Bulgarian Journal of Meteorology and Hydrology

Национален институт по метеорология и хидрология

National Institute of Meteorology and Hydrology

Bul. J. Meteo & Hydro 24/1 (2020) 18-39

Missing data imputation for hourly PM_{10} concentrations in Sofia stations

Nadya Neykova, Plamen Neytchev

National Institute of Meteorology and Hydrology,
Tsarigradsko shose 66, 1784 Sofia, Bulgaria

Abstract: In this paper several imputation methods for filling missing data of hourly PM_{10} concentrations for 5 stations in Sofia city for the period 01.01.2017 - 29.02.2020 are discussed and compared. The generalized additive models (GAMs) with gamma and Weibull distributions and standard methods such as mean value, last & next, running mean, linear interpolation are used. The mean absolute error (MAE), root mean squared error (RMSE), normalized root mean square error (NRMSE) and correlation coefficient are used in order to assess the quality of the imputation. The results show that GAMs methodology provides high quality data imputation.

Keywords: missing data imputation, PM_{10}, generalized additive models (GAMs), gamma and Weibull distributions

Възстановяване на липсващи часови данни за $ФПЧ_{10}$ за станции в София

Надя Нейкова*, Пламен Нейчев

Национален институт по метеорология и хидрология,
Бул. Цариградско шосе 66,1784 София, България

Резюме: В статията са разгледани и сравнени различни методи за възстановяване на липсващи часови стойности за ФПЧ10 за 5 станции в град София за периода 01.01.2017 г. - 29.02.2020 г. Използвани са обобщени адитивни модели (GAMs) с гама и Вейбул разпределения, както и стандартни методи за възстановяване на данни

* nadya.neykova@meteo.bg

като средна стойност, последна и следваща, плъзгаща средна стойност, линейна интерполация. Средната абсолютна грешка (MAE), средно квадратичната грешка (RMSE), нормираната средно квадратична грешка (NRMSE) и коефициентът на корелация са използвани за оценка на качеството на възстановяване. Получените резултати показват, че методологията, основана на GAMs, осигурява високо качество на възстановяване на липсващите данни.

Ключови думи: възстановяване на липсващи данни, $ФПЧ_{10}$, обобщени адитивни регресионни модели (GAMs), гама и Вейбул разпределения

1. ВЪВЕДЕНИЕ

Присъствието на завишени концентрации на атмосферни замърсители във въздуха е сериозен проблем поради вредните ефекти, които те оказват върху човешкото здраве, растителността и екосистемите. По тези причини ежедневно се извършват измервания в автоматичните измервателни станции (АИС) на системата за мониторинг на качество на атмосферния въздух (КАВ) на МОСВ. Те измерват атмосферни замърсители като фини прахови частици ($ФПЧ_{10}$ и $ФПЧ_{2.5}$), серен диоксид, азотен диоксид, въглероден оксид, озон и други. В град София има 6 пункта за контрол на качеството на атмосферния въздух - Надежда, Хиподрума, Дружба, Павлово, Младост, Копитото. В редиците от данни на замърсителите често се наблюдават липсващи данни. В някои случаи липсват данни за отделни часове, в други липсват за цели дни. Проблемите с липсващи стойности възникват почти във всяка област, където измерванията се съхраняват в бази от данни. Причините за това са различни, но често са свързани с повреда в измервателната апаратура. Липсващите стойности създават проблеми, понеже широко използваните статистически програмни процедури за анализ на данни не могат да работят, дори ако има една липсваща стойност в някой ред от матрицата на данните. Поради това данните се филтрират, за да останат само наблюденията без липсващи стойности. Филтрирането на наблюдения с липсващи стойности не е добра стратегия, тъй като размерността на данните може да бъде редуцирана значително. По-добър подход е липсващите стойности да бъдат възстановени (заменени) с подходящи стойности. Възстановяването на тези данни е важна и актуална задача, тъй като това ще подобри тяхното качество и съответно получените резултати при анализа им.

Настоящото изследване предлага възстановяване на липсващи часови данни за $ФПЧ_{10}$ за станциите Надежда, Хиподрума, Дружба, Павлово и Младост в град София за периода 01.01.2017 - 29.02.2020 г. чрез използване на обобщени адитивни регресионни модели (Generalized Additive Models - GAMs). За предиктори в моделите за всяка от станциите са използвани сплайн функции от часови данни за $ФПЧ_{10}$ от останалите софийски станции. Използвани са и други методи за възстановяване на липсващите стойности като средна стойност, последна и

следваща (last & next), плъзгаща средна стойност и линейна интерполация. Резултатите допринасят за подобряване качеството на данните за ФПЧ$_{10}$ и предоставят по-големи възможности при избор на станция или времеви период, които са от интерес за различни научни изследвания.

2. СЪСТОЯНИЕ НА ИЗСЛЕДВАНИЯТА

Обзор на съществуващите пакети за възстановяване на липсващи данни на времеви редици в програмната среда R са разгледани в Moritz and Bartz-Beielstein (2017), но голямата част от тези пакети се отнасят за едномерни и многомерни Гаусово разпределени времеви редове, докато разпределението на ФПЧ$_{10}$ не е Гаусово. Различни методи за възстановяване на данни за атмосферни замърсители са разгледани в статиите на Junninen et al. (2004), Junger and Leon (2015), Zakaria and Noor (2018), Sukatis et al. (2019). Тези методи са подходящи за едномерни и многомерни Гаусово или дискретни (бинарно и номинално) разпределени времеви редове. Поради липсата на универсален софтуер за възстановяване на ФПЧ$_{10}$ данни, които са дясно скосено разпределени, в настоящата статия се използват GAMs модели и софтуерни процедури на R за възстановяване на този тип данни.

3. МЕТОДИЧЕН ПОДХОД

В настоящото изследване са използвани GAMs модели, както и стандартните методи, широко използвани в практиката за възстановяване на данни като: аритметична средна стойност, последна и следваща (last & next), плъзгаща средна стойност, линейна интерполация. Резултатите от възстановяването на тези методи са сравнени чрез статистическите мерки: средна абсолютна грешка (MAE), средно квадратична грешка (RMSE), нормирана средно квадратична грешка (NRMSE) и коефициент на корелация.

3.1. GAMs модели

GAMs представляват широк клас от нелинейни Гаусови и негаусови непараметрични регресионни модели, чиито предиктори участват под формата на сплайн функции или друг тип непараметрични функции за разлика от класическия модел на множествена линейна регресия, в който предикторите участват линейно. Използването на сплайн функция от предиктор предоставя възможности за определяне на функционални зависимости между предиктант и предиктор, дори когато корелационният коефициент между тях не е статистически значим. Корелационният коефициент е мярка за линейна зависимост между две величини и отсъствието на линейна зависимост между тях не изключва съществуването

на строга функционална зависимост. Методологията на моделирането с GAMs се развива интензивно през последните 30 години. Наличието на софтуер за моделиране дава възможности за рутинно приложение в практиката на GAMs с различни непрекъснати и дискретни негаусови разпределения като гама, Вейбул, Стюдент, инверсно Гаусово, бинарно, Поасоново, полиномно и много други разпределения. В програмната среда R са предоставени повече от 30 пакета, реализиращи GAMs. За провеждане на необходимите пресмятания за определяне на оценки на GAMs моделите са използвани пакетите gamlss и mgcv, достъпни в програмната среда R (R Core Team, 2019). Оценяването на неизвестните параметри на GAMs моделите чрез тези пакети се основава на метода на максималното правдоподобие с пенализация. Повече подробности за методологията на моделиране с GAMs и тези пакети могат да бъдат намерени в Hastie and Tibshirani (1990), Stasinopoulus et al. (2017) и Wood (2017).

За постигане на целите на изследването са използвани гама и Вейбул разпределени GAMs модели, тъй като часовите данни за $ФПЧ_{10}$ се характеризират с дясно скосено разпределение. Използвани са плътностите на гама и Вейбул разпределенията със следната параметризация:

$$g(y,a,b)=\begin{cases}\dfrac{b^a\, y^{(a-1)}\exp(-by)}{\Gamma(a)}, & y>0\\ 0 & y=0,\end{cases}$$

$$w(y;a,b)=\begin{cases}\dfrac{a\, y^{(a-1)}\exp(-(y/b)^a)}{b^a}, & y>0\\ 0 & y=0,\end{cases}$$

където $a > 0$ и $b > 0$ са неизвестни константи (параметри).

Съгласно методологията на GAMs, връзката между стойностите на предиктанта $y(t)$ и предикторите $x_1(t), ..., x_p(t), z_1(t), ..., z_q(t)$ за $t=1, ..., T$ и неизвестните константи (параметри) a и b на гама и Вейбул разпределенията се осъществява чрез използване на експоненциалната функция като свързваща плътността с предикторите:

$$a_t=\exp\left(\alpha_0+\alpha_1x_1(t)+\cdots+\alpha_px_p(t)+f_1\left(x_{p+1}(t)\right)+\cdots+f_u\left(x_{p+u}(t)\right)\right)$$

$$b_t=\exp\left(\beta_0+\alpha_1z_1(t)+\cdots+\alpha_qz_q(t)+g_1\left(z_{q+1}(t)\right)+\cdots+g_v\left(z_{q+v}(t)\right)\right)$$

където, $\alpha_0,\cdots,\alpha_p,\beta_0,\cdots,\beta_q$ са неизвестни константи, $f_1,\cdots,f_u,g_1,\cdots,g_v$ са неизвестни функции, характеризиращи нелинейните зависимости между предиктанта и предикторите. По този начин във всеки момент на времето t се дефинира едно динамично разпределение, което е функция на параметри a_t и b_t, които са функции на динамичните предиктори. Използването на експоненциалната функция като свързваща предикторите с a_t и b_t гарантира положителност на a_t и b_t

при оценяването им по данните, вместо да се решава по-сложната оптимизационна задача с ограничението $a_t>0$ и $b_t>0$ за всеки момент t.

В общия случай, предикторите $z_1(t),\cdots,z_q(t)$ съвпадат или са подмножество на $x_1(t),\cdots,x_p(t)$. Някои от предикторите $x(t)$ и $z(t)$ могат да бъдат лагове на $y(t)$ - например $x_1(t)=y(t-1),\cdots,\ x_l(t)=y(t-l)$, когато данните са времеви редици. За отчитането на сезонни трендове на предиктанта $y(t)$ някои от предикторите могат да са елементите на краен ред на Фурие, например първата хармоника $x_1(t)=\sin(2\pi t/365.25),\ x_2(t)=\cos(2\pi t/365.25)$.

Понеже функциите $f_1,\cdots,f_r,g_1,\cdots,g_s$ са неизвестни, вместо тях се използват техни апроксимации, например сплайн функции с възли, които са равномерно разположени в дефиниционните интервали на съответните предиктори:

$$a_t=\exp\left(\alpha_0+\alpha_1x_1(t)+\cdots+\alpha_px_p(t)+s_1\left(x_{p+1}(t)\right)+\cdots+s_u\left(x_{p+u}(t)\right)\right)$$

$$b_t=\exp\left(\ \beta_0+\alpha_1z_1(t)+\cdots+\alpha_qz_q(t)+s_1\left(z_{q+1}(t)\right)+\cdots+s_v\left(z_{q+v}(t)\right)\right)$$

Пример за такива функции са регресионните кубични сплайн функции с k възела, които се дефинират като:

$$s_i(x)=\theta_0+\theta_1x+\theta_2x^2+\theta_3x^3+\sum_{l=1}^{k}\theta_j(x-\tau_l)_+^3,$$

където $\theta_0,\cdots,\theta_k$ са (k+4) неизвестни коефициенти (параметри), които се оценяват по данните, $\tau_1,\cdots,\tau_k$ са възлите на сплайна, удовлетворяващи условието $\tau_1<\cdots<\tau_k$ и

$$(x-\tau_l)_+^3=\begin{cases}(x-\tau_l)^3, & x>\tau_l\\ 0, & x\le\tau_l.\end{cases}$$

Привеждането на примера с регресионните кубични сплайн функции е продиктувано от чисто методичен характер, тъй като GAMs методологията се възприема по-лесно. От алгоритмична гледна точка на числените методи и софтуерна реализация се използват гладки сплайн функции, B, P, изотропни, тензорни или друг тип сплайн функции (Wood, 2017).

За определяне на оценките на неизвестните коефициенти $\alpha_0,\alpha_1,\cdots,\alpha_p$, $\beta_0,\beta_1,\cdots,\beta_q$ и на неизвестните коефициенти на апроксимационните сплайн функции $s_1,s_2,\cdots$ се използва методът на максималното правдоподобие. За целите на проекта се максимизира функцията на правдоподобие:

$$\max_{\alpha,\beta,\theta}L(\alpha,\beta,\theta)=\max_{\alpha,\beta,\theta}\prod_{i=l}^{\mathrm{T}}f(y(t),a_t,b_t),$$

където $f(y(t),a_t,b_t)$ е плътността на гама или Вейбул разпределенията.

Значимостта на оценките на неизвестните коефициенти в моделите на GAMs се основава на Т-теста на Стюдент. Проверката за значимост на модел срещу алтернативен модел, които са йерархично вложени (предикторите на единия модел са подмножество на предикторите в алтернативния), се основава на логаритъма на отношенията на правдоподобие на двата модела, който е асимптотично χ^2 разпределен. Чрез логаритъма на отношенията на правдоподобие на двата модела се дефинира функцията на отклоненията в GAMs методологията като разликата от логаритмите на правдоподобията на двата модела. Функцията на отклоненията се редуцира до остатъчната сума от квадрати в регресионните модели с Гаусово разпределение на грешката. Когато моделите не са йерархично вложени са използвани информационните критерии на Акайке (AIC) или Шварц (BIC) (Stasinopoulus et al., 2017; Wood, 2017).

3.2. Едномерни подходи за възстановяване на липсващи стойности

3.2.1. Възстановяване с аритметична средна стойност

Този подход се използва за възстановяване на липсващата стойност x_{t_1+1} в момент $t_1 + 1$ чрез средната стойност $\bar{x}_{t_1+1}$ на наличните стойности x_{t_1} и x_{t_1+2} в моментите t_1 и $t_1 + 2$:

$$\bar{x}_{t_1+1} = \frac{x_{t_1} + x_{t_1+2}}{2}.$$

Един от най-лесните начини за възстановяване на липсващи стойности е като се заменят липсващите стойности в даден час със средната стойност на измерените стойности за предишния и следващия час. Подходяща оценка е при единично липсващи стойности.

3.2.2. Възстановяване чрез „последна и следваща“ (last & next) стойности

Този подход се използва за възстановяване на липсващи последователности от една, две, три и повече стойности с последната налична стойност, $x_{t_{\text{последна}}}$, след която започва последователността на липсващите стойности или със следващата налична стойност, $x_{t_{\text{следваща}}}$, след последователността на липсващите стойности:

$$\hat{x}_{t_{\text{м}}} = \begin{cases} x_{t_{\text{последна}}}, & t_{\text{м}} \le t_{\text{последна}} + \dfrac{t_{\text{следваща}} - t_{\text{последна}}}{2} \\ x_{t_{\text{следваща}}}, & t_{\text{м}} > t_{\text{последна}} + \dfrac{t_{\text{следваща}} - t_{\text{последна}}}{2}, \end{cases}$$

където $t_{\text{м}}$ е моментът на възстановяване на съответната липсваща стойност, удовлетворяващ условието $t_{\text{последна}} < t_{\text{м}} < t_{\text{следваща}}$, $t_{\text{последна}}$ и $t_{\text{следваща}}$ са

съответите моменти на последната и следващата налични стойсти, ограждащи последователността от липсващи стойности.

3.2.3. Възстановяване чрез линейна интерполация

Този подход се използва за възстановяване на последователност от няколко липсващи стойности. За целта се построява права линия през последната, x_{t_1} и първата, x_{t_2}, измерени стойности, които ограждат последователността от липсващи стойности в моментите t_1 и t_2:

$\hat{x}_{t_м} = x_{t_1} + \theta(t_м - t_1)$, където $\theta = \frac{x_{t_2} - x_{t_1}}{t_2 - t_1}$ и $t_1 < t_м < t_2$.

3.2.4. Възстановяване чрез плъзгаща се средна

Този подход се използва за възстановяване на липсващи последователности от една, две или три стойности, като липсващата стойност се замества със средната стойност.

3.2.5. Мерки за качеството на възстановените данни

За оценка на методите за възстановяване на данни са пресметнати следните широко използвани статистически мерки:

1. Средната абсолютна грешка (MAE) се пресмята по формулата:

$$\mathrm{MAE} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} |\mathrm{P_i} - \mathrm{O_i}|,$$

където n е броят на наблюденията, O_i и P_i са измерените и възстановените стойности. Средната абсолютна грешка (MAE) варира от 0 до безкрайност и възстановяването е перфектно, когато MAE е равно на 0.

2. Средно квадратичната грешка (RMSE) се пресмята по формулата:

$$\mathrm{RMSE} = \left(\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{P_i} - \mathrm{O_i})^2 \right)^{\frac{1}{2}},$$

където n е броят на наблюденията, O_i и P_i са измерените и възстановените стойности. Средно квадратичната грешка варира от 0 до безкрайност и възстановяването е перфектно, когато RMSE е равно на 0.

3. Нормираната средно квадратична грешка (NRMSE) се пресмята по формулата:

$$\mathrm{NRMSE} = \left(\frac{\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{P_i} - \mathrm{O_i})^2}{\frac{1}{\mathrm{n}-1} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{O_i} - \overline{\mathrm{O}})^2} \right)^{\frac{1}{2}},$$

където n е броят на наблюденията, O_i и P_i са измерените и възстановените стойности, а $\overline{O}$ е средната стойност на измерените стойности.

4. Коефициентът на корелация (r) се пресмята по формулата:

$$r = \frac{1}{n}\sum_{i=1}^{n}\frac{(P_i-\overline{P})(O_i-\overline{O})}{\overline{\sigma}_P\overline{\sigma}_O},$$

където $\overline{P}$ и $\overline{O}$ са средните стойности, а $\overline{\sigma}_P$ и $\overline{\sigma}_O$ са стандартните грешки, пресметнати по измерените и възстановени стойности O_i и P_i за i = 1, ..., n. Възстановяването се счита за перфектно при r = 1

4. РЕЗУЛТАТИ

4.1. Първични статистически характеристики на часовите данни за $ФПЧ_{10}$

Използвани са часови данни за $ФПЧ_{10}$ за периода 01.01.2017 - 29.02.2020 г. от АИС в град София, разположени в кв. Дружба, Надежда, Хиподрума, Павлово и Младост, представени на картата на Фиг. 1. Тези данни са налични в НИМХ, като част от архива на системата за ранно предупреждение за замърсяване на въздуха с $ФПЧ_{10}$ в гр. София. Времевите редици от измервани стойности за $ФПЧ_{10}$ от АИС се характеризират с дясно скосено разпределение, което е представено на хистограмите на Фиг. 2 по данни от АИС Дружба и Павлово. Тези редици от данни притежават ярко изразен дневен, месечен и сезонен ход, което се вижда от бокс-плотовете на Фиг. 3 за данните от АИС Дружба.

Fig. 1. Map of the stations in Sofia city situated in Druzhba, Nadezhda, Hipodruma, Pavlovo and Mladost.

Фиг. 1. Карта на АИС в град София, разположени в кв. Дружба, Надежда, Хиподрума, Павлово и Младост.

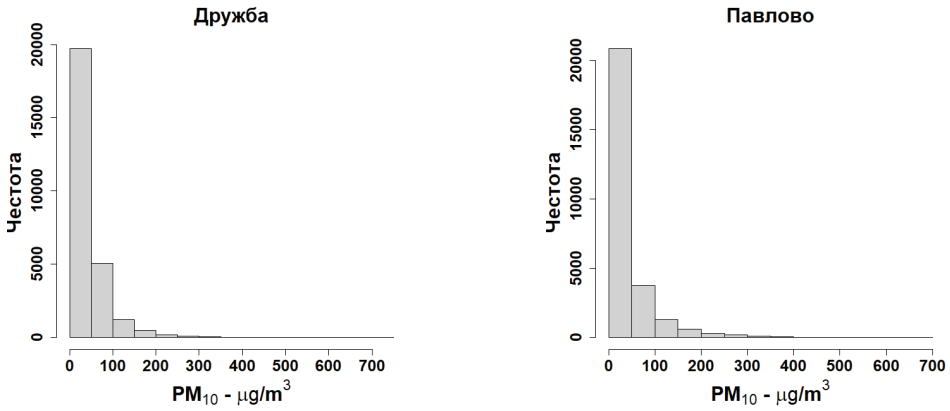
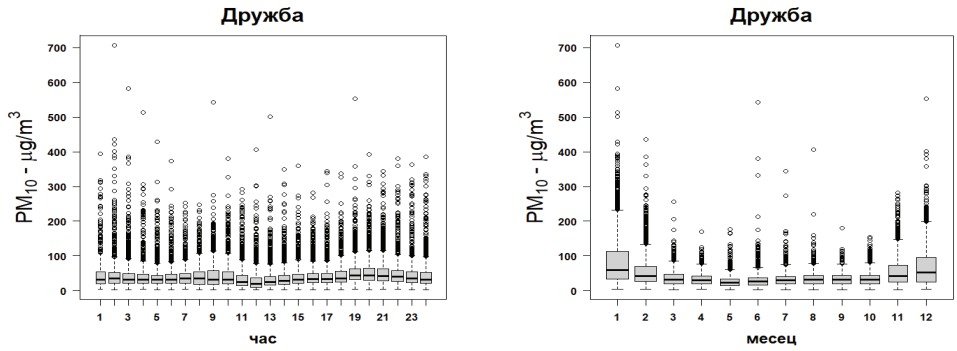


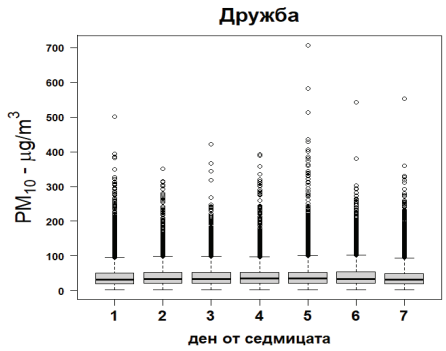
Fig. 2. Histogram of the observed hourly PM_{10} concentrations for Druzhba and Pavlovo stations for the period 01.01.2017г. - 29.02.2020.

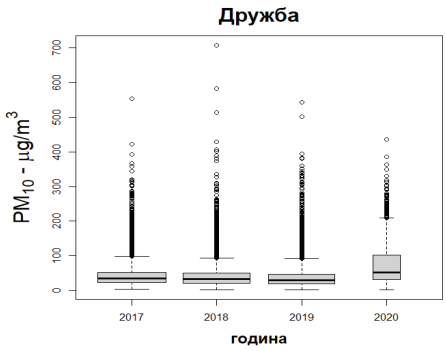
Фиг. 2. Хистограма на измерените часови концентрации на ФПЧ$_{10}$ за АИС Дружба и Павлово за периода 01.01.2017 - 29.02.2020 г.

Забелязва се, че часовите стойности на ФПЧ$_{10}$ са по-високи през студеното полугодие, отколкото през топлото. Бокс-плотовете показват, че екстремните стойности на замърсяване са измерени през тъмната част от денонощието през студеното полугодие, като през 2017 г. са измерени рекордно високи стойности. Аналогични са резултатите за ФПЧ$_{10}$ от АИС, разположени в кв. Надежда, Хиподрума, Павлово и Младост.



a) b)





c) d)

Fig. 3. Hourly (a), monthly (b), day of the week (c) and annual (d) distributions of PM_{10} concentrations for Druzhba station for the period 01.01.2017г. - 29.02.2020.

Фиг. 3. Часови (a), месечни (b), ден от седмицата (c) и годишни (d) разпределения на концентрациите на $ФПЧ_{10}$ за АИС Дружба през периода 01.01.2017г. - 29.02.2020 г.

Матрицата от данни за $ФПЧ_{10}$ от петте АИС за периода 01.01.2017 - 29.02.2020 г. се състои от 27720 реда и 5 стълба. Броят на липсващите стойности на $ФПЧ_{10}$ е даден в Таблица 1.

Table 1. Missing PM_{10} data values for the period 01.01.2017 - 29.02.2020

Таблица 1. Брой липсващи стойности на $ФПЧ_{10}$ за периода 01.01.2017 - 29.02.2020 г.

АИС	Младост	Дружба	Павлово	Хиподрума	Надежда	Общо
Липсващи стойности	844	1062	1174	1644	3485	8209

В Таблица 2 са дадени корелационните коефициенти, пресметнати по измерените стойности за $ФПЧ_{10}$ между всеки две АИС, в сроковете, в които няма липсващи стойности.

Table 2. Correlation coefficients based on complete data (without missing data)

Таблица 2. Корелационни коефициенти по пълните данни (без липсващи стойности)

	Дружба	Павлово	Надежда	Хиподрума	Младост
Дружба	1.000	0.692	0.694	0.718	0.593
Павлово	0.692	1.000	0.704	0.851	0.690
Надежда	0.694	0.704	1.000	0.755	0.586
Хиподрума	0.718	0.851	0.755	1.000	0.691
Младост	0.593	0.690	0.586	0.691	1.000

В Таблица 3 са представени описателни статистики, отнасящи се до разпределението на липсващите стойности в АИС Дружба. В таблицата е включена информация за дължината на редицата от измерени стойности на ФПЧ$_{10}$ (27720), броя на липсващи стойности (1062), процента на липсващи стойности (3.83 %).

Table 3. Druzhba station missing values descriptive statistics

Таблица 3. Статистика на липсващите стойности за АИС в Дружба

Дружба
Дължина на редицата от измерени стойности на ФПЧ$_{10}$ - 27720
Брой на липсващи стойности NA - 1062
Процент на липсващи стойности - 3.83 % (1062 от 27720)
Блокова статистика: Блок 1 (от 1 до 6930) съдържа: 213 липсващи стойности (3.07%) Блок 2 (от 6931 до 13860) съдържа: 272 липсващи стойности (3.92%) Блок 3 (от 13861до 20790) съдържа: 219 липсващи стойности (3.16%)“ Блок 4 (от 20791 до 27720) съдържа: 358 липсващи стойности (5.17%)”
Единични липсващи стойности се срещат 681 пъти
Последователности от 2 липсващи стойности се срещат 16 пъти
Последователности от 3, 4, 5, 6, 7, 8, 11, 12, 17, 25, 26, 37, 41, 43, 49 и 55 липсващи стойности стойности се срещат по 1 път.

Матрицата от данните условно можем да разделим на 4 групи (блока) с равен брой редове (срокове), което дава информация за разпределението на липсващите стойности в отделните блокове. Следващите редове на таблицата показват, че единичните липсващи стойности се срещат в данните 681 пъти, последователностите от 2 липсващи стойности се срещат 16 пъти, докато последователностите от 3, 4, 5, 6, 7, 8, 11, 12, 17, 25, 26, 37, 41, 43, 49 и 55 липсващи стойности се срещат по 1 път. Аналогични описателни статистики са получени и за останалите АИС. Тези статистики показват, че броят на единичните липсващи стойности преобладава във всички станции. Големият процент на липсващите стойности в АИС Надежда (12.6%) вероятно се дължи на повреда в апаратурата, поради наличието на дълги последователности от 67, 74, 98, 173, 230, 759 и 931 срокове, в които липсват измервания.

4.2. Структура на липсващите стойности

Стандартен подход за възстановяване на единични и последователности от 2 и 3 липсващи стойности е използването на средна стойност и линейна интерполация. В Таблица 4 са представени корелационни коефициенти, пресметнати по запълнените данни след линейна интерполация с максимална дължина на

последователност от 3 липсващи стойности. Корелационните коефициенти са статистически значими, което прави данните потенцилно добри предиктори.

Table 4. Correlation coefficients based on filled data after linear interpolation imputation to gap-fill up to 3 missing values

Таблица 4. Корелационни коефициенти на запълнените данни след линейна интерполация с максимална дължина на последователност до 3 липсващи стойности

	Дружба	Павлово	Надежда	Хиподрума	Младост
Дружба	1.000	0.695	0.686	0.678	0.593
Павлово	0.695	1.000	0.707	0.791	0.683
Надежда	0.686	0.707	1.000	0.696	0.582
Хиподрума	0.678	0.791	0.696	1.000	0.657
Младост	0.593	0.683	0.582	0.657	1.000

В резултат от прилагането на тези подходи е получена Таблица 5, в която е дадена структурата на липсващите стойности в матрицата от данни. В клетката, съответстваща на 1-ви ред и 1-ви стълб на тази таблица, стойността 23370 представлява броят на редовете без липсващи стойности едновременно за всички станции, поради което на 1-ви ред стълбовете от 2 до 6 са отбелязани с 1, а в клетката от 7-ми стълб (Брой АИС с NAs) е записана стойността 0, т.е. нула липсващи стойности. В клетката, съответстваща на 1-ви стълб и 2-ри ред на таблицата е дадена стойността 2593, която показва броя на редовете без липсващи стойности за станциите, които са отбелязани с 1, а за станциите, отбелязани с 0 липсват стойности за тези редове. В случая в станция Надежда липсват стойности за 2593 реда, в които обаче останалите станции нямат липсващи стойности. Аналогична е интерпретацията на стойността 844 в клетката, формирана от 1-ви стълб и 3-ти ред. Тази стойност дава броя на липсващите стойности в АИС Хиподрума. Стойността 52 в клетката, формирана от 1-ви стълб и 4-ти ред, показва броя на липсващите наблюдения в едни и същи срокове в АИС Надежда и Хиподрума, поради което в последната клетка на 4-ти ред е поставена стойност 2. Аналогична е интерпретацията на стойностите в останалите клетки. Стойностите 35 и 5, съответно в клетките, формирани от 1-ви и 7-ми стълб в предпоследния ред на таблицата, показват, че в един и същи срок са регистрирани 35 на брой реда, в които има липсващи стойности във всичките пет АИС, включени в изследването. Стойностите, поместени в последния ред, означават броя на липсващите стойности в съответните АИС.

Table 5. Data pattern after linear interpolation imputation to gap-fill up to 3 missing values

Таблица 5. Структура на данните след възстановяване с линейна интерполация с максимална дължина на последователност до 3 липсващи стойности

Брой редове	Младост	Дружба	Павлово	Хиподрума	Надежда	Брой АИС с NAs
23370	1	1	1	1	1	0
2593	1	1	1	1	0	1
844	1	1	1	0	1	1
52	1	1	1	0	0	2
315	1	1	0	1	1	1
110	1	1	0	1	0	2
262	1	0	1	1	1	1
26	1	0	1	1	0	2
5	1	0	1	0	1	2
17	1	0	0	1	1	2
65	0	1	1	1	1	1
20	0	1	1	1	0	2
6	0	1	1	0	0	3
35	0	0	0	0	0	5
Брой **NAs**	**126**	**346**	**477**	**943**	**2842**	**4734**

4.3. Модели за възстановяване

Структурата на липсващите стойности в Таблица 5 задава насоката за възстановяването им чрез подходящи статистически модели. По-конкретно: 1) липсващите 2593 стойности в АИС Надежда могат да бъдат възстановени чрез моделите по наличните измерени стойности в същите срокове на останалите АИС; 2) липсващите 844 стойности в АИС Хиподрума могат да бъдат възстановени чрез моделите и наличните измерени стойности в същите срокове на останалите АИС; 3) липсващите 52 стойности в АИС Хиподрума и Надежда могат да бъдат възстановени чрез моделите и наличните данни от АИС Дружба, Младост и Павлово. Аналогично се интерпретират останалите редове от Таблица 5. В първия ред няма липсващи стойности, докато последния ред показва 35 реда, в които липсват стойности за всички станции, които не могат да бъдат възстановени по съседни станции. Структурата на липсващите данни, формирана за разглеждания период, ни показва, че за да възстановим липсващите данни е необходимо да създадем 12 подходящи модела, включващи за предиктори наличните данни в съответните станции.

Създаването на моделите налага разбиване на матрицата от данни на пълни и непълни. По пълните данни са създадени 12-те модела. За възстановяването на липсващите стойности са използвани подходящи подмножества на матрицата от непълните данни, съответни на структурата на Таблица 5. За целта от матрицата на данните за $ФПЧ_{10}$ с размерност 27720x5 се създават две основни подматрици - на пълните данни с размерност 23370x5 и на данните с липсващи стойности с размерност 4350x5, където 4350 = 2593 + 844 + 52 + 315 + 110 + 262 + 26 + 5+17+65+20+6+35 е броят на редовете с липсващи стойности от Таблица 5. Следвайки структурата на липсващите стойности се създават гама и Вейбул разпределените GAMs модели по матрицата от пълните данни. Резултатите от оценяването на всеки модел се съхраняват в обектен (изходен) файл, който се използва за предсказване по нови данни на предикторите. Възстановяването на липсващите стойности се провежда последователно с функцията prediction() на библиотеките mgsv или gamlss. За целта се формират нови данни, които представляват подмножества от матрицата на непълните данни, съответстваща на модела за възстановяване. Липсващите стълбове в тези нови данни, се възстановяват по останалите стълбове на новите данни с помощта на функцията prediction() и обектния файл, съответен на модела за възстановяване. Създадените модели са статистически значими и обясняват повече от 88% от дисперсията на данните ($R^2 \geq 0.88$).

4.4. Проверка за качеството на методите за възстановяване на данни

За да оценим качеството на възстановените данни е необходимо процент от пълните данни да бъде заменен с липсващи стойности по подходящ начин, за да може след това да бъдат сравнени възстановените с реалните данни. След това се пресмятат статистическите мерки по възстановените по моделите и реалните данни. Резултатите от сравнителния анализ са представени в табличен и графичен вид. За целта използваме матрицата от пълните данни с размерност 23370x5, от която е формирана тестова извадка с последователни срокове с размерност 1300x5 и обучаваща извадка с размерност 22070x5. Част от данните в тестовата извадка, съответни на АИС Дружба, Павлово, Младост, Надежда и Хиподрума се заменят с последователности от липсващи стойности NA, след което се възстановяват с GAMs моделите и процедурите за възстановяване на липсващи стойности – „последна и следваща“ и линейна интерполация. За оценка на качеството на възстановяване на тези методи са пресметнати съответните статистически мерки: MAE, RMSE, NRMSE и коефициент на корелация между възстановени и измерени стойности. Този процес на заместване с липсващи стойности, възстановяване с изброените методи и пресмятане на статистическите мерки се повторя неколкократно. Получените резултатите са осреднени и са представени в Таблици 6 - 9. Възстановените данни по метода „последна и следваща“ са трансформирани допълнително с процедура на плъзгащи се средни.

Table 6. Statistical measures between observed and recovered PM_{10} data based on GAMs model with gamma distribution

Таблица 6. Статистически мерки между измерени и възстановени ФПЧ$_{10}$ с гама разпределен GAMs модел

GAMs - гама	Дружба	Павлово	Младост	Надежда	Хиподрума
MAE	3.148	6.924	10.694	7.878	5.986
RMSE	13.431	23.332	32.691	24.162	20.151
NRMSE	0.263	0.266	0.383	0.230	0.206
Коеф. на корелация	0.965	0.967	0.924	0.923	0.978

Table 7. Statistical measures between observed and recovered PM_{10} data based on GAMs model with Weibull distribution

Таблица 7. Статистически мерки между измерени и възстановени ФПЧ$_{10}$ с Вейбул разпределен GAMs модел

GAMs - Вейбул	Дружба	Павлово	Младост	Надежда	Хиподрума
MAE	4.549	7.664	9.838	8.045	5.655
RMSE	16.740	24.768	30.960	24.821	19.208
NRMSE	0.328	0.282	0.363	0.236	0.197
Коеф. на корелация	0.944	0.959	0.932	0.971	0.980

От резултатите в Таблици 6 и 7 се вижда, че стойностите на грешките на процедурите по възстановяване с гама и Вейбул разпределени GAMs модели са много по-малки от съответните грешки на процедурата по възстановяване с „последна и следваща“ и линейна интерполация, които са дадени в Таблици 8 и 9, съответно.

Table 8. Statistical measures between observed and recovered PM_{10} data based on “last & next” method

Таблица 8. Статистически мерки между измерени и възстановени ФПЧ$_{10}$ с „последна и следваща“

„Последна и следваща“	Дружба	Павлово	Младост	Надежда	Хиподрума
MAE	44.931	34.381	11.021	44.618	37.170
RMSE	65.461	52.349	34.259	69.885	57.839
NRMSE	0.767	0.613	0.401	0.819	0.678
Коеф. на корелация	0.882	0.852	0.918	0.758	0.843

Table 9. Statistical measures between observed and recovered PM_{10} data based on linear interpolation

Таблица 9. Статистически мерки между измерени и възстановени $ФПЧ_{10}$ с линейна интерполация

линейна интерполация	Дружба	Павлово	Младост	Надежда	Хиподрума
MAE	44.984	34.400	11.026	44.530	37.049
RMSE	65.580	52.386	34.177	69.764	57.693
NRMSE	0.768	0.614	0.400	0.817	0.676
Коеф. на корелация	0.881	0.852	0.919	0.759	0.845

От матрицата на пълните данни се формира и втора тестова извадка с последователни срокове с размерност 250x5, която съответства на редове от 16751 до 17000 на матрицата от пълните данни и съответната ѝ обучаваща извадка с размерност 23120x5. Втората тестова извадка се използва за визуализиране на реалните и възстановени данни. Във втората тестова извадка 36 последователно измерени стойности с индекси от 51 до 86 в първи стълб, съответен на АИС Дружба, са заменени с последователност от 36 NA, като останалите стълбове са използвани за възстановяване със съответния модел за Дружба, създаден по втората обучаваща извадка. По този начин, пълните данни в стълбовете на матрицата 250x5, съответни на АИС Павлово, Младост, Надежда и Хиподрума са използвани за възстановяване на 36 липсващи стойности в стълба, съответен на АИС Дружба. Аналогично, последователно са заменени и възстановени данните за останалите АИС чрез съответните им модели, създадени по втората обучаваща извадка. Резултатите от възстановяването с различните методи - гама и Вейбул разпределени GAMs модели, както и „последна и следваща“ и линейната интерполация са представени на Фиг. 4-8. За всяка АИС са дадени 4 плота, съответни на използваните процедури за възстановяване на липсващи стойности. На всички плотове измерените стойности са дадени със сини и зелени точки, докато червените точки съответстват на възстановените с използваните методи. Зелените точки съответстват на измерените реални стойности, които са заменени с 36 липсващи стойности, за да бъдат възстановени по 4-те процедури и след това сравнени. На Фиг. 4 са дадени плотове за АИС Дружба представящи възстановените данни по описаните методи: a) „последна и следваща“ и изглаждане с плъзгаща средна стойност; b) линейна интерполация; c) гама разпределен GAMs модел; d) Вейбул разпределен GAMs модел. Аналогично се интерпретират Фиг. 5-8 за останалите АИС.

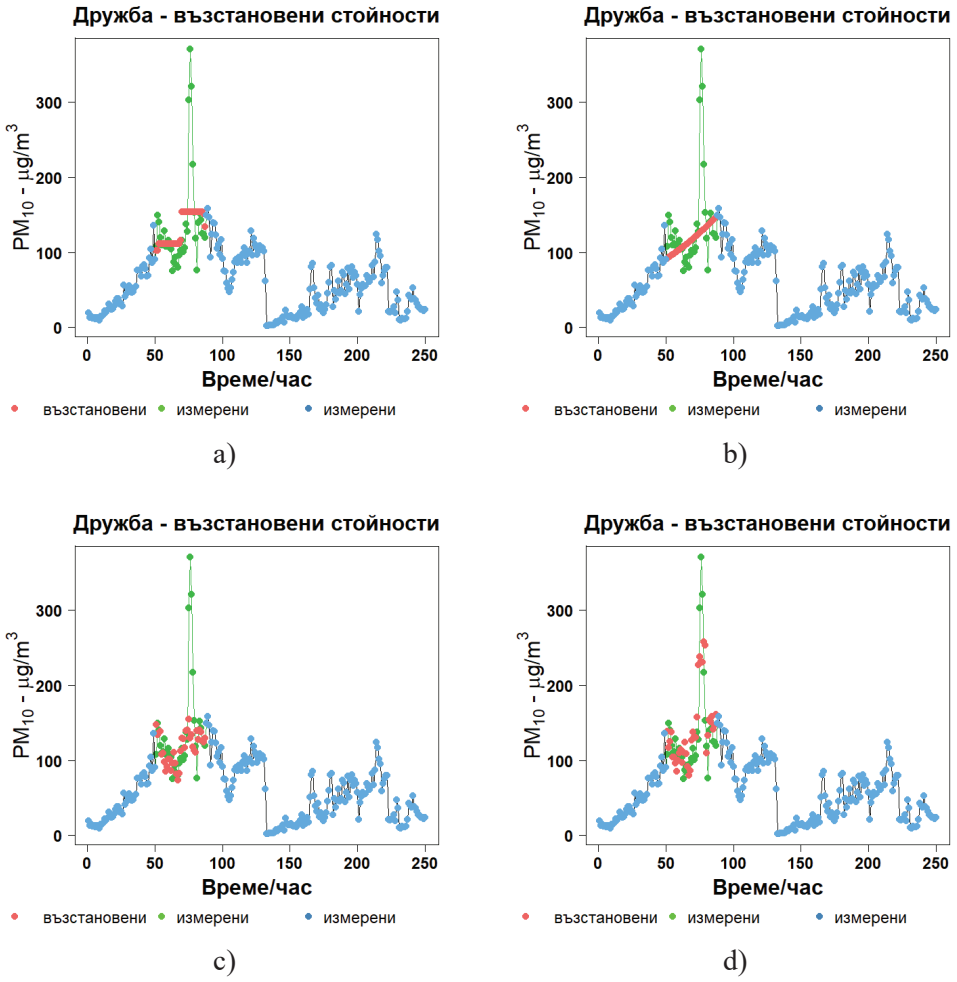


Fig. 4. Station Druzhba: the observed values are marked with blue and green dots, the red dots are based on the following imputation methods: a) "last & next" and running mean; b) linear interpolation; c) GAMs with gamma distribution; d) GAMs with Weibull distribution

Фиг. 4. АИС Дружба: измерените стойности са означени със сини и зелени точки, червените точки са възстановени по методите на: а) „последна и следваща" и изглаждане с плъзгаща средна стойност; b) линейна интерполация; c) гама разпределен GAMs модел; d) Вейбул разпределен GAMs модел.

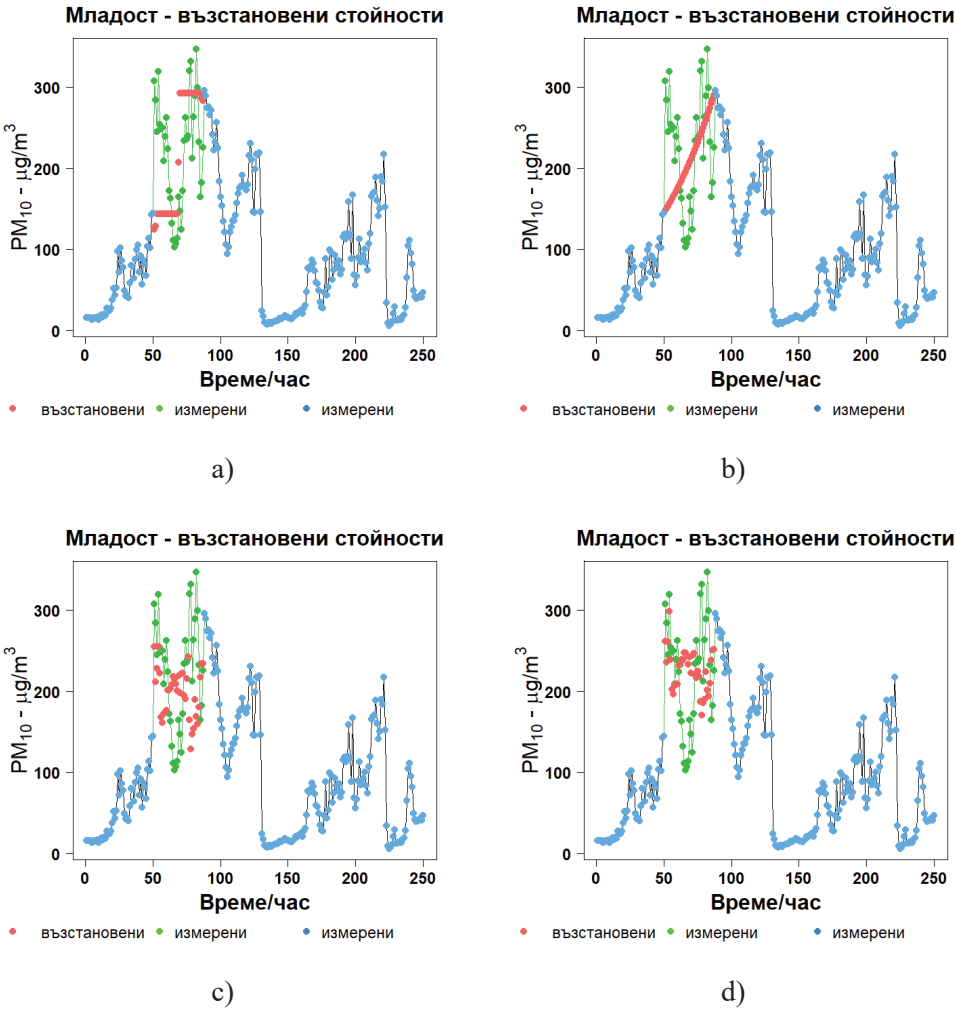


Fig. 5. Station Mladost: the same as for **Fig. 4**

Фиг. 5. АИС Младост: същото като на **Фиг. 4**

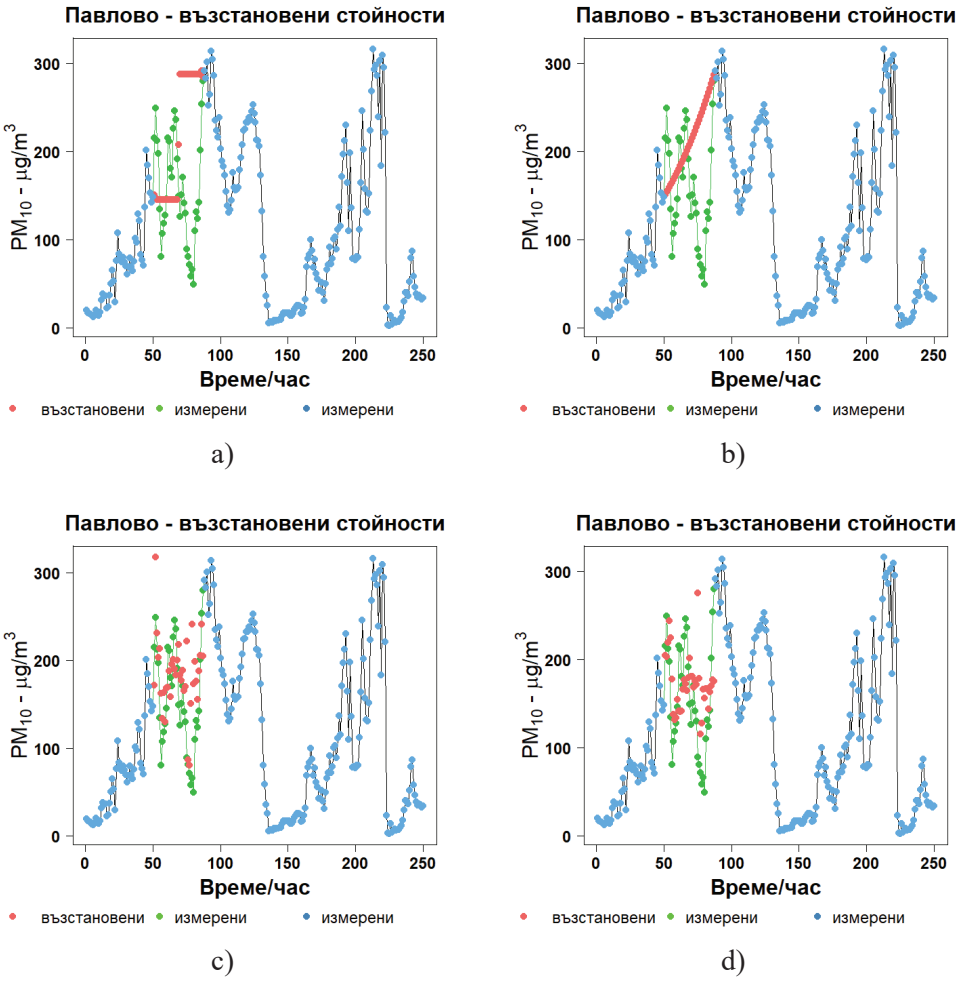


Fig. 6. Station Pavlovo: the same as for **Fig. 4**

Фиг. 6. АИС Павлово: същото като на **Фиг. 4**

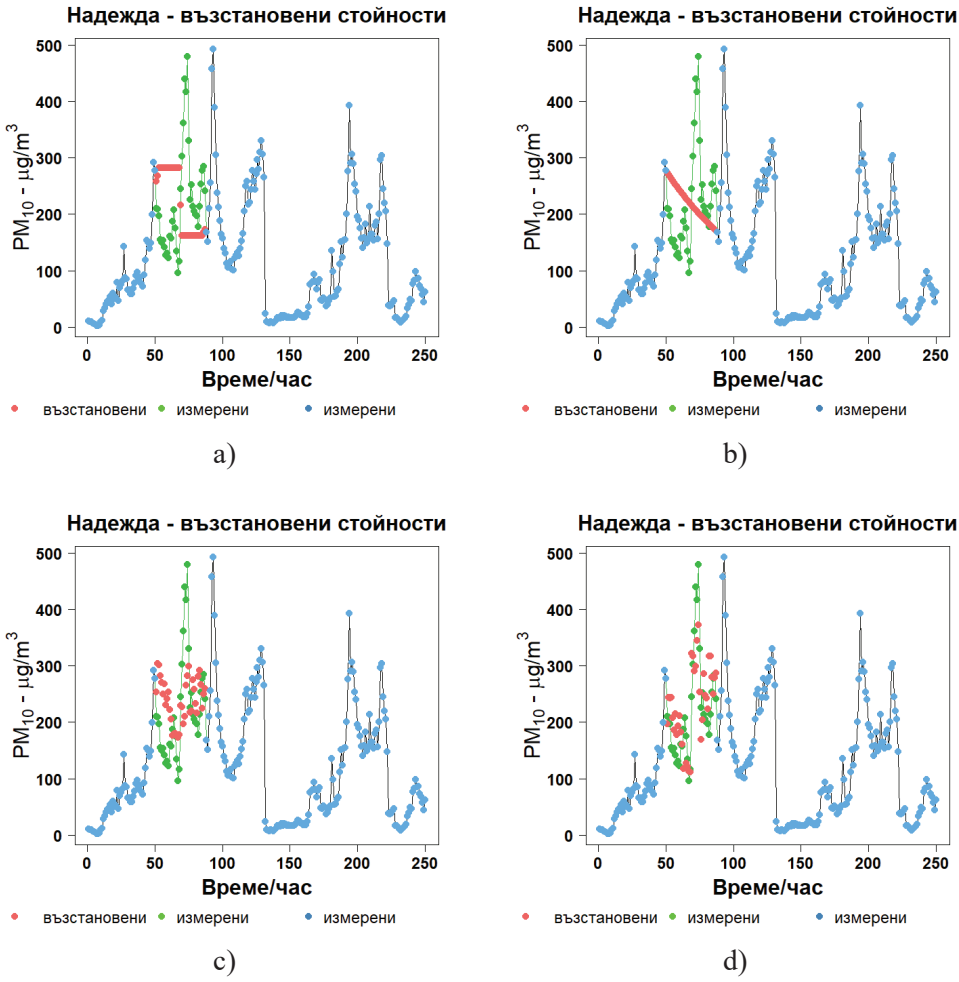


Fig. 7. Station Nadezhda: the same as for **Fig. 4**

Фиг. 7. АИС Надежда: същото като на **Фиг. 4**

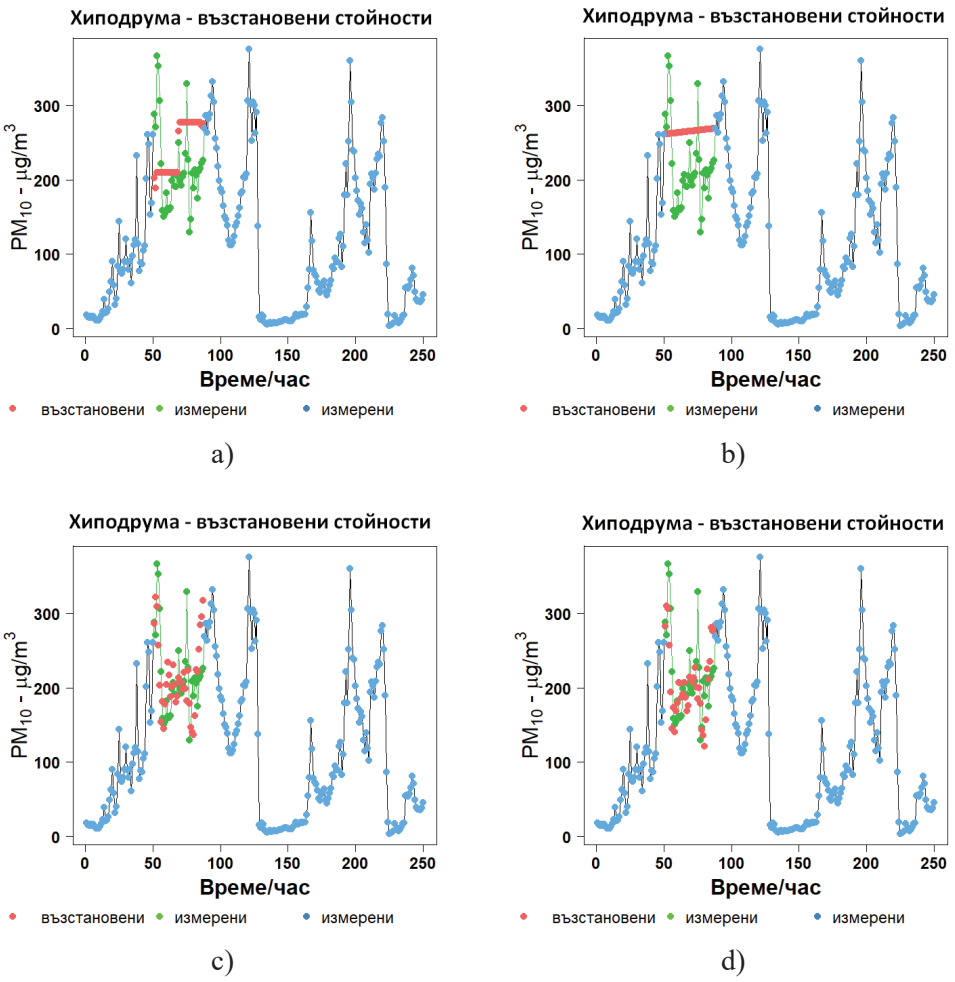


Fig. 8. Station Hipodruma: the same as for **Fig. 4**

Фиг. 8. АИС Хиподрума: същото като на **Фиг. 4**

5. ЗАКЛЮЧЕНИЕ

GAMs методологията позволява успешно възстановяване на липсващите данни в разглежданите от нас случаи. Широко използваните методи за възстановяване със средна стойност, „последна и следваща“ и линейна интерполация са лесни

за използване и са подходящи при последователности от липсващи стойности в кратки поредни срокове. В сравнение с тях GAMs методите се представят много по-добре и възстановените данни са с високо качество. GAMs методите с гама и Вейбул разпределения са приложими и за други типове дясно скосено разпределени данни, като например повечето атмосферни замърсители, което може да бъда задача за бъдещи изследвания.

БЛАГОДАРНОСТИ

Това изследване е финансирано от Национална програма „Млади учени и постдокторанти“ на Министерството на образованието и науката (МОН), договор №: ЧР-04-14/31.03.2020 г.

ЛИТЕРАТУРА

Hastie, T.J. and Tibshirani, R.J. (1990), Generalized Additive Models. Boca Raton, Florida: Chapman & Hall/CRC.

Junger, W.L., de Leon, A.P. (2015), Imputation of missing data in time series for air pollutants. Atmos. Environ. 102, 96-104.

Junninen, H., Niska, H., Tuppurainen, K. and Ruuskanen, J. (2004), Methods for imputation of missing values in air quality data sets. Atmos. Environ. Vol. 38(18), 2895-2907.

Moritz, S. and Bartz-Beielstein, T. (2017), imputeTS: Time series missing value imputation in R. The R Journal, vol. 9(1), pp 207-218.

R Core Team (2019), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V. and Bastiani, F. (2017), Flexible Regression and Smoothing Using GAMLSS in R, CRC Press.

Sukatis, F. F.,Noor, N.M., Zakaria, N. A., Ul-Saufie, A. Z. and Suwardi, A. (2019), Estimation of missing values in air pollution dataset by using various imputation methods. International Journal of Conservation Science, vol. 10(4), pp.791-804.

Wood, S.N. (2017), Generalized Additive Models: An Introduction with R, 2nd edition, CRC Press.

Zakaria,N.A. and Noor, N., M. (2018), Imputation methods for filling missing data in urban air pollution data for Malaysia. Urbanism. Arhitectura. Constructii, vol. 9, pp. 159-166.