

УДК: 519.246.85; 551.510.534

УЧЕТ АВТОКОРРЕЛЯЦИИ В ЗАДАЧЕ ЛИНЕЙНОЙ РЕГРЕССИИ НА ПРИМЕРЕ АНАЛИЗА ОБЩЕГО СОДЕРЖАНИЯ NO₂ В АТМОСФЕРЕ

© 2019 г. А.Н. Груздев

*Институт физики атмосферы имени А.М. Обухова РАН
Россия, 119017 Москва, Пыжевский пер., 3
E-mail: a.n.gruzdev@mail.ru*

Поступила в редакцию 06.12.2017 г.
После доработки 16.04.2018 г.

Предложен метод учета сериальной корреляции (автокорреляции) данных в задаче линейной регрессии, позволяющий учитывать автокорреляцию на больших масштабах. Остаточный ряд представляется в виде процесса авторегрессии, порядок которой k может быть намного больше 1, и автокорреляционная функция процесса рассчитывается путем решения системы уравнений Юла-Уокера. По автокорреляционной функции конструируется автокорреляционная матрица, которая входит в формулы для оценок регрессионных коэффициентов и их стандартных ошибок. Эффективность метода демонстрируется на основе множественного регрессионного анализа данных 26-летних измерений общего содержания NO₂ на Звенигородской научной станции Института физики атмосферы. Оценки регрессионных коэффициентов и их ошибки зависят от порядка авторегрессии k . Сначала ошибка увеличивается с увеличением k . Затем она достигает максимума и после этого начинает уменьшаться. В случае NO₂ ошибка в максимуме возрастает более чем вдвое по сравнению с ее начальным значением. Уменьшение ошибки после достижения максимума прекращается, если k достигает значения, при котором авторегрессионный процесс позволяет описать важные особенности автокорреляционной функции остаточного ряда. Получены оценки сезонно зависимых трендов и эффектов воздействия на NO₂ природных факторов, таких как 11-летний солнечный цикл, квазидвухлетняя цикличность, Североатлантическое колебание и других.

Ключевые слова: линейная регрессия, сериальная корреляция, NO₂

DOI: <https://doi.org/10.31857/S0002-351553173-82>

ВВЕДЕНИЕ

Одно из важных свойств многих геофизических процессов – память. Так, наличие дальней памяти значительно влияет на способность предсказания и/или детектирования долговременных изменений [1]. В регрессионном анализе временных рядов память может проявляться через автокорреляцию (сериальную корреляцию) остаточного ряда, и наличие автокорреляции влияет на погрешности регрессионных коэффициентов [2].

Известны несколько эффективных в практическом плане методов учета сериальной корреляции в оценках ошибок регрессионных коэффициентов. Они основаны на предположении о том, что остаточный ряд есть авторегрессионный (АР) процесс первого порядка. В АР модели k -го порядка текущее значение процесса

выражается в виде линейной комбинации k предыдущих значений процесса [3]. В АР процессе 1-го порядка текущее значение зависит только от одного, непосредственно предшествующего ему значения.

Наличие автокорреляции ведет к сокращению числа независимых значений, то есть, к уменьшению числа степеней свободы. В [4] оно выражено через АР коэффициент (равный коэффициенту автокорреляции при единичном сдвиге) и получена поправка к дисперсии остаточного ряда. Этот способ был, например, использован при оценке тренда температуры поверхности океана в [5]. В [6] предложена несколько иная аппроксимация дисперсии остаточного ряда с учетом автокорреляции, также выраженная через коэффициент АР процесса 1-го порядка. Данный метод был применен, например, в [7] при оценках трен-

дов NO_2 . В [8] предложена итерационная процедура учета сериальной корреляции в случае, когда остаточный ряд является АР процессом первого порядка (см. также [9]). Предполагается, что итерационный процесс сходится к глобальному минимуму суммы квадратов остаточного ряда. Эта процедура использовалась также в предположении, что остаточный ряд является АР процессом 2-го порядка [10]. Такой подход использовался и для учета сериальной корреляции в недавних оценках озонных трендов [11].

Автокорреляционная (АК) функция АР процесса 1-го порядка экспоненциально затухает [3]. Однако АК функция действительного остаточного ряда может значительно отличаться от экспоненты. Поэтому в рамках АР представления может потребоваться больший порядок авторегрессии для описания более сложного поведения остаточного ряда. Это означает, что память остаточного ряда проявляется на большем масштабе, чем предполагается АР процессом 1-го порядка, и это характерно для разных геофизических временных рядов.

Цель настоящей работы состоит в разработке метода учета сериальной корреляции на больших масштабах изменчивости остаточного ряда, который может быть представлен АР процессом высокого порядка. Эффективность метода демонстрируется на примере анализа данных 26-летних наземных спектрометрических измерений общего содержания NO_2 на Звенигородской научной станции ИФА им. А.М. Обухова РАН.

1. МЕТОД

Линейную регрессионную модель удобно записать в матричной форме [2]:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1)$$

где \mathbf{Y} – вектор n наблюдений (анализируемый ряд) размера $(n \times 1)$, \mathbf{X} – матрица предикторов (независимых переменных) размера $(n \times p)$, p – число предикторов, $\boldsymbol{\beta}$ – вектор неизвестных коэффициентов (искомых коэффициентов регрессии) размера $(p \times 1)$, $\boldsymbol{\varepsilon}$ – вектор ошибок (остатков) длины n (остаточный ряд). Переопределенная система уравнений (1) ($n > p$) решается стандартным методом наименьших квадратов. Если остаточный ряд $\boldsymbol{\varepsilon}$ подчиняется гауссову распределению с нулевым средним и дисперсией σ^2 , то стандартная ошибка искомых коэффициентов находится одновременно с решением системы [2].

1.1. Учет автокорреляции (сериальной корреляции)

Приведем, следуя [2], основные соотношения для решения задачи (1), то есть выражения для среднеквадратичных оценок коэффициентов регрессии и их стандартных ошибок. Пусть $\mathbf{V}(\boldsymbol{\varepsilon})$ есть ковариационная матрица остаточного ряда размера $(n \times n)$, такая что $\mathbf{V}(\boldsymbol{\varepsilon}) = \mathbf{V}\sigma^2$, где σ^2 – дисперсия некоррелированной компоненты $\boldsymbol{\varepsilon}$. Если ошибки (члены остаточного ряда) $\boldsymbol{\varepsilon}$ некоррелированы, то $\mathbf{V}(\boldsymbol{\varepsilon}) = \mathbf{I}\sigma^2$, где \mathbf{I} – единичная матрица, и матрица $\mathbf{V}(\boldsymbol{\varepsilon})$ есть просто диагональная дисперсионная матрица. Однако для геофизических данных типично, когда ошибки сериально коррелированы и, по крайней мере, некоторые недиагональные элементы матрицы $\mathbf{V}(\boldsymbol{\varepsilon})$ отличны от нуля.

Согласно [2], существует невырожденная симметричная матрица \mathbf{P} , такая что

$$\mathbf{P}^T \mathbf{P} = \mathbf{P} \mathbf{P} = \mathbf{P}^2 = \mathbf{V}, \quad (2)$$

где \mathbf{P}^T – транспонированная матрица для матрицы \mathbf{P} . Если остаточный ряд $\boldsymbol{\varepsilon}$ можно представить в виде АР процесса, то матрица \mathbf{V} будет его АК матрицей [2]. Например, в случае АР процесса первого порядка элемент V_{ij} матрицы \mathbf{V} есть $\rho^{|i-j|}$, где ρ – коэффициент автокорреляции при единичном запаздывании [3] (см. также приложение А в [9]).

Обозначим через \mathbf{e} вектор некоррелированных ошибок, ассоциированных с вектором коррелированных ошибок $\boldsymbol{\varepsilon}$. Тогда, согласно [2],

$$\mathbf{e} = \mathbf{P}^{-1}\boldsymbol{\varepsilon} \quad (3)$$

и искомый вектор \mathbf{b} , элементы которого являются обобщенными среднеквадратичными оценками элементов вектора регрессионных коэффициентов $\boldsymbol{\beta}$, есть:

$$\mathbf{b} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{Y}. \quad (4)$$

При этом вектор квадрата стандартной ошибки $\boldsymbol{\delta}_b^2$ оценок \mathbf{b} состоит из диагональных элементов матрицы $(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1}\sigma^2$:

$$\boldsymbol{\delta}_b^2 = \text{diag}(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1}\sigma^2. \quad (5)$$

Формулы (4) и (5) описывают решение регрессионной задачи (1), обобщенное на случай сериальной корреляции. В отсутствие сериальной корреляции $\mathbf{V}^{-1} = \mathbf{I}$ и формулы (4), (5) принимают стандартный вид [2].

1.2. Метод решения

Как будет показано в разделе 4, низкий порядок АР представления остаточного ряда может быть недостаточным для учета долговременной корреляции. Вместо оценочных формул [4–7] и итерационной процедуры [8–11] мы предлагаем метод оценки регрессионных коэффициентов и их стандартных ошибок с использованием обобщенных формул (4), (5). Ключевым моментом является конструирование АК матрицы \mathbf{V} АР процесса, аппроксимирующего поведение остаточного ряда. Известно, что АК функция АР процесса и коэффициенты авторегрессии связаны рекуррентными соотношениями, образующими систему уравнений Юла–Уокера (Yule–Walker) [3].

Рассмотрим два способа нахождения матрицы \mathbf{V} . Пусть АР процесс имеет порядок k . АК функцию остаточного ряда для задержек от 1 до k рассчитываем обычным путем и, решая систему уравнений Юла–Уокера, по значениям АК коэффициентов находим k коэффициентов АР процесса. Затем по рекуррентным соотношениям рассчитываем остаток (хвост) АК функции (для задержек от $k + 1$ до n). В результате квадратная матрица \mathbf{V} размера $(n \times n)$ полностью определена.

В более сложном, но более оптимальном способе используем алгоритм Бёрга (Burg) для расчета АР коэффициентов, который основан на принципе максимальной энтропии [12]. По значениям АР коэффициентов, путем решения системы уравнений Юла–Уокера относительно АК коэффициентов, находим k значений коэффициентов автокорреляции, соответствующих АР процессу, а затем рассчитываем прежним способом хвост АК функции.

Таким образом, регрессионная задача (1) решается в два шага. На первом шаге решение находится стандартным методом наименьших квадратов (в предположении $\mathbf{V} = \mathbf{I}$). Остаточный ряд представляется в виде АР процесса порядка k , который может значительно отличаться от 1. В алгоритме Бёрга k может достигать трети длины ряда. АР коэффициенты и значения соответствующей АК функции для задержек в интервале от 1 до k рассчитываются одним из описанных способов. Затем рассчитывается хвост АК функции и определяется матрица \mathbf{V} .

На втором шаге в соответствии с (4) и (5) рассчитываются оценки регрессионных коэффициентов и их стандартные ошибки. При необходимости с помощью (3) может быть найден вектор

некоррелированных ошибок \mathbf{e} . Матрица \mathbf{P} есть решение квадратного матричного уравнения (2) и находится с помощью преобразования матрицы \mathbf{V} к жордановой форме.

В настоящей работе для анализа данных использован метод с применением алгоритма Бёрга.

2. ИСПОЛЬЗОВАННЫЕ ДАННЫЕ И РЕГРЕССИОННАЯ МОДЕЛЬ

Для демонстрации предложенного метода воспользуемся данными измерений общего содержания NO_2 на Звенигородской научной станции в 1990–2016 гг. Станция расположена в сельской местности в 50 км западнее Москвы и входит в состав Сети по обнаружению изменений состава атмосферы (Network for the Detection of Atmospheric Composition Change – NDACC). Измерения выполняются в утренние и вечерние сумерки по рассеянной из зенита солнечной радиации в видимой области спектра. Метод наблюдений и восстановления содержания NO_2 описан в [13–15], а данные находятся в свободном доступе по адресу <ftp://ftp.cpc.ncep.noaa.gov/ndacc/station/zvenigor/>. Анализируемое общее содержание NO_2 есть интегральное содержание NO_2 в вертикальном столбе атмосферы выше ее пограничного слоя. В настоящей работе использованы данные вечерних измерений. Результаты измерений подверглись статистической отбраковке по способу [16], и по ним рассчитаны среднемесячные значения содержания NO_2 . Они приведены на рис. 1а.

В модель множественной линейной регрессии для NO_2 включены следующие предикторы: линейный член, зональная скорость экваториального ветра на изобарической поверхности 20 гПа и среднее зональных скоростей на поверхностях 40 и 50 гПа (<http://www.geo.fu-berlin.de/en/met/ag/strat/produkte/qbo/index.html>) – в качестве индексов квазидвухлетней цикличности (КДЦ), индекс Ниньо 3.4 (<https://www.esrl.noaa.gov/psd/data/correlation/nina34.data>) и смещенный во времени индекс Ниньо 3.4 – для учета эффектов Эль-Ниньо – Южного колебания (ЭНЮК), индекс Североатлантического колебания (САК) (<https://crudata.uea.ac.uk/cru/data/nao/nao.dat>), индекс солнечной активности F10.7 (<https://www.ngdc.noaa.gov/stp/space-weather/solar-data/solar-features/solar-radio/>), среднее по северному полушарию значение оптической толщи стратосферного аэрозоля на длине волны 550 нм (<https://data.giss.nasa.gov/modelforce/strataer/>) – для учета воздействия

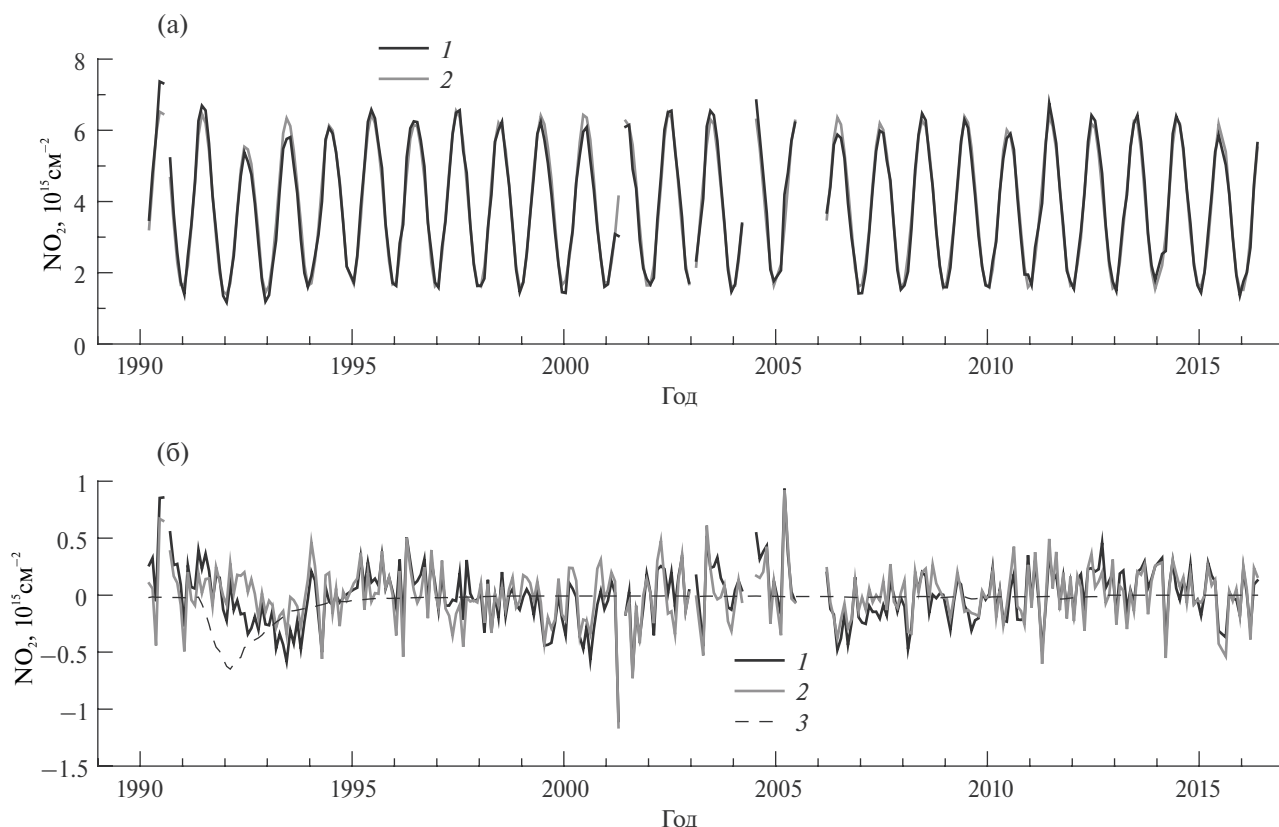


Рис 1. а – Среднемесячные значения общего содержания NO_2 по данным наблюдений (1) и их аппроксимация с помощью модели множественной линейной регрессии (2). б – Коррелированная (1) и некоррелированная (2) ошибки и изменение содержания NO_2 под действием вулканического аэрозоля (3). Некоррелированная ошибка соответствует порядку авторегрессии $k = 50$.

продуктов извержения вулкана Пинатубо на NO_2 . Эти предикторы представляют в регрессионной модели основные важные процессы, оказывающие влияние на изменчивость стратосферных примесей. В различных комбинациях они использовались при анализе трендов NO_2 [7, 16–21].

Два индекса КДЦ использованы для учета задержки отклика NO_2 на воздействие экваториальной КДЦ [22–23]. Вариации скоростей ветра на изобарической поверхности 20 гПа и в слое 40–50 гПа в экваториальной стратосфере в высокой степени ортогональны между собой (коэффициент корреляции $\text{КК} \sim -0.01$). Использование двух ортогональных индексов КДЦ также типично при анализе озонных трендов [10–11].

Аналогичная идея использована и для учета эффекта ЭНЮК. Значения второго индекса запаздывают относительно значений оригинального индекса Ниньо 3.4 на 14 месяцев, так что два индекса ортогональны между собой ($\text{КК} \sim 0.01$).

Стоит отметить, что не все предикторы являются попарно независимыми. В наибольшей степени это относится к аэрозольной оптиче-

ской толще (АОТ). Форму ее временной зависимости с обратным знаком можно представить по штриховой кривой на рис. 1б. Максимум толщи, обусловленный продуктами извержения вулкана Пинатубо, приходится на начало 1992 г. Поэтому изменения АОТ в период 1990–2016 гг. имеют трендовую составляющую. Кроме того, максимум АОТ пришелся на окончание фазы максимума солнечного цикла, и уменьшение АОТ и солнечного индекса F10.7 в последующие пять лет происходили синхронно, за счет чего между этими предикторами есть корреляционная связь. Обусловленный этим $\text{КК} \sim 0.1$ для периода 1990–2016 гг. Корреляционная связь с АОТ может возникнуть при наложении максимума АОТ на экстремум другого предиктора, как это случилось с индексом Ниньо 3.4 ($\text{КК} = -0.27$). Однако исключать АОТ из регрессионной модели не целесообразно, так как извержение вулкана Пинатубо привело к значительному влиянию на стратосферное содержание NO_2 [7, 13, 16–18, 20, 24].

Наличие корреляции между предикторами приводит к корреляции между оценками регрессионных коэффициентов (4), и значения пар-

ных КК определяются недиагональными элементами матрицы $(X^T V^{-1} X)^{-1}$ [2]. Корреляция между предикторами отражается и на ошибках регрессионных коэффициентов (см. (5)).

Все искомые коэффициенты в регрессионной модели представлены в виде разложения в ряд Фурье по парам синус-косинус, соответствующим годовой и полугодовой гармоникам, с целью учета годового хода NO_2 , сезонной зависимости трендов NO_2 и сезонной зависимости воздействия других предикторов на NO_2 (см. также [10–11]). Годовой ход NO_2 на Звенигородской станции в основном обусловлен годовой и полугодовой гармониками [13].

3. РЕЗУЛЬТАТЫ АПРОБАЦИИ МЕТОДА НА ПРИМЕРЕ NO_2

Рисунок 1а показывает, что содержание NO_2 , предсказываемое регрессионной моделью, хорошо соответствует наблюдаемому содержанию NO_2 (КК = 0.99). Это связано в первую очередь с большой амплитудой годового хода NO_2 , которая намного превышает амплитуды межмесячных и межгодовых вариаций. С исключенным годовым ходом из наблюдаемых и модельных данных КК = 0.63.

Как показывает штриховая кривая на рис. 1б, значительное уменьшение содержания NO_2 в 1991–1992 гг. есть результат влияния на NO_2 продуктов извержения вулкана Пинатубо (в июне 1991 г.). Оно обусловлено гетерогенными химическими процессами на поверхности стратосферного сульфатного аэрозоля [25]. Однако стоит отметить, что индекс АОТ недооценивает вулканический эффект в 1993 г. (ср. две кривые на рис. 1а и сплошную черную кривую на рис. 1б). Согласно наблюдениям, восстановление NO_2 происходило медленнее, чем предсказывается моделью с АОТ в качестве предиктора.

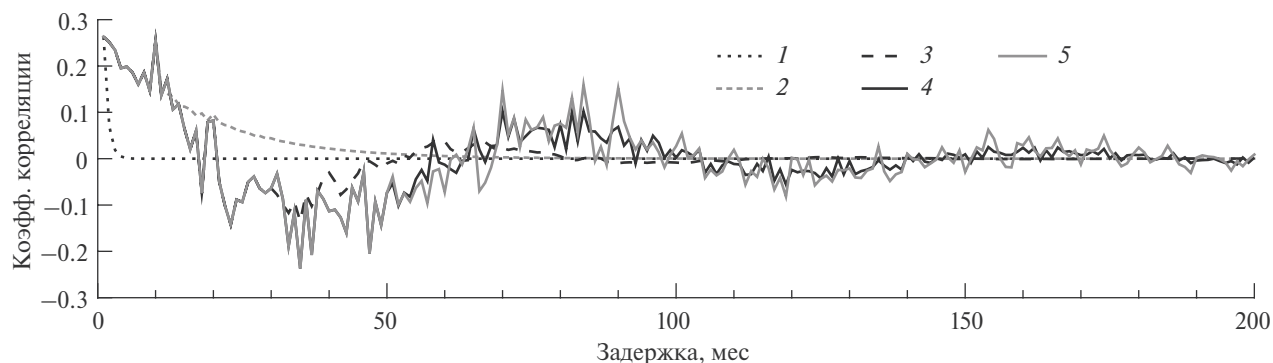


Рис. 3. Автокорреляционная функция АР процесса, аппроксимирующего временной ряд коррелированной ошибки, при значениях порядка авторегрессии k , равных 1 (1), 10 (2), 30 (3) 50 (4) и 90 (5).

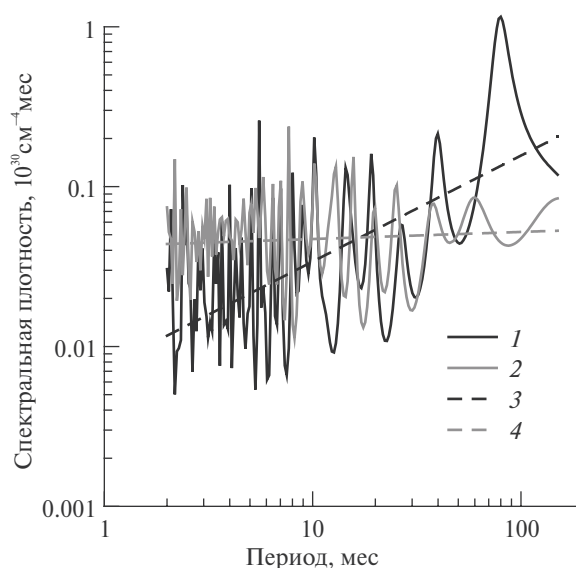


Рис. 2. Спектры коррелированной (1) и некоррелированной (2) ошибок и степенная аппроксимация спектров (3 и 4 соответственно). Некоррелированная ошибка соответствует порядку авторегрессии $k = 50$.

Коррелированная ошибка (остаток) ε и некоррелированная ошибка e , соответствующая порядку авторегрессии 50, приведены на рис. 1б, а на рис. 2 сравниваются спектры ошибок (рассчитаны с помощью метода максимальной энтропии [12]). Спектр коррелированной ошибки указывает на наличие в остаточном ряде ε вариаций с периодом около 80 мес. Спектр имеет общий наклон; черная штриховая прямая – степенная аппроксимация с показателем степени 0.7. В отличие от него спектр некоррелированной ошибки следует степенному закону с показателем, близким к нулю, что является теоретической величиной для спектра белого шума.

На рис. 3 показаны АК функции АР процессов, аппроксимирующих временной ряд коррелированной ошибки при различных значениях

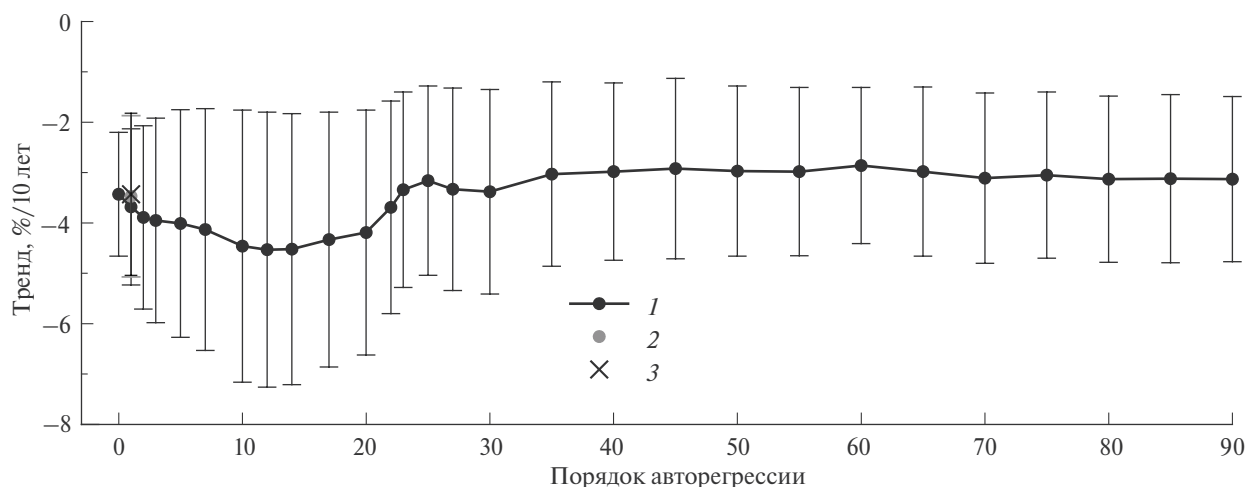


Рис. 4. Годовое значение тренда NO_2 и его 95% доверительный интервал в зависимости от порядка АР процесса, аппроксимирующего временной ряд коррелированной ошибки, оцененные с помощью нового метода (1), и оценки тренда и его 95% доверительного интервала, рассчитанные с помощью методов [8] (2) и [6] (3).

порядка авторегрессии k . АК функция, соответствующая значению $k = 90$ (сплошная серая кривая), демонстрирует колебания с периодом около 80 мес, что согласуется с периодом, проявившимся в спектре на рис. 2. АК функция для $k = 1$ экспоненциально затухает, а в случае $k = 10$ затухание проявляется на задержках более 10 мес. При значениях $k = 30$ АК функция содержит колебания, но их период пока меньше 80 мес. Значение $k = 50$ достаточно, чтобы разрешить 80-месячные вариации, и соответствующая АК функция, совпадая в диапазоне k от 1 до 50 мес с функцией для $k = 90$, в целом следует ей и при больших задержках.

Для демонстрации влияния порядка авторегрессии на оценки коэффициентов регрессии, на рис. 4 приведены годовые оценки линейного тренда NO_2 и их 95%-е доверительные интервалы в зависимости от порядка k АР процесса. Значение $k = 0$ соответствует случаю без учета сериальной корреляции. Рисунок 4 показывает, что величина k влияет и на значения, и на ошибки регрессионных коэффициентов. Оценки тренда при разных значениях k изменяются примерно на треть его величины; однако различия находятся в пределах 95%-х доверительных интервалов. Ошибка оценки тренда значительно, более чем вдвое, увеличивается с увеличением k при небольших значениях k , достигает максимума при $k = 12 \div 15$ и постепенно уменьшается, пока k не достигает 50. При $k > 50$ оценки тренда и их ошибки остаются примерно постоянными.

На рис. 4 серым цветом показана также оценка тренда и доверительный интервал, расчи-

танные по методу [8] (итерационная процедура). Крестиком показан тренд и доверительный интервал, оцененный по методу [6]. Эти оценки близки между собой, а значения трендов отличаются от оценки тренда, полученного по нашему методу при $k = 1$. Эти значения трендов практически (для метода [6] в точности) совпадают со значением тренда при $k = 0$, то есть в отсутствие сериальной корреляции.

На рис. 5 приведены оценки линейных трендов NO_2 и величин изменений содержания NO_2 , обусловленных различными предикторами, полученные при $k = 50$. На рисунках нанесены годовые, месячные и сезонные значения. Процентные значения рассчитаны относительно многолетнего среднегодового значения содержания NO_2 . Линейный тренд на рис. 5а отражает изменение содержания NO_2 за 10 лет. Эффект 11-летнего солнечного цикла (рис. 5б) соответствует изменению содержания NO_2 при изменении уровня солнечной активности в 11-летнем солнечном цикле от минимума к максимуму. Вулканический эффект на рис. 5в означает изменение содержания NO_2 , соответствующее максимуму АОТ. Эффект КДЦ на рис. 5г есть изменение содержания NO_2 при изменении экваториальной зональной скорости ветра от максимальной восточной до максимальной западной фазы. Эффекты САК и ЭНЮК (рис. 5д, е) — это изменения содержания NO_2 , обусловленные изменениями индексов САК и Ниньо 3.4 соответственно, на величины их среднеквадратичных отклонений. Эффект ЭНЮК приведен в виде отклика NO_2 на изменение только одного индекса Ниньо 3.4, так как отклик на другой,

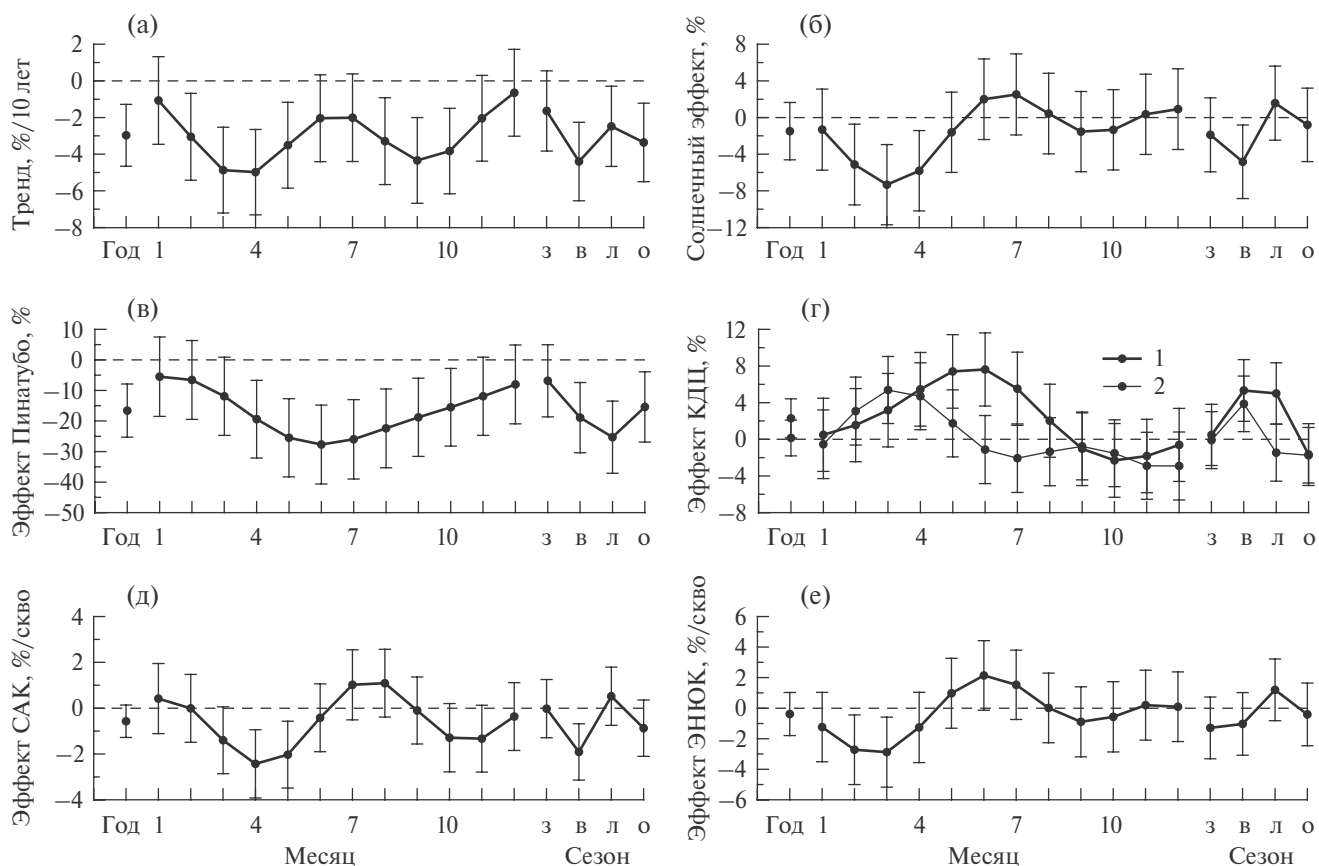


Рис. 5. Годовая (точки в левых частях рисунков), месячные (кривые в средних частях рисунков) и сезонные (кривые в правых частях рисунков) оценки линейного тренда (а) и величин эффектов солнечного цикла (б), извержения вулкана Пинатубо (в), КДЦ (г), САК (д) и ЭНЮК (е) и их 95% доверительные интервалы при порядке авторегрессии $k = 50$. Эффект КДЦ соответствует отклику NO_2 на скорость экваториального ветра на уровне 20 гПа (1) и в слое 40–50 гПа (2).

ортогональный ему индекс статистически незначим на уровне 95%.

Согласно рис. 5а, линейный тренд NO_2 сезонно зависим с максимальными по модулю значениями весной (до -6% за 10 лет) и осенью. Годовой тренд составляет около -3% за 10 лет и также статистически значим.

В то время как годовая оценка эффекта солнечной активности в NO_2 невелика и статистически незначима, значительный эффект получен для периода с конца зимы по середину весны (рис. 5б). Так, содержание NO_2 в марте во время фазы минимума солнечной активности на 8% больше, чем во время фазы максимума. Относительно среднемесячных значений содержания NO_2 эффект солнечного цикла в феврале и марте составляет около -10% .

Максимальный абсолютный эффект извержения вулкана Пинатубо, около -30% , получен для летнего сезона (рис. 5в). Однако если изменения содержания NO_2 нормировать на среднемесячные его значения, то вулканический

эффект в меньшей степени зависит от сезона и варьирует вблизи значения -20% .

Эффект КДЦ в NO_2 выражен преимущественно в весенне-летний период (рис. 5г). Его величина составляет летом около 8% относительно среднегодового содержания NO_2 . Разный сезонный ход двух кривых на рис. 5г указывает на то, что фазовый сдвиг эффекта КДЦ в NO_2 относительно экваториальной КДЦ изменяется с сезоном. Например, эффект в NO_2 в июне совпадает по фазе со скоростью ветра на 20 гПа. Эффект КДЦ в апреле запаздывает относительно КДЦ экваториальной скорости ветра на уровне 20 гПа на одну восьмую периода КДЦ (~ 3.5 мес), поскольку апрельские значения двух кривых на рис. 5г почти одинаковы, а фаза КДЦ экваториального ветра распространяется сверху вниз. Летние значения содержания NO_2 во время западной (положительной) фазы экваториального ветра на уровне 20 гПа больше, чем во время восточной фазы. Весенние значения содержания NO_2 во время западной фазы КДЦ

во всем слое 20–50 гПа в целом больше, чем во время восточной фазы.

Небольшой, отрицательный по величине, статистически значимый эффект САК в NO_2 получен для весеннего и, на пределе статистической значимости, осеннего периодов (рис. 5д). Изменения содержания NO_2 весной оцениваются в 2–3% при изменении индекса САК на величину его среднеквадратичного отклонения.

Отрицательный эффект ~3% получен для эффекта ЭНЮК в конце зимы – начале весны, а эффект противоположного знака, но на пределе статистической значимости, проявился в июне (рис. 5е).

Следует отметить, что оценки линейных трендов NO_2 , эффекта 11-летнего солнечного цикла и вулканического эффекта в NO_2 на Звенигородской станции, полученные методом множественной линейной регрессии, были представлены и в более ранних работах [16, 18, 26]. Знаки оценок из настоящей работы и из указанных работ совпадают. Годовая оценка тренда NO_2 , полученная с помощью нового метода, близка к наиболее поздней оценке в [26], хотя и несколько меньше ее по модулю, но в два раза меньше (по модулю) более ранних оценок в [16, 18]. Вулканический эффект по результатам настоящей работы близок к эффектам, полученным в [16, 18, 24] по данным вечерних измерений, но в полтора раза меньше приведенных в [27, 28] ранних оценок, полученных без использования множественной регрессии. Для эффекта солнечного цикла в вечерних данных об NO_2 в [24] получено статистически значимое отрицательное значение, почти вдвое превышающее по модулю статистически незначимую годовую оценку, полученную в настоящей работе. Причины отмеченных количественных расхождений могут быть разными. Имеют значение как длина анализируемого ряда, так и различия использованных регрессионных моделей, в частности, выбор предиктора для описания влияния вулканического аэрозоля.

4. ЗАКЛЮЧЕНИЕ

В работе предложен метод учета сериальной корреляции данных в задаче линейной регрессии. Одно из достоинств метода – то, что он позволяет учесть автокорреляцию остаточного ряда, которая может проявляться на больших масштабах. Эффективность метода продемонстрирована на основе анализа данных 26-летних

измерений общего содержания NO_2 на Звенигородской научной станции с применением модели множественной линейной регрессии.

По результатам работы можно сделать некоторые общие выводы, которые могут быть справедливыми при анализе других данных, так как долговременная корреляция присуща различным геофизическим параметрам.

Оценки регрессионных коэффициентов и их ошибки зависят от порядка k АР процесса, аппроксимирующего коррелированный остаточный ряд. В целом, чем больше масштаб автокорреляции, тем больший порядок авторегрессии требуется.

Сначала ошибка увеличивается с увеличением k . Затем она может достичь максимума и после этого начать уменьшаться. В случае с NO_2 на Звенигородской станции ошибка более чем удваивается в своем максимуме по сравнению с ее начальным значением.

Уменьшение ошибки может прекратиться при больших значениях k , если k достигнет значения, при котором АР процесс этого порядка обеспечит воспроизведение важных особенностей автокорреляционной функции остаточного ряда. Однако если ошибка не достигает максимума в диапазоне допустимых значений k , то следует полагать, что использованная регрессионная модель не вполне адекватно описывает поведение наблюдаемой переменной и должна быть усовершенствована (см. также [2]).

В случае с использованными данными об NO_2 максимум ошибки пришелся на $k \sim 10$, а затем она уменьшалась по достижении k значения ~50. В диапазоне k от 50 до 90 регрессионные коэффициенты и их ошибки сохраняли примерно постоянные значения.

Модель множественной линейной регрессии для данных наблюдений общего содержания NO_2 включает сезонно зависимый годовой тренд и сезонно зависимые эффекты воздействия на NO_2 11-летнего солнечного цикла, экваториальной КДЦ, САК, ЭНЮК и продуктов извержения вулкана Пинатубо. Годовые и сезонно зависимые оценки вклада этих эффектов в долговременную и межгодовую изменчивость NO_2 получены с учетом автокорреляции ошибки (остаточного ряда) до временного масштаба 50 мес.

Представленный в работе метод применим к анализу самых разных данных. Количественные характеристики зависимости регрессион-

ных коэффициентов и их ошибок от порядка AP процесса для них будут своими.

Источник финансирования. Работа выполнена при финансовой поддержке РФФИ (проект № 16-05-00663) и программ Российской академии наук. Данные измерений содержания NO₂ архивированы NDACC. Данные об индексе солнечной активности F10.7 представлены для свободного доступа Национальным геофизическим центром NOAA (NOAA National Geophysical Data Center). Данные о зональной скорости экваториального стратосферного ветра подготовлены Свободным берлинским университетом (Freie Universität Berlin). Индекс Ниньо 3.4 представлен Лабораторией исследований земной системы NOAA (NOAA-ESRL Physical Sciences Division). Данные об индексе САК представлены Климатическим отделом Университета Восточной Англии (Climatic Research Unit, University of East Anglia). Данные об аэрозольной оптической толщине подготовлены Годдардовским институтом космических исследований НАСА (NASA Goddard Institute for Space Studies).

СПИСОК ЛИТЕРАТУРЫ

- Beran J.* Statistics for long memory processes. New York, USA: Chapman & Hall, 1994. 315 p.
- Дранер Н.Р., Смит Г.* Прикладной регрессионный анализ. М.: Издательский дом "Вильямс", 2007. 912 с.
- Бокс Дж., Дженкинс Г.* Анализ временных рядов. Прогноз и управление. Выпуск 1. М.: Мир, 1974. 400 с.
- Bartlett M.S.* Some aspects of the time correlation problem in regard to test of significance // *J. Roy. Stat. Soc.* 1935. V. 98. № 3. P. 536–543.
- Folland C.K., Salinger M.J., Jiang N., Rayner N.A.* Trends and variations in South Pacific island and ocean surface temperatures // *J. Climate.* 2003. V. 16. № 17. P. 2859–2874.
- Weatherhead E.C., Reinsel G.C., Tiao G.C. et al.* Factors affecting the detection of trends: Statistical considerations and applications to environmental data // *J. Geophys. Res.* 1998. V. 103. № D14. P. 17149–17161.
- Hendrick F., Mahien E., Bodeker G.E. et al.* Analysis of stratospheric NO₂ trends above Jungfraujoch using ground-based UV-visible, FTIR, and satellite nadir observations // *Atmos. Chem. Phys.* 2012. V. 12. № 18. P. 8851–8864.
- Cochrane D., Orcutt G.H.* Application of least squares regression to relationships containing auto-correlated error terms // *J. Amer. Stat. Association.* 1949. V. 44. № 245. P. 32–61. DOI: 10.1080/01621459.1949.10483290.
- Tiao G.C., Reinsel J.C., Xu D., Pedrick J.H., Zhu X., Miller A.J., DeLuisi J.J., Mateer C.L., Wuebbles D.J.* Effects of autocorrelation and temporal sampling schemes on estimates on trend and spatial correlation // *J. Geophys. Res.* 1990. V. 95. № D12. P. 20507–20517.
- Bodeker G.E., Boyd I.S., Matthews W.A.* Trends and variability in vertical ozone and temperature profiles measured by ozonesondes at Lauder, New Zealand: 1986–1996 // *J. Geophys. Res.* 1998. V. 103. № D22. P. 28661–28681.
- Harris N.R.P., Hassler B., Tummon F. et al.* Past changes in the vertical distribution of ozone – Part 3: Analysis and interpretation of trends // *Atmos. Chem. Phys.* 2015. V. 15. P. 9965–9982.
- Kay S.M., Marple S.L.* Spectrum analysis – A modern perspective. // *Proc. IEEE.* 1981. V. 69. № 11. P. 1380–1419.
- Елохов А.С., Груздев А.Н.* Измерения общего содержания и вертикального распределения NO₂ на Звенигородской научной станции // *Изв. РАН. Физика атмосферы и океана.* 2000. Т. 36. № 6. С. 831–846.
- Gruzdev A.N., Elokhov A.S.* Validation of Ozone Monitoring Instrument NO₂ measurements using ground based NO₂ measurements at Zvenigorod, Russia // *Internat. J. Remote Sensing.* 2010. V. 31. № 2. P. 497–511.
- Gruzdev A.N., Elokhov A.S.* Variability of stratospheric and tropospheric nitrogen dioxide observed by visible spectrophotometer at Zvenigorod, Russia // *Internat. J. Remote Sensing.* V. 32. № 11. P. 3115–3127.
- Груздев А.Н.* Широтная зависимость вариаций стратосферного содержания NO₂ // *Изв. РАН. Физика атмосферы и океана.* 2008. Т. 44. № 3. С. 345–359.
- Liley J.B., Johnston P.V., McKenzie R.L., Thomas A.J., Boyd I.S.* Stratospheric NO₂ variations from at Lauder, New Zealand // *J. Geophys. Res.* 2000. V. 105. № D9. P. 11633–11640.
- Gruzdev A.N.* Latitudinal structure of variations and trends in stratospheric NO₂ // *Internat. J. Remote Sensing.* 2009. V. 30. № 15. P. 4227–4246.
- Werner R., Valev D., Atanassov A., Guineva V., Kirillov A.* Analysis of variations and trends of the NO₂ slant column abundance obtained by DOAS measurements at Stara Zagora and at NDACC European mid-latitude stations in comparison with subtropical stations // *J. Atmos. Sol.-Terr. Phys.* 2013. V. 99. P. 134–142.
- Боровский А.Н., Арабов А.Я., Голицын Г.С., Груздев А.Н., Еланский Н.Ф., Елохов А.С., Мохов И.И., Савиных В.В., Сеник И.А., Тимажев А.В.* Вариации общего содержания диоксида азота в атмосфере на Северном Кавказе // *Метеорология и гидрология.* 2016. № 2. С. 29–44.

21. Yela M., Gil-Ojeda M., Navarro-Comas M. et al. Hemispheric asymmetry in stratospheric NO₂ trends // Atmos. Chem. Phys. 2017. V. 17. № 21. P. 13373–13389.
22. Груздев А.Н. Квазидвухлетние вариации общего содержания NO₂ // ДАН. 2011. Т. 438. № 5. С. 678–682.
23. Агеева В.Ю., Груздев А.Н. Сезонные особенности квазидвухлетних вариаций стратосферного содержания NO₂ по результатам наземных измерений // Изв. РАН. Физика атмосферы и океана. 2017. Т. 53. № 1. С. 74–85.
24. Груздев А.Н. Оценка эффектов извержения вулкана Пинатубо в стратосферном содержании O₃ и NO₂ с учетом вариаций уровня солнечной активности // Оптика атмосферы и океана. 2014. Т. 27. № 6. С. 506–514.
25. Koike M., Jones N.B., Matthews W.A., Johnston P.V., McKenzie R.L., Kinnison D., Rodriguez J. Impact of Pinatubo aerosols on the partitioning between NO₂ and HNO₃ // Geophys. Res. Lett. 1994. V. 21. № 7. P. 597–600.
26. Gruzdev A.N., Arabov A.Ya. Borovsky A.N., Elan-sky N.F., Elokhov A.S., Golitsyn G.S., Mokhov I.I. Trends in stratospheric column NO₂ in mid-latitudes of the European part of Russia // Quadrennial Ozone Symposium, 4–9 September 2016, Edinburgh, United Kingdom. Presentations, http://presentations.copernicus.org/QOS2016-218_presentation.pdf
27. Elokhov A.S., Gruzdev A.N. Measurements of column contents and vertical distribution of NO₂ at Zvenigorod Scientific Station // Proc. SPIE. 1998. V. 3583. P. 547–554.
28. Gruzdev A.N., Elokhov A.S. Ground-based spectrometric measurements of vertical distribution and column abundance of NO₂ at Zvenigorod, Russia // Proc. SPIE. 2005. V. 5832. P. 292–299. DOI: 10.1117/12.619837

Accounting for autocorrelation in a linear regression problem on an example of analysis of atmospheric column NO₂ content

A.N. Gruzdev

*Obukhov Institute of Atmospheric Physics, RAS
Pyzhevsky per., 3, Moscow, 119017, Russia
E-mail: a.n.gruzdev@mail.ru*

Received: 06.12.2017

Revised version received: 16.04.2018

A method is proposed for taking into account a serial correlation (an autocorrelation) of data in a linear regression problem, which allows accounting for the autocorrelation on long scales. A residual series is presented as an autoregressive process of an order, k , that can be much larger than 1, and the autocorrelation function of the processes is calculated by solving the system of the Yule–Walker equations. Given the autocorrelation function, the autocorrelation matrix is constructed which enters the formulas for estimates of regression coefficients and their errors. The efficiency of the method is demonstrated on the base of the multiple regression analysis of data of 26-year measurements of the column NO₂ contents at the Zvenigorod Research Station of the Institute of Atmospheric Physics. Estimates of regression coefficients and their errors depend on the autoregression order k . At first the error increases with increasing k . Then it approaches its maximum and thereafter begins to decrease. In the case of NO₂ at the Zvenigorod Station the error more than doubled in its maximum compared to the beginning value. The decrease in the error after approaching the maximum stops if k approaches the value such that the autoregressive process of this order allows accounting for important features of the autocorrelation function of the residual series. Estimates have been obtained of seasonally dependent linear trends and effects on NO₂ of nature factors such that the 11-year solar cycle, the quasi-biennial oscillation, the North Atlantic Oscillation and other.

Keywords: linear regression, serial correlation, NO₂