Д.С. Жуков, С.К. Лямин

D.S. Zhukov, S.K. Lyamin

Проблемы идентификации розового шума в исторических данных

Problems of pink noise identification in historical data

Исследование выполнено при финансовой поддержке Министерства образования и науки России в рамках государственного задания вузам в сфере научной деятельности (проект № 33.956.2014/К), а также при поддержке РФФИ (проект № 14-06-00093а «Приложение теории самоорганизованной критичности к изучению исторических процессов»)

Аннотация, abstract: Приложение теории самоорганизованной критичности в социо-гуманитарных исследованиях возможно лишь после идентификации розового шума в социальных процессах. Однако формализованные знания о таких процессах обычно представляют собой относительно небольшие наборы точек данных. В этой статье предпринята попытка ответить на следующие вопросы. Возможна ли идентификация розового шума на незначительном количестве точек данных? Какое минимальное количество точек данных необходимо, чтобы мы могли корректно обнаружить розовый шум?

Application of the theory of self-organized criticality in social research is possible in the case of detection of pink noise in social processes. Usually, however, these processes are represented by small sets of data points. The article attempts to answer the following questions. Is it possible to identify the pink noise on a small number of data points? What is minimum number of data points is necessary so that we can detect pink noise correctly?

Ключевые слова, keywords: самоорганизованная критичность, розовый шум, социальные процессы, история, self-organized criticality, pink noise, social processes, history

Авторы, authors: Жуков Дмитрий Сергеевич – Тамбовский государственный университет им. Г.Р. Державина, кандидат исторических наук, доцент кафедры международных отношений и политологии, ineternatum@mail.ru

Лямин Сергей Константинович – Тамбовский государственный университет им. Г.Р. Державина, кандидат исторических наук, доцент кафедры Российской истории, laomin@mail.ru

Zhukov, Dmitry S. – Tambov State University, Tambov, Russian Federation, PhD, associate professor of the International Relations and Political Science Department, ineternatum@mail.ru

Lyamin, Sergey K. – Tambov State University, Tambov, Russian Federation, PhD, associate professor of the Russian History Department, laomin@mail.ru

УДК 303.094

Статья поступила: 07.03.2015

Статья принята к печати: 09.03.2015

© Д.С. Жуков, С.К. Лямин, 2015

Проблема

Для интерпретации исторических и прочих социальных процессов в духе теории самоорганизованной критичности требуется, как правило, идентифицировать розовый шум в исследуемых процессах. Довольно часто при этом мы сталкиваемся с необходимостью определить тип сигнала/процесса по небольшому количеству точек данных. Возникает вопрос: где граница, за которой мы не можем уверенно идентифицировать розовый шум?

Известно, что при сокращении количества точек данных сигнал довольно часто «белеет» – становится всё более и более похожим на случайный набор данных, на белый шум. Естественно, сам тип исследуемого процесса при этом не меняется – просто растет неопределённость в наших идентификационных процедурах. И дело здесь не в пороках конкретного инструментария. Это фундаментальный эффект. Какой бы упорядоченный процесс мы не взяли, его небольшой отрывок может быть случайным сочетанием событий. Собственно, если мы имеем одну точку данных, то она может принадлежать к любому типу процесса и быть частью абсолютно любой закономерности.

Но нас интересует не фундаментальные проблемы, а практический вопрос: на скольких точках мы можем применять спектральный анализ для идентификации розового шума по величине α в формуле

$$S = v\frac{1}{f^{\alpha}} \qquad (1)$$

где f – частота; S – мощность; v – коэффициент, выражающий соотношение единиц S и f; α – показатель степени. П. Бак полагает, что «степень α может принимать значения от 0 до 2» [Бак, 2014: 69]. Идеальный розовый шум – *1/f*-шум – имеет α=1.

Эксперименты

Для ответа на этот вопрос рассмотрим, насколько сильно при сокращении количества данных «белеет» сигнал, то есть как сильно растёт неопределённость. Мы взяли несколько образцов «хорошего» искусственно сгенерированного розового шума (для этих образцов заведомо точно известно, что α=1 [Little etc., 2007]); и исследовали, какую величину α мы обнаружим, взяв из образца разные отрезки (разное количество точек данных). Для этого было проведено 160 вычислительных экспериментов. Детально применяемые для обнаружения розового шума идентификационные процедуры описаны в нашей статье [Жуков и др., 2014].

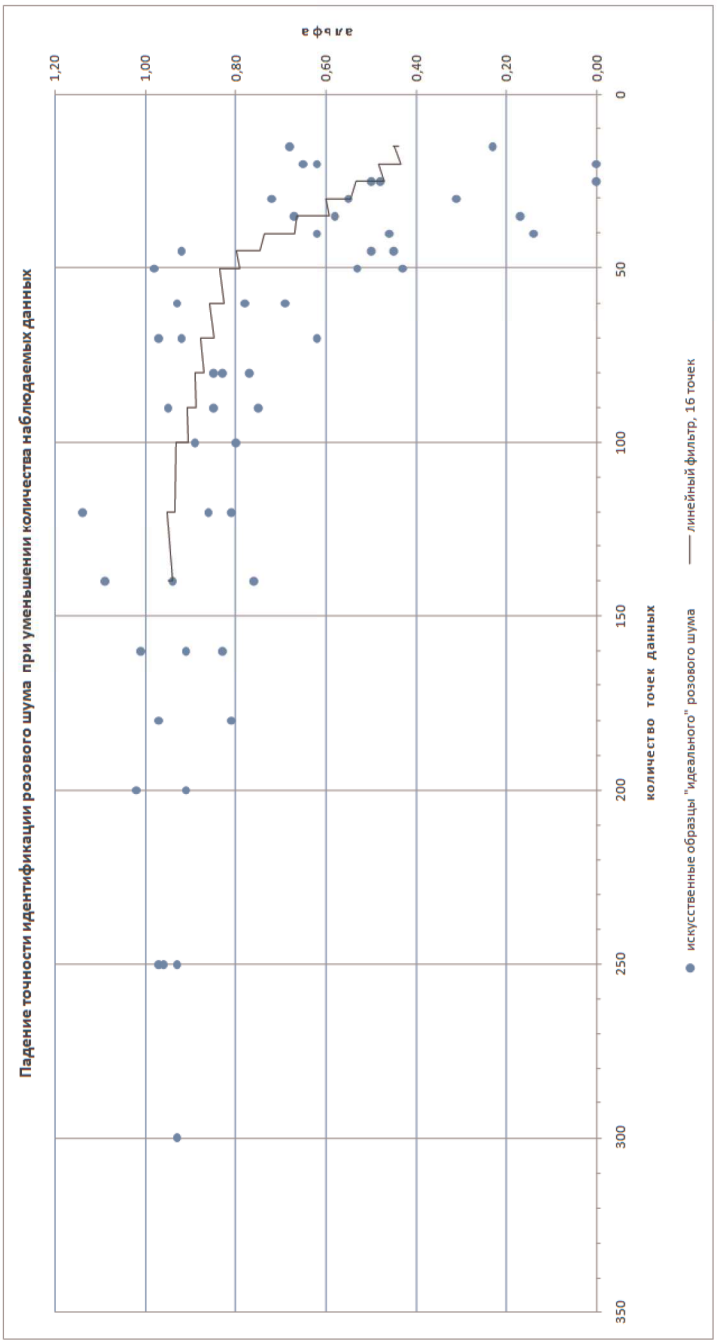

Диаграмма 1. Результаты изучения искусственных образцов розового шума, от 300 точек

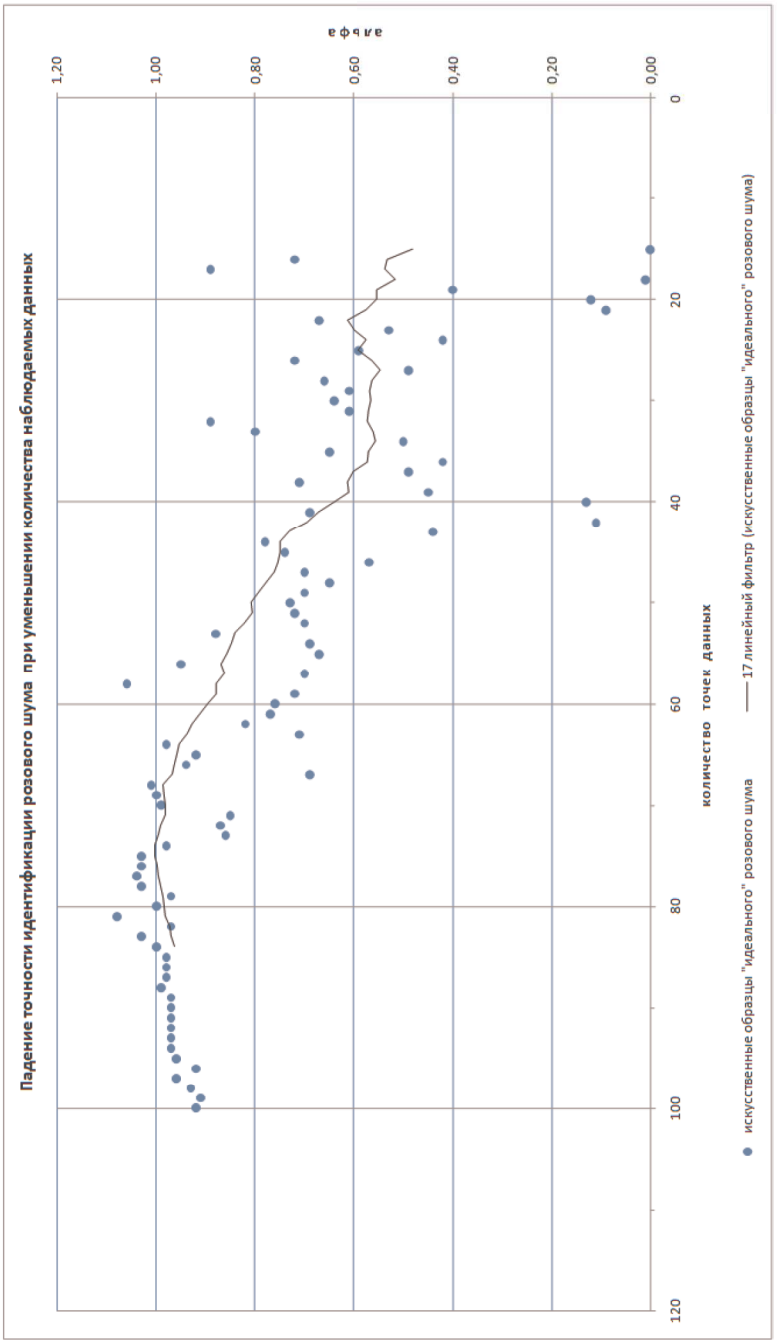

Диаграмма 2. Результаты изучения искусственных образцов розового шума, от 100 точек

Результаты

На диаграммах 1 и 2 по оси *x* откладывается количество точек данных, участвующих в конкретном эксперименте, а по оси *y* – величина α, полученная в данном эксперименте. Координаты каждой точки (то есть результаты конкретного эксперимента), таким образом, указывают на количество данных и величину α в данном эксперименте.

На диаграммах видно, чем α меньше, тем выше ошибка наших идентификационных процедур. Например, при α=0,4, ошибка составляет 60% от реального значения α, которое, как мы доподлинно знаем, равно 1.

Значительную часть вычислительных экспериментов мы поводили в диапазоне менее 100 точек данных – нас интересует более всего именно этот диапазон, так как именно 100 точками ограничена значительная часть временных рядов, которыми располагают историки. Эксперименты в этом диапазоне специально отражены на диаграмме 2; а на диаграмме 1 представлены эксперименты с более обширными рядами данных (до 300 точек), но тоже с упором на диапазон менее 100 точек.

Обсуждение и выводы

1. Как и предполагалось, с уменьшением числа точек данных, уменьшается и величина α. Кроме того, с уменьшением числа точек возрастает разброс значений α. Оба эти наблюдения говорят о том, что точность наших идентификационных процедур уменьшается. Это ожидаемо.

1.1. Никакой чётко определённой границы, после которой точность идентификации резко падала бы, нет. Точность падает плавно.

2. Как оказалось, точность наблюдений падает не катастрофически. На обеих диаграммах видно, что принятые идентификационные процедуры позволяют определить розовый шум и величину α с погрешностью до 20% уже при наличии 75 точек данных и более. А при наличии более 100 точек уменьшается разброс данных (и погрешность вычисления величины α).

3. В промежутке от 75 до 40-50 точек данных мы можем уверенно идентифицировать «хороший» розовый шум, т.к. идентификационные процедуры выдают 50 % ошибки, занижая в ряде случаев настоящую величину α. (Например, на 50 точках идеальный розовый шум может выглядеть как шум с α=0,5. Конечно, перед нами может быть сигнал с реальным, а не занижснным значсниси α=0,5. То есть при таком количестве данных мы не сможем отличить его от идеального розового шума с α=1. Но ведь мы относим к розовому шуму сигналы и α=0,5, и α=1).

Оговорка 3.1. В промежутке от 75 до 40-50 точек данных мы в ряде случаев не сможем идентифицировать недостаточно «хороший» розовый шум (например α=0,65), так как большая погрешность может снизить его α практически до нуля.

Оговорка 3.2. В промежутке от 75 до 40-50 точек данных мы не сможем с высоким уровнем доверия относиться к вычисленной конкретной величине α. Ина-

че говоря, мы будем видеть розовый шум, но вряд ли сможем сказать, насколько он «розовый» или «бело-розовый». Это затруднит сравнение различных процессов по конкретной величине α.

4. В промежутке от 40-50 до 20 точек данных мы можем не заметить даже идеальный розовый шум, так как идентификационная процедура способна выдать до 100 % ошибки, и мы не сможем даже хороший образец розового шума отличить от белого шума.

Оговорка 4.1. Как видно из диаграмм, во многих случаях даже на 20 точках данных (и, тем более, на 40 точках) мы можем получить ошибку в пределах 60%. Это позволяет нам предполагать, например, что если α исследуемого сигнала равно 0,4, то мы имеем дело с сигналом, α которого может быть в реальности равной от 0,4 до приблизительно 1. То есть, если мы имеем от 40-50 до 15 точек данных, то мы можем выдвигать обоснованные гипотезы о наличии розового шума, но не можем делать доказательные утверждения.

5. 15 точек данных является системным минимумом для спектрального анализа.

Посмотрим, на какое количество точек реальных данных опираются зарубежные коллеги, исследуя и обнаруживая розовый шум и степенные законы. Робертс и Тьюкотт [Roberts, Turcotte, 1998] приводят примеры исследований по следующим временным рядам:

– 82 (количество войн, исследованных Ричардсоном),

– 105 (количество войн, исследованных Ричардсоном в другой работе),

– 119 (количество войн, исследованных Леви),

– 118 (количество войн, исследованных Смолом и Зингером),

– 4284 (количество пожаров в США за определённый период на определённой территории),

– 120 (количество пожаров в США за определённый период на определённой территории),

– 164 (количество пожаров в США за определённый период на определённой территории),

– 298 (количество пожаров в США за определённый период на определённой территории).

Пиколи и коллеги [Picoli etc. 2014] обнаруживают эффекты СОК в следующих наборах данных:

– 3000 (количество конфликтных событий в Ираке),

– 123 (количество конфликтных событий в Афганистане),

– 2143 (количество конфликтных событий в Северной Ирландии).

Как видим, в значительном числе случаев размер исследуемых рядов колеблется около ста точек данных. Социальная реальность редко предоставляет в руки исследователей более массивные данные.

Наши расчёты показали, что для простой идентификации розового шума (без претензии на высокую точность вычисления величины α) достаточно меньшего количества точек – до 40-50. А для выдвижения обоснованных гипотез по поводу наличия

розового шума при удаче может потребоваться ещё меньше точек – даже до 20.

А если в идентификации участвовало 75 или даже более 100 точек данных, мы можем не только уверенно и доказательно обнаружить розовый шум, но и с высоким уровнем доверия относиться к вычисленной величине α.

Литература

Бак П. Как работает природа: теория самоорганизованной критичности. М.: УРСС, 2013.

Жуков Д.С., Канищев В.В., Лямин С.К. Возможности использования теории самоорганизованной критичности в изучении демографических процессов в российском позднем аграрном обществе // Историческая информатика. 2014. №1. С. 70-91.

Little M.A., McSharry P.E., Roberts S.J., Costello D.A.E., Moroz I.M. Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection // BioMedical Engineering OnLine. 2007. № 6.

Picoli, S., M. del Castillo-Mussot, H. V. Ribeiro, E. K. Lenzi, and R. S. Mendes. Universal bursty behaviour in human violent conflicts // Sci. Rep. 2014. Vol. 4. P. 1–3.

Roberts D.C., Turcotte D.L. Fractality and Self-Organized Criticality of Wars // Fractals. 1998. Vol. 6. № 4. P. 351–358.