



ВЫБОР ВРЕМЕННЫХ МАСШТАБОВ ПРИ ПОСТРОЕНИИ ЭМПИРИЧЕСКОЙ МОДЕЛИ

М.В. Сысоева, Т.В. Диканев, И.В. Сысоев

В работе рассматривается задача оптимального учёта временных масштабов исходного временного ряда при построении эмпирических прогностических моделей, используемых далее для оценки причинности по Грейнджеру. Предлагается для сравнительно коротких временных рядов, характерных, в частности, для медико-биологических приложений, использовать комбинацию из дальности прогноза и лага – различных по величине. Для подбора дальности прогноза и лага построены автоматизированные методики. Предложенный подход апробируется на большом числе различных эталонных систем, по итогам формулируются конкретные рекомендации по выбору дальности прогноза.

Ключевые слова: Реконструкция моделей по временным рядам, оптимальное вложение, анализ связанности, нелинейная причинность по Грейнджеру.

Введение

Анализ связанности систем по их временным рядам является одним из актуальных направлений современной науки. Классическим подходом для этой цели является метод причинности по Грейнджеру [1]. Основная идея этого метода заключается в построении прогностической модели, и если данные из первого временного ряда помогают точнее предсказывать поведение второго, то считается, что первая система влияет на вторую. Обзор различных вариантов метода причинности по Грейнджеру и сходных мер см. в [2, 3].

В настоящее время метод причинности по Грейнджеру используется в различных областях: в эконометрике (например, [4]), в нейрофизиологии (например, [5–7]), климатологии [8]. Одна из главных проблем при этом – удачный выбор структуры модели, в том числе – способа вложения.

В данной работе модели строились в виде нелинейного отображения последования вида

$$x_{n+P} = f_x(x_n, x_{n-l}, \dots, x_{n-(D-1)l}) + \xi_n, \quad (1)$$

где f_x – полином общего вида от D переменных [9], P – дальность прогноза, l – лаг, ξ – остатки модели, по смыслу являющиеся шумом. Коэффициенты модели подбирались методом наименьших квадратов.

При построении модели часто ограничиваются единственным лагом (он же и дальность прогноза), равным единице. Однако такой подход нередко приводит к малоэффективным и громоздким моделям с большой размерностью и огромным числом коэффициентов. В ряде работ, например, в [10] предлагаются алгоритмы выбора оптимальных величин и количества временных лагов. Но недостатком этих подходов является требование большого объёма данных: нужно иметь длинные временные ряды или их ансамбли. В данной работе предлагается компромиссный подход к учёту имеющихся в системе временных масштабов – использовать две различные величины: дальность прогноза и лаг. Таким образом, по сравнению с введением двух различных лагов можно уменьшить вычислительные затраты на подбор и использовать сравнительно короткие временные ряды, что типично, например, для медико-биологических приложений.

Цели работы – показать эффективность предложенного подхода и, основываясь на анализе результатов для набора численных примеров, сформулировать критерии выбора параметров моделирования, связанных с учётом временных масштабов: дальности прогноза P и лага l .

1. Описание метода и эталонных систем

Для анализа причинности по Грейнджеру вначале строится индивидуальная модель (1), учитывающая точки только из одного ряда $\{x_n\}_{n=1}^N$, влияние на который оценивается. Затем – совместная модель, которая учитывает точки из обоих рядов $\{x_n\}_{n=1}^N$ и $\{y_n\}_{n=1}^N$:

$$x_{n+P} = f_{xy}(x_n, x_{n-l}, \dots, x_{n-(D_1-1)l}, y_n, y_{n-l}, \dots, y_{n-(D_2-1)l}) + \eta_n, \quad (2)$$

где f_{xy} – полином общего вида от $(D_1 + D_2)$ переменных, P – дальность прогноза, l – лаг, η – остатки модели. Для обеих моделей рассчитываются среднеквадратичные ошибки аппроксимации ε_x для индивидуальной и ε_{xy} для совместной модели, коэффициенты подгоняются методом наименьших квадратов.

Коэффициент улучшения прогноза, характеризующий причинность по Грейнджеру, выражается через эти ошибки

$$G = \frac{\varepsilon_x^2 - \varepsilon_{xy}^2}{\varepsilon_x^2}. \quad (3)$$

Влияние дальности прогноза и лагов на причинность по Грейнджеру исследуется с помощью тестовых примеров, поскольку в них известны все параметры (размерность, порядок полинома), а направленность и силу связи (коэффициент k) мы можем задать сами.

В качестве тестовых примеров использовались следующие связанные системы.

1. Процессы авторегрессии первого порядка

$$\begin{aligned} x_{n+1} &= \alpha x_n + k y_n + \xi_n, \\ y_{n+1} &= \beta y_n + \eta_n, \end{aligned} \quad (4)$$

где $\alpha = 0.99$, $\beta = 0.98$.

2. Процессы авторегрессии второго порядка

$$\begin{aligned}x_{n+1} &= \alpha_1 x_n + \alpha_2 x_{n-1} + ky_n + \xi_n, \\y_{n+1} &= \beta_1 y_n + \beta_2 y_{n-1} + \eta_n,\end{aligned}\tag{5}$$

где $\alpha_1 = \beta_1 = 1.99$, $\alpha_2 = \beta_2 = -0.99$.

3. Системы ФитцХью–Нагумо

$$\begin{aligned}\frac{dx_1}{dt} &= x_1(a - x_1)(x_1 - 1) - x_2 + I_a + ky_2, \\ \frac{dx_2}{dt} &= bx_1 - \gamma x_2, \\ \frac{dy_1}{dt} &= y_1(a - y_1)(y_1 - 1) - y_2 + I_a, \\ \frac{dy_2}{dt} &= by_1 - \gamma y_2,\end{aligned}\tag{6}$$

$a = 0.8$, $b = 0.008$, $\gamma = 0.0033$, $I_a = 0.84$. Уравнения интегрировались методом Эйлера с шагом $h = 0.5$, поскольку в систему вводился динамический шум со среднеквадратичным отклонением $\sigma_n = 0.02\sigma_s$, где σ_s – среднеквадратичное отклонение сигнала. Устойчивость алгоритма тестировалась путём проверки на меньших шагах; далее временной ряд перевыбирался с шагом $\Delta t = 3h$, чтобы получить примерно 10^2 отсчётов на характерном периоде.

4. Системы Рёсслера

$$\begin{aligned}\frac{dx_1}{dt} &= -(x_2 + x_3) + ky_1, & \frac{dx_2}{dt} &= x_1 + a_x x_2, & \frac{dx_3}{dt} &= b_x - x_3(x_1 - c_x), \\ \frac{dy_1}{dt} &= -(y_2 + y_3), & \frac{dy_2}{dt} &= y_1 + a_y y_2, & \frac{dy_3}{dt} &= b_y - y_3(y_1 - c_y),\end{aligned}\tag{7}$$

где $a_x = 0.398$, $b_x = 2$, $c_x = 4$, $a_y = 0.2$, $b_y = 0.2$, $c_y = 5.7$ для хаотического режима и $a_x = 0.3$, $b_x = 0.2$, $c_x = 1.5$, $a_y = 0.25$, $b_y = 0.2$, $c_y = 2$ для периодического. В периодическом режиме система интегрировалась методом Эйлера с шагом $h = 0.0001$ и динамическим шумом ~ 2 , данные перевыбирались с интервалом $\Delta t = 500h$; в хаотическом – методом Рунге–Кутты 4-го порядка с шагом $h = 0.005$ без динамического шума, и данные перевыбирались с шагом $\Delta t = 10h$.

5. Системы Лоренца

$$\begin{aligned}\frac{dx_1}{dt} &= a_x(x_2 - x_1) + ky_1, & \frac{dx_2}{dt} &= x_1(b_x - x_3) + x_2, & \frac{dx_3}{dt} &= x_1x_2 - c_x x_3, \\ \frac{dy_1}{dt} &= a_y(y_2 - y_1), & \frac{dy_2}{dt} &= y_1(b_y - y_3) + y_2, & \frac{dy_3}{dt} &= y_1y_2 - c_y y_3,\end{aligned}\tag{8}$$

где $a_x = 10$, $b_x = 46$, $c_x = 83$, $a_y = 10.01$, $b_y = 47$, $c_y = 83$ для хаотического режима. Система интегрировалась методом Рунге–Кутты 4-го порядка с шагом $h = 0.001$ без динамического шума, и данные перевыбирались с шагом $\Delta t = 4h$.

Во всех случаях к сгенерированным временным рядам добавлялись реализации измерительного нормального некоррелированного шума со среднеквадратичным отклонением $\sigma_m = 0.01\sigma_s$. От каждой системы записывались короткие временные ряды длиной 2000 отсчётов или приблизительно 20–30 характерных периодов. Такая длина рядов была выбрана, поскольку она типична для разного рода физиологических данных, например, сигналов электроэнцефалограмм, магнитоэнцефалограмм, электрокардиограмм и др.

2. Выбор оптимального лага и дальности прогноза

Причинность по Грейнджеру показывает максимально корректные результаты, если индивидуальная модель подобрана наилучшим образом, поскольку иначе улучшение прогноза может быть обусловлено наличием во втором сигнале информации, присутствующей также и в первом, но недостаточно учтённой. Например, такое может быть, если на самом деле воздействие одностороннее, но направлено в сторону, противоположную предполагаемой, либо обе рассматриваемые системы находятся под некоторым общим внешним воздействием. Поэтому оптимальный лаг l подбирается по критерию минимизации ошибки аппроксимации именно индивидуальной модели ε_x^2 , при этом осуществляется перебор значений в определённом диапазоне, начиная от единицы. Такой подбор осуществляется при каждой дальности прогноза. На графиках зависимости оптимальных лагов от дальности прогноза (рис. 1) как для периодических, так и для хаотических процессов наблюдаются склоны и горизонтальные участки. Склоны на графиках соответствуют ситуациям, когда оптимальным является учёт точки, отстоящей от предсказываемой на определенное расстояние.

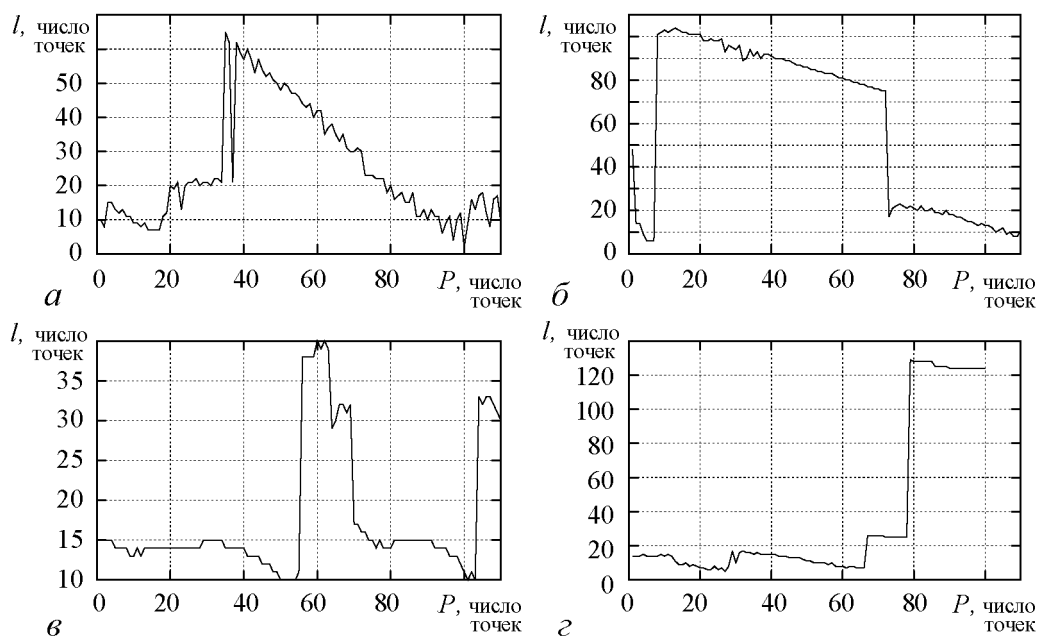


Рис. 1. Зависимость оптимального лага l от дальности прогноза P . Типичные картинки для периодических процессов – системы ФитцХью–Нагумо (а) и Рёсслера (б), для хаотических процессов – Рёсслера (в) и Лоренца (г). Везде порядок полинома $S = 5$

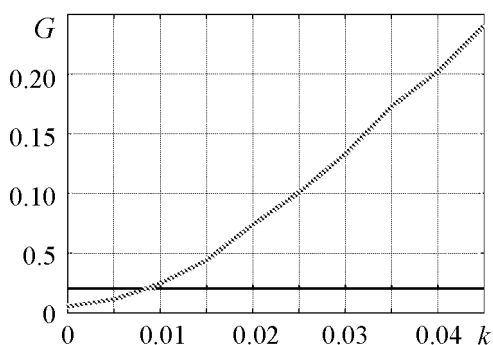


Рис. 2. Зависимость коэффициента улучшения прогноза G от коэффициента связи k при фиксированной дальности прогноза P . Чёрная горизонтальная линия – 95-е по величине значение G для $k = 0$ (проверка значимости); серая линия – среднее значение улучшения прогноза для каждого коэффициента связи

Для малых дальностей прогноза, как правило, лаг берётся такой, чтобы захватить точку, лежащую через интервал Z , соответствующий нулю автокорреляционной функции. Для больших дальностей прогноза оптимальным оказывается лаг, выбираемый так, чтобы захватить точку, лежащую через характерный период $P + l = T$ или через два характерных периода $P + 2l = T$. Горизонтальные участки говорят о том, что в векторе состояния необходимы точки, которые лежат на определенном расстоянии друг от друга (Z или T).

Методика исследования зависимости причинности по Грейнджеру от дальности прогноза заключается в следующем.

1. Генерируются 100 пар рядов без связи ($k = 0$), для каждой пары рассчитывается улучшение прогноза G , 95-е по величине значение выводим на график (чёрная горизонтальная линия на рис. 2). Таким образом осуществляется проверка значимости на уровне 5%.

2. Генерируются ряды с однонаправленной связью. Для каждого значения k делаются 100 пар рядов, для каждой пары рассчитывается G , и вычисляется среднее значение. На график для каждого значения силы связи выводится среднее значение улучшения (серая линия на рис. 2).

3. Если среднее значение G для данной силы связи k лежит выше 95-процентного уровня (то есть на рисунке серая линия лежит выше чёрной), то мы наблюдаем значимое улучшение прогноза. Обозначим пересечение этих двух линий \bar{k} . Чем раньше пересечение, то есть чем меньше \bar{k} , тем выше чувствительность метода.

4. Описанное выше делается для различных значений дальности прогноза. Строятся графики зависимости порога чувствительности от дальности прогноза $\bar{k}(P)$. Закрашенная на рис. 3 серым область показывает, в каком диапазоне значений коэффициента связи обнаруживается значимая связь. Закрашенная белым область – где, согласно методу связь считается незначимой.

5. Подобные графики зависимости порога чувствительности от дальности прогноза $\bar{k}(P)$ строятся для связи в правильную сторону (рис. 3, а, в) и в неправильную сторону (рис. 3 б, г).

Оптимальная дальность прогноза определяется по двум критериям.

- *Чувствительность метода*: значимая связь в правильную сторону обнаруживается при как можно меньшем значении коэффициента связи.
- *Специфичность метода*: значимая связь в неправильную сторону должна быть как можно меньше.

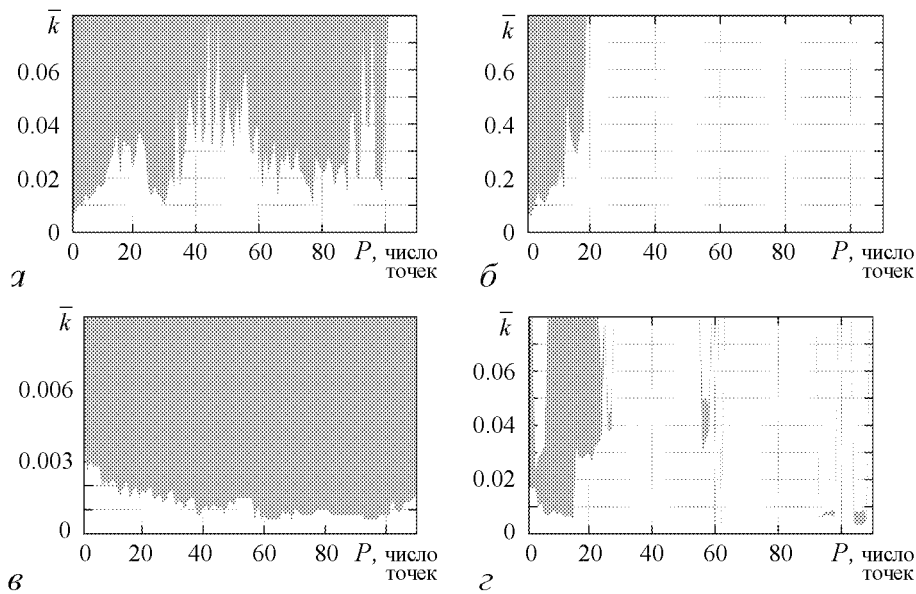


Рис. 3. Зависимость порога чувствительности \bar{k} от дальности прогноза P . В правильную сторону (слева) и в неправильную (справа). Сверху типичная картинка для периодических, снизу для хаотических процессов. Серая область – диапазон значений коэффициента связи, в котором обнаруживается значимая связь. Белая область – диапазон значений коэффициента связи, в котором обнаруживается незначимая связь

Если ориентироваться на чувствительность метода, то получаются следующие результаты.

Для процессов авторегрессии первого порядка с ростом дальности прогноза чувствительность метода падает, то есть предпочтительными являются единичные дальности прогноза. Это согласуется с тем фактом, что такие сигналы не имеют выделенного временного масштаба.

Для периодических режимов (процессов авторегрессии второго порядка, систем ФитцХью–Нагумо, систем Рёсслера в периодическом режиме) практически для всех дальностей прогноза чувствительность метода одинакова, причём чуть лучше чувствительность метода для единичных дальностей прогноза. Только при дальности прогноза, равной или характерному периоду T или его половине, чувствительность падает, поскольку улучшение предсказательной способности индивидуальной модели ухудшает работу метода.

Для хаотических режимов систем Рёсслера и Лоренца чувствительность метода на всех дальностях прогноза примерно одинакова. Чуть лучше метод работает при дальностях прогноза, равных половине характерного периода $T/2$ и характерному периоду T . И чуть хуже – при единичных дальностях прогноза.

Критерий качества модели невозможно построить, анализируя только чувствительность метода. Поэтому анализируется также специфичность метода, то есть ищется связь в неправильную сторону.

Для процессов авторегрессии первого и второго порядка специфичность метода очень хорошая: связь в заведомо неверную сторону не обнаруживается для всех дальностей прогноза в широком диапазоне значений коэффициента связи.

Для систем ФитцХью–Нагумо связь обнаруживается для малых дальностей прогноза; для систем Ресслера – для малых дальностей прогноза и для дальностей прогноза в районе половины характерного периода $P = T/2$.

Для хаотических процессов: для систем Рёсслера связь обнаруживается для малых дальностей прогноза, а также для дальностей прогноза в диапазоне от половины характерного периода $T/2$ до характерного периода T ; для систем Лоренца – в диапазоне от $T/2$ до T .

Подводя итог, можно порекомендовать брать дальность прогноза, равную четверти характерного периода $P = T/4$. Хотя в отдельных примерах это может быть не самым оптимальным выбором, в среднем оказывается, что в таком случае мы получаем хорошую специфичность метода при достаточно хорошей чувствительности. То есть вероятность ложных выводов относительно направления связи, определяемой методом Грейнджера, мала. Из общих теоретических соображений этот вывод может быть подтверждён тем, что четверть характерного периода примерно соответствует нулю автокорреляционной функции – расстоянию между отсчётами, на котором они статистически линейно не связаны, таким образом, последовательно включённые в вектор состояния точки несут максимально возможное количество информации (по крайней мере, для линейных моделей). Вследствие этого индивидуальная модель лучше описывает наблюдаемый сигнал, снижается вероятность получения ложных положительных результатов о наличии связей, что подтверждается результатами численных экспериментов.

Заключение

В работе исследуется эффективность оригинального подхода к выбору структуры эмпирических прогностических моделей, применяемых для анализа причинности по Грейнджеру. Существенная новизна состоит в том, что предлагается использовать при построении модели два различных параметра: дальность прогноза и лаг. Это позволяет учесть несколько временных масштабов, используя недлинный временной ряд. Для построенных таким образом моделей на основе анализа большого набора тестовых примеров сформулированы основные критерии выбора предложенных параметров, опирающиеся на анализ чувствительности и специфичности метода причинности по Грейнджеру.

Показано, что популярное среди исследователей значение дальности прогноза и лага, равное единице, оказывается далеко не оптимальным, поскольку построенные таким образом модели часто имеют очень низкую специфичность – показывают связь в заведомо ложную сторону. Оптимальной оказывается дальность прогноза порядка четверти характерного периода (или минимума автокорреляционной функции), именно при таком значении P достигается наилучшее сочетание специфичности и чувствительности. При этом лаг часто выбирается таким образом, чтобы в модели использовалась точка, отстоящая от предсказываемой на характерный период.

Работа выполнена при поддержке РФФИ: грант № 12-02-00377 и № 11-02-00599.

Библиографический список

1. *Granger C.W.J.* Investigating causal relations by econometric models and cross-spectral methods // *Econometrica*. 1969. Vol. 37, № 3. P. 424.
2. *Gourévitch B., Le Bouquin-Jeannés R., Faucon G.* Linear and nonlinear causality between signals: Methods, examples and neurophysiological applications // *Biological Cybernetics*. 2006. Vol. 95. P. 349.
3. *Rosenblum M.G. and Pikovsky A.S.* // *Physical Review E*. 2001. Vol. 64. 045202.
4. *Abhyankar A.* Linear and nonlinear Granger causality: Evidence from the U.K. Stock index futures markets // *The Journal of Futures Markets*. 1998. Vol. 18, № 5. P. 512.
5. *Bernasconi C., König P.* On the directionality of cortical interactions studied by structural analysis of electrophysiological recordings // *Biol. Cybern.* 1999. Vol. 81. P. 199.
6. *Smirnov D.A., Barnikol U.B., Barnikol T.T., Bezruchko B.P., Hauptmann C., Buehrle C., Maarouf M., Sturm V., Freund H.-J., and Tass P.A.* The generation of Parkinsonian tremor as revealed by directional coupling analysis // *Europhysics Letters*. 2008. Vol. 83. 20003.
7. *Сысоев И.В., Караваев А.С., Наконечный П.И.* Роль нелинейности модели в диагностике связей при патологическом треморе методом грейнджеровской причинности // *Изв. вузов. ПНД*. 2010. Т. 18, № 4. С. 81.
8. *Smirnov D.A., Mokhov I.I.* From Granger causality to long-term causality: Application to climatic data // *Physical Review*. 2009. Vol. E80. 016208.
9. *Chaos and its Reconstruction / Eds. G. Gouesbet, G. Meunier-Guttin-Cluzel, O. Merlaud.* Nova Science Publishers, New York, 2003.
10. *Vlachos I., Kugiumtzis D.* Nonuniform state-space reconstruction and coupling detection // *Physical Review E*. 2010. Vol. 82. 016207.

*Саратовский государственный
университет им. Н.Г. Чернышевского*

*Поступила в редакцию
После доработки*

*21.11.2011
20.03.2012*

SELECTING TIME SCALES FOR EMPIRICAL MODEL CONSTRUCTION

M.V. Sysoeva, T.V. Dikanov, I.V. Sysoev

The task is considered of taking into account the multiple time scales of original time series, with these time series being used for Granger causality estimation. It is proposed to use the combination of prediction length and lag, different in value, that could be fruitful for comparatively short times series, e. g. of medical-biological nature. The automated methods are constructed to select lag and prediction length values. The proposed approach is tested on a set of examples – ethalon systems. Based on this consideration the concrete proposal for prediction length value is formulated.

Keywords: Model reconstruction from time series, optimal embedding, coupling analysis, nonlinear Granger causality.



Сысоева Марина Вячеславовна – родилась в Саратове (1987). Окончила Лицей № 37 (2005) и факультет нано- и биомедицинских технологий СГУ (2011). Магистр техники и технологии. В настоящее время – аспирант кафедры динамического моделирования и биомедицинской инженерии. Научные интересы – анализ биологических сигналов методами математической статистики.

410012 Саратов, ул. Астраханская, 83
Саратовский государственный университет им. Н.Г. Чернышевского
E-mail: bobrichkek@mail.ru



Диканев Тарас Викторович – родился в Саратове (1979). Окончил Лицей прикладных наук (1997). Окончил факультет нелинейных процессов СГУ (2002). Кандидат физико-математических наук (2005). До 2011 года – доцент кафедры динамического моделирования и биомедицинской инженерии. В настоящее время – инженер компании Huawei Technologies. Научные интересы – реконструкция моделей по временным рядам, анализ временных рядов биологических систем. Автор более 40 публикаций.

Российское представительство компании
Huawei Technologies, Москва
E-mail: tvdikanev@yandex.ru



Сысоев Илья Вячеславович – родился в Саратове (1983). Окончил Лицей прикладных наук (1999) и факультет нелинейных процессов СГУ (2004). Защитил диссертацию на соискание учёной степени кандидата физико-математических наук (2007). Работал на кафедре электроники, колебаний и волн (2005–2007). В настоящее время – доцент базовой кафедры динамического моделирования и биомедицинской инженерии. Научные интересы – исследование сигналов биологической природы методами нелинейной динамики, исследование эффективности и модернизация подходов к анализу сигналов. Автор более 40 публикаций.

410012 Саратов, ул. Астраханская, 83
Саратовский государственный университет им. Н.Г. Чернышевского
E-mail: ivssci@gmail.com