при използването на Big Data в различни екологични и индустриални сектори, все още липсват утвърдени методологични решения за съвместна обработка на оперативни и климатични данни, особено при наличие на липсващи данни, различна структура и несинхронизирани измервания. Налице са добре разработени подходи за анализ на полуструктурирани данни и за използване на IoT платформи, но отсъства цялостна рамка, която да комбинира предварителна обработка, запълване на липсващи стойности, унифициране и интегриран количествен анализ на реални данни от ПСОВ и МС.

Въз основа на установените нерешени аспекти, настоящото изследване си поставя за цел да разработи аналитичен подход за предварителна обработка, уеднаквяване и количествен анализ на хетерогенни екологични и метеорологични данни с липсващи стойности. Целта е да се извлекат значими зависимости и да се изградят модели, приложими за управление и оптимизация на процесите в ПСОВ.

II.2. Системи за пречистване на отпадни води и метеорологични влияния като източници на данни за анализ (ГЛАВА II)

Данните, използвани в дисертационния труд, са предоставени от два източника - ПСОВ и МС на НИМХ в гр. Кърджали, за периода 2021-2022 година. Първичната информация от ПСОВ включва измервания на 15 параметъра: дебит, температура, рН, БПК₅, ХПК, концентрации на азот, фосфор и твърди вещества както на входа, така и на изхода от пречиствателния процес. Данните от МС на НИМХ обхващат 44 атмосферни и метеорологични показателя, сред които температура на въздуха, атмосферно налягане, относителна влажност, количество валежи и други.

Измерванията в ПСОВ се извършват два до три пъти седмично, което води до наличие на липсващи стойности във времевите редове. За да се постигне съпоставимост с метеорологичните данни, стойностите са възстановени и приведени до дневен формат. Методите за импутация са разгледани подробно в глава трета на дисертационния труд. Измерванията в МС на НИМХ се регистрират осем пъти дневно в съответствие със синоптичния график. Тези данни са агрегирани чрез осредняване по дати, за да бъдат приведени към същата дневна резолюция, съпоставима с тази на обработените данни от ПСОВ.

Макар двата източника да отразяват реални измервания в рамките на един и същ географски регион, те се различават съществено по структура, честота на измерване, формат и пълнота на информацията. В ПСОВ данните се записват в хоризонтална структура (по дати), докато при НИМХ те са организирани вертикално - по часови интервали. Тази несъгласуваност, заедно с различните формати и липсата на синхрон, налага предварителна трансформация и унифициране с оглед осигуряване на тяхната съвместимост за последващия анализ и моделиране.

Допълнително предизвикателство представлява фактът, че и двата набора от данни първоначално са налични само на хартиен носител. Данните от НИМХ съдържат специфични синоптични символи и означения. Информацията от ПСОВ включва както автоматично отчитани параметри, така и резултати от лабораторни анализи. Това затруднява използването на специализиран софтуер за автоматизирана дигитализация (OCR). Поради това дигитализацията е извършена ръчно, с проверка за точност и логическа консистентност.

След ръчното въвеждане и структуриране, данните от двата източника са синхронизирани по дати и приведени в унифициран формат. Липсващите стойности в данните от ПСОВ са възстановени чрез специализирани методи, а пълният масив е подложен на нормализация и контрол за допустимост на стойностите. Извършена е проверка за отсъствие на дублирани или физически невъзможни записи. В резултат е създаден обединен масив от 59 количествени променливи, който осигурява надеждна база за съвместни анализи.

Представените в глава втора източници на данни и предприетите стъпки по тяхната дигитализация, първична обработка и унифициране формират основата на аналитичната рамка на дисертационното изследване. Направената подготовка осигурява хомогенност, съпоставимост и аналитична надеждност на входните данни -

необходимии условия за постигане на валидни резултати от последващите анализи на взаимовръзките между технологични и климатични променливи. Методологичните решения, приложени в тази фаза, определят надеждността на изведените модели и зависимости, представени в следващите глави.

II.3. Интегриран подход за анализ и запълване на липсващи данни в системите за мониторинг на отпадни води (ГЛАВА III)

Трета глава поставя акцент върху необходимостта от възстановяване на липсващите стойности в данните от ПСОВ, които представляват едно от основните предизвикателства при анализа на големи екологични масиви. Наличието на пропуски може съществено да компрометира точността, валидността и интерпретируемостта на резултатите. При ПСОВ - Кърджали стойностите на редица технологични параметри не се отчитат ежедневно, което води до пропуски в измерванията, включително дни с липсващи стойности за всички показатели.

Поради необходимостта от висока плътност на данните при моделиране на зависимости между климатични и технологични параметри, възстановяването на липсващите стойности е ключов етап от предварителната обработка. Размерът и системността на липсите изключват възможности като изтриване или игнориране, което налага използване на специализирани импутационни методи, целящи максимално запазване на вътрешната структура и информативност на масива. Неподходящата импутация може да доведе до изкривяване на зависимостите и некоректни изводи (Dębski, 2020), поради което тази част от дисертацията служи като критична връзка между данните и последващия количествен анализ.

В тази връзка са приложени три математически подхода:

- **Кубична интерполация на Ермит (PCHIP)** - използва функция, която се стреми да поддържа същата монотонност като тази на данните. Посредством части от кубични полиноми за интерполация между дадени точки се запазва непрекъснатостта на първата производна на функцията. Методът е подходящ при плавно променящи се данни, като същевременно избягва изкуствени екстремуми;

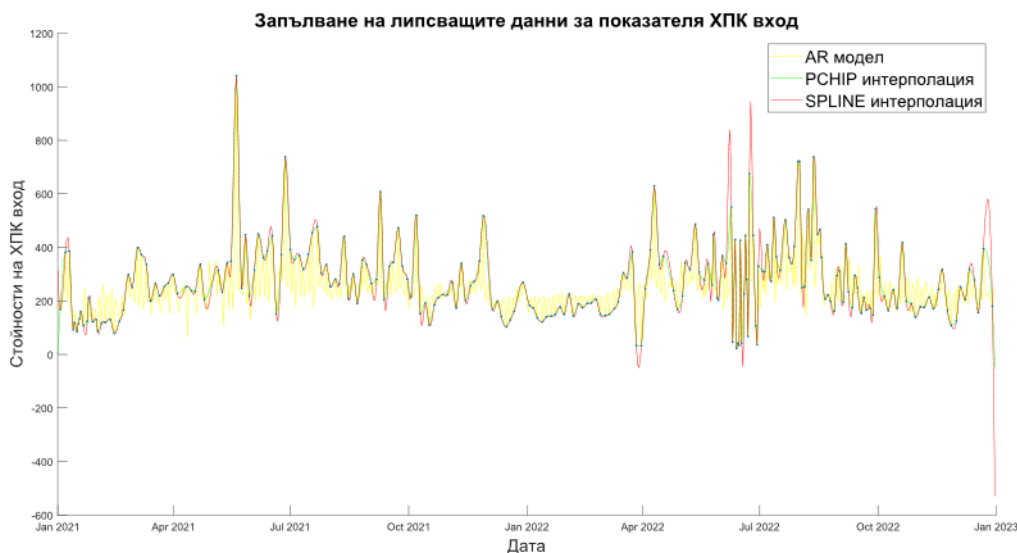
- **Сплайн-интерполация (SPLINE)** - използва последователни кубични полиноми с непрекъснати първи и втори производни, които формират гладка крива през всички налични точки. Въпреки високата гладкост, в контекста на екологични данни с ясно изразени преходи този подход може да създаде изкуствени екстремуми или нереалистични отрицателни стойности;

- **Авторегресионен модел (AR)** - всяка стойност се оценява като линейна комбинация от предходни стойности. Параметрите на AR модела са оценени чрез класически статистически техники, като метода на най-малките квадрати (OLS) и уравненията на Yule-Walker. Моделът улавя сезонност и вътрешни зависимости, показвайки висока ефективност при променливи с ясно изразени колебания (Ivanova et. al, 2024).

Всеки от методите е приложен поотделно върху показателите от ПСОВ. За графична визуализация на поведението на методите са използвани два примерни показателя - ХПК и Общ азот на входа на ПСОВ. Резултатите, представени на Фиг. 1 и 2, показват съществена разлика в характера на възстановените стойности.



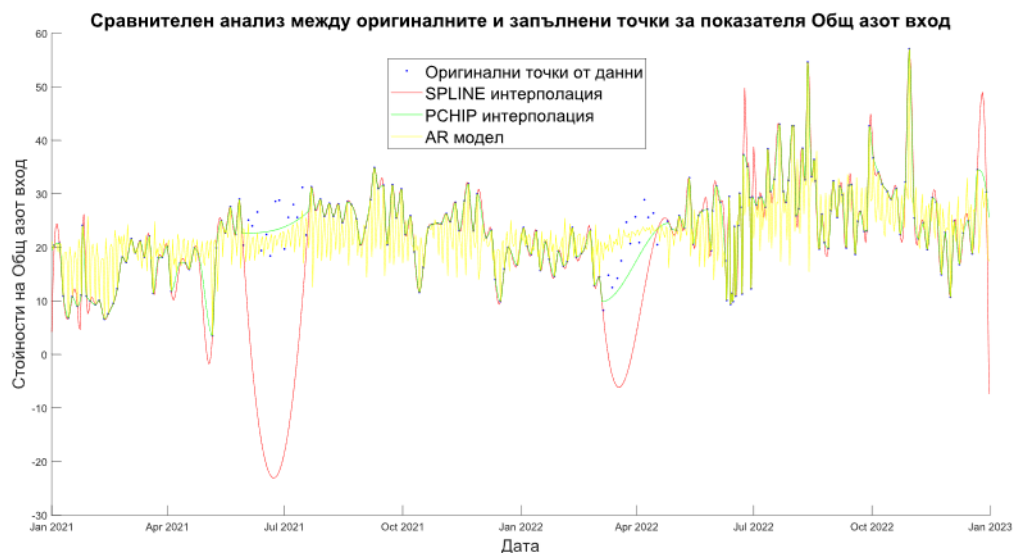
Фигура 1. Резултати от запълването на липсващите стойности чрез PCHIP, SPLINE и AR за показателя Общ азот на входа на ПСОВ



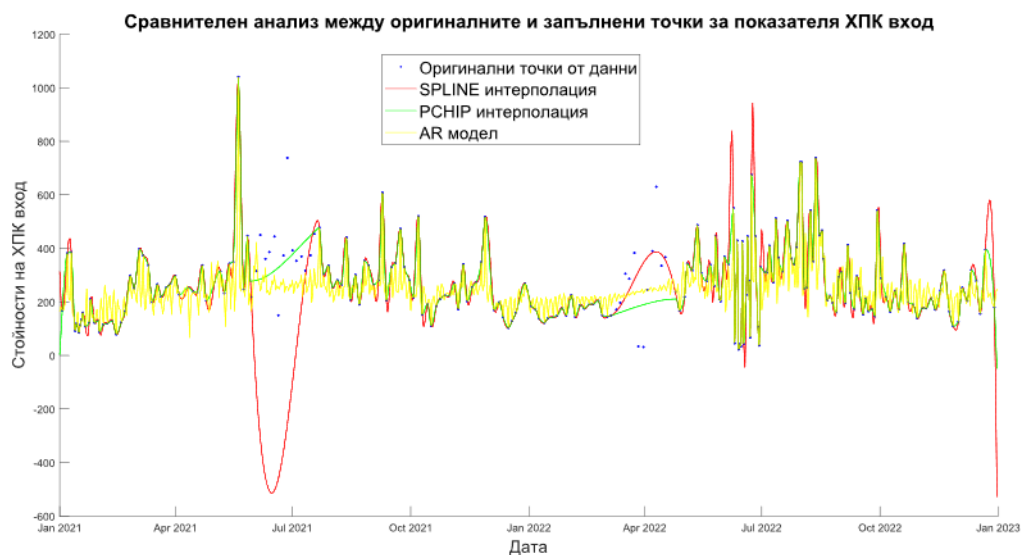
Фигура 2. Резултати от запълването на липсващите стойности чрез PCHIP, SPLINE и AR за показателя ХПК на входа на ПСОВ

Поради наблюдаваните отклонения при прилагането и на трите метода, включително създаването на изкуствени екстремуми (видими на графиките), и липсата на валидационен набор, се налага прилагането на допълнителен експериментален подход. В попълнените масиви умишлено са занулени по два интервала, за които има

известни реални данни. Всеки от трите метода е приложен за възстановяване на тези стойности, като получените резултати са сравнени с реалните наблюдения. На Фиг. 3 и 4 са представени резултатите от експеримента. Графиките илюстрират възстановените стойности от всеки метод, като са показани както оригиналните точки, така и запълнените данни след зануляването.



Фигура 3. Сравнение между оригиналните данни и методите за запълване на липсващи данни за показателя Общ азот вход при два занулени интервала от данните



Фигура 4. Сравнение между оригиналните данни и методите за запълване на липсващи данни за показателя ХПК вход при два занулени интервала от данните

За количествена оценка на точността са изчислени средна абсолютна грешка (MAE), корен от средноквадратична грешка (RMSE) и средна абсолютна процентна грешка

(MAPE). Резултатите от сравнението са представени в Табл. 1 и позволяват директно сравнение по променлива и метод.

Таблица 1. Оценка на ефективността на приложените модели за запълване на липсващи данни

Изследвани параметри		Средна абсолютна грешка (MAE)			Корен квадратен от средноквадратична грешка (RMSE)			Средна абсолютна процентна грешка (MAPE) [%]		
		<i>SPLINE</i>	<i>PCHIP</i>	<i>AR</i>	<i>SPLINE</i>	<i>PCHIP</i>	<i>AR</i>	<i>SPLINE</i>	<i>PCHIP</i>	<i>AR</i>
Входни показатели	ХПК	41,67	15,50	7,66	165,30	52,78	33,53	18,29	7,44	4,02
	БПК5	25,47	6,74	3,79	90,73	28,53	18,53	22,74	4,76	9,63
	pH	0,03	0,01	0,01	0,10	0,04	0,03	0,36	0,16	0,11
	Общ фосфор	3,93	1,03	0,65	14,05	3,25	2,27	21,77	5,93	4,07
	Общ азот	3,23	0,48	0,30	10,21	1,54	1,09	14,42	2,13	1,57
	Неразтв. вещества	24,44	7,22	3,55	82,19	26,54	16,76	14,00	3,82	2,04
Изходни показатели	ХПК	3,94	0,83	0,35	16,40	3,35	1,70	16,81	3,70	1,37
	БПК5	0,63	0,18	0,11	2,20	0,65	0,43	19,65	17,00	11,69
	pH	0,08	0,02	0,01	0,23	0,06	0,04	1,01	0,25	0,17
	Общ фосфор	0,72	0,32	0,29	2,34	1,20	1,13	34,94	7,16	6,28
	Общ азот	0,40	0,19	0,19	1,30	0,66	0,68	15,82	7,96	7,60
	Неразтв. вещества	1,45	0,42	0,22	5,59	1,45	0,88	37,19	15,91	5,79

При почти всички показатели AR моделът демонстрира най-ниски стойности на грешка, включително при силно вариращи променливи. Методът PCHIP също показва добри резултати, особено при по-гладки редове, но в някои случаи изглажда прекалено данните. SPLINE е с най-слаба ефективност, като въвежда изкуствени екстремуми и физически невъзможни стойности, включително под нулата.

На базата на визуалния и количествения анализ, **AR моделът е избран като основен метод за запълване на липсващите стойности** в пълния набор от данни, използван за последващите анализи. Изборът е направен с оглед запазване на вътрешната динамика и физическата реалистичност на редовете. Методът осигурява висока точност, стабилност при показатели с изразена сезонност и минимизиране на отклоненията спрямо реалните стойности.

II.4. Аналитичен подход за моделиране на зависимости между атмосферни фактори и екологични показатели (ГЛАВА IV)

Глава 4 разглежда аналитични изследвания, проведени върху обединените данни от ПСОВ и МС на НИМХ в гр. Кърджали. Целта на аналитичния процес е не само да бъдат описани основните статистически характеристики на изследваните променливи, но и да се идентифицират зависимости между атмосферните условия и параметрите на пречистване на отпадните води. Получените резултати биха могли да подпомогнат процесите на управление и прогнозиране на пречиствателните системи при различни климатични условия.

II.4.1. Корелационен анализ

Първоначално е извършен корелационен анализ за определяне на линейните зависимости между климатичните и технологичните параметри. Изчислена е пълната корелационна матрица, в която ясно се открояват статистически значими зависимости. Например, между температурата на отпадната вода и максималната температура на въздуха е наблюдавана много силна положителна корелация ($r \approx 0.98$), както и с минималната температура на въздуха ($r \approx 0.92$), което потвърждава прякото влияние на климатичните условия върху температурния режим на отпадните води.

Освен това се установяват силни връзки между показателите на входа и изхода на ПСОВ. Например, корелацията между ХПК на входа и на изхода е $r \approx 0.93$, а между БПК₅ на входа и изхода - $r \approx 0.87$. Наблюдавани са и високи положителни корелации между концентрациите на общ азот и общ фосфор както на входа ($r \approx 0.9$), така и на изхода ($r \approx 0.87$), което отразява технологичната взаимозависимост между тези показатели.

Идентифицирани са и случаи на отрицателна корелация, като например между атмосферното налягане и БПК₅ на изхода ($r \approx -0.53$), както и между относителната влажност и рН на изхода ($r \approx -0.62$), което подчертава сложността на взаимодействията между климатичните и технологичните процеси. Повечето от останалите зависимости са слаби или статистически незначителни, което мотивира преминаването към по-детайлни времеви анализи.

II.4.2. Крос-корелационен анализ

За надграждане на резултатите от корелационния анализ и с цел идентифициране не само на силата на зависимостите, но и времевия интервал, в който те се проявяват най-ясно, е извършен крос-корелационен анализ. Изчисленията са проведени за различни времеви отмествания (лагове) - в диапазон от ± 7 и ± 14 дни, като е установено, че най-съществени промени в корелационните зависимости настъпват в рамките на ± 7 дни. Поради това окончателният анализ е фокусиран върху този времеви прозорец.

Крос-корелационният анализ разкрива съществени зависимости между атмосферни и технологични променливи при отчитане на времеви лагове. Например, температурата на постъпващата вода в ПСОВ показва много силна зависимост с минималната

температура на въздуха при лаг -1 ден ($r \approx -0.93$), което показва, че климатичните условия влияят почти незабавно върху температурния режим на водата на входа на станцията. Общият азот на входа демонстрира отрицателна зависимост с обема на постъпилата вода, като максималният ефект се проявява със закъснение от 4 дни ($r \approx -0.6$), което предполага разреждане на концентрацията вследствие на повишен дебит.

Допълнително се наблюдават зависимости между химичните показатели на входа и изхода при различни времеви измествания. Това потвърждава, че част от взаимовръзките между атмосферните и технологичните параметри не се проявяват моментално, а с известно закъснение, което е характерно явление за екологични и технологични системи.

Други двойки променливи не показват съществени подобрения на зависимостта при времево отместване, което насочва към отсъствие на отложени ефекти или необходимост от алтернативни модели. Крос-корелационният анализ разширява разбирането за взаимодействията в системата, като установява отложени ефекти между определени климатични и технологични параметри, които могат да послужат като основа за динамично моделиране.

От общо 59 количествени променливи, включени в обединения масив, 44 произхождат от метеорологичната станция. Сред тях са установени както силно корелиращи помежду си показатели, така и такива, които не показват ясно установима връзка с параметрите от ПСОВ. На база на резултатите от корелационния и крос-корелационния анализ, броят на атмосферните променливи е редуциран до 17, а общия масив - до 32 показателя, които впоследствие са използвани за изграждане на факторния модел.

II.4.3. Факторен анализ

В настоящото изследване е проведен факторен анализ с цел идентифициране на латентни зависимости в обединения масив от екологични и технологични променливи, както и редуциране на размерността на данните без съществена загуба на информация. Разгледани са няколко варианта на анализа с различен брой фактори - два, три, четири и пет, като изборът на оптималната структура е направен на базата на съвкупност от критерии: критерий на Кайзер (собствена стойност > 1), графична интерпретация чрез Scree Plot и оценка на обяснената дисперсия от всеки фактор. Основната цел е да се открият водещите фактори, които оказват влияние върху качеството на водата, и да се създаде компактен и интерпретируем модел.

Факторният анализ е извършен чрез метода на главните компоненти (Principal Component Analysis - PCA), последван от Varimax ротация с цел постигане на по-ясна структура и разпределение на факторните товари.

Въпреки че критерият на Кайзер позволява извличане на до осем фактора, резултатите показват, че при увеличаването им над четири броят на дублиращите се компоненти с висока взаимна корелация нараства, което влошава стабилността и интерпретируемостта на модела. Допълнително е анализирана и Scree Plot графиката

(Фиг. 5), където се наблюдава отчетлив спад в обяснената дисперсия след четвъртия фактор. Тази „точка на огъване“ потвърждава, че включването на допълнителни фактори не води до съществено увеличение на обяснената вариация в данните.



Фигура 5. Scree Plot - обяснена дисперсия по фактори

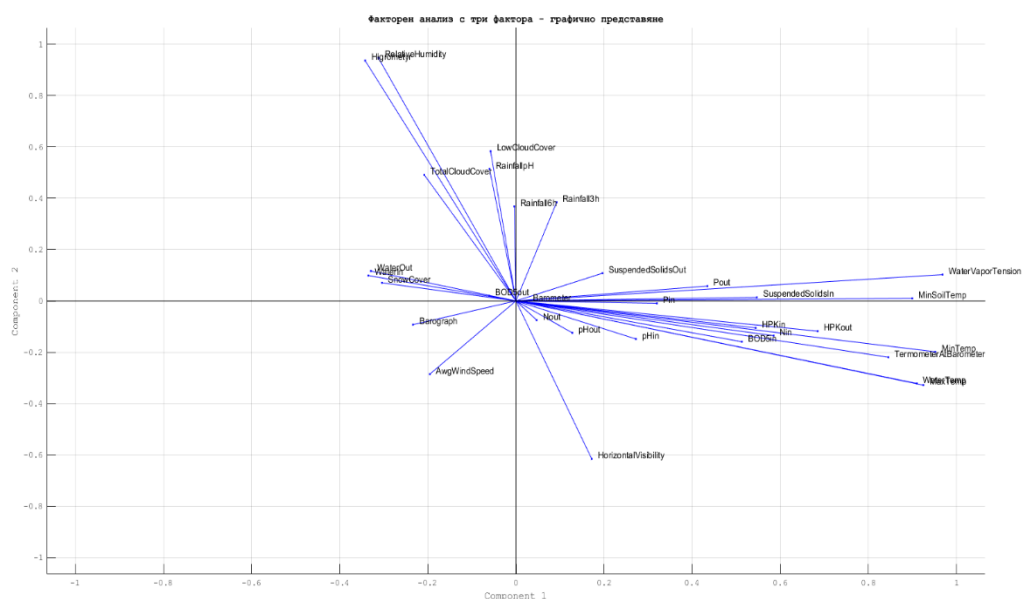
На база на резултатите от използваните критерии е избран трифакторен модел, който обяснява 55.02% от общата дисперсия в данните и осигурява баланс между информативност, интерпретируемост и структурна стабилност. Всеки от факторите показва ясно обособена семантична структура и логическа консистентност:

- **Фактор 1 /Температурно-химичен/:** Характеризира се с високи положителни факторни товари за температурните показатели и основните химични параметри на водата. Тази конфигурация потвърждава, че факторът обединява температурните условия (на въздуха и водата) и концентрациите на вещества като БПК₅, ХПК, азот и фосфор, измерени на входа и изхода на ПСОВ.

- **Фактор 2 /Атмосферно-метеорологичен/:** Включва високи товари за атмосферните показатели като относителна влажност, облачност и валежи и хигрометрични измервания. Това потвърждава, че факторът отразява метеорологичната динамика в разглеждания регион.

- **Фактор 3 /Хидравличен/:** Групира променливи, свързани с обема и дебита на отпадните води, което предполага влияние на хидравличните процеси върху функционирането на пречиствателната система.

Факторните товари на отделните променливи в рамките на трите фактора са визуализирани на Фиг. 6. Пълната таблица с числовите стойности е включена в основния текст дисертацията, като в автореферата са представени само най-характерните и интерпретируеми стойности (Табл. 2).



Фигура 6. Графично представяне на факторния анализ с три фактора

Таблица 2. Основни променливи с най-високи факторни товари в модела с три фактора

Показател	Фактор 1	Фактор 2	Фактор 3
Температура на постъпващата вода в ПСОВ	0,9098	-0,3217	-0,0966
Обем на постъпилата вода в ПСОВ	-0,3350	0,0987	0,9342
Общ азот на входа на ПСОВ	0,5846	-0,1354	-0,3429
БПК5 на входа на ПСОВ	0,5125	-0,1590	0,1862
ХПК на входа на ПСОВ	0,5439	-0,1052	0,0506
Неразтворени вещества на входа на ПСОВ	0,5466	0,0138	0,1428
Обем на отпадна вода на изход на ПСОВ	-0,3292	0,1164	0,9294
ХПК на изхода на ПСОВ	0,6848	-0,1179	-0,0507
Максимална температура на въздуха	0,9244	-0,3281	-0,0932
Минимална температура на въздуха	0,9515	-0,1982	-0,1046
Пъргавина на водната пара	0,9680	0,1023	-0,0949
Относителна влажност	-0,3122	0,9470	-0,0049
Минимална темп. на повърхността на почвата	0,8992	0,0097	-0,1479
Ниска облачност	-0,0575	0,5833	0,0284
Шестчасов валеж	0,0923	0,3840	0,0778

Факторният модел не само позволява редукция на размерността на наблюдаваните променливи, но също така поставя теоретичната основа за формулиране и емпирично тестване на изследователски хипотези, разгледани в следващата част на дисертационния труд. Идентифицираните латентни структури - температурно-химичен, атмосферно-метеорологичен и хидравличен фактор - обясняват над 55% от вариацията

в обединения масив и притежават ясно интерпретируемо съдържание. Резултатите от анализа потвърждават, че динамиката на пречиствателните процеси се формира под въздействието както на вътрешни технологични параметри, така и на външни климатични влияния. Извлечените фактори ще бъдат използвани за проверка на устойчивостта на установените зависимости и за оценка на възможностите за опростено моделиране на системата.

II.4.4. Формулиране и проверка на статистически хипотези на база резултатите от факторния анализ

II.4.4.1. Хипотеза 1: Влияние на Температурно-химичния фактор (Φ_1) върху процесите в ПСОВ

- **Нулева хипотеза (H_0):** Температурно-химичният фактор (Φ_1) има най-голямо влияние върху процесите на пречистване на отпадни води в ПСОВ.

- **Алтернативна хипотеза (H_1):** Температурно-химичният фактор (Φ_1) няма значимо влияние върху процесите на пречистване на отпадни води в ПСОВ.

Фактор 1 (температурно-химичен) обединява температурни показатели, рН, ХПК, БПК₅, както и концентрации на общ азот и общ фосфор. Методологичният подход за проверка на хипотезата включва количествено сравнение на относителния дял от общата дисперсия, обяснявана от всеки от факторите, получени при прилагането на факторния анализ. Освен това е извършен корелационен анализ между факторните оценки по Φ_1 и стойностите на съответните химични променливи, за да се установи степента на връзка между тях.

Резултатите от анализа показват, че Φ_1 обяснява 55.02% от общата дисперсия, което го прави най-съществения от трите извлечени фактора. За сравнение, Φ_2 (атмосферно-метеорологичен) обяснява 26.87%, а Φ_3 (хидравличен) – 18.11%. Тези стойности са представени в Табл. 3.

Таблица 3. Процент на обяснена дисперсия от всеки фактор

Фактор	Обяснена дисперсия	Обяснена дисперсия [%]	Кумулативна дисперсия
Фактор 1	7,8505	55,0171	55,0171
Фактор 2	3,8347	26,8735	81,8905
Фактор 3	2,5841	18,1095	100,0000

Корелационният анализ потвърждава очакваната логика - температурно-химичният фактор показва най-силна връзка с показатели, които характеризират както степента на замърсяване, така и ефективността на пречиствателния процес.

Най-висока положителна корелация с Φ_1 е отчетена при ХПК на изхода на ПСОВ ($r = 0.7632$), което е обосновано с факта, че този показател отразява както входното

органично натоварване, така и степента на неговото разграждане. Високи стойности на корелация се наблюдават и при Общ азот вход ($r = 0.7021$), ХПК вход ($r = 0.6688$) и БПК5 вход ($r = 0.6139$), което показва, че $\Phi 1$ улавя предимно термохимичната натовареност на входа.

По-слаби връзки са отчетени при изходните стойности на Общ фосфор ($r = 0.4529$), рН ($r = 0.1391$), Общ азот ($r = 0.1206$) и особено БПК5 ($r = -0.0241$). Тези променливи са в по-голяма степен повлияни от вътрешни биологични и технологични процеси в системата, отколкото от първоначалните условия. Резултатите потвърждават, че температурно-химичният фактор оказва съществено влияние върху динамиката на пречиствателните процеси и може да бъде използван като обобщена латентна променлива в бъдещи модели за оценка и прогнозиране на ефективността.

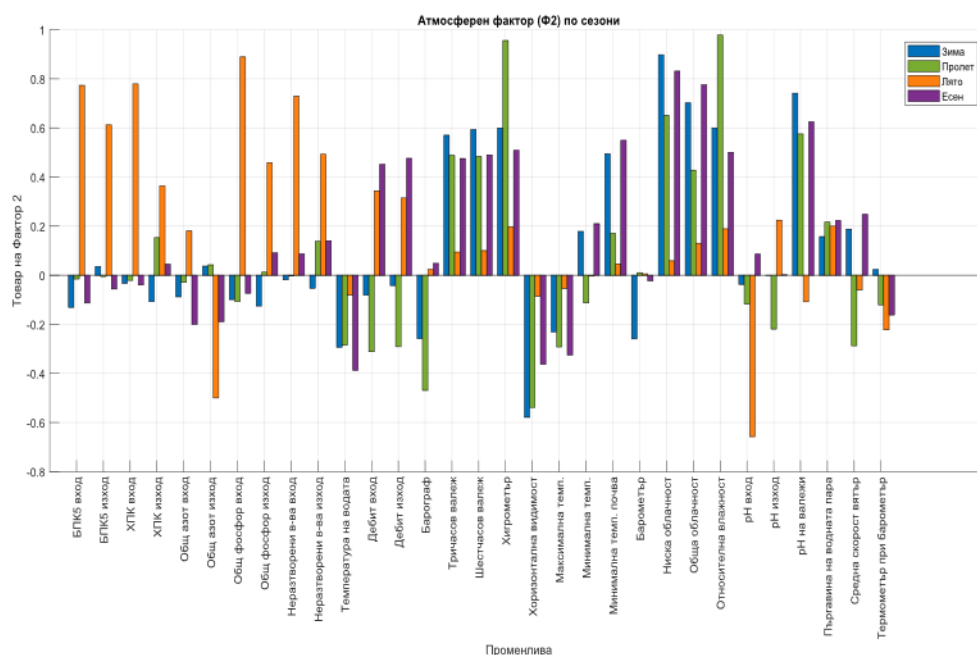
Извод: Потвърждава се H_0 . Температурно-химичният фактор ($\Phi 1$) е основен компонент, който интегрира ключовите вътрешни параметри на ПСОВ. Той оказва най-съществено влияние върху качеството на отпадната вода и притежава потенциал за използване при бъдещо моделиране и прогнозиране на ефективността на пречиствателните процеси.

II.4.4.2. Хипотеза 2: Анализ на влиянието на Атмосферния фактор ($\Phi 2$) върху ПСОВ през преходните сезони

- **Нулева хипотеза (H_0):** Атмосферният фактор ($\Phi 2$) оказва най-силно влияние върху процесите в ПСОВ през преходните сезони — пролет и есен.
- **Алтернативна хипотеза (H_1):** Атмосферният фактор ($\Phi 2$) няма значимо влияние върху процесите в ПСОВ през преходните сезони.

Целта на Хипотеза 2 е да се провери дали факторът, който обединява атмосферно-метеорологичните променливи ($\Phi 2$), има сезонно диференцирано въздействие и дали неговото влияние се усилва именно през преходните периоди на годината. Предположението е, че повишената нестабилност на атмосферните условия през пролетта и есента - температурни флуктуации, валежи, променлива облачност и влажност - оказва по-съществено въздействие върху динамиката на пречиствателните процеси.

За проверка на хипотезата е извършен факторен анализ по сезони, при който пълният масив от данни е разделен на четири сезонни подмножества - зима, пролет, лято и есен. За всяко множество е изчислен отделен факторен модел, като са проследени факторните оценки за $\Phi 2$ и съответните факторни товари на свързаните атмосферни променливи, които отразяват интензивността на атмосферния компонент през съответния сезон (Фиг. 7).



Фигура 7. Сезонна вариация на факторните товари по Фактор 2 (атмосферен фактор)

Резултатите показват, че основните показатели, формиращи Ф2 - относителна влажност, ниска облачност, валежи, атмосферно налягане - проявяват най-високи факторни товари именно през пролетта и есента.

Наблюдаваната сезонна диференциация е от съществено значение за управлението на ПСОВ, тъй като подчертава чувствителността на системата към климатични въздействия в условия на повишена атмосферна нестабилност. Това потвърждава необходимостта от прилагане на адаптивни модели, които да отчитат сезонната специфика при прогнозиране и контрол на пречиствателните процеси.

Извод: **Потвърждава се H_0** - атмосферният фактор (Ф2) оказва най-силно влияние върху пречиствателните процеси именно през преходните сезони. Това наблюдение има пряко приложение в оптимизацията на мониторинговите стратегии и разработването на адаптивни модели за управление на системи с висока зависимост от външни климатични условия.

II.4.4.3. Хипотеза 3: Анализ на въздействието на Хидравличния фактор (Ф3) върху концентрациите на азот и фосфор

- **Нулева хипотеза (H_0):** Промените в хидравличния фактор (Ф3) оказват значително влияние върху ефективността на биологичното пречистване и концентрациите на азот и фосфор в отпадната вода в ПСОВ.

- **Алтернативна хипотеза (H_1):** Промените в хидравличния фактор (Ф3) не оказват значимо влияние върху ефективността на биологичното пречистване и концентрациите на азот и фосфор.

Целта на тази хипотеза е да се установи дали Ф3, интерпретиран като хидравличен компонент, има статистически значима връзка с концентрациите на общ азот (N) и общ фосфор (P) както на входа, така и на изхода на ПСОВ. Предполагането е, че променливостта на дебита, температурата на водата и други свързани параметри оказват влияние върху ефективността на отстраняване на азот и фосфор.

За проверка на хипотезата са използвани следните подходи:

- **корелационен анализ** между факторните оценки по Ф3 и стойностите на общ азот и общ фосфор на входа и изхода. Резултатите са показани в Табл. 4:

Таблица 4. Корелации между Фактор 3 и концентрациите на азот и фосфор

Показател	Корелация с Ф3
Общ азот на входа на ПСОВ	-0,6422
Общ азот на изхода на ПСОВ	-0,7230
Общ фосфор на входа на ПСОВ	0,4243
Общ фосфор на изхода на ПСОВ	0,1438

Най-силна е отрицателната зависимост между Ф3 и концентрацията на общ азот на изхода ($r = -0.7230$), което показва, че повишеното хидравлично натоварване се свързва с по-ниска ефективност при отстраняване на азот. Подобна тенденция се наблюдава и при входния азот. При фосфора обаче корелациите са по-слаби – особено на изхода, където влиянието на Ф3 не е статистически убедително.

- **регресионен анализ**, който оценява предсказващата способност на Ф3 спрямо същите показатели. Резултатите са обобщени в Табл. 5:

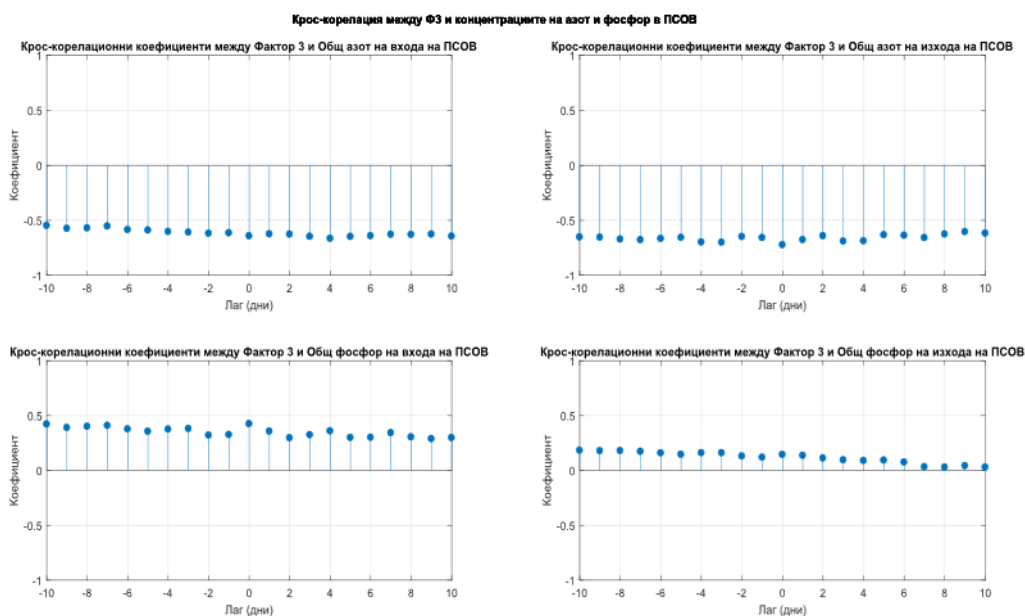
Таблица 5. Линейни регресионни модели на химичните показатели по Ф3

Показател	Наклон по Ф3	Стандартна грешка	t - стойност	p – стойност	R^2	RMSE	Заклучение
Общ азот на входа на ПСОВ	-1,1228	0,0703	-15,9620	7,9332e-44	0,5230	4,1800	Силна, значима връзка
Общ азот на изхода на ПСОВ	-0,6029	0,0302	-19,9410	2.8317e-60	0,4120	1,8000	Много силна връзка
Общ фосфор на входа на ПСОВ	0,7979	0,0894	8,9265	2.2183e-17	0,1800	5,3200	Умерена връзка
Общ фосфор на изхода на ПСОВ	0,0544	0,0196	2,7699	0,0059	0,0207	1,1700	Много слаба връзка

Най-високата обяснена дисперсия се наблюдава при модела „Общ азот вход - ФЗ“ ($R^2 = 0.5230$), което означава, че над 52% от колебанията на общия азот на входа могат да бъдат обяснени чрез ФЗ. Моделът „Общ азот изход - ФЗ“ също показва висока степен на обяснена дисперсия ($R^2 = 0.4120$). При „Общ фосфор изход - ФЗ“, въпреки че регресионната връзка е статистически значима ($p = 0.0059$), стойността на $R^2 = 0.0207$ показва, че моделът обяснява едва 2,1% от вариацията. Всички модели са статистически значими ($p < 0.05$), но силата на връзката при отделните показатели е различна.

- **крос-корелационен** анализ за проследяване на наличието на времеви лагове между ФЗ и стойностите на азот и фосфор (Фиг. 8).

Крос-корелационният анализ допълва резултатите, като показва максимум при нулев лаг за общия азот на изхода и лаг от +4 дни за азота на входа. Това предполага, че хидравличният ефект върху изходящия азот се проявява почти мигновено, докато при входящия - със закъснение. По отношение на общия фосфор, връзките с ФЗ са по-слабо изразени. Тези резултати предполагат, че влиянието на хидравличния фактор върху фосфора е по-ограничено и вероятно се модерира от допълнителни химико-биологични процеси в системата.



Фигура 8. Крос-корелационни зависимости между Фактор 3 и концентрациите на азот и фосфор в ПСОВ

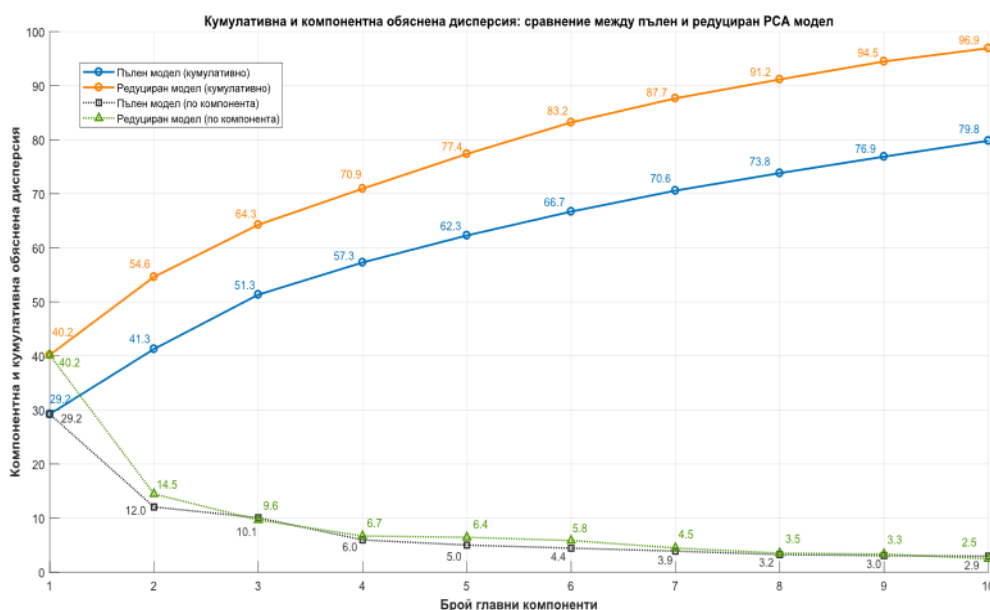
Обобщение: Хипотезата се **потвърждава по отношение на азота** - факторът, интерпретиран като хидравличен компонент, има ясно изразено и статистически значимо влияние върху концентрациите на общ азот както на входа, така и на изхода на ПСОВ. **По отношение на фосфора** обаче установените зависимости са по-слаби и неубедителни, поради което **не може категорично да се потвърди значимо влияние на хидравличния фактор** върху промените в концентрациите на фосфор в системата.

II.4.4.4. Хипотеза 4: Влияние на редуцията на броя на наблюдаваните променливи върху достоверността на анализа и информационната цялост

- **Нулева хипотеза (H_0):** Намаляването на броя на наблюдаваните променливи не води до съществена загуба на информация в анализа.
- **Алтернативна хипотеза (H_1):** Намаляването на броя на наблюдаваните променливи води до съществена загуба на информация.

В рамките на четвъртата хипотеза е изследвана възможността чрез редуциране на броя на наблюдаваните променливи да се постигне по-компактен и интерпретируем модел, без това да води до загуба на обяснителна способност. Основната идея се основава на резултатите от факторния анализ, при който са идентифицирани три ясно разграничими латентни фактора - температурно-химичен, атмосферно-метеорологичен и хидравличен. Въз основа на най-високите факторни товари е формиран редуциран масив, съдържащ 15 променливи - по пет за всеки от факторите.

Сравнителният анализ между пълния набор от 32 променливи и редуцирания поднабор е извършен чрез прилагане на метода на главните компоненти (PCA), като резултатите са представени графично на Фиг. 9. В нея са изобразени както компонентната, така и кумулативната обяснена дисперсия по компоненти за двата модела. Графиката показва, че при редуцирания масив от 15 променливи се обяснява по-висок дял от общата вариация още в първите компоненти - например, до петата компонента се достигат 77.36% обяснена дисперсия, спрямо 62.27% за пълния модел с 32 променливи. От шестата компонента нататък при редуцирания модел е под 1%, което потвърждава, че информационното съдържание е успешно кондензирано. Това демонстрира, че чрез подходяща селекция на показатели, базирана на факторните товари, може да се изгради по-компактен и по-интерпретируем аналитичен модел без загуба на съществена информация.



Фигура 9. Обяснена дисперсия от първите главни компоненти

Извод: **Получените резултати потвърждават H_0 .** Намалването на броя на наблюдаваните променливи - чрез елиминиране на показатели с ниски факторни товари и дублираща информация - не води до загуба на значима информация. Напротив, редуцираният модел постига по-висока обяснителна способност с по-малко компоненти и значително по-ясна интерпретируемост. Това доказва, че внимателната селекция на променливи въз основа на факторни зависимости е ефективен аналитичен инструмент за оптимизация на измерването, съхранението и анализа на данни в системи с висока параметрична плътност.

III. ОБОБЩЕНИ ИЗВОДИ И ЗАКЛЮЧЕНИЕ

В настоящия дисертационен труд е разработена цялостна аналитична рамка за обработка и анализ на хетерогенни данни, постъпващи от ПСОВ и МС в гр. Кърджали. Изследването доказва, че чрез прилагане на съвременни количествени методи могат да се идентифицират значими зависимости между екологични и атмосферни променливи, които са от съществено значение за ефективното управление на водните ресурси.

Основната цел на дисертацията - създаване на интегрирана методология за синхронизиране, обработка и анализ на реални оперативни данни - е постигната чрез прилагане на цялостен последователен аналитичен подход. Той включва: дигитализация на хартиени дневници, преобразуване и осредняване на метеорологични данни до дневна резолюция, запълване на липсващи стойности в технологичните редове от ПСОВ чрез сравнителен анализ на три импутационни метода (AR, PCHIP, SPLINE), както и изграждане на обединен масив чрез уеднаквяване и синхронизация на данните от двата източника по време и структура.

Върху получения обединен масив от 59 променливи са приложени корелационен и крос-корелационен анализ за идентифициране на моментни и отложени зависимости между атмосферни и технологични параметри. Чрез тях са елиминирани силно корелирани и несъотнесими променливи от метеорологичния блок. Последващият факторен анализ позволява редуциране на размерността чрез извличане на латентни компоненти и изграждане на интерпретируем модел, устойчив на структурна излишност. В резултат са обособени три основни фактора: температурно-химичен (Φ_1), атмосферно-метеорологичен (Φ_2) и хидравличен (Φ_3), които отразяват водещите влияещи механизми в системата.

На базата на извлечената факторна структура са формулирани четири изследователски хипотези. Три от тях проверяват устойчивостта и интерпретируемостта на факторите - чрез оценка на влиянието на Φ_1 върху качеството на водата, сезонната динамика на Φ_2 и ефекта на Φ_3 върху концентрациите на азот и фосфор. Четвъртата хипотеза изследва възможността за редукция на броя на наблюдаваните променливи чрез подбор на тези с най-високи факторни товари - с резултат, който надхвърля първоначалните очаквания: предложеният поднабор от 15 променливи обяснява над 77% от общата вариация в данните, като същевременно предлага по-компактен и интерпретируем модел.

Предложената методология доказва своята приложимост в условия на непълни, несъгласувани и полуструктурирани данни и може да бъде адаптирана към други индустриални и екологични сценарии. Това отваря възможности за интердисциплинарни изследвания, основани на реални наблюдения и устойчиви количествени оценки.

В заключение, дисертационният труд обогатява съвременните методологични решения за обработка и анализ на Big Data в екологичен контекст и предлага практически приложима рамка за мониторинг, прогнозиране и управление на пречиствателни процеси при динамично променящи се климатични условия. Получените резултати могат да служат като основа за изграждане на системи за ранно предупреждение и оптимизация на оперативния контрол в управлението на водни ресурси.

IV. ОГРАНИЧЕНИЯ НА ИЗСЛЕДВАНЕТО

1. *Локализираност на данните* - Изследването се базира на данни от една конкретна пречиствателна станция и метеорологична станция в гр. Кърджали. Въпреки че резултатите са представителни за анализирания период и условия, обобщаването им към други географски региони или климатични зони изисква допълнителни изследвания.

2. *Ограничен времеви обхват* - Данните обхващат конкретен двугодишен период (2021-2022 г.), но не включват екстремни събития като наводнения, продължителни суши или технологични аварии. Такива събития биха могли да окажат съществено влияние върху функционирането на системата и следва да бъдат предмет на отделен анализ.

3. *Тип на пречиствателната станция* - Изследваната ПСОВ е изградена като проточна биологична система, при която динамиката на процесите следва определена хидравлична логика и технологичен режим. Резултатите и изводите от анализа са тясно свързани с този тип система. При прилагане към станции с различна конфигурация (напр. периодични, химични или мембранни) се изисква преоценка на приложимостта на използваните модели.

4. *Липса на биологични показатели* - Данните от ПСОВ не включват микробиологични или токсикологични показатели. Включването на такива параметри би позволило по-пълна оценка на екосистемното въздействие и качеството на пречистването.

5. *Ограничения на честотата на измерване* - Първичните данни от НИМХ са с висока времева резолюция (осем измервания на ден), докато тези от ПСОВ са събирани с по-ниска и нерегулярна честота. Необходимото осредняване и синхронизиране води до известна загуба на информация за динамични процеси.

6. *Методологични компромиси при обработката на липсващи стойности* - Приетите допускания при запълване на липсващи стойности (вкл. третиране на липсващ

валеж като 0) са обосновани в контекста на системата, но могат да доведат до отклонения, ако се прилагат към други системи без допълнителна адаптация.

V. ЛИТЕРАТУРА, ПОСОЧЕНА В АВТОРЕФЕРАТА

1. Anastasova, Y. (2019). Technologies for ensuring data quality and security in industrial information systems. Journal of Mining and Geological Sciences, 62(4), 5-8. Available on: <https://mgu.bg/wp-content/uploads/2025/02/Journal-4-HUM-2019-5-8.pdf>
2. Dębski, R. (2020). Real-time interpolation of streaming data. Computer Science, 21(4). <https://doi.org/10.7494/csci.2020.21.4.3932>
3. Dimitrov, A., **Ivanova, N.**, Nesheva, R., & Yanev, N. (2021). Preparation of industrial data for implementation in big data. Annual of University of Mining and Geology "St. Ivan Rilski", 64, 151-154.
4. Fadlallah, H., Kilany, R., Dhayne, H., Haddad, R. E., Haque, R., Taher, Y., & Jaber, A. (2023). Context-aware Big Data Quality Assessment: A scoping review. Journal of Data and Information Quality, 15(3), 1–33. <https://doi.org/10.1145/3603707>
5. **Ivanova, N.** (2023). Classification of basic methods for data analysis in industry. Annual of university of mining and geology "St. Ivan Rilski", 66, 228-232. <https://doi.org/10.5281/zenodo.8335379>
6. **Ivanova, N.**, Yanev, N., & Vasileva, S. (2024). Utilization of Basic Methods for Filling Missing Data from Wastewater Treatment Plants. 5th International Conference on Communications, Information, Electronic and Energy Systems (CIEES), 1-6. IEEE. doi: 10.1109/CIEES62939.2024.10811430
7. Shrestha, N. (2021). Factor Analysis as a Tool for Survey Analysis. American Journal of Applied Mathematics and Statistics, 9(1), 4-11. DOI: 10.12691/ajams-9-1-2

VI. ПРИНОСИ НА ДИСЕРТАЦИОННИЯ ТРУД

Научни приноси

1. Предложени са допълнителни критерии за усъвършенстване на интегрираната методология за унифициране на хетерогенни екологични и метеорологични данни, включваща синхронизация на измерванията по дати, стандартизация на променливите и запълване на липсващи стойности чрез специализирани импутационни методи. Методологията е приложена успешно върху реални оперативни данни от ПСОВ и метеорологична станция и може да бъде адаптирана към други типове хетерогенни данни.
2. Адаптирани са статистически подходи за извършване на факторен анализ в контекста на екологични системи, чрез използване на метода на главните компоненти

(PCA) и Varimax ротация. Извлечени са стабилни латентни фактори с ясно интерпретируемо съдържание – температурно-химичен, атмосферно-метеорологичен и хидравличен, което разкрива дълбочинно залегналата структура в сложни екологични данни.

3. **Разработена е експериментална процедура за сравнителна оценка на методи за запълване на липсващи стойности**, основана на зануляване на избрани интервали с известни стойности от данните и последващо възстановяване чрез PCNIP, SPLINE и авторегресивен модел. Точността на методите е оценена с метриците MAE, RMSE и MAPE, което осигурява обективен критерий за избор на подход при импутация.

Научноприложни приноси

4. **Валидирана е хипотеза за съществуването на отложени (лагови) зависимости между атмосферни и водни показатели**, чрез прилагане на крос-корелационен анализ. Установени са значими зависимости със закъснение от няколко дни, което обогатява разбирането на динамиката в пречиствателните системи и има потенциал за разработване на прогнозни модели.

5. **Създадена е аналитична база от зависимости между климатични и технологични променливи**, която позволява по-задълбочено моделиране на поведението на пречиствателната система при различни атмосферни условия и оперативни натоварвания.

6. **Предложен е подход за оптимизация на мониторинговата мрежа в ПСОВ**, чрез обобщение на факторни товари и анализ на обяснена дисперсия. Извлечената редукция е съставена от 15 променливи, които запазват 77,36% от общата обяснена дисперсия. Това позволява значително намаляване на броя на наблюдаваните показатели без съществена загуба на информация и създава предпоставки за оптимизация на измервателните мрежи в ПСОВ.

7. **Поставена е основа за прилагане на прогнозни модели**, базирани на факторни оценки и атмосферни данни, което би подпомогнало оперативното управление и би повишило ефективността на пречиствателните процеси чрез по-информирано вземане на решения.

8. **Предложени са оригинални методи за обработка и анализ на Big Data в контекста на ПСОВ**, чрез съчетаване на техники от статистическия анализ и машинното обучение, което разширява съществуващите изследвания и дава основа за разработване на практически приложими алгоритми за анализ на реални данни от непрекъснат мониторинг.

9. **Систематизирана и апробирана е методология за интегриран анализ на екологични и технологични данни**, която обединява процесите на унифициране, редукция на размерността чрез факторен анализ и проверка на хипотези. Тази рамка е

приложима за широк спектър от изследвания и демонстрира как съчетаването на класически статистически техники с Big Data подходи може да доведе до значими интердисциплинарни резултати.

VII. СПИСЪК НА ПУБЛИКАЦИИТЕ ПО ДИСЕРТАЦИОННИЯ ТРУД

1. Dimitrov, A., **Ivanova, N.**, Nesheva, R., & Yanev, N. (2021). Preparation of industrial data for implementation in big data. Annual of University of Mining and Geology "St. Ivan Rilski", 64, 151-154.
2. **Ivanova, N.**, & Nesheva, R. (2022). Software products for risk assessment and management of occupational accidents and diseases in industry, Annual of University of Mining and Geology "St. Ivan Rilski", 65, 131-134.
3. **Ivanova, N.** (2023). Classification of basic methods for data analysis in industry. Annual of University of Mining and Geology "St. Ivan Rilski", 66, 228-232. <https://doi.org/10.5281/zenodo.8335379>
4. **Ivanova, N.**, Yanev, N., & Vasileva, S. (2024). Utilization of Basic Methods for Filling Missing Data from Wastewater Treatment Plants. 5th International Conference on Communications, Information, Electronic and Energy Systems (CIEES), 1-6. IEEE. doi: 10.1109/CIEES62939.2024.10811430

SUMMARY

Dissertation Title: "Exploration of Approaches for the Unification of Big Data from Diverse Environmental Sources for Enhanced Analysis"

Author: Nikolina Yordanova Ivanova

This dissertation develops a comprehensive analytical framework for the unification and processing of heterogeneous environmental data collected from a wastewater treatment plant (WWTP) and a meteorological station in the town of Kardzhali. The study addresses practical and methodological challenges related to missing data, inconsistent formats, and different measurement frequencies, aiming to improve the applicability of such data in analytical modelling.

The research begins with the digitization of handwritten records and the transformation of meteorological data to daily resolution. Missing values in WWTP data are imputed using and comparing three techniques: Piecewise Cubic Hermite Interpolation (PCHIP), cubic spline interpolation (SPLINE), and an autoregressive model (AR). The AR method is selected for further use based on a quantitative comparison using MAE, MAPE, and RMSE.

A consolidated dataset of 59 variables is constructed by synchronizing the two sources. Correlation and cross-correlation analysis are applied to identify instantaneous and lagged relationships, leading to the reduction of the dataset to 32 variables by eliminating only meteorological variables that were highly correlated or irrelevant. A factor analysis is then conducted to reduce dimensionality and extract latent components. Three interpretable factors are identified: temperature-chemical, atmospheric-meteorological, and hydraulic.

Four hypotheses are formulated to assess the structure and utility of the resulting analytical model. Three of them evaluate the stability and explanatory power of the factors across different seasons and parameters. The fourth hypothesis investigates the possibility of reducing the number of monitored variables. The findings confirm that a subset of 15 selected indicators retains over 77% of the total explained variance, offering a more compact and interpretable model.

The proposed methodology is applicable in real-world conditions with incomplete, inconsistent, and semi-structured data, and can be extended to other environmental or industrial scenarios. The results contribute to the methodological basis for Big Data processing and have practical implications for monitoring, forecasting, and decision-making in wastewater management under variable climatic conditions.