Пловдивски университет „Паисий Хилендарски“ Факултет
по математика и информатика
Катедра „Приложна математика и моделиране“

Атанас Вълев Иванов

Статистическо моделиране на качеството на въздуха

Автореферат

за присъждане на образователна и научна степен “Доктор” в
област на висше образование: 4. Природни науки, математика и информатика
Професионално направление: 4.5. Математика
Докторска програма: Математическо моделиране и приложение на математиката

Научни ръководители:

проф. д-р Снежана Георгиева Гочева-Илиева

доц. д-р Дойчин Тодоров Бояджиев

Пловдив

2015

Дисертационният труд е обсъден и насрочен за защита на разширен катедрен съвет на катедра „Приложна математика и моделиране” при Факултет по математика и информатика на Пловдивски университет „Паисий Хилендарски”, град Пловдив, проведен на 12.01.2015 г.

Дисертационният труд е с общ обем от 134 страници и включва увод, четири глави, приложения и библиография, състояща се от 115 източника. Списъкът на авторските публикации включва 4 заглавия, от които 1 на български език и 3 на английски език.

Защитата на дисертационния труд ще се състои на ..... г. в Заседателна зала на Нова сграда на Пловдивски университет „Паисий Хилендарски” на открито заседание на научно.

Номерацията на формулите, цитиранията, примерите, таблиците и фигурите съвпада с тяхната номерация в дисертационния труд.

Материалите по защитата са на разположение на интересуващите се в секратариат на ФМИ, нова сграда на ПУ, каб. 330, всеки ден от 8:30 до 17:00 часа.

Автор на дисертациония труд: Атанас Вълев Иванов

Заглавие: Статистическо моделиране на качеството на въздуха

Съдържание

Обща характеристика на дисертационния труд 4

Глава 1. Използвани методи за моделиране и изследване на качеството на въздуха 6

Трансформации на начални данни 6

Факторен анализ 6

Стохастични ARIMA и SARIMA методи 6

Моделиране с обобщена PathSeeker (GPS) регуляризираща регресия 7

Глава 2.1. Времеви анализ на качеството на въздуха в град Кърджали 9

Описателна статистика на данните за град Кърджали 9

Построяване на SARIMA модели и анализ 10

Глава 2.2. Времеви анализ на качеството на въздуха в град Шумен 10

Описание на данните за град Шумен 10

Групиране на замърсителите чрез факторен анализ 13

АРИМА модели с отчитане на метеорологичните редове 14

Глава 3. ВРЕМЕВИ АНАЛИЗ НА ЕЖЕДНЕВНИТЕ ЗАМЪРСИТЕЛИ НА ВЪЗДУХА В ГРАД БЛАГОЕВГРАД 15

Използвани данни 15

Трансформация на данните 15

Факторен анализ 17

Построяване и диагностика на SARIMA моделите за трансформираните променливи 18

Глава 4. ПРИЛОЖЕНИЕ НА SARIMA И GPS РЕГРЕСИЯ ЗА ИЗСЛЕДВАНЕ ЗАМЪРСЯВАНЕ НА ВЪЗДУХА С PM10 НА ГРАД ШУМЕН 21

Използвани данни 21

Трансформация на данните 22

Резултати от SARIMA модела 22

Сравнение на прогнозните резултати от SARIMA и GPS-RuleLearner моделите 24

Библиография 27

Обща характеристика на дисертационния труд

Актуалност на проблемите

В днешно време качеството на атмосферния въздух и проблемите, свързани с това, са от все по-голямо значение за човешкото здраве и цялостното състояние околната среда.

В България систематично се наблюдават 12 типа атмосферни замърсители с помощта на повече от 36 автоматизирани станции на Изпълнителната агенция за околна среда, която управлява и координира дейностите, свързани с контрола и опазването на околната среда в страната. Следят се и препоръчителните нива на замърсяване според европейските и националните норми [30, 31, 32, 71]. Съответните данни се препращат в реално време към Европейската агенция за опазване на околната среда. На местно ниво редовно се изготвят доклади за качеството на атмосферния въздух за отделните райони на страната.

Констатират се редица систематично превишения на замърсяванията на въздуха за населените места в почти цяла България [3, 5, 6, 30, 31, 32]. Най-сериозно са замърсени населените места в България с фини прахови частици (PM10). Сред най-проблемните градове са: Перник (най-замърсен за Европа през 2011 г.), Пловдив, София, Благоевград и др. Значителни превишавания на нивата от серен диоксид (SO2) и азотни окиси (NO2, NO, NOx) се наблюдават в Димитровград, Стара Загора и др.

Проблеми на математическото моделиране на замърсяването на въздуха

В световната литература съществува огромен брой изследвания, свързани със замърсяването на въздуха във всичките му аспекти и са представени резултати от емпирични модели. Сред нерешените проблеми в областта на моделирането на качеството на въздуха в България са: 1) Обработка на наличните реални данни и моделиране на нивата на концентрациите на основни замърсители, и в частност тези замърсители, които са проблемни в отделни градове и области; 2) Анализиране на данните и моделите за извличане на съществена информация от тях за индентифициране на взаимовръзки между отделни замърсители, и зависимости с отчитане на метеорологичните, екологични и други фактори; 3) Приложение на методите на математическото моделиране за прогнозиране с цел превенция от замърсване на въздуха в България; 4) Интерпретация на моделите и получените резултати.

Цели и задачи на дисертациония труд

Обект на изследване: Основни замърсители на чистотата на въздуха в населени места в България.

Хипотеза: Натрупаните големи масиви от експериментални данни съдържат съществена информация и скрити взаимовръзки, чието разкриване е възможно с помощта на статистическо моделиране и подходящи методи.

Основна цел: Създаване, изследване и приложение на статистически модели на основни замърсители на въздуха в населени места в България за идентифициране на комплексни замърсявания от общи източници, откриване на зависимости и получаване на краткосрочни прогнози за превенция и поддържане на качеството на въздуха.

Задачи:

1) С помощта на едномерен стохастичен сезонен ARIMA (SARIMA) метод да се построят и изследват модели на замърсяванията на въздуха с PM10 и SO2 на град Кърджали и да се получат краткосрочни прогнози.

2) Да се проведе времеви анализ на 6 основни замърсители (NO2, NO, O3, PM10, SO2, NOx) на град Шумен чрез построяване и приложение на стохастични ARIMA модели с отчитане на 6 метеорологични фактори.
3) За данните от 6-те замърсителя на град Шумен да се намери групиране чрез многомерен факторен анализ и да се приложат получените факторни времеви редове за построяване на ARIMA модели и получаване на краткосрочни прогнози.
4) Да се намери групиране на 6-те основни замърсители на въздуха (NO2, NO, O3, PM10, SO2, NOx) на град Благоевград чрез многомерен факторен анализ и да се интерпретират резултатите.
5) Да се построят стохастични SARIMA модели за изследване на динамиката и предсказване на замърсителите на въздуха на град Благоевград, и да се проведе детайлен анализ, и подбор на моделите.
6) Да се проведе статистически анализ на времевия ред за замърсяване на въздуха с PM10 на град Шумен с отчитане на 6-те метеорологични фактори, чрез построяване, изследване и сравнение на стохастични SARIMA модели и модели на базата на обобщена Pathseeker (GPS) регуляризираща регресия и предсказващи техники. Да се приложат моделите за краткосрочни прогнози.

Структура и обем на дисертациония труд

Дисертационният труд е структуриран в увод, 4 глави, заключение и литература с общ обем от 134 печатни страници.

Глава 1 няма приносен характер. Тя съдържа преглед на състоянието на проблема, описание на използваните в дисертацията математически методи и техники и формулира целите и задачите на изследванията.

В глави 2 са временните редове за замърсявания с прахови частици (PM10) и серен диоксид (SO2) в зависимост от времето се изследват чрез SARIMA модели. В същата глава е Проведен е анализ на временни редове и предсказване на замърсяването на въздуха за 6 замърсителя (азотен диоксид(NO2), азотен оксид(NO), озон(O3), финни прахови частици (PM10), серен диоксид(SO2), азотни оксиди(NOx) с факторен анализ и ARIMA с използване на метеорологични параметри.

В глава 3 се изследват временни редове за замърсяването на въздуха на 6-те въздушни замърсителя в град Благоевград с приложение на факторен анализ и SARIMA метод.

В глава 4 се изследват данни за концентрациите на фини прахови частици PM10 за град Шумен. Концентрациите превишават многократно допустимите норми. Целта е да се построят съответни емпирични статистически модели с отчитане на метеорологичните параметри на базата на два различни подхода - стохастичен SARIMA и General Pathseeker регресия, да се сравни способността им да предсказват реалните нива на замърсяване с PM10, както и приложимостта им за краткосрочно прогнозиране.

В заключението е представено кратко резюме, списък на публикациите по дисертационния труд и апробации, и са систематизирани претенциите за научните приноси на докторанта в областта на статистическото изследване на атмосферни замърсявания. Литературата включва 115 източника.

Дисертационният труд е разработен към катедра “Приложна математика и моделиране” на Факултета по математика и информатика на Пловдивския университет “Паисий Хилендарски” в периода 2011-2014 година.

Кратко съдържание на дисертациония труд

Глава 1. Използвани методи за моделиране и изследване на качеството на въздуха

Трансформации на начални данни

В дисертацията е използвана трансформацията на Йео-Джонсън, която обобщава известната степенна трансформация на [12], за случая на нулеви и отрицателни стойности.

Степенната трансформация на Йео-Джонсън, за преобразуване на дадена променлива *x* има вида [100]:

$$\psi_{YJ}(\lambda, x) = \begin{cases} \left\{(x+1)^{\lambda} - 1\right\} / \lambda & x \geq 0, \lambda \neq 0 \\ \log(x+1) & x \geq 0, \lambda = 0 \\ -\left\{(-x+1)^{2-\lambda} - 1\right\} / (2-\lambda) & x < 0, \lambda \neq 2 \\ -\log(-x+1) & x < 0, \lambda = 2 \end{cases}, \qquad \lambda \in [-2, 2] \tag{1.1}$$

За определяне на константата $\lambda \in [-2, 2]$ във формула (1.1) сме разработили съответен код на система Wolfram Mathematica [99, 107] за променлива *x*, при зададена стъпка на λ. При подбор на оптималната стойност на λ сравняваме оптималната стойност на показателя на Колмогоров-Смирнов за нормално разпределение на данните.

Факторен анализ

Методът се използва и за изясняване на подобността и разликите в многомерно множество от данни. Математическите основи на метода могат да бъдат намерени в редица книги и статии [51]. Наличието на високи коефициенти (над 0.5) в корелационната матрица за данните често показва наличието на мултиколинеарност (т.е. детерминантата е близка до 0). При провеждане на факторен анализ, силната корелация на две или повече променливи се представят чрез обща латентна (изкуствена) променлива, наречена фактор. По време на този процес се губи част от информацията, но се откриват взаимовръзки между групираните променливи. При факторния анализ се прилагат различни методи за извличане и трансформиране (ротация) на факторите. Друг ключов момент е определянето на броя фактори, които да заместят първоначалните независими променливи. Изборът на конкретен метод, трансформация и брой фактори зависи от изследователя.

Процедурата се състои от 5 основни стъпки:

(1) изчисляване на корелационната матрица

(2) тест за адекватност на факторния анализ

(3) извличане на факторите, избор на оптимален брой фактори

(4) въртене на факторите

(5) изчисляване и съхраняване на коефициентите на факторните променливи за провеждане на следващи анализи.

Стохастични ARIMA и SARIMA методи

Едномерният ARIMA (авторегресионен анализ, тенденция и плаващи средни) е название на обобщен клас от методи, въведени от Бокс и Дженкинс през 1976 ([13]). ARIMA намира широко приложение в екологичната статистика за оценка и предсказване поведението на времеви редове. Общ вид на ARIMA и SARIMA модел се представя с линейно или нелинейно диференчно

уравнение. Включва брой параметри (p,d,q), съответно (P,D,Q)(p,d,q)s, където: p – авторегресия (AR), P – сезонен AR, d – тенденция (I), D – сезонен I параметър, q– плаващи средни (MA), Q – сезонни MA. s – сезонност, цикличност.

Нека стойностите за времето да бъдат t = 1, 2, 3, ..., n, където n е общият брой наблюдения във времевия ред. С X_i е стойността на променлива от времевия ред (замърсител) в момент t = i, а L е операторът за забавяне, действащ по формулите:

$$LX_i = X_{i-1},\ L^2X_i = X_{i-2},\dots\ ,L^kX_i = X_{i-k},\ \ 0 \le i-k \le n. \tag{1.2}$$

Статичният времеви ред има фиксирана средна стойност и няма тенденция, т.е. *d* = 0. При по-сложни случаи би могла да съществува тенденция, която е представена под формата на полином (линеен, квадратен или от по-висока степен, до *d*). При наличие на ясно изразена сезонна компонента в едномерен временен ред, зависещ единствено от времето t, ARIMA се обобщава към SARIMA.

Едномерният мултипликативен SARIMA (p, d, q) (P, D, Q)s модел има следния общ вид:

$$\varphi_p(L)\Phi_P\left(L^s\right)(1-L)^d\left(1-L^s\right)^D Y_t = \theta_q(L)\Theta_Q\left(L^s\right)a_t + c, \quad a_t \sim WN(\sigma^2) \tag{1.7}$$

където φ_p, Φ_p са диференчни полиноми за несезонни и сезонни авторегресионни параметри, а θ_q, Θ_q са параметри на несезонни и сезонни плаващи средни, означени с изразите

$$\varphi_p(L) = \left(\sum_{j=0}^{p}\phi_j L^j\right),\ \Phi_P\left(L^s\right) = \left(\sum_{j=0}^{P}\Phi_j L^{j+s}\right),\ \theta_q(L) = \left(\sum_{j=0}^{q}\theta_j L^j\right),\ \ \Theta_Q\left(L^s\right) = \left(\sum_{j=0}^{Q}\Theta_j L^{j+s}\right)$$

a_t е белият шум в момента *t*, *c* е константа. Приема се, че временният ред е стационарен (плосък). Ако е нестационарен, той може да бъде трансформиран в такъв чрез диференчни разлики от някакъв ред *d*, което е означено с оператор $(1-L)$ в (1.7). Трябва да припомним, че стационарният процес има постоянна средна стойност и дисперсия. Необходимо е да отбележим, че параметрите на модела (1.7) отговарят на определени изисквания.

Когато временният ред Y се влияе от няколко входящи временни реда $X_1, X_2, \dots, X_r$, дефинирани за същите стойности на времето, SARIMA моделът се обобщава до модел с трансферни функции от диференчен вид [13, 14]

$$Y_t = \sum_{j=1}^{r}\frac{\omega_j(L)}{\delta_j(L)}L^{b_j}X_{jt} + \frac{\theta_a(L)}{\Phi_a(L)}a_t \tag{1.8}$$

където $\omega_j(L), \delta_j(L), \theta_a(L), \Phi_a(L)$ са диференчни полиноми от даден ред на *L*, които могат да включват несезонни, сезонни членове, b_j са неотрицателни цели числа за забавянията.

Построяването на ARIMA/SARIMA модел изисква предварително индентифициране на параметрите (p,d,q) или съответно (P,D,Q)(p,d,q)$_s$, което се провежда чрез изследване на ACF (автокорелационни функции) и PACF (частични автокорелационни функции). За по-детайлно изследване наличието на тренд и неговия порядък (стойност на d), може да се използват специфични статистически тестове, например на Дики-Фулър и негово обобщение, за идентифициране на единичен корен (корен =1) на характеристичното уравнение на модела.

Качествата на модела се оценяват по определени статистически критерии, като коефициент на детерминация (стационарен и нестационарен), средноквадратични грешки, включително информационни критерии като BIC (Bayersian или Schwarz Information Criterion) [80] Swartz и други. При избора на модел, докато се апроксимират данните, трябва да се вземат предвид следните общи изисквания:

- Остатъците на модела трябва да имат относително малка дисперсия;

- Остатъците трябва да имат нормално разпределение
- Моделът трябва да удовлетворява принципа за простота [51].

Моделиране с обобщена PathSeeker (GPS) регуляризираща регресия

Регуляризиращата GPS регресия е статистически метод от ново поколение, разработен от Джером Х. Фридман през 2008 [36]. Методът е предвиден за обработка на непрекъснати или двоични данни и произвежда няколко пътеки (класове) за избор на многомерна линейна регресия или модели за логистична регресия, които са по-ефективни от всички други видове регресионни модели [87]. Чрез GPS се построяват висококачествени линейни модели под формата на нормална многомерна регресия. Методът не осигурява автоматично откриване на нелинейности, взаимодействия между предиктори или функция за обработка на липсващи данни. Първоначалният модел, създаден чрез TreeNet е съвкупност от стотици или дори хиляди малки дървета. Много от тях обикновено са еднакви или имат много подобна структура. Компресията на TreeNet модела може да бъде получена с помощта на ISLE алгоритъм чрез отстраняване на излишните дървета. Коефициентите на моделите се коригират чрез GPS метода. Накрая RuleLearner алгоритъмът се използва като техника за последваща обработка, при което се избира подмножеството от възли с най-голямо влияние, намалявайки сложността на TreeNet модела.

За да въведем GPS метода, разглеждаме извадка от n наблюдения, дадени като

$$\left\{\mathrm{Y_i}, \mathbf{X}_i\right\}_{\mathrm{i}=1}^{\mathrm{n}} = \left\{Y_i, X_{i1}, ..., X_{im}\right\}_{i=1}^{n}, \tag{1.9}$$

където *Y* е откликът, а $\mathbf{X} = \left\{X_1, ..., X_m\right\}$ са предикторните променливи.

GPS моделът се търси във вида [36]

$$\hat{Y} = a_0 + a_1 X_1 + ... + a_m X_m \tag{1.10}$$

за предсказване на *Y* по дадено $\mathbf{X}$, а $\mathbf{a} = \left\{a_0, a_1, ..., a_m\right\}$ са параметрите. За да се получат по-равномерни оценки $\hat{\mathbf{a}}$ на $\mathbf{a}$, в регуляризиращите методи се решава оптималната задача

$$\hat{\mathbf{a}}_\gamma(\lambda) = \arg\min_{\mathbf{a}} \left[\hat{L}(\mathbf{a}) + \lambda P(\mathbf{a})\right] \tag{1.11}$$

В (1.11) $L(\mathbf{a})$ е емпирична функция на грешката, например:

$$L(\mathbf{a}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2, \tag{1.12}$$

а $P(\mathbf{a})$ е коригираща (наказателна, ограничителна) функция, представена във вида

$$P_\gamma(\mathbf{a}) = \sum_{i=1}^{m}\left|a_m\right|^\gamma, \quad \gamma \in [0, 2]. \tag{1.13}$$

Ключов избор в (1.11) е параметърът за регуляризация . λ, който се използва при подбора на модел. Параметърът γ обобщава и комбинира три добре познати метода. При $\gamma = 0$ се получава методът на компактна еластичност, при $\gamma = 1$ е Lasso метод (метод на минималното абсолютно свиване и селектиращ оператор) [95], а при $\gamma = 2$-хребетна регресия [46]. Последните 2 метода се обобщават от фамилията еластични методи [103] с избор на коригираща функция

$$P(\mathbf{a}) = P_\beta(\mathbf{a}) = \sum_{j=1}^{m}(\beta - 1)a_j^2 / 2 + (2 - \beta)\left|a_j\right|, \;\; 1 \le \beta \le 2, \tag{1.14}$$

където β се нарича коефициент на еластичност.

В GPS метода се обобщават всички горни методи, като $0 \le \beta \le 2$. Друг съществен момент е, че не се решава оптимизационната задача (1.11) на всяка стъпка, а коригиращата функция

$P(\mathbf{a})$ се избира така, че да удовлетворява за всички компоненти на $\mathbf{a}$ условието

$$\left\{\partial P(\mathbf{a}) / \partial \left|a_j\right| > 0\right\}_{j=1}^{m}. \tag{1.15}$$

Програмно GPS на SPM се реализира като много бърз стъпков алгоритъм със специализирана процедура за подбор на променливи [36, 37, 87].

Глава 2.1. Времеви анализ на качеството на въздуха в град Кърджали

Постановка на задачата

В тази глава е направен времеви статистически анализ на замърсяването на въздуха за град Кърджали, като се акцентира на основните замърсители, влияещи на човешкото здраве - прахови частици (PM_{10}) и серен диоксид (SO_2). Използвани са резултати на почасови данни от мониторинга за 4 последователни години, от 2008 година до 2011 година.

Конкретните задачи на изследването са: 1) Провеждане на времеви анализ с помощта на метода ARIMA и избор на адекватни параметрични модели; 2) Математическа оценка и анализ на построените модели; 3) Приложение на моделите за краткосрочна прогноза;

Описателна статистика на данните за град Кърджали

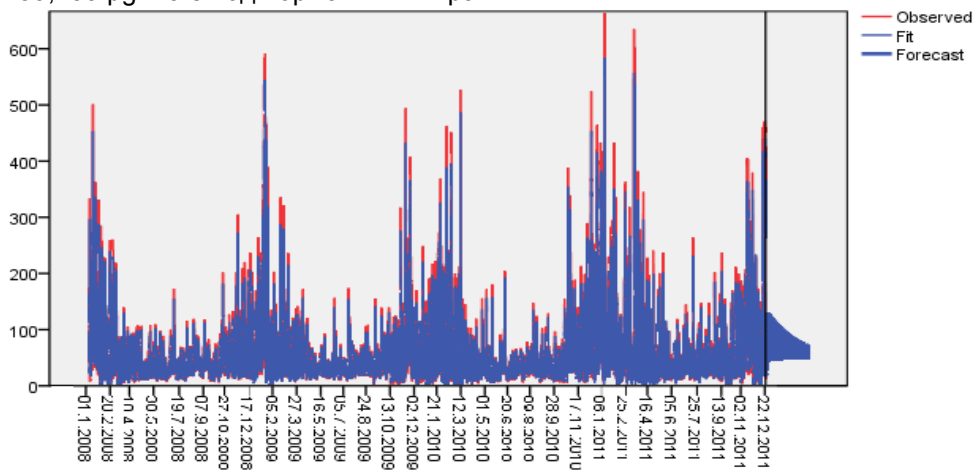
Резултатите от първичната обработка са показани в Таблица 2.1. Ще отбележим, че има 14 броя липсващи данни за SO2, съставляващи под 1% от целия наблюдаван времеви ред.

Таблица 2.1. Описателна статистика на данните за град Кърджали[1].

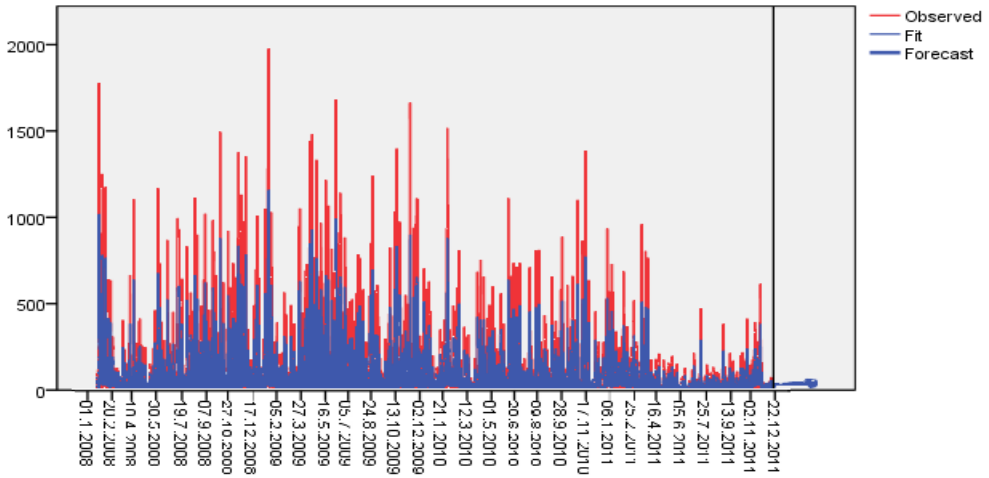
Променлива	Брой случаи	Мини-мум	Макси-мум	Средна стой-ност	Станд. откло-нение	Диспер-сия	Аси-метрия	Ексцес
PM10 (μg/m^3)	35016	-0.401	663.832	**45.821**	46.239	2138.035	3.913	23.288
SO2 (μg/m^3)	35003	-2.462	1977.232	39.439	89.937	8088.658	7.352	78.716

[1]Стандартна грешка на асиметрия=0.013, стандартна грешка на ексцес=0.026.

Максималната стойност от 663.8 μg/m3 на PM10 превишава над 13 пъти среднодневната норма от 50 μg/m3 [32]. Такъв тип превишения не са единични. Характерът на концентрацията с течение на времето (по часове) е показан на Фигура 2.1. От таблица 2.1 виждаме, че средната стойност от 45,821 за PM10 превишава годишната норма от 40 μg/m3 съгласно българското и европейското законодателство [6, 110]. Съответно, от същата таблица 2.1 виждаме, че максималната стойност от 1977,232 μg/m3 на SO2 също превишава 13,181 пъти среднодневната норма, а средната стойност от 39,439 μg/m3 е под нормативния праг.



Фиг. 2.1. Наблюдавани почасови стойности на PM10 за град Кърджали, моделни стойности от SARIMA и предсказани стойности за 14 дни напред.



Фиг. 2.2. Наблюдавани почасови стойности на SO2 за град Кърджали, моделни стойности от SARIMA и предсказани стойности за 14 дни напред.

Построяване на SARIMA модели и анализ

Бяха подбрани SARIMA моделите, дадени на Таблица 2.2.

Таблица 2.2 SARIMA модели на замърсителите на PM10 и SO2 за град Кърджали.

Променлива	Модел SARIMA	Качества на модела		Ljung-Box Q(18)			Брой отдалечени случаи
		Стац. R^2	R^2	Statistics	DF	Sig.	
SO2, μg/m^3	$(4,0,10)(2,0,1)_{24}$	0.521	0.521	5.874	1	0.015	0
PM10, μg/m^3	$(4,0,10)(2,0,1)_{24}$	0.822	0.822	27.411	1	0.000	0

Може да се отбележи много добрата статистическа значимост (Sig.=0.000 и Sig.=0.015) на моделите и за двата замърсителя, разглеждана при ниво α=0.05. Коефицентът на детерминация за качеството на приближението на модела за PM10 е R^2=0.822, което означава че моделът описва 82% от данните. Съответно за модела за SO2 имаме R^2=0.521.

Глава 2.2. Времеви анализ на качеството на въздуха в град Шумен

Постановка на задачата

Търсим математически модели, описващи поведението на концентрациите на прахови частици (PM10), серен диоксид (SO2), азотен диоксид (NO2), азотен оксид (NO), азотни оксиди (NOx) и озон (O3), съотнесени спрямо метеорологичните данни. Данните са изразени в единици масова концентрация на замърсителите в μg/m^3, за NOx - частици на милион.

Използват се и следните 6 метеорологични променливи: атмосферно налягане (PRESS), посока на вятъра (DVP), приземна температура на въздуха (TEMP), слънчева радиация (RADST), относителна влажност (UMR) и скорост на вятъра (VVG).

Цели на изследването са: 1) Провеждане на времеви анализ на отделните замърсители в зависимост от използване на метеорологичните редове; 2) Групиране на замърсителите чрез факторен анализ; 3) Използване на факторните (групираните) редове за построяване на модели; 4) Оценка и анализ на моделите; 5) Провеждане на краткосрочни прогнози.

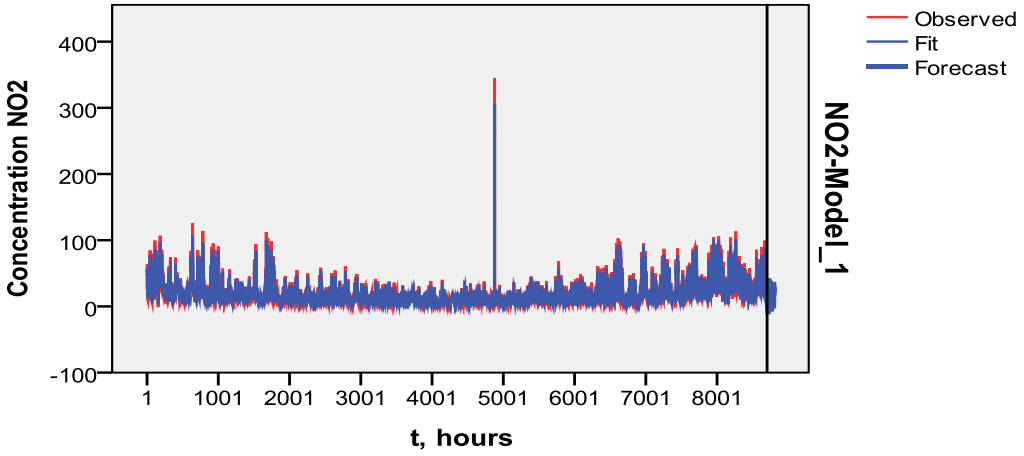
Описание на данните за град Шумен

Общият обем на данните включва от 8546 до 8702 случая, по часове. Наблюдаваният максимален брой липсващи стойности (за всеки параметър) е под 2% и няма повтаряща се тенденция. Поради тази причина не беше приложена специална процедура за обработка на липсващите данни. Освен това данните се моделират без да се премахват големи отклонения (outliers).

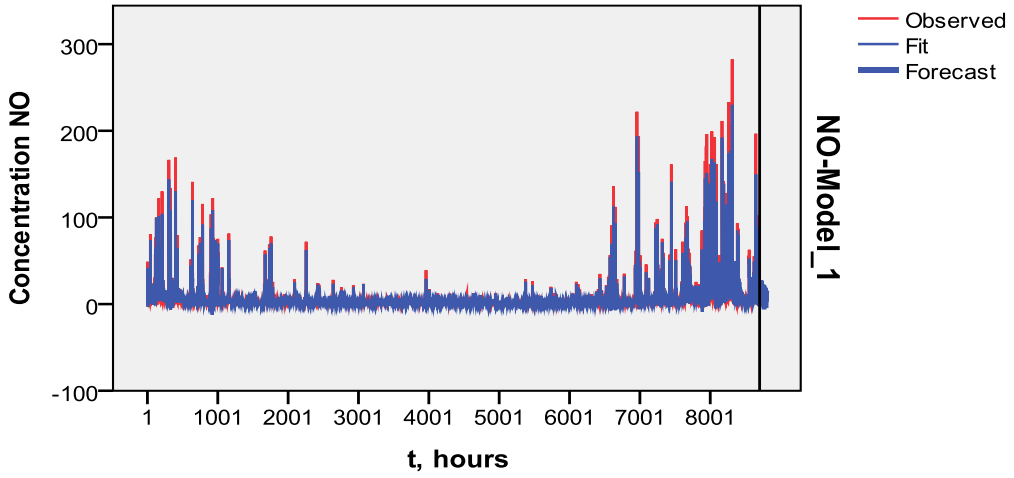
Таблица 2.4. Описателна статистика на наблюдаваните замърсители и метеоролигчни данни на град Шумен.

Променлива	N	Минимум	Максимум	Средна	Стд.откл.	Асиметрия	Ексцес
NO2, μg/m^3	8546	1.229	344.618	18.419	17.221	2.781	19.469
NO, μg/m^3	8546	1.299	282.535	6.010	18.164	6.510	54.188
O3, μg/m^3	8661	3.201	1445.880	59.384	34.928	7.344	285.792
PM10, μg/m^3	8699	0.794	530.790	47.368	52.192	3.251	13.512
SO2, μg/m^3	8662	0.354	1091.003	4.8159	14.114	54.276	4065.040
NOx, μg/m^3	8546	0.507	261.447	14.447	21.525	4.455	26.421
PRESS, hPa	8798	956.714	1000.333	980.678	6.556	0.167	0.076
DVP, degrees	8798	33.626	299.665	182.462	48.610	0.090	-1.032
TEMP, ^{o}C	8798	-12.065	32.381	9.386	9.529	0.142	-0.914
RADSTW/m^2	8798	0.000	1086.172	172.556	252.206	1.655	1.620
UMR, %	8798	13.065	93.627	62.891	18.062	-0.444	-0.828
VVG, m/s	8798	0.039	7.596	2.218	0.782	1.496	3.719

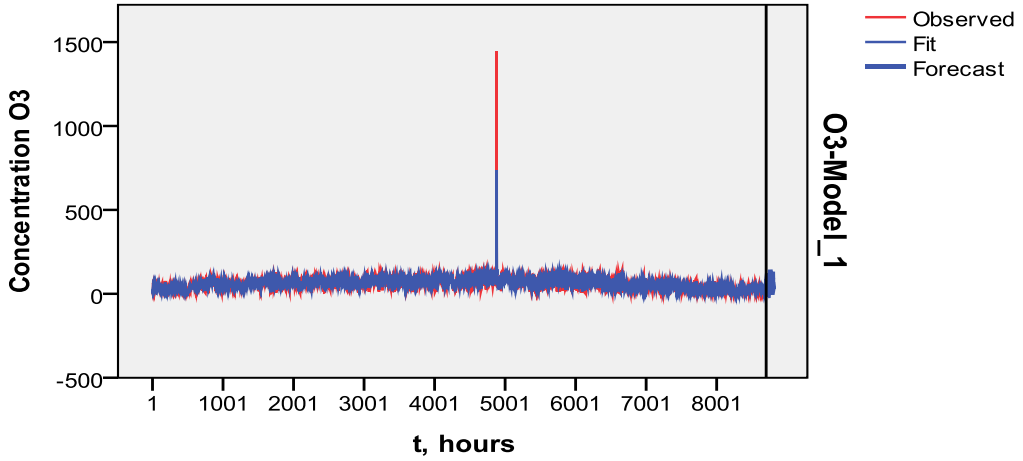
Подробният анализ на резултатите от Таблица 2.4 показва, че за някои променливи (NO2, PM10 и TEMP) съответните стандартни отклонения и средни стойности са почти равни. Освен това порядъкът на стандартното отклонение на неотрицателните параметри NO и SO2 е повече от три пъти по-големи от съответните средни стойности. Това е индикатор за висока чувствителност към неопределеностите при тези параметри.



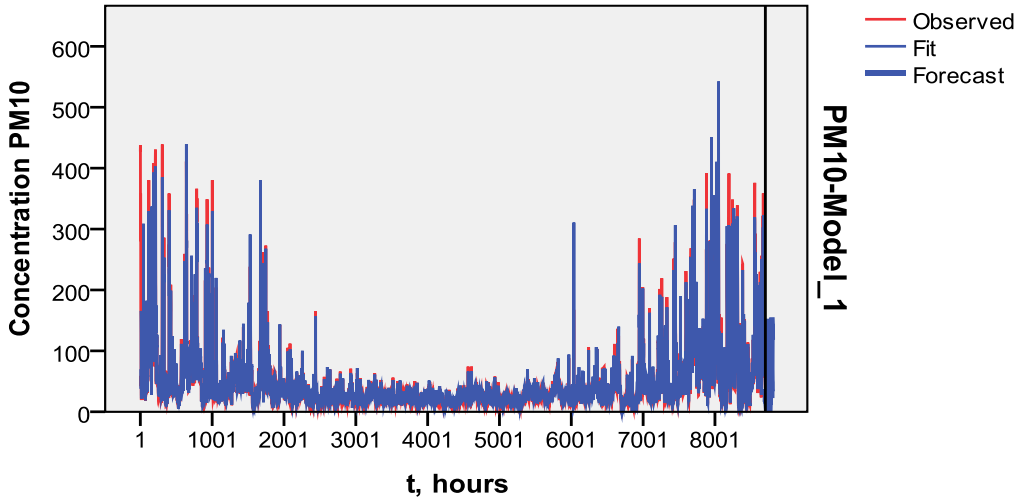
Фиг. 2.6. Наблюдавани, описани и предсказани данни на *NO*2 за град Шумен.



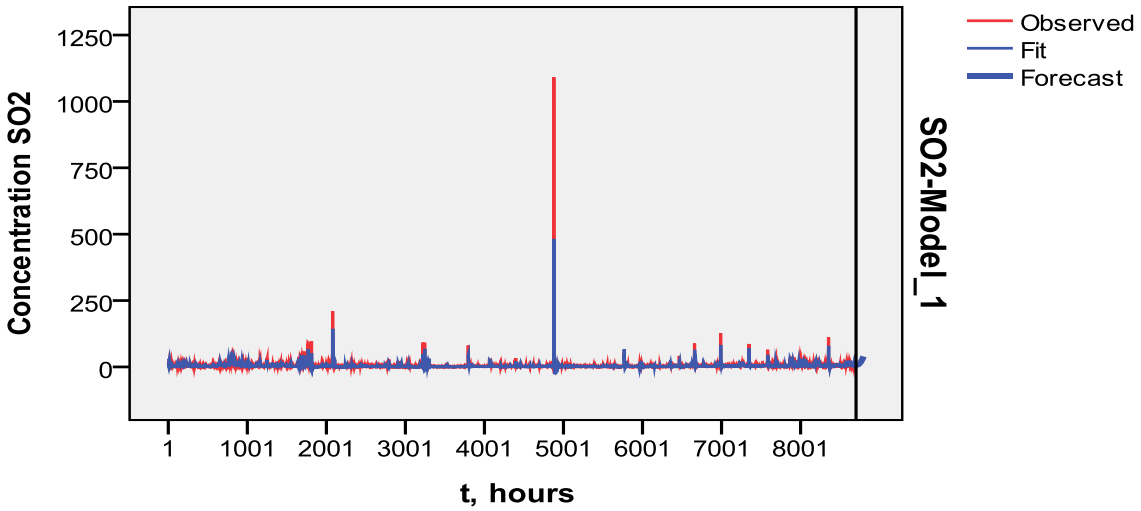
Фиг. 2.7 Наблюдавани, описани и предсказани данни на *NO* за град Шумен.



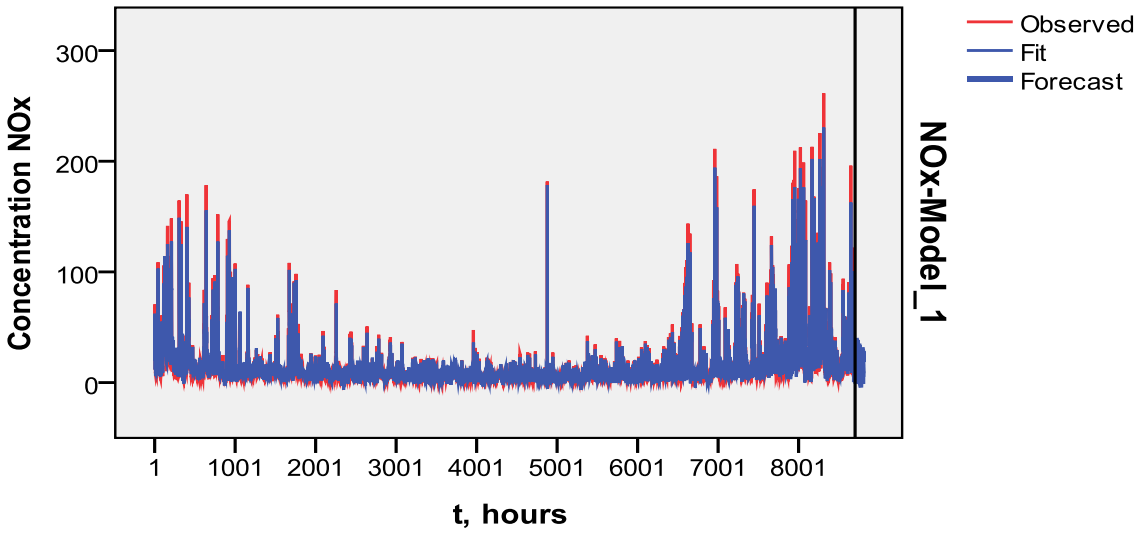
Фиг. 2.8. Наблюдавани, описани и предсказани данни на *O*3 за град Шумен.



Фиг. 2.9. Наблюдавани, описани и предсказани данни на *PM*10 за град Шумен.



Фиг. 2.10. Наблюдавани, описани и предсказани данни на *SO*2 за град Шумен.



Фиг. 2.11. Наблюдавани, описани и предсказани данни на *NO*x за град Шумен.

Групиране на замърсителите чрез факторен анализ

Факторният анализ е приложен за изследваните 6 замърсителя в 5 стъпки:

1) Съответната корелационна матрица за 6-те замъсители е дадена в Таблица 2.10. От нея се вижда, че детерминантата е малка ($7.21*10^{-6}$) и има големи статистически значими корелационни коефициенти (дадени в получерни цифри). Може да се добави, че O3 корелира отрицателно с другите наблюдавани замърсители. Всичко това е индикатор, че провеждането на факторния анализ е препоръчително.
2) Обикновено при факторния анализ, измерването на адекватността става с помощта на статистическите индекси на Кайзер-Майер-Олкин (KMO) за адекватността на извадката и тестът за сферичност на Бартлет, където KMO трябва да бъде по-висок от 0.5, а теста за сферичност на Бартлет да е статистически значим (sig<0.05). В нашия случай KMO =0.590, а тестът на Бартлет е значим при Sig. =0, което показва, че съществува връзка между някои променливи и се препоръчва провеждането на факторен анализ. Таблица 2.5, например, показва силната корелация между NO и NOx, за които се очаква да бъдат групирани в един фактор. Най-големите корелационни коефиценти в таблица 2.5 са в удебелен шрифт.
3) Факторите са извлечени с помощта на PCA метод [54] Jolifee, в резултат на което са получени 4 фактора.
4) Методът за факторно въртене Promax се оказа най-подходящият. За нашите данни осигурява рязко разграничение между факторите в сравнение с Varimax и други методи за въртене. Promax въртенето е добре известна техника в екологията и климатологията [78]. За нашите данни, той се използва за въвеждането на 4 фактора в неортогонална координатна система. Получените факторни тегла са дадени в Таблица 2.6, където теглата под 0.3 са пропуснати, в съответствие с [21] (виж също Глава 1).
5) Важен резултат от факторния анализ е изчисляването на стойностите на факторните (латентни) променливи. В нашия случай резултатните нови променливи са означени по-нататък съответно с F1, F2, F3 и F4.

Замърсителите са ясно разделени по групи. Изпълнено е изискването всяка променлива да участва съществено само в един фактор и много малко в другите.
Получените четири фактора описват 97.43% от общата изменчивост на данните. Частичният принос на факторите е съответно: 32% за F1, 32% за F2, 22% за F3 и 14% за F4.

Таблица 2.5. Корелационна матрица на замърсителите[1].

Променлива	NO2	NO	O3	PM10	SO2	NOx
NO2	1	0.653	-0.482	0.805	0.302	0.859
NO	**0.653**	1	-0.348	0.686	0.068	0.948
O3	-0.482	-0.348	1	-0.468	0.357	-0.436
PM10	**0.805**	**0.686**	-0.468	1	0.125	0.800
SO2	0.302	0.068	0.357	0.125	1	0.173
NOx	**0.859**	**0.948**	-0.436	**0.800**	0.173	1

[1]Всички корелационни коефициенти са значими със Sig.= 0.001.

Таблица 2.6. Матрица с теглата на шесте замърсителя в 4 факторни групи[1].

Променлива	**Фактор F1**	**Фактор F2**	**Фактор F3**	**Фактор F4**
NO	1.102			
NOx	0.806			
PM10		1.108		
NO2		0.550		
O3			-1.020	
SO2				1.001

[1] Използвани са: PCA метод и Promax въртене с Кайзер нормализация. Теглата по-малки от 0.3 са пропуснати.

АРИМА модели с отчитане на метеорологичните редове

За всеки от изследваните 6 замърсителя на град Шумен бяха построени голям брой ARIMA модели с предиктори – описани в началото на параграфи 2.2 и таблица 2.4.
Избраните най-добри модели, със съответните статистики са представени в таблица 2.7
От таблица 2.7 се виждат високите стойности на статистическите коефициенти на детерминация R^2 за 5 замърсителя (NO2, NO, O3, SO2, NOx) и ниска стойност R^2=0,307 за PM10. Също R^2 от колона 4 на таблица 2.7 показва, че избраните модели обясняват от 67% до 88% от данните. Всички модели са статистически значими и със статистически значими параметри. Конкретните параметри на моделите са дадени в таблица 2.6.

Таблица 2.7. ARIMA модели на замърсителите за град Шумен и групираните факторни променливи.

Променлива	**Брой предиктори**	**Статистики на модела**			**Ljung-Box Q(18)**			**Трансформация**	**Брой отдалечени**	**ARIMA модел**
		Стац. R^2	**R^2**	**RMS**	**Индекс**	**DF**	**Sig.**			
NO2	5	0.763	0.763	8.383	34.895	8	0.000	-	0	(3,0,7)
NO	6	0.673	0.673	10.402	77.127	9	0.000	-	0	(3,0,9)
O3	6	0.738	0.738	17.903	24.405	9	0.002	-	0	(3,0,7)
PM10	6	0.307	0.878	18.011	28.173	13	0.009	Nat log	0	(2,1,9)
SO2	6	0.922	0.922	13.218	65.318	6	0.000	-	0	(5,3,7)
NOx	6	0.747	0.747	10.832	68.109	13	0.000	-	0	(3,0,9)
F1	6	0.706	0.706	0.543	83.183	13	0.000	-	0	(3,0,9)
F2	6	0.851	0.851	0.383	42.380	9	0.000	-	0	(2,0,7)
F3	6	0.795	0.795	0.543	41.011	12	0.000	-	0	(3,0,7)
F4	6	0.921	0.921	0.919	54.505	11	0.000	-	0	(5,3,7)

Изводи към глава 2

В 2.1 са изследвани почасови данни за замърсяването с прахови частици под 10 микрона PM_{10} и серен диоксид SO_2 на гр. Кърджали за период от 4 години. За моделиране на концентрациите на PM_{10} и SO_2 са построени едномерни стохастични SARIMA модели. Установи се, че избраният подход на моделиране е адекватен, като получените модели са с много добри статистически показатели. Предсказаните стойности от моделите имат добро съвпадение с измерените концентрации. Получени са краткосрочни прогнози за период от 14 дни напред.
В 2.2 са изследвни концентрациите на 6 замърсители на въздуха в зависимост от 6 метеорологични променливи, описващи качеството на атмосферния въздух в град Шумен. Бе

установено, че данните за 6-те замърсителя корелират силно. Бе приложен факторен анализ и те бяха групирани в 4 добре разграничени фактора, което се обяснява с наличие на общи източници на замърсяване. След това същите данни са моделирани с ARIMA методологията. Построени са модели за отделните 6 замърсители с участие на 6-те метеорологични реда, с много добри качества. Аналогочни модели са получени и за четирите факторни променливи. Моделите са приложени за краткосрочни прогнози за срок от 72 часа напред, като резултатите показаха много добро съвпадение, сравнени с известни измервания, неучаствали в моделирането. Всички получени математически модели са в добро съгласие с докладите за околната среда [115].

Глава 3. ВРЕМЕВИ АНАЛИЗ НА ЕЖЕДНЕВНИТЕ ЗАМЪРСИТЕЛИ НА ВЪЗДУХА В ГРАД БЛАГОЕВГРАД

Постановка на задачата

Основната цел на настоящото проучване е установяването на зависимост с варирането в нивата на замърсяването на въздуха и възможните комбинирани въздействия в рамките на една година, както и разкриването на техните източници. Приложени са два статистически подхода - факторен анализ и SARIMA.

Въпросите, разгледани в главата са: 1) Групиране на наблюдаваните замърсители с помощта на факторен анализ и извличане на факторни променливи, обясняващи комбинираното въздействие на замърсяването; 2) Провеждане на анализ на временни редове чрез определяне на съответните параметрични сезонни ARIMA модели (базирани на почасови данни) на замърсителите; 3) Анализ и диагностика на построените модели; 4) Приложение на модели за краткосрочно предсказване; 5) Интерпретация на резултатите и определяне на условията, на които се дължи надвишаването на пределно допустимите норми според националните и европейските стандарти за разглежданите замърсители на въздуха.

Използвани данни

Наблюдаваните са почасови данни за въздушните замърсители за град Благоевград в периода септември 2011 г. - август 2012 г. са дадени в таблица 3.1.

Таблица 3.1: Описателна статистика на наблюдаваните въздушни замърсители за град Благоевград[1].

Променлива	Минимум	Максимум	Средна	Станд.Откл.	Асиметрия	Ексцес	K-S тест
NO2, μg/m^3	0.121	166.6	21.70	27.07	1.90	4.79	0.137
NO, μg/m^3	0.000	179.1	5.75	10.50	5.14	41.02	0.293
O3, μg/m^3	1.078	174.5	61.90	64.52	0.37	-0.53	0.043
PM10,μg/m^3	0.467	555.5	**46.87**	38.90	2.87	11.77	0.213
SO2, μg/m^3	0.000	321.5	17.44	25.33	3.60	21.83	0.212
NOx, ppb	1.184	188.2	16.00	24.64	2.81	12.27	0.188

[1] Станд. грешки на коефициентите на асиметрия са 0.026, а станд. грешки на коефициентите на ексцес са 0.052.

Трансформация на данните

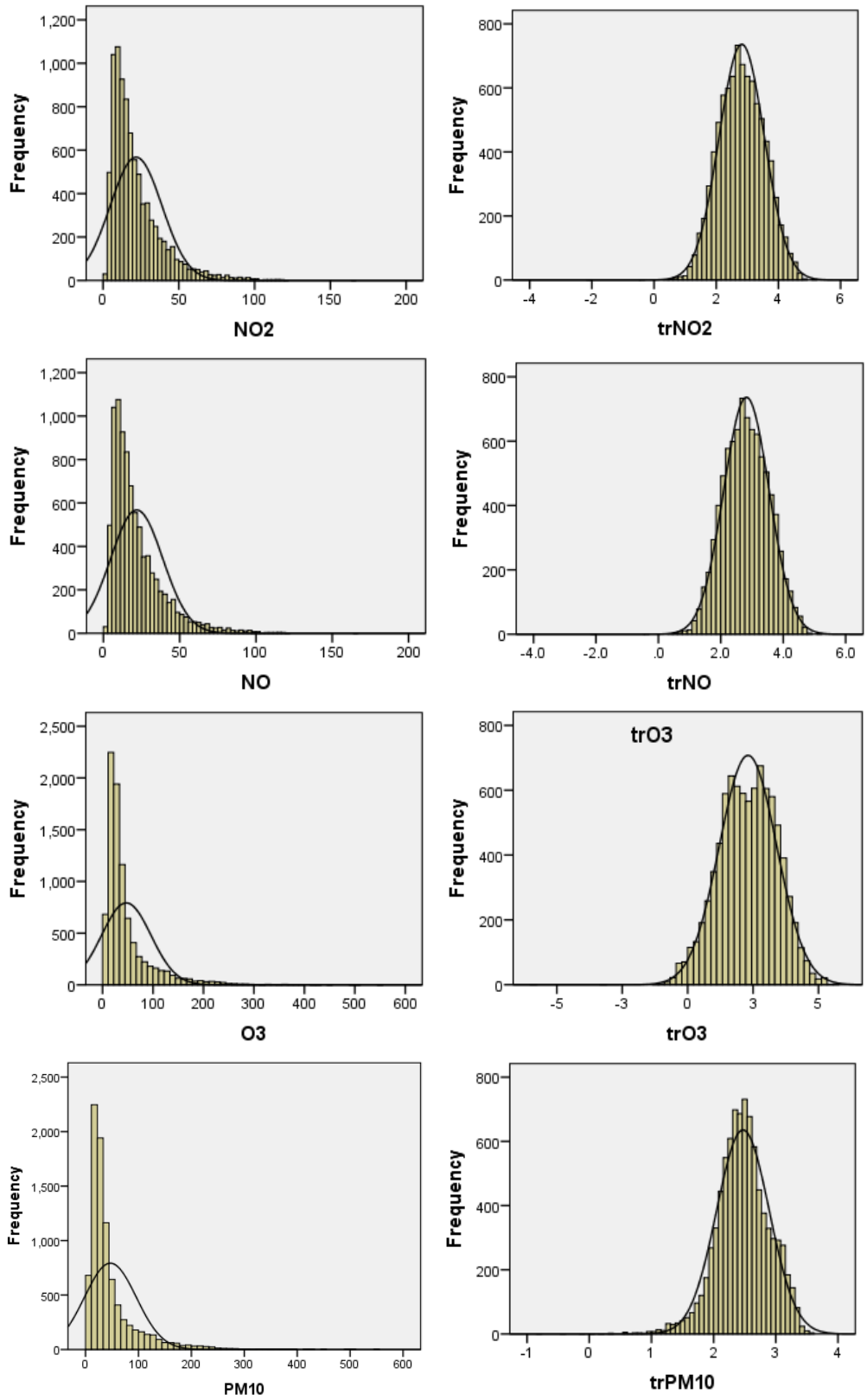
Ефектът от прилагането на трансформацията на Йео-Джонсън е показан иа на сравнителните честотни графики от фигури 3.7-3.12.

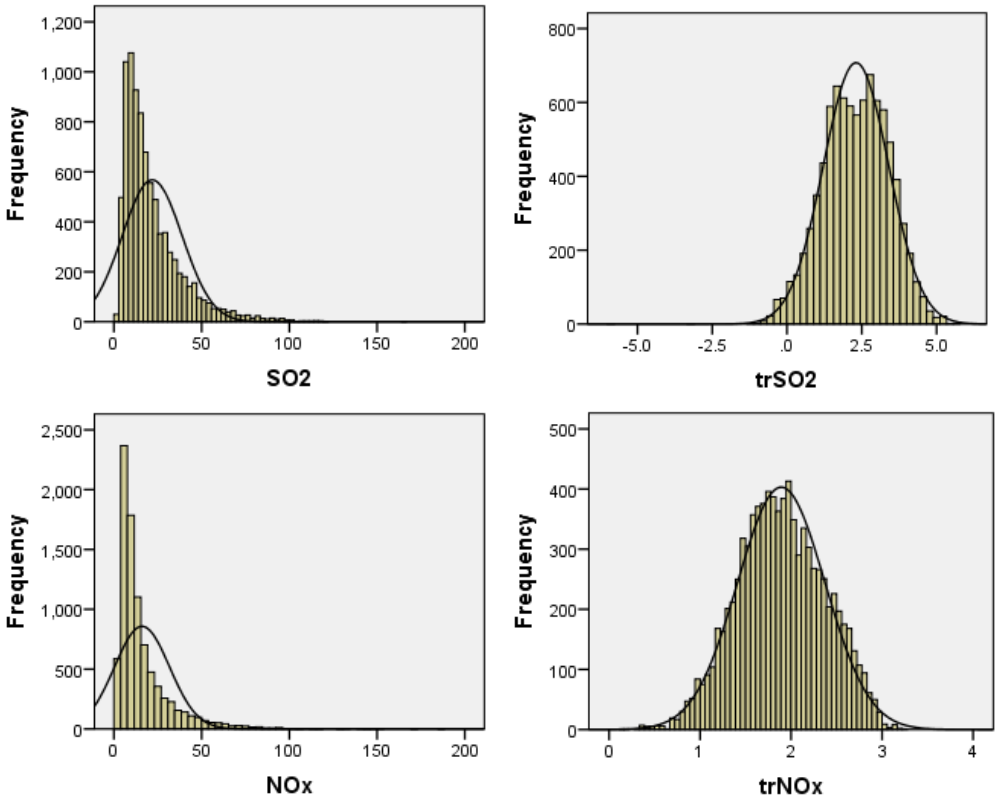
Таблица 3.2. Описателна статистика на трансформираните данни за град Благоевград[1,2].

Трансф. променлива	Коефициент λ	Минимум	Максимум	Средно	Станд. откл.	Асиметрия	Ексцес	K-S тест
trNO2	**0**	**-2.12**	**5.12**	**2.82**	**0.73**	**0.01**	**-0.18**	**0.02**
trNO	**-0.2**	**-9.80**	**3.23**	**0.75**	**0.97**	**-0.28**	**1.58**	**0.02**
trO3	**0.8**	**-1.25**	**76.44**	**31.72**	**15.33**	**0.14**	**-0.68**	**0.03**
trPM10	**-0.2**	**-0.82**	**3.59**	**2.47**	**0.42**	**-0.46**	**1.58**	**0.03**
trSO2	**0**	**-5.95**	**5.78**	**2.30**	**1.09**	**-0.17**	**0.04**	**0.03**
trNOx	**-0.2**	**0.17**	**3.25**	**1.89**	**0.48**	**-0.04**	**-0.34**	**0.02**

[1] Стандартната грешка на коефициентите на асиметрия е 0.026, стандартната грешка на коефициентите на ексцес е 0.052.
[2] Lilliefors корекцията е използвана за теста на Колмогоров-Смирнов (K-S test) в SPSS.

Хистограмите преди и след трансформациите са показани на следващите фигури.





Факторен анализ

Факторният анализ е приложен за изследваните 6 замърсителя. Съответната корелационна матрица с корелация на Пиърсън е дадена в Таблица 3.3. От нея се вижда, че детерминантата е малка ($8{,}53*10^{-7}$) и има големи статистически значими корелационни коефициенти.

Таблица 3.3: Корелационна матрица на шесте замърсителя на въздуха (не трансформирани) на град Благоевград.

		NO2 µg/m^3	NO µg/m^3	O3 µg/m^3	PM10 µg/m^3	SO2 µg/m^3	NOx ppb
Корелация по Пирсън	NO2	1					
	NO	**0.63**	1				
	O3	-0.56	-0.41	1			
	PM10	**0.84**	**0.629**	-0.45	1		
	SO2	0.35	0.15	-0.07	0.33	1	
	NOx	**0.91**	**0.90**	-0.54	**0.82**	0.28	1

[a] Детерминанта = $8.53*10^{-7}$. Значимостта на всички корелационни коефициенти е 0.000.

Факторите са извлечени с помощта на метода на главните елементи (PCA метод), като са получени 3 фактора. Методът за факторно въртене Promax се оказа най-подходящият. За нашите данни осигурява рязко разграничение между факторите в сравнение с Varimax и други популярни методи за въртене. Promax въртенето е добре известна техника в екологията и климатологията [78]. За нашите данни, то се използва за въвеждането на 3 фактора в неортогонална координатна система. Получените факторни тегла са дадени в Таблица 3.4, където теглата под 0.3 са игнорирани.

Замърсителите са ясно разделени в 3 групи. Изпълнено е изискването всяка променлива да участва само в един фактор.

Получени са следните фактори:
F1={NO2, NO, NOx, PM10}, *F*2={*O3*}, *F*3={*SO*2}. (3.1)
Тези три фактора описват 90.74% от общата изменчивост на данните. Частичният принос на факторите е съответно: 45% за F1, 27% за F2, 19% за F3.

Таблица 3.4: Моделна матрица на факторния анализ[a,b].

Променливи	Компонента		
	Factor F1	Factor F2	Factor F3
NO2, μg/m^3	1.075		
NO, μg/m^3	0.979		
NOx, ppb	0.775		
PM10, μg/m^3	0.703		
O3, μg/m^3		-1.027	
SO2, μg/m^3			1.014

Extraction method: Principal Component Analysis; Rotation method: Promax with Kaiser Normalization; Rotation converged in 5 iterations.[b] Factor loadings less than 0.3 are omitted.

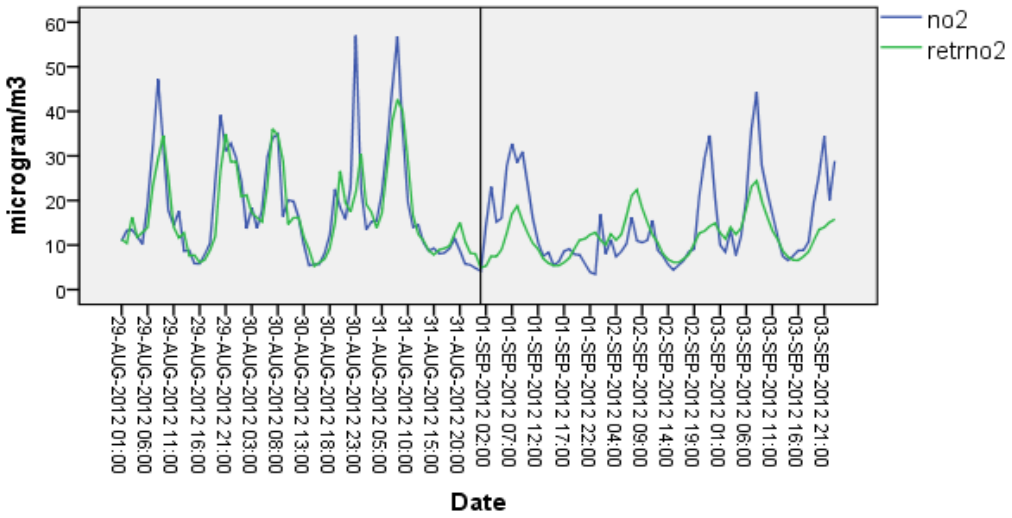
Построяване и диагностика на SARIMA моделите за трансформираните променливи

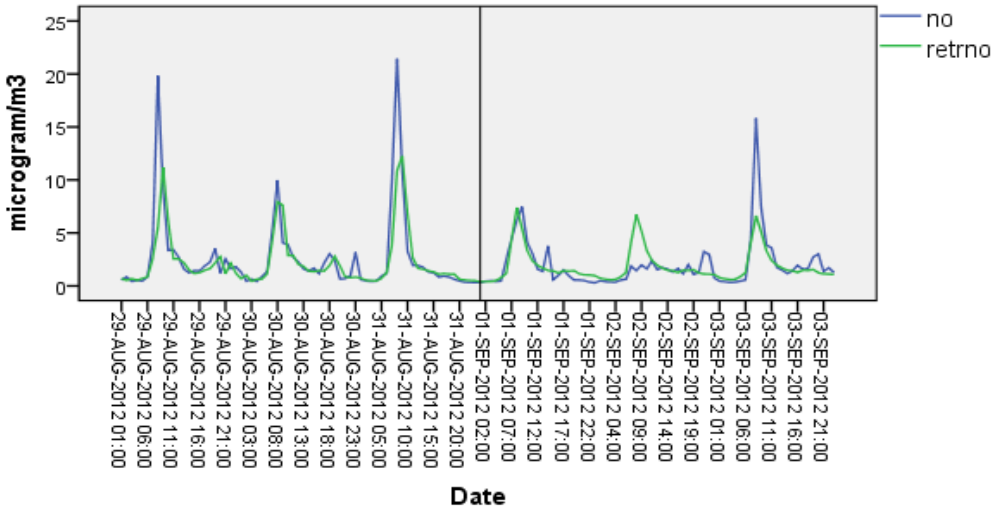
Обикновено, за да се определи дали трансформираната променлива представя стационарен процес, се прави тест за единични корени. Приложихме допълнения тест на Дики-Фулър (ADF) [24, 79] за всяка от извадките с трансформирани данни за замърсителите с помощта на софтуера EViews 7 [33]. Имайки предвид, че RMSE е добър показател за точността, който се използва за сравняване на грешката при предсказване на различните модели за определена променлива, а не между променливите. При изследването не се взема предвид MASE. В допълнение към BIC критерия, стандартните грешки (RMSE) и средна абсолютна процентна грешка (MAPE) се използват като по-важни статистики за приближение на модела към данните. Резултатите за статистиката на модела, както и обичайно използваните мерки за приближение, като стационарния R^2, R^2 и др. При построяване на SARIMA моделите се задават различни стойности на параметрите (p,d,q)(P,D,Q)24 и се изчисляват коефицентите на избрания модел със статитическите им индекси. Бяха построени няколкостотин модела, от които са подбрани оптималните в следващата таблица.

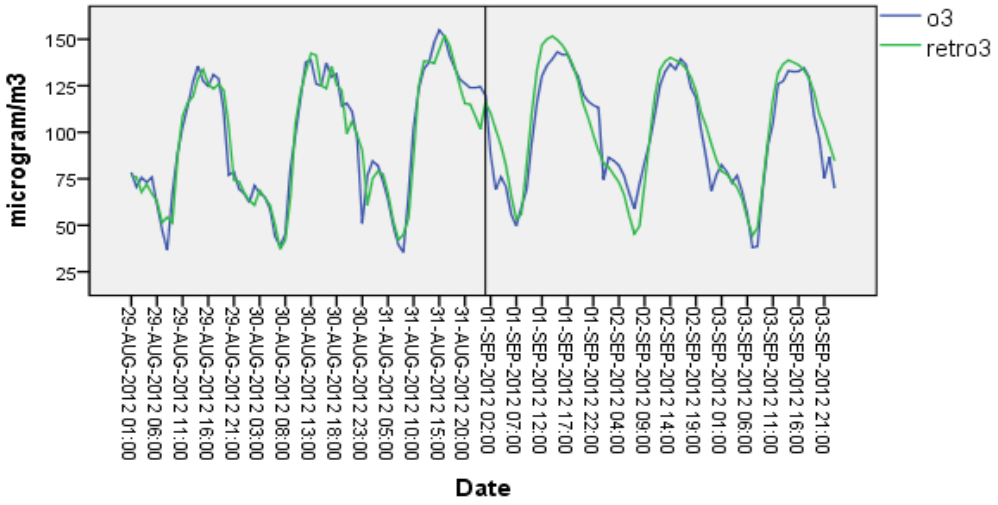
Изборът на модели е направен в съответствие със следните комбинирани критерии 1) Минимален BIC; 2) Минимална RMSE; 3) Максимална стойност на R^2; 4) Значимост на ниво 0.05;5) Минимална MAPE.
Някои от другите кандидат модели, сортирани според този критерии са дадени под избраните оптимални модели.
Разпределенията на остатъците в рамките на 5% доверителни интервали и нормалността на остатъците във всички кандидат модели бяха проверени преди да бъде приложена процедурата за предложените комбинирани критерии.

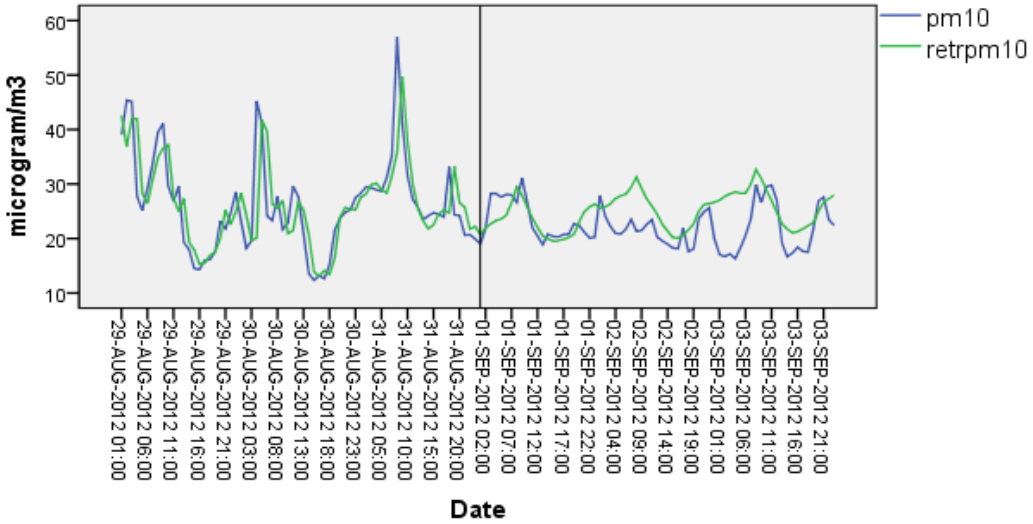
Трансф. Променливи	SARIMA модел	Стац.R^2	R^2	RMSE	MAPE	Норм. BIC	Sig.
trNO2	$(2,0,1)(2,0,1)_{24}$	0.841	0.841	0.290	8.461	-2.467	0.003
trNO	$(1,0,7)(2,0,1)_{24}$	0.783	0.783	0.451	225.84	-1.579	0.003
trO3	$(2,1,1)(1,1,1)_{24}$	0.437	0.941	3.740	13.813	2.643	0.000
trPM10	$(4,0,4)(1,0,1)_{24}$	0.888	0.888	0.142	4.170	-3.897	0.000
trSO2	$(2,0,2)(2,0,1)_{24}$	0.887	0.887	0.369	20.492	-1.988	0.000
trNOx	$(3,0,2)(2,0,1)_{24}$	0.841	0.841	0.191	8.595	-3.302	0.009

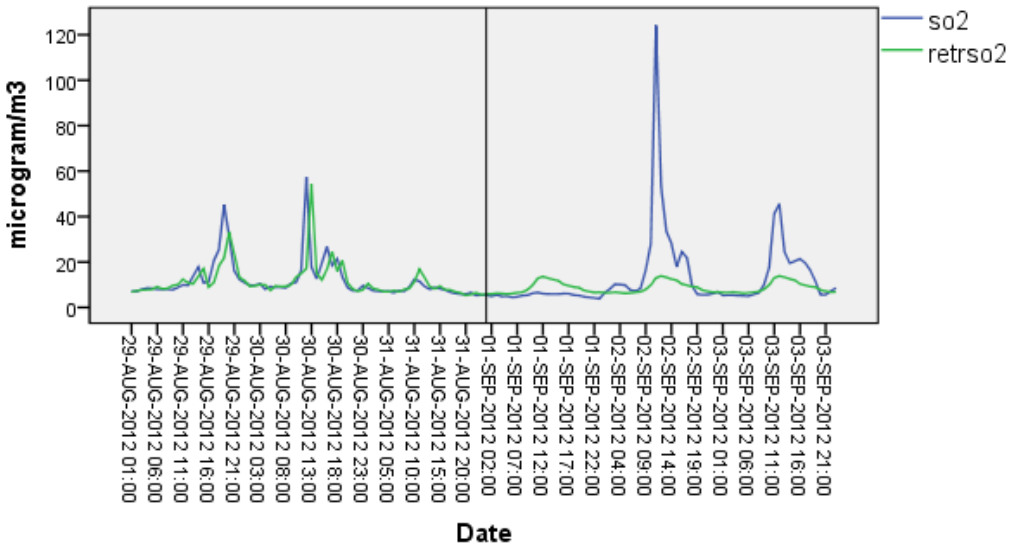
Сравнение на наблюдаваните данни за последните 72 часа (вляво от вертикалната линия) и предсказаните данни за замърсителите използвайки външен тест за сравнение с истински данни за 72 часа (вдясно от вертикалната линия) за ретрансформираните данни за съответните SARIMA модели от горната таблица.

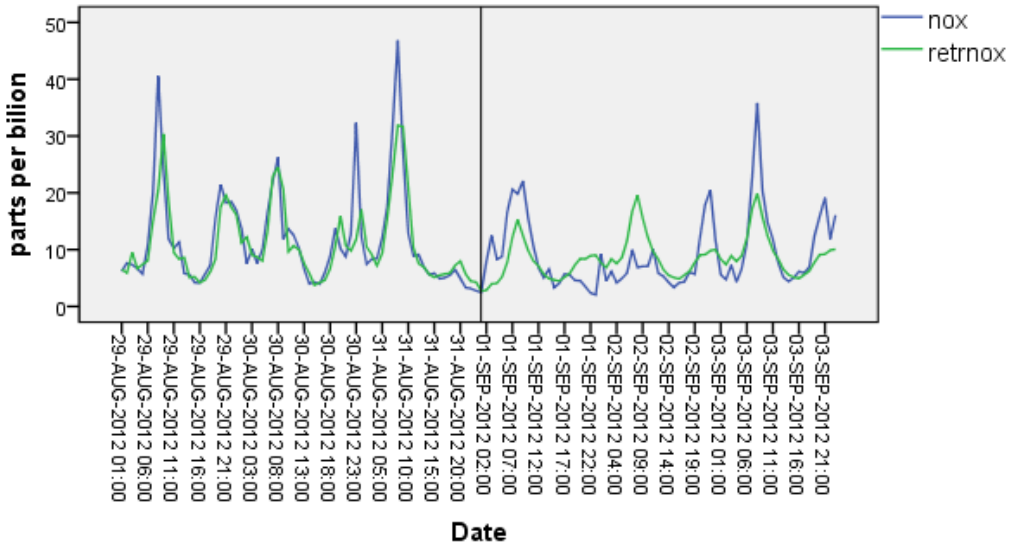







pm10
retrpm10
microgram/m3
60
50
40
30
20
10
29-AUG-2012 01:00
29-AUG-2012 06:00
29-AUG-2012 11:00
29-AUG-2012 16:00
29-AUG-2012 21:00
30-AUG-2012 03:00
30-AUG-2012 08:00
30-AUG-2012 13:00
30-AUG-2012 18:00
30-AUG-2012 23:00
31-AUG-2012 05:00
31-AUG-2012 10:00
31-AUG-2012 15:00
31-AUG-2012 20:00
01-SEP-2012 02:00
01-SEP-2012 07:00
01-SEP-2012 12:00
01-SEP-2012 17:00
01-SEP-2012 22:00
02-SEP-2012 04:00
02-SEP-2012 09:00
02-SEP-2012 14:00
02-SEP-2012 19:00
03-SEP-2012 01:00
03-SEP-2012 06:00
03-SEP-2012 11:00
03-SEP-2012 16:00
03-SEP-2012 21:00
Date


so2
retrso2
microgram/m3
120
100
80
60
40
20
0
29-AUG-2012 01:00
29-AUG-2012 06:00
29-AUG-2012 11:00
29-AUG-2012 16:00
29-AUG-2012 21:00
30-AUG-2012 03:00
30-AUG-2012 08:00
30-AUG-2012 13:00
30-AUG-2012 18:00
30-AUG-2012 23:00
31-AUG-2012 05:00
31-AUG-2012 10:00
31-AUG-2012 15:00
31-AUG-2012 20:00
01-SEP-2012 02:00
01-SEP-2012 07:00
01-SEP-2012 12:00
01-SEP-2012 17:00
01-SEP-2012 22:00
02-SEP-2012 04:00
02-SEP-2012 09:00
02-SEP-2012 14:00
02-SEP-2012 19:00
03-SEP-2012 01:00
03-SEP-2012 06:00
03-SEP-2012 11:00
03-SEP-2012 16:00
03-SEP-2012 21:00
Date


nox
retrnox
parts per bilion
50
40
30
20
10
0
29-AUG-2012 01:00
29-AUG-2012 06:00
29-AUG-2012 11:00
29-AUG-2012 16:00
29-AUG-2012 21:00
30-AUG-2012 03:00
30-AUG-2012 08:00
30-AUG-2012 13:00
30-AUG-2012 18:00
30-AUG-2012 23:00
31-AUG-2012 05:00
31-AUG-2012 10:00
31-AUG-2012 15:00
31-AUG-2012 20:00
01-SEP-2012 02:00
01-SEP-2012 07:00
01-SEP-2012 12:00
01-SEP-2012 17:00
01-SEP-2012 22:00
02-SEP-2012 04:00
02-SEP-2012 09:00
02-SEP-2012 14:00
02-SEP-2012 19:00
03-SEP-2012 01:00
03-SEP-2012 06:00
03-SEP-2012 11:00
03-SEP-2012 16:00
03-SEP-2012 21:00
Date

Изводи към глава 3

Изследвани са едномерни времени редове за 6 замърсителя на атмосферния въздух в град Благоевград. Приложени са два статистически подхода за описание на реалните данни в рамките на период от една година, на базата на почасови измервания. Данните са обработени с помощта на факторен анализ, което позволи тяхното групиране в 3 основни фактора, идентифициращи комбинираното замърсяване. Това се обяснява с наличието на общи източници на замърсителите в получените групи.
Основната част от резултатите са получени с помощта на прилагането на SARIMA метод и са построени голям брой модели. По-конкретно, резултатите показват, че PM10 надвишава официалните национални и европейски норми, ето защо състоянието на този екологичен индикатор е силно обезпокоителен. Моделите са приложени за краткосрочно прогнозиране в рамките на период от 72 часа, като резултатите демонстрират много добри характеристики при сравнение с реалните данни. Най-добрите модели са получени и избрани на базата на информационния критерий BIC и други често използвани критерии за оценка. Като цяло беше показано, че факторният анализ и SARIMA подходът са много подходящи инструменти за изследване на нивата на замърсяване на въздуха в малки населени места с цел подпомагане на ежедневния контрол и прогнозиране по отношение на качеството на въздуха.

Глава 4. ПРИЛОЖЕНИЕ НА SARIMA И GPS РЕГРЕСИЯ ЗА ИЗСЛЕДВАНЕ ЗАМЪРСЯВАНЕ НА ВЪЗДУХА С PM10 НА ГРАД ШУМЕН

Постановка на задачата
Конкретните цели на изследването са:
1) Провеждане на анализ на временните редове с помощта на стохастични SARIMA модели, вземащи предвид периодичността (по часове) и зависимостта на праховите частици PM10 от 6 метеорологични променливи.
2) Анализ и оценка на получените модели.
3) Прилагане на моделите за краткосрочно прогнозиране.
4) Провеждане на анализ на временните редове с помощта на стохастични GPS модели, вземащи предвид периодичността (по часове) и зависимостта на праховите частици PM10 от 6 метеорологични променливи.
5) Анализ и оценка на получените модели.

Прилагане на моделите за краткосрочно прогнозиране.
Използвани данни и описание на региона

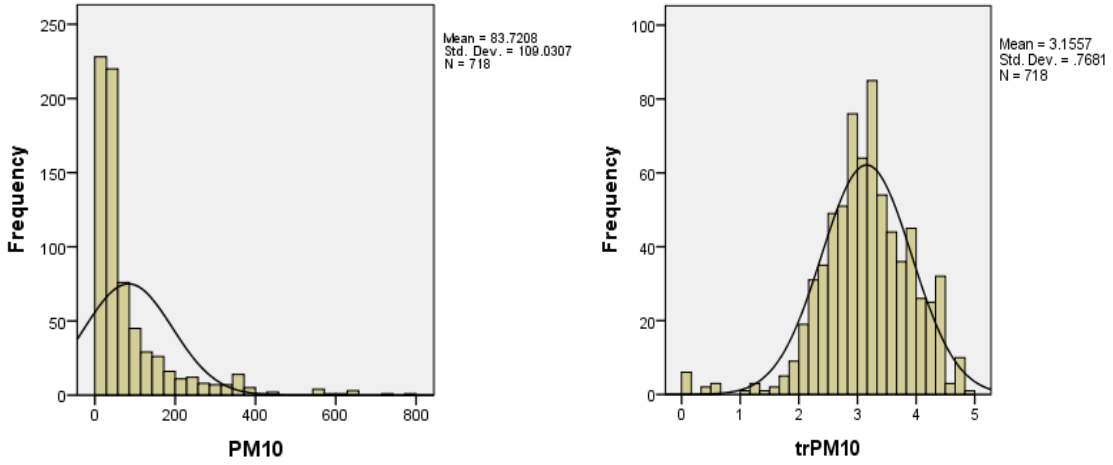
В настоящата глава са анализирани данните за концентрациите на прахови частици (PM10) като един от основните замърсители в град Шумен в периода от 28.11.2012 г. до 27.12.2012 г., вземайки предвид почасови данни. Данните са изразени в единици масова концентрация на PM10 в $\mu g/m^3$. Следните 6 метеорологични променливи също бяха разгледани: атмосферно налягане (PRESS), посока на вятъра (DVP), температура на въздуха на нивото на земята (TEMP), слънчева радиация (RADST), относителна влажност (UMR) и скорост на вятъра (VVG). Общият обем данни е от 718 до 816 случая, по часове. Максималният брой липсващи стойности (за всеки параметър) е по-малък от 1% и е без съществено значение. Таблица 4.1 показва основната описателна статистика за 7-те наблюдавани редове, приемайки, че замърсителят PM10 е зависима променлива, а шестте метеорологични променливи са независими.

Таблица 4.1. Описателна статистика за наблюдаваните концетрации на PM10 и 6 метеорологични данни за град Шумен. Последният ред показва трансформираната променлива на PM10.

Променлива	Брой	Мин.	Макс.	Средно	Стд. Откл.	Асимет рия	Стд. гр.аси метри я	Ексцес	Стд. гр. ексцес
*PM*10, μg/m^3	718	0.0	797.5	83.7	109.0	2.845	0.091	9.895	0.182
PRESS, hPa	816	956.6	994.5	978.4	8.2	-0.035	0.086	-0.621	0.171
DVP, degrees	816	78.2	285.9	190.5	43.3	-0.278	0.086	-0.829	0.171
TEMP, ^{o}C	816	-11.4	17.7	-0.5	5.6	0.746	0.086	0.624	0.171
RADST, W/m^2	816	9.7	484.8	70.7	98.8	2.084	0.086	3.461	0.171
UMR, %	816	27.8	93.1	71.3	12.4	-0.668	0.086	-0.096	0.171
VVG, m/s	816	1.3	5.9	2.5	1.0	1.034	0.086	0.133	0.171
trPM10	718	0.0	4.87	3.2	0.77	-0.675	0.091	1.987	0.182

Трансформация на данните

За тази цел най-напред използваме отново степенната трансформация на Йео-Джонсън [100] (виж Глава 1, формула (1.1)). Определената задоволителна стойност на ламбда за PM10, при която разпределението на данните е най-близко до нормалното с най-ниска вариация е $\lambda_{YJ} = -0.1$. Съответните статистически индекси за trPM10 са дадени на последния ред на Таблица 4.1. Трансформираните данни са означени като trPM10. От фигури 4.2 и 4.3 се вижда значителното подобрение в разпределението на данните.



Фиг. 4.2-3. Честотни таблици на разпределнието на PM_{10} и $trPM_{10}$

Резултати от SARIMA модела

SARIMA(2,1,1)(1,0,1)$_{24}$ моделът беше избран за трансформирания ред на PM10. Диференциращият член от първа степен (d = 1) беше намерен от обобщения тест на Дики-Фулър с помощта на софтуера EViews [33]. За определянето на предварителните входящи параметри (несезонни, сезонни и циклични), бяха изчислени и оценени над 50 кандидат модела. Беше установено, че входните величини PRESS, RADST и DVP не оказват значително влияние върху стойностите на trPM10. Статистическите индекси на избрания най-добър модел са дадени в Таблица 4.2.

Таблица 4.2. Обощени статистически данни на избрания SARIMA модел $(2,1,1)(1,0,1)_{24}$.[(*)]

Про мен лива	Статистики на приближението на модела					Ljung-Box Q(18)			Outlie rs
	Стац. R^2	R^2	RMSE	MAE	Норм. BIC	Stat.	DF	Sig.	
trPM10	0.511	0.952	0.170	0.123	-3.320	24.6	13	0.026	14

[(*)] The predictor parameters (*p*, *d*, *q*) are as follows: *TEMP* (0,2,0), *UMR*(0,1,0), (0,1,0), *VVG natural Log* (0,1,0), *PRESS natural Log* (0,0,0) and *RADST* (0,0,0), *DVP* was omitted. All other predictor parameters are 0.

Резултати от GPS-RuleLearner модел за PM10

приложение на обобщената регуляризираща GPS регресия, съчетана с методи за интелигентен анализ на данни. Построените модели се използват за предсказване на бъдещи замърсявания в кратък времеви период и се сравняват с реални измервания, които не са включени в моделите. GPS модели са построени за трансформирания временен ред trPM10 и 6 входящи метеорологични реда. Приложен е GPS в комбинация със седем предсказващи техники. Първо без изведен първоначалният регресионен модел на trPM10 и наборът от предиктори с помощта на TreeNet. След което са построени GPS моделите в комбинация с допълнителни SPM техники. Таблица 4.4 показва получените най-добри модели и основните им характеристики. Избран е GPS-RuleLearner_RawPredictors, който се представя по-добре от всички други модели по отношение на R^2. Съответната еластичност е хребетна Lasso еластичност 1.1.

Показатели на построеният GPS модел за трансформираните PM10 (trPM10) данни.

Модел	Обуч. R^2	Обуч. бр. коеф.	% комп.	Тест R^2	Тест бр. коеф.	% комп.	Еластичност
RuleLearner_RawPredictors	**0.96**	208	89.4%	0.90	208	89.6%	Ridged Lasso (1.1)

Индексите за статистическа адекватност на най-добрия GPS модел са дадени в Таблица 4.5.

Таблица 4.5. **К**ачеството на съвпадение за обучителната и тестовата извадка на избрания GPS модел RuleLearner_RawPredictors за trPM10.

Критерий	Обучителна извадка	Тестова извадка
AIC	-1558.5	-10.2
BIC	-793.6	517.3
RMSE	0.163	0.254
MAD	0.125	0.193
MRAD	0.045	0.060
MSE	0.028	0.053
R^2	**0.959**	**0.898**

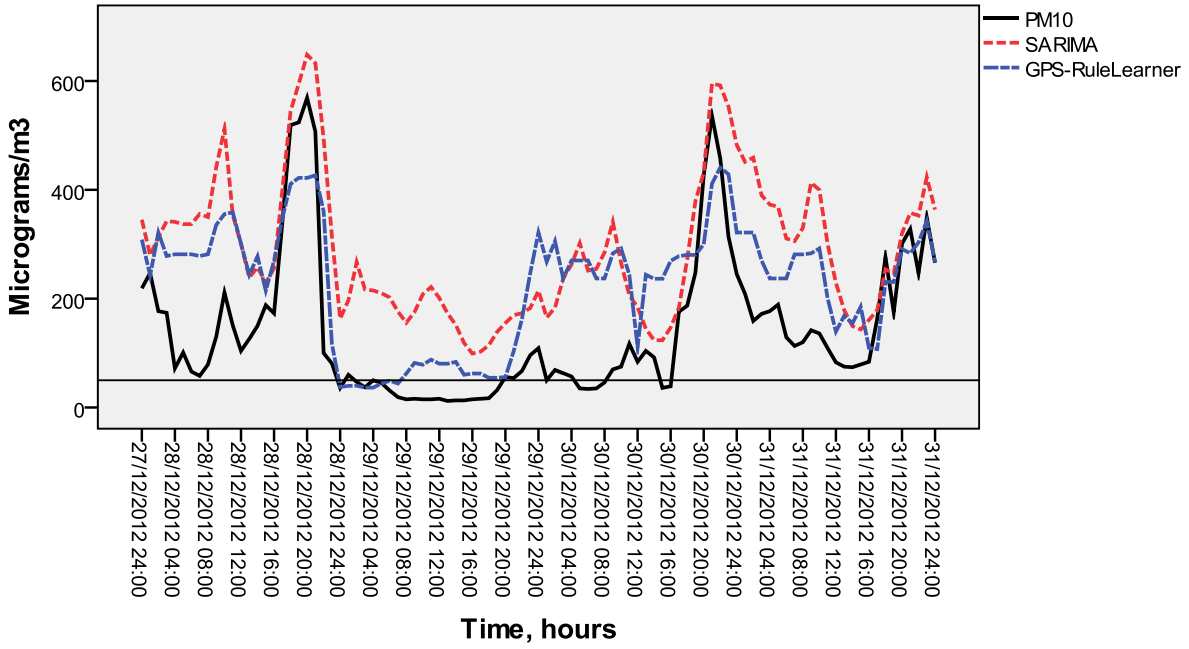
В таблица 4.6 са дадени относителните коефициенти на влияние на съответните предсказващи променливи в общия регресионен модел. Най-силно е влиянието на променливата VVG (натурален логаритъм) – скорост на вятъра (100 относителни единици).

Таблица 4.6. Относително влияние на променливите в модела RuleLearner_RawPredictors.

Променлива	Относително тегло (принос) в модела
VVG	100
DAYHOUR	64.90672
PRESS	49.65270
HOUR_A	40.91189
TEMP	22.40314
UMR	19.00693
DAY	14.15690
DVP	10.99046
RADST	2.06065

Сравнение на прогнозните резултати от SARIMA и GPS-RuleLearner моделите

резултатите, получени със SARIMA и GPS-RuleLearner моделите за краткосрочно прогнозиране на 96 часов (4-дневен) бъдещ период, който не е използван в процеса на моделиране. За тази цел, метеорологичните параметри за следващите 96 часа са добавени до n = 816 времеви точки, а предсказаните стойности или концентрациите на PM10 са изчислени с помощта на моделите. Предсказванията, направени чрез трите модела са сравнени с измерените концентрации на PM10 и показани на Фигура 4.8. Наблюдава се като цяло много добро качество на предсказването и по двата подхода – със SARIMA и с GPS регресията, съчетана с предсказващите методи. Трябва да се отбележи доброто предвиждане на пиковете стойности на замърсяване с PM10 (на 28 и 30 декември 2012 г. вечерта), като SARIMA моделът дава най-високи (надвишаващи реалните) стойности, а GPS моделът дава предсказвания под реалните стойности. По-слаби резултати показват и двата модела за ниските стойности (в частност – в средата на наблюдавания интервал), като отново известно преимущество показва GPS моделът.



Фиг. 4.8. Предсказаните и реалните стойности на PM10 концентрацията в град Шумен за 96 часов период. Хоризонталната черта указва допустимия среднодневен лимит от 50 μg/m^3.

Изводи към Глава 4

Направено е сравнително статистическо проучване с 2 метода на концентрацията на прахови частици PM10 в град Шумен за период от 1 година на базата на почасови данни и с отчитане на 6 метеорологични фактори. Построени са стохастичен модел по метода SARIMA и регресионен модел с GPS регресия, съчетана с предсказващи техники за интелигентен анализ на данни. Получените емпирични модели показват много добри качества на съвпадане със съществуващите данни. Приложени са за 4-дневно предсказване на бъдещи замърсявания, като отново дават много добри резултати. Може да се заключи, че двата метода са перспективни за краткосрочни прогнози и превенция от замърсяване на чистотата на въздуха за малки и средни населени места. Вторият подход се използва за първи път в литературата в областта на екологията.

ЗАКЛЮЧЕНИЕ

Главният извод, който можем да направим е, че целта на настоящия дисертационен труд и поставените задачи са постигнати. Връзката между основните получени резултати, поставената цел и задачи, приносите и публикациите на дисертационния труд са дадени в следната таблица.

Таблица. Връзка на основните резултати с целите и публикациите.

Принос	Цел	Задача	Параграф	Публикации
1	**1**	**1**	**2.1**	**[P1]**
2	**1**	**2**	**2.2.1, 2.2.2, 2.2.4**	**[P2]**
3	**1**	**3**	**2.2.1, 2.2.2, 2.2.3**	**[P2]**
4	**1**	**4**	**3.2.1, 3.2.2, 3.2.3**	**[P3]**
5	**1**	**5**	**3.2.1, 3.2.2, 3.2.4, 3.2.5, 3.2.6**	**[P3]**
6	**1**	**6**	**4.1, 4.2, 4.3**	**[P4]**

Перспективи за бъдеща работа

Разработените методи и подходи могат да се прилагат на сходни временни редове от областта на екологията и за други населени места. Бъдещите изследвания следва да бъдат насочени за към:

- Математическо моделиране на чистотата на въздуха за големи региони (части от България) и определяне на възможни общи тенденции и различия.
- Прилагане на нови методи, вкл. методи за машинно обучение, методи на невронни мрежи и др.
- Провеждане на анализи на замърсители, които показват най-големи превишавания спрямо официалните национални и европейски законови норми.

Апробация

Резултатите от дисертацията са част от изследванията на следните проекти:

1. Научен проект НИ11 ФМИ-004 „Разработка и приложение на иновативни ИКТ за провеждане на качествени конкурентоспособни научни изследвания и цялостно осъвременяване процеса на обучение във ФМИ“, Фонд „Научни изследвания“ при ПУ „П. Хилендарски“, 2011/2012 г.;
2. Научен проект НИ13 ФМИ-002/19.03.2013 „Интеграция на ИТ в научните изследвания по математика, информатика и педагогика на обучението“, Фонд „Научни изследвания“ при ПУ „Паисий Хилендарски“, 2013/2014 г.

Част от резултатите са докладвани на следните конференции и семинари:

[D1] Fifth Conference of the Euro-American Consortium for Promoting the Application of Mathematics in Technical and Natural Sciences (AMiTaNS'2013), Albena, Bulgaria, June 24-29, 2013.
[D2] Fourth Conference of the Euro-American Consortium for Promoting the Application of Mathematics in Technical and Natural Sciences, (AMiTaNS'12), St.St. Constantine and Helena, Varna, Bulgaria, June 11-16, 2012.
[D3] IX Национална научно-техническа конференция с международно участие “ЕКОЛОГИЯ И ЗДРАВЕ 2012”, Пловдив, 17 май 2012.
[D4] Семинар по Изчислителна статитика, ФМИ на ПУ, 16.10.2012.
[D5] Семинар по Изчислителна статитика, ФМИ на ПУ, 10.05.2012.
[D6] Семинар по Изчислителна статитика, ФМИ на ПУ, 26.04.2012.
[D7] Семинар по Изчислителна статитика, ФМИ на ПУ, 10.04.2012.

Резюме на получените резултати

1) За моделиране на концентрациите на PM10 и SO2 в 4-годишен период за почасови данни за град Кърджали са построени и изследвани едномерни стохастични SARIMA модели. Получено е добро съвпадение с реалните данни и прогнози за 14-дневен бъдещ период.
2) Построени са и са изследвани стохастични ARIMA модели на 6 основни замърсителя на въздуха на град Шумен (NO2, NO, O3, PM10, SO2, NOx) на почасови данни от период от 15 месеца с отчитане на 6 метеорологични фактора. Моделите са приложени за предсказване.
3) Установено е групиране на 6 замърсителя на въздуха на град Шумен в 4 групи с помощта на факторен анализ, дължащи се на общи източници на замърсяването. Построени са модели с факторните променливи.
4) Построени са и са изследвани голям брой едномерни стохастични SARIMA модели за 6 основни замърсители на въздуха на град Благоевград. Чрез трансформация на данните е подобрено разпределението към нормалното. С процедура за избор на оптимални модели са получени много добри резултати в съвпадение с реалните данни, а също и предсказването напред във времето.
5) С помощта на факторен анализ е получено групиране на 6-те изследвани замърсители на град Благоевград в 3 основни фактора, за идентификация на общите източници на замърсяване.
6) За нивото на замърсяване с PM10 на град Шумен са построени, анализирани и сравнени голям брой модели с помощта на 2 метода: SARIMA с метеорологични данни и GPS регресия, комбиниран с предказващи техники за интелигентен анализ на данни. Получените модели имат много добри характеристики. Те са приложени за 96 часови прогнози напред във времето, като са получени отлични предсказващи характеристики. Методът на GPS регресия е приложен за пръв път за моделиране на екологични данни в литературата.

Благодарности

Изказвам своята благодарност и признателност към научните ми ръководители проф. д-р Снежана Гочева-Илиева и доц. д-р Дойчин Бояджиев за тяхната постоянна ангажираност към моята работа и за всеотдайната им подкрепа. Благодаря на ръководството на Факултета по математика и информатика, в лицето на декана проф. д-р Асен Рахнев, за съдействието и подкрепата в процеса на цялостната ми работа. Изказвам благодарност и на членовете на катедра „Приложна математика и моделиране“, които по един или друг начин са били съпричастни към обучението ми като докторант и са ми оказвали помощ при подготовката на този труд.

Публикации по дисертационния труд

[P1] Иванов А., Д. Войникова, С. Гочева-Илиева, Д. Бояджиев, Времеви анализ на ежедневните замърсители на въздуха в град Кърджали, Сб. трудове на IX Национална научно-техническа конференция с международно участие “ЕКОЛОГИЯ И ЗДРАВЕ 2012”, 17 май 2012, Акад. изд. АУ, Пловдив, стр. 505-510, 2012.

[P2] Ivanov A., D. Voynikova, S. Gocheva-Ilieva, D. Boyadzhiev, Parametric time series analysis of daily air pollutants of city of Shumen, Bulgaria”, Proceedings of the Fourth Conference of the Euro-American Consortium for Promoting the Application of Mathematics in Technical and Natural Sciences, St.St. Constantine and Helena, Varna, Bulgaria, June 11-16, 2012 (AMiTaNS'12), Memorial Volume devoted to Prof. Christo I. Christov, Ed. M. D. Todorov, American Institute of Physics, *AIP Conf. Proc.* vol. 1487, pp. 386-396; ISBN: 978-0-7354-1099-2. DOI: http://dx.doi.org/10.1063/1.4758982. **SCImagoJR=0.168**

[P3] Gocheva-Ilieva S. G., A. V. Ivanov, D. S. Voynikova, D. T. Boyadzhiev, Time Series Analysis and Forecasting for Air Pollution in Small Urban Area: an SARIMA and Factor Analysis Approach, *Stochastic Environmental Research and Risk Assessment*, vol. 28 (4), pp. 1045-1060, 2014, Springer, DOI: 10.1007/s00477-013-0800-4, ISSN 1436-3240. **Impact Factor = 2.673 (2014)**

[P4] Ivanov A. V., S. G. Gocheva-Ilieva, Short-Time Particulate Matter PM10 Forecasts Using Predictive Modeling Techniques, Proc. of the Fifth Conference of the Euro-American Consortium for Promoting the Application of Mathematics in Technical and Natural Sciences (AMiTaNS'2013), Albena, Bulgaria, June 24-29, 2013. Ed. M. Todorov, American Institute of Physics, *AIP Conf. Proc*. vol. 1561, pp. 209-218, 2013. ISBN: 978-0-7354-1189-0, http://dx.doi.org/10.1063/1.4827230. **SCImagoJR = 0.162**

Библиография

[1] Abdul-Wahab S. A., C. S. Bakheit, S. M. Al-ALawi, Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. Environ Modell & Softw, 20(5), 1263-1271 (2005). ISSN 1364-8152.

[2] Agirre-Basurko E., G. Ibarra-Berastegi, I. Madariaga, Regression and multilayer perceptron-based models to forecast hourly O3 and NO2 levels in the Bilbao area, Environmental Modelling & Software, 21 (4), 430–446, April 2006. ISSN 1364-8152.

[3] Air Pollution, GreenFacts (Particulate Matter, Nitrogen Dioxide, Ozone, Nitrogen oxides (NOx)), http://www.greenfacts.org/en/digests/air-pollution.htm, http://www.greenfacts.org/en/particulate-matter-pm/, http://www.greenfacts.org/en/nitrogen-dioxide-no2/index.htm,

[4] Air Pollution, World Health Organization, http://www.who.int/topics/air_pollution/en/

[5] Ambient (outdoor) air pollution in cities database 2014, WHO, http://www.who.int/phe/health_topics/outdoorair/databases/cities/en/

[6] Air Quality Standards. European Commission. Environment. http://ec.europa.eu/environment/air/quality/standards.htm. 2014.

[7] Akaike H., A new look at the statistical model identification, IEEE Transactions on Automatic Control, 19 (6), 716-723 (1974). ISSN 0018-9286.

[8] Al-Madfai H., A. J. Geens, D. G. Snelson, Modelling the Multi-year Maximum Daily PM10 Concentration in Edinburgh: an Application of the Variability Decomposition Transfer Function Model. In Air Pollution XVIII, eds. C. A. Brebbia, J. W. S. Longhurst, WIT Transactions on Ecology and the Environment, WIT Press, vol. 136, 349-356 (2010). doi: 10.2495/AIR100311. ISSN 1743-3541.

[9] Azid A., H. Juahir, M. E. Toriman et al., Prediction of the Level of Air Pollution Using Principal Component Analysis and Artificial Neural Network Techniques: a Case Study in Malaysia, Water, Air, & Soil Pollution, Springer Intern. Publishing, pp. 225-2063, July 2014. DOI 10.1007/s11270-014-2063-1. Print ISSN 0049-6979, Online ISSN 1573-2932.

[10] Blagoevgrad, Wikipedia. http://en.wikipedia.org/wiki/Blagoevgrad 2014.

[11] Blifford I. H. Jr, G. O. Meeker, A factor analysis model of large scale pollution. Atmos Environ, 1, 147–157 (1967). doi:10.1016/0004-6981(67)90042-X. ISSN 1352-2310.

[12] Box G. E. P., D. R. Cox, An analysis of transformations. J Royal Stat Soc, Ser B, 26, 211–252, 1964. Wiley, ISSN 0035-9246.

[13] Box G. E. P., G. M. Jenkins, Time series analysis, forecasting and control, revized edn. Holden Day, San Francisco, 1976. ISBN 0816211043.

[14] Box G. E. P., G. M. Jenkins, G. S. Reinsel, Time Series Analysis, Forecasting and Control, 3rd edn., Prentice-Hall, Inc., New Jersey, 1994. ISBN 0130607746.

[15] Brunelli U., V. Piazza, L. Pignato, F. Sorbello, S. Vitabile, Two-days ahead prediction of daily maximum concentrations of SO2, O3, PM10, NO2, CO in the urban area of Palermo, Italy. Atmos Environ, 41(11), 2967–2995 (2007). doi:10.1016/j.atmosenv.2006.12.013. ISSN 1352-2310.

[16] Burnham K. P., D. R. Anderson, Model selection and inference: a practical information-theoretic approach, 2nd edn. Springer, New York, 2002. ISBN-13: 978-0387953649

[17] Chan, T. W. and M. Mozurkewich, Application of absolute principal component analysis to size distribution data: identification of particle origins, Atmos. Chem. Phys. 7, 887–897 (2007). ISSN 1680-7316.

[18] Chatfield C. Time-series forecasting. Chapman & Hall CRC, Boca Raton, 2000. ISBN 1584880635.

[19] Chatfield C., The analysis of time series: an introduction. Chapman & Hall/CRC, Boca Raton, 1996. ISBN 0-412-71640-2

[20] Comparison of statistical packages, http://en.wikipedia.org/wiki/Comparison_of_statistical_packages, посетен на 18.12.2014 г.

[21] Computation with SOLO Power Analysis. BMDP Statistical Software Inc., Los Angeles, 1993.

[22] Davidson R., J.G. MacKinnon, Estimation and Inference in Econometrics, Oxford, 1993. ISBN 0-19-506011-3.

[23] Di'az-Robles L. A., J. C. Ortega, J. S. Fu, G. D. Reed, J. C. Chow, J. G. Watson, J. A. Moncada-Herrera, A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: the case of Temuco, Chile. Atmos Environ, 42(35), 8331-8340 (2008). doi:10.1016/j.atmosenv.2008.07.020 ISSN 1352-2310.

[24] Dickey, D. A., W. A. Fuller, Distribution of the Estimators for Autoregressive Time Series with a Unit Root, Journal of the American Statistical Association, 74(366a), 427-431 (1979). DOI: 10.1080/01621459.1979.10482531. ISSN 0162-1459.

[25] Dimov I., R. Georgieva, S. Ivanovska, T. Ostromsky, Z. Zlatev, Studying the sensitivity of pollutants' concentrations caused by variations of chemical rates. J Comput Appl Math, 235, 391-402 (2010). doi:10.1016/j.cam.2010.05.041 ISSN 0377-0427.

[26] Dimov I., R. Georgieva, T. Ostromsky, Z. Zlatev, Advanced algorithms for multidimensional sensitivity studies of large-scale air pollution models based on Sobol sequences. Computers and Mathematics with Applications, 65(3), 338-351 (2013). doi:10.1016/j.camwa.2012.07.005 ISSN 0898-1221.

[27] Directive 2008/50/EC of the European Parliament and of the council of 21 May 2008 on ambient air quality and cleaner air for Europe, Official Journal of the European Union, L 152/1, (2008). http:// eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:152: 0001:0044:EN:PDF. 2014.

[28] Donev E, K Zeller, A Avramov, Preliminary background ozone concentrations in the mountain and coastal areas of Bulgaria, Environmental Pollution, 117(2) 281-286 (2002). ISSN 0269-7491.

[29] Dueñas C., M. C. Fernández, S. Cañete, J. Carretero, E. Liger, Stochastic model to forecast ground-level ozone concentration at urban and rural areas. Chemosphere, 61(10), 379-1389 (2005). Doi: 10. 1016/j.chemosphere.2005.04.079 ISSN 0045-6535.

[30] EEA Daily Bulletin for air quality in the country. EEA— Executive environment agency, National system for realtime air quality control in Bulgaria. http://pdbase.government.bg/airq/ bulletin-en.jsp.

[31] EEA Reports, Bulletins. EEA—Executive environment agency. http://eea.government.bg/en/output/index.html. 2014

[32] EEA (European Environment Agency): Particulate matter (PM10): annual mean concentrations in Europe. http://www.eea.europa.eu/themes/air/interactive/pm10. 2014.

[33] EViews for Windows, http://www.eviews.com, посетен на 18.12.2014 г.

[34] Friedman J. H., B. E. Popescu, Importance sampled learning ensembles. Stanford University, Department of Statistics. Technical Report, (2003). http://www-stat.stanford.edu/~jhf/ftp/isle.pdf

[35] Friedman J. H., B. E. Popescu, Predictive Learning Via Rule Ensembles, (2005). http://www-stat.stanford.edu/~jhf/ftp/RuleFit.pdf, 2014.

[36] Friedman J. H., Fast sparse regression and classification. Stanford University, Technical Report, 2008; and Int J Forecasting, 28(3), 722-738 (2012). Elsevier, http://dx.doi.org/10.1016/j.ijforecast.2012.05.001. ISSN: 0169-2070.

[37] Friedman J. H., Greedy function approximation: A gradient boosting machine. 1999 Reitz Lecture. Annals of Statistics, 29(5), 1189-1232 (2001). ISSN 0090-5364.

[38] Ganev K. D. Syrakov, M. Prodanova, E. Atanasov, T. Gurov, A. Karaivanova, N. Miloshev, H. Chervenkov, Grid Computing for Air Quality and Environmental: Studies in Bulgaria, EnviroInfo 2009 (Berlin), Environmental Informatics and Industrial Environmental Protection: Concepts, Methods and Tools, Shaker Verlag, pp. 147-155, 2009. ISBN: 978-3-8322-8397-1

[39] Gardner M. W., S. R. Dorling, Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. Atmos Environ, 33(5), 709-719 (1999). doi:10.1016/S1352- 2310(98)00230-1. ISSN 1352-2310.

[40] Gocheva-Ilieva S. G., A. V. Ivanov, D. S. Voynikova, D. T. Boyadzhiev, Time Series Analysis and Forecasting for Air Pollution in Small Urban Area: an SARIMA and Factor Analysis Approach, Stochastic Environmental Research and Risk Assessment, vol. 28 (4), 1045-1060 (2014), Springer, DOI: 10.1007/s00477-013-0800-4, ISSN: 1436-3240 (Print) 1436-3259 (Online) http://link.springer.com/article/10.1007/s00477-013-0800-4

[41] Gocheva-Ilieva S. G., Application of the Generalized Pathseeker regularized regression, Mathematics and Education in Mathematics, Rusev P. (Ed.), vol. 2014, Proc. 43th Spring Conference of the Union of Bulgarian Mathematicians, 2-6 April 2014, Borovetz, pp. 34-43, 2014. ISSN 1313-3330

[42] Gocheva-Ilieva S. G., Using Generalized PathSeeker Regularized Regression for Modeling and Prediction of Laser Generation, (submitted).

[43] He, H. D., W. Z.Lu, Y.Xue, Prediction of PM10 concentrations at urban traffic intersections using semi-empirical box modelling with instantaneous velocity and acceleration, Atmospheric Environment, 43, 6336–6342 (2009). ISSN 1352-2310.

[44] He H. D., W. Z. Lu, Y. Xue, Prediction of particulate matter at street level using artificial neural networks coupling with chaotic particle swarm optimization algorithm, Building and Environment, 78, 111-117 (2014). ISSN 0360-1323.

[45] Henry R. C., G. M. Hidy, Multivariate analysis of particulate sulfate and other air quality variables by principal components—part I: annual data from Los Angeles and New York. Atmos Environ 13(11), 1581-1596 (1979). doi:10.1016/0004-6981(79)90068-4. ISSN 1352-2310.

[46] Hoerl A. E., R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics, 42 (1), 80-86 (1970). doi:10.1080/00401706.1970.10488634. ISSN 0040-1706.

[47] http://en.wikipedia.org/wiki/ISO/IEC_17025, посетен на 18.12.2014 г.

[48] http://www.salford-systems.com/products/spm, посетен на 18.12.2014 г.

[49] Ivanov A., D. Voynikova, S. Gocheva-Ilieva, D. Boyadzhiev, Parametric time series analysis of daily air pollutants of city of Shumen, Bulgaria", Proceedings of the Fourth Conference of the Euro-American Consortium for Promoting the Application of Mathematics in Technical and Natural Sciences, St.St. Constantine and Helena, Varna, Bulgaria, June 11-16, 2012. AMiTaNS '12 Memorial Volume devoted to Prof. Christo I. Christov, Ed. M. D. Todorov, American Institute of Physics AIP Conf. Proc. vol. 1487, pp. 386-396; ISBN: 978-0-7354-1099-2. doi: http://dx.doi.org/10.1063/1.4758982, SCImagoJR=0.168.

[50] Ivanov A. V., S. G. Gocheva-Ilieva, Short-Time Particulate Matter PM10 Forecasts Using Predictive Modeling Techniques, Proc. of Fifth Conference of the Euro-American Consortium for Promoting the Application of Mathematics in Technical and Natural Sciences (AMiTaNS'2013), Albena, Bulgaria, June 24-29, 2013. Ed. M. Todorov, American Institute of Physics AIP Conf. Proc. vol. 1561, pp. 209-218, 2013. ISBN: 978-0-7354-1189-0, http://dx.doi.org/10.1063/1.4827230 SCImagoJR: 0.162.

[51] Izenman, A., Modern multivariate statistical techniques, Springer, 2008.

[52] Jacobson M. Z., Fundamentals of atmospheric modeling, 2nd edn. Cambridge Univ. Press, Cambridge, 2005. ISBN 0521548659, 9780521548656.

[53] Jian L., Y. Zhao, Y. P. Zhu, M. B. Zhang, D. Bertolatti, An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China. Sci Total Environ, 426, 336-345 (2012). doi:10.1016/j.scitotenv.2012.03.025. ISSN 0048-9697.

[54] Jolliffe I. T., Principal component analysis, 2nd edn. Springer, New York, 2002. ISBN 387-95442-2.

[55] Kaplunovsky A. S., Factor analysis in environmental studies. HAIT J Sci Eng B 2(1-2), 54-94 (2005). ISSN 1565-4990.

[56] Kim J., C. W. Mueller, Factor analysis: statistical methods and practical issue. Sage, Beverly Hills, 1986. ISBN-10: 0803911661.

[57] Kim S. E., A. Kumar, Accounting seasonal nonstationarity in time series models for short-term ozone level forecast. Stoch Environ Res Risk Asses, 19, 241-248 (2005). doi:10.1007/s00477-004-0228-y. ISSN: 1436-3240.

[58] Kim S. E., Tree-based threshold modeling for short-term forecast of daily maximum ozone level. Stoch Environ Res Risk Asses, 24(1), 19-28 (2010). doi:10.1007/s00477-008-0295-6. ISSN: 1436-3240.

[59] Kumar R., A. E. Joseph, Air pollution concentrations of PM2.5, PM10 and NO2 at Ambient and Kerbsite and their correlation in Metro City—Mumbai. Environ Monit Asses, 119(1-3), 191-199 (2006). doi:10.1007/s10661-005-9022-7. ISSN 0167-6369.

[60] Kumar U., V. K. Jain, ARIMA forecasting of ambient air pollutants (O3, NO, NO2 and CO). Stoch Environ Res Risk Asses, 24, 751-760 (2010). Doi:10.1007/s00477-009-0361-8. ISSN: 1436-3240.

[61] Lengyel A., K. Heberger, L. Paksy, O. Banhidi, R. Rajko, Prediction of ozone concentration in ambient air using multivariate methods. Chemosphere, 57(8), 889-896 (2004). Doi:10.1016/j. chemosphere.2004.07.043. ISSN 0045-6535.

[62] Liddle A. R. Information criteria for astrophysical model selection. Cornell University Library, arHiv.org. (2008). http://arxiv.org/PS_ cache/astro-ph/pdf/0701/0701113v2.pdf .

[63] Lima E. A. P., E. C. Guimaraes, S. A. Pozza, M. A. S. Barrozo, J. R. Coury, A Study of Atmospheric Particulate Matter in a City of the Central Region of Brazil Using Time-series Analysis. Int J Environ Eng 1(1), 80-94 (2009). doi: 10.1504/IJEE.2009.026444. ISSN online: 1756-8471, print: 1756-8463.

[64] List of atmospheric dispersion models, Wikipedia. http://en.wikipedia.org/wiki/List_of_atmospheric_dispersion_models

[65] Liu P. W. G., Simulation of the Daily Average PM10 Concentrations at Ta-Liao with Box-Jenkins Time series Models and Multivariate Analysis. Atmos Environ, 43, 2104-2113 (2009). doi:10.1016/j.atmosenv.2009.01.055. ISSN 1352-2310.

[66] McBerthouex P., L. C. Brown, Statistics for environmental engineers. CRC Press, Boca Raton, 1994. ISBN 1420056638.

[67] Milionis A. E., T. D. Davies, Regression and stochastic models for air pollution. I. Review, comments and suggestions. Atmos Environ 28, 2801-2810 (2002). doi:10.1016/1352-2310(94)90083-3. ISSN 1352-2310.

[68] Milionis A. E., T. D. Davies, Box-Jenkins univariate modelling for climatological time series analysis: an application to the monthly activity of temperature inversions. Int J Climatol, 14, 569-579 (1994). ISSN 089908418.

[69] Muñoz E., M. L. Martin, I. J. Turias, M. J. Jimenez-Come, F. J. Trujillo, Prediction of PM10 and SO2 exceedances to control air pollution in the Bay of Algeciras, Spain. Stoch Environ Res Risk Asses, 3, 1409-1420 (2014). ISSN: 1436-3240 (Print) 1436-3259 (Online).

[70] Nagendra S. M. S., M. Khare, Artificial neural network approach for modelling nitrogen dioxide dispersion from vehicular exhaust emissions. Ecological Modelling, 190, 99-115 (2006). ISSN 0304-3800.

[71] National System for Environmental Monitoring, Bulgaria (2013). http://eea.government.bg/en/nsmos/index.html, 2014.

[72] Nisbet R., J. Elder, G. Miner, Handbook of Statistical Analysis and Data Mining Applications, Burlington, MA: Academic Press Elsevier Inc., 2009. ISBN 9780123747655.

[73] Pankratz A., Forecasting with univariate Box-Jenkins models: concepts and cases. Wiley, New York, 1983. ISBN-10: 0471090239.

[74] Petelin D., A. Grancharova, J. Kocijan, Evolving Gaussian process models for prediction of ozone concentration in the air, Simul Model Pract Th (EUROSIM 2010), 33, pp. 68-80, (2013). doi:http:// dx.doi.org/10.1016/j.simpat.2012.04.005. ISSN 1569-190X.

[75] Pires J. C. M., F. G. Martins, S. I. V. Sousa, M. C. M. Alvim-Ferraz, M. C. Pereira, Selection and validation of parameters in multiple linear and principal component regressions. Environ Modell & Softw, 23(1) 50-55 (2008). DOI: 10.1016/j.envsoft.2007.04.012. ISSN 1364-8152.

[76] Polydoras G. N., J. S. Anagnostopoulos, G. Ch. Bergeles, Air quality predictions: dispersion model vs Box-Jenkins stochastic models. An implementation and comparison for Athens, Greece. Appl Therm Eng, 18(11), 1037-1048 (1998). doi:10.1016/S1359-4311(98) 00016-7. ISSN 1359-4311.

[77] Prodanova M., J. L. Perez, D. Syrakov, R. San Jose, K. Ganev, N. Miloshev, S. Roglev, Application of mathematical models to simulate an extreme air pollution episode in the Bulgarian city of Stara Zagora, Applied Mathematical Modelling, 32(8) 1607-1619 (2008). ISSN 1307-904X.

[78] Richman M. B., Rotation of principal components. J Climatol, 6, 293-335 (1986). ISSN 0899-8418.

[79] Said E. S., D. A. Dickey, Testing for unit roots in autoregressive- moving average models of unknown order. Biometrika 71(3), 599-607 (1984). doi:10.1093/biomet/71.3.599. ISSN 0006-3444.

[80] Schwarz G., Estimating the dimension of a model, Annals of Statistics, 6, 461-464 (1978). doi:10.1214/aos/1176344136. ISSN 0090-5364.

[81] Sharma P., A. Chandra, S. C. Kaushik, Forecasts using Box- Jenkins models for the ambient air quality data of Delhi City. Environ Monit Assess, 157(1-4), 105-112 (2009). doi:10.1007/s10661- 008-0520-2. ISSN 0167-6369.

[82] Sillman S., The relation between ozone, NOx and hydrocarbons in urban and polluted rural environments. Atmos Environ, 33(12) 1821-1845 (1999). Doi: 10.1016/s1352-2310(98)00345-8. ISSN 1352-2310. http://www-personal.umich.edu/~sillman/web-publications/Sillmanreview99.pdf

[83] Simeonov P., L. Bocheva, T. Marinova, Severe convective storms phenomena occurrence during the warm half of the year in Bulgaria (1961–2006), Atmospheric Research, 93(1–3) 498-505 (2009). ISSN 0169-8095.

[84] Slini T., A. Kaprara, K. Karatzas, N. Moussiopoulos, PM10 Forecasting for Thessaloniki, Greece. Environ Modell Softw, 21(4), 559-565 (2006). http://dx.doi.org/10.1016/j.envsoft.2004.06.011. ISSN 1364-8152.

[85] Slini Th., K. Karatzas, N. Moussiopoulos, Statistical analysis of environmental data as the basis of forecasting: an air quality application. Sci Total Environ, 288, 227-237 (2002). doi:10.1016/S0048- 9697(01)00991-3. ISSN 0048-9697.

[86] Sousa S. I. V., F. G. Martins, M. C. M. Alvim-Ferraz, M. C. Pereira, Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. Environ Modell & Softw, 22(1), 97–103 (2007). 1364-8152.

[87] SPM Users Guide. Introducing Generalized PathSeeker, San Diego: Salford Systems, 2013.

[88] SPSS IBM Statistics, http://www-01.ibm.com/software/analytics/spss/products/statistics/, 2014 г.

[89] Statheropoulos M., N. Vassiliadis, A. Pappa, Principal component and canonical correlation analysis for examining air pollution and meteorological data, Atmos Environ. 32, 1087-1095 (1998). ISSN 1352-2310.

[90] Steinberg D., M. Golovnya, Introduction to Modern Regression: From OLS to GPS® to MARS® http://info.salford-systems.com/share-your-datamining-video-contest, August, 2013

[91] Steinberg D., M. Golovnya, Tree Ensembles and Extensions An Overview of TreeNet, RandomForests ISLE Model Compression & RuleLearner, http://info.salford-systems.com/jsm-2013-ctw2 2014.

[92] Steinberg D., M. Golovnya, Tree Ensembles: The Power of Post-Processing, San Diego, Salford Systems, 2012. http://media.salford-systems.com/video/tutorial/advances_in_gradient_boosting_the_power_of_post_processing.pdf, 2014.

[93] Syrakov D., K. Ganev, M. Prodanova, N. Miloshev, G. Jordanov, E. Katragkou, D. Melas, A. Poupkou, K. Markakis, Background Pollution Forecast over Bulgaria, Large-Scale Scientific Computing, Lecture Notes in Computer Science, I. Lirkov, S. Margenov, J. Waśniewski (eds.), Springer , Vol. 5910, pp. 531-537, 2010. Series ISSN 0302-9743, ISBN 978-3-642-12534-8.

[94] Tabachnik B. G., L. S. Fidell, Using multivariate statistics, 5th edn., Pearson Int. Edition, Boston, 2005. ISBN 0205459382.

[95] Tibshirani R., Regression shrinkage and selection via the lasso, J Royal Stat Soc B, 58 (1), 267-288 (1996). Wiley, ISSN 0035-9246.

[96] Ul-Saufie A. Z., A. S. Yahaya, N. A. Ramli, N. Rosaida, H. A. Hamid, Future daily PM10 concentrations prediction by combining regression models and feedforward backpropagation models with principle component analysis (PCA). Atmos Environ, 77, 621–630 (2013). http://dx.doi.org/10.1016/j.atmosenv.2013.05.017. ISSN 1352-2310.

[97] Vesely V., J. Tonner, Z. Hrdlivckova, J. Michalek, M. Kolar, Analysis of PM10 Air Pollution in Brno Based on Generalized Linear Model with Strongly Rank-deficient Design Matrix. Environmetrics, 20(6), 676-698 (2009). doi: 10.1002/env.971. ISSN 1180-4009.

[98] Wei W. W. S., Time Series Analysis, Univariate and Multivariate Methods. 2nd edn., Pearson Education, Boston, 2006. ISBN-10: 0321322169.

[99] Wolfram Mathematica, http://www.wolfram.com/mathematica/, посетен на 18.12.2014 г.

[100] Yeo I. K., R. A. Johnson, A new family of power transformations to improve normality or symmetry. Biometrika, 87(4), 954-959 (2000). doi:10.1093/biomet/87.4.954. ISSN 0006-3444.

[101] Yuanyuan C., S. Runhe , S. Shijie, G. Wei, Ensemble and enhanced PM10 concentration forecast model based on stepwise regression and wavelet analysis, Atmos Environ, 74, 346–359 (2013). http://dx.doi.org/10.1016/j.atmosenv.2013.04.002. ISSN 1352-2310.

[102] Zickus M., A. J. Greig, M. Niranjan, Comparison of Four Machine Learning Methods for Predicting PM10 Concentrations in Helsinki, Finland. Water, Air, and Soil Pollution: Focus 2, 717-729 (2002). doi: 10.1023/A:1021321820639. ISSN 0049-6979.

[103] Zou H., T. Hastie, Regularization and variable selection via the elastic net, J Royal Statist Soc B 67, 301-320 (2005). Wiley, ISSN 0035-9246.

[104] Zwozdziak A., L. Samek, I. Sowka, L. Furman, M. Skretowicz, Aerosol pollution from small combustors in a village. Scientific World Journal, Hindawi, 2012, Article ID 956401 (2012). doi:10.1100/2012/956401. ISSN 1537-744X.

[105] Вундев Д. Л., Записки по приложна статистика 1, СУ "Св. Кл. Охридски", С., 2003. http://www.fmi.uni-sofia.bg/fmi/statist/Personal/Vandev/lectures/applstat1.pdf, на 18.12.2014 г.

[106] Вундев Д. Л., Записки по приложна статистика 2, СУ "Св. Кл. Охридски", С., 2003. http://www.fmi.uni-sofia.bg/fmi/statist/Personal/Vandev/lectures/applstat2.pdf, на 18.12.2014 г.

[107] Гочева-Илиева С. Г., Въведение в система Mathematica, ЕксПрес, Габрово, 2009. ISBN: 978-954-490-079-3.

[108] Гочева-Илиева С. Г., Приложна математика, Университетско издателство „Паисий Хилендарски" (електронно издание), 2013. ISBN 978-954-423-847-6. http://fmi.uni-plovdiv.bg/gocheva/PM-lekcii.pdf, посетен на 18.12.2014 г.

[109] Иванов А., Д. Войникова, С. Гочева-Илиева, Д. Бояджиев, Времеви анализ на ежедневните замърсители на въздуха в град Кърджали, Сб. тр. IX Нац. научно-техн. конф. с межд. уч. "Екология и здраве", 17 май, 2012, Пловдив, стр. 505-510.

[110] Качество на атмосферния въздух, Нормативни актове. Министерство на околната среда и водите, http://www.moew.government.bg/?show=top&cid=119, на 18.12.2014 г.

[111] Кулина Хр. Н., М. В. Такев, Разработка на учебни материали по многомерен статистически анализ на данни от областта на маркетинга с помощта на софтуерния пакет SPSS, Научна сесия на СУБ - Пловдив, 2013. Научни трудове на СУБ-Пловдив, Серия В, Техника и технологии, том XI, стр. 52-59, 2014. ISSN 1311-9192.

[112] Магнус Я. Р., П. К. Катышев, А. А. Пересецкий, Эконометрика. Начальный курс, 8 изд., Москва, Дело, 2007. ISBN 978-5-7749-0473-0

[113] Програма за намаляване на нивата на замърсителите в атмосферния въздух и достигане на установените норми за вредните емисии в град Благоевград. http://eea.government.bg/bg/prev_nsmos/air/roukav/obshtini3/blagoevgrad.pdf

[114] Програма за намаляване на нивата на замърсителите в атмосферния въздух и достигане на установените норми за вредните емисии в град Кърджали. http://eea.government.bg/bg/prev_nsmos/air/roukav/obshtini2/KAV-Kurdjali.doc

[115] Програма за намаляване на нивата на замърсителите в атмосферния въздух и достигане на установените норми за вредните емисии в град Шумен. http://eea.government.bg/bg/prev_nsmos/air/roukav/obshtini2/KAV-Shumen.doc