Вычисление спектров гармонических колебаний при помощи Метода Наименьших Квадратов.

А.В.Антипин a1_mail@inbox.ru

Для определения периода и фазы колебаний, присутствующих в экспериментальных данных, разработан оригинальный компьютерный алгоритм, использующий метод полного перебора моделей, получаемых Методом Наименьших Квадратов (МНК).

Отличительная черта алгоритма – возможность использовать исходные данные с пропусками, с произвольно расположенными по оси X отсчётами, а также отсутствие необходимости предварительной подготовки данных в силу возможности задавать произвольные модели для МНК.

Calculation of harmonic oscillation spectra using the Least Squares Method.

A.V. Antipin

To determine the period and phase of the oscillations present in the experimental data, an original computer algorithm has been developed using the method of full iteration of models obtained by the Least Squares Method (LSM).

A distinctive feature of the algorithm is the ability to use source data with omissions, with samples arbitrarily arranged along the X axis, as well as the absence of the need for preliminary data preparation due to the possibility of setting arbitrary models for LSM.

Одним из основных способов обработки экспериментальных данных является их анализ на присутствие гармонических колебаний, который осуществляется с помощью метода Фурье анализа. Однако, часто исходные данные неравномерно распределены по оси X (по оси параметра). В этом случае применяются разные сглаживающие методы, например, интерполяция, после чего и производится Фурье анализ для выбираемых, в т.ч. и в областях интерполирования, точек.

Нам представляется, что проведение предварительной обработки, аналогичной интерполяции, оправдано только для не слишком важных данных. Интерполяции поинтервально, безусловно, улучшает ситуацию, но хотелось бы иметь нечто более интуитивно приемлемое.

В связи с таким подходом, при возникновении в нашей практике необходимости определить периоды явно гармонического колебания (рис.1), важность которого оценивалась в тот момент как крайне высокая, мы пошли по следующему пути.

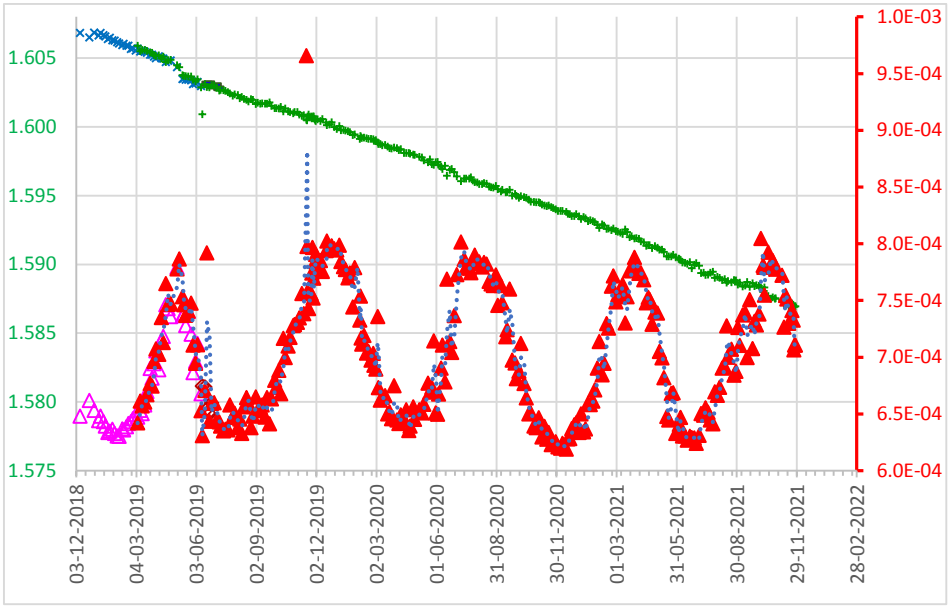

Рис.1

Регистрация наших данных, в силу длительного периода наблюдений, велась неравномерно, т.е. через схожие, но достаточно произвольные промежутки времени. Т.о., исходные данные содержат измерения, промежутки между которыми составляют от 1 суток до недель. Наиболее часто промежутки между последующими измерениями составляют 3 и 4 дня. Такая вариативность и неравномерность дат измерений привела к тому, что использование метода Фурье «в лоб», оказалось невозможным, т.к. последний требует равномерности отсчётов по времени.

Решая этот вопрос – поиск математически корректного метода определения периода колебаний для **не равноотстоящих по времени данных**, мы остановились на компьютерном вычислительном методе **тотального перебора** математических моделей с последующим анализом полученных данных. Т.е. на вычислении моделей Методом Наименьших Квадратов (МНК) во **ВСЕХ** заданных диапазонах (для нескольких параметров моделей), сравнении получающихся невязок и выборе модели с минимальной невязкой.

Предварительно необходимо отметить следующее. В качестве демонстрационного примера для описываемого ниже алгоритма, используется массив исходных данных с целочисленными значениями параметра по оси X. Значения параметра являются просто номером дня, относительно первой даты наблюдений. Учитывая, что **ВСЕ** экспериментальные данные **ВСЕГДА** определяются с помощью приборов, имеющих определённую точность, нетрудно распространить описываемый способ на данные, полученные для «непрерывного» параметра. Например, принимая за **один шаг в расчётах**, 1/2…1/3 **приборной ошибки**. В связи с этим замечанием, на адаптации описываемого способа к данным, непрерывным по оси X, мы не останавливаемся, считая такую адаптацию очевидной.

Демонстрационные данные, в нашем случае, являются массивом значений среднеквадратичного отклонения напряжения в зависимости от даты. Перебор осуществлялся для модели: функция= **sin(φ),** значения периода: **от 2-х до 1000 суток**, сдвиг ноля (фазы) параметра **φ: от 0 до 400 суток** для каждого из значений периода, т.е. для каждого числа: 2, 3, …, 1000.

По результатам такого тотального вычисления моделей, для каждого периода определялись экстремальные, т.е. Лучшая и Худшая модели по критерию невязки модели относительно исходных данных. Выбирались модели с минимальной – т.е. Лучшая и максимальной – Худшая, невязками.

Для отобранных моделей запоминались невязки, значения периодов, для каждого периода - сдвиг ноля и коэффициенты модели, полученные в МНК (см. ниже).

После выполнения таких расчётов, мы получаем результаты, основываясь на которых, легко можем определять значения периода и фазы для Лучшей модели, т.е. для модели, максимально близкой к исходным данным.

Тут автор не может не отметить, что даже бытовые компьютеры сегодня, имеют такие высочайшие показатели вычислительной мощности, которые не могут не восхищать исследователя. Автор долгое время занимается математическим программированием и должен признаться, что испытывает почти мистическое чувство непостижимости (по человеческим меркам) от скорости происходящих вычислений: когда программа, длительность кода которой автору прекрасно известна и по тем же человеческим меркам – необозрима и практически невыполнима человеком – просто «перемалывается» на глазах в компьютерной цифровой мельнице.

Например, если бы наши расчёты для поиска Лучшей модели производил человек, пусть и пользуясь калькулятором, это заняло бы у него время, порядка 500…1000 лет! Бытовой компьютер (3ГГц) выполняет эти расчёты за время, менее минуты…

Итак, на основе **Метода Наименьших Квадратов**, описанного в книге [1], как программа **SVD,** нами был написан код, реализующий вышеописанную идею. Безусловно, наш алгоритм является, скорее, демонстрацией идеи, чем завершённым методом и может и должен быть доработан.

Конкретная последовательность **нашей** реализации алгоритма состояла в следующем:

1. Берутся исходные данные: точки (t_j,y_j); **j=1…N**.
   (*В нашем случае **N**=307. Длительность периода наблюдений в днях – 1050 дней*).
2. Исходные данные нормируются. Т.е.: $e_j = y_j/a$, где **a** – среднее по всем **N** данным.
   Здесь мы должны сделать замечание, что нормировка наших данных происходила **по историческим причинам**. В общем случае очевидно, что т.к. модели в МНК могут быть произвольными и, в частности, включать в себя как учёт среднего, так и вычитание тренда, то **предварительная подготовка данных необязательна**.

3. Для каждого n_L и k_m**, Методом Наименьших Квадратов** ищется Модель в виде:
   $\sin[(t_j + k_m) * (2\pi/n_L)]$ и вычисляется её невязка **d** с исходными данными.
   Смысл присутствия в модели члена **sin()** абсолютно прозрачен – т.к. мы ищем гармоники в данных. В частности, в наших данных мы видим синусообразную кривую и подбираем её аппроксимацию в виде единственного синуса.
   Т.к. нам неизвестна ни фаза (т.е. ни точка t_k , где **sin=0**), ни период синуса n_L (который мы и ищем), то мы **В ДВУХ ЦИКЛАХ ТОТАЛЬНО** перебираем **периоды** и **фазы**.
   Перебор величины **периода** синуса осуществляется в цикле по числу n_L. Для **n** =1 период равен 1 дню, для **n** =2 – два дня и т.д. Т.о. для **n** =N, период синуса равен N дням.
   Перебор **фазы** осуществляется числом k_m , смысл которого – СДВИГ графика синуса **влево** по оси **n** (играющей роль оси **X**) на k_m дней. Т.е., пусть для некого периода, при **k**=0, значение синуса в точке, например, n_j= 7, равно **R.** Тогда в процессе перебора фазы по **k**, такую же величину **R** синус будет иметь при **k**=1 в точке **n**=6, при **k**=2 в **n**=5, при **k**=3 в **n**=4,

и т.д. Т.к. исходные данные «неподвижны», одна и та же (для каждого n_L) модель сдвигается влево и, соответственно, невязка её с исходными данными изменяется.
Невязка вычисляется обычным образом: как сумма квадратов разности между исходными точкам {$e(n_j)$} и значениями Модели в тех же n_j ($j=1...N$).
Т.е. $d= \Sigma_{j=1}^{j=N} (e(n_j) - Модель(\text{в точке } n_j))^{**}2$
(*В нашем случае* n*=2…1000,* k*=0…400. Используемая нами невязка, также по историческим причинам, равна КОРНЮ из указанной невязки* d*).*

4. Т.к. мы вычисляем Модели для множества значений **периодов** и **фаз**, то для каждого значения **периода** определяются две экстремальные невязки в зависимости от **фазы (сдвига):** Максимальная (т.е. Худшая) и Минимальная (т.е. Лучшая).
5. Все полученные данные для КАЖДОЙ модели: значение периода, значение смещения, Худшая и Лучшая невязки для каждого значений периода (которые отобраны из всего массива невязок для всех сдвигов для этого конкретного периода), а также коэффициенты этой модели, полученные в МНК, могут выводиться в результирующий файл **REZ-1**.
(*В нашем случае размер этого файла составляет (при распечатке всех «блоков сдвига» - см. ниже) до 10МБ, время счёта* ***ВСЕХ*** *вариантов на обычном стационарном бытовом компьютере с процессором 3 ГГц - менее минуты*).

6. Полученный файл забирается в **EXCEL** и строятся графики насчитанных Худшей и Лучшей невязки, в зависимости от n_L (играющей роль переменной **X**).
(*График невязок в зависимости от периода в нашем случае - Рис.2).*

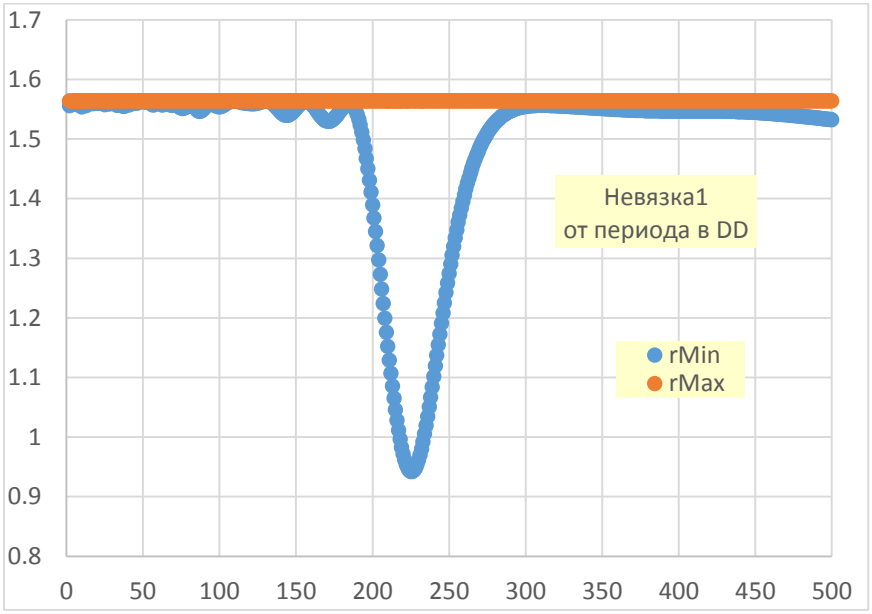

Рис.2 График ***Лучшей*** *(****rMin****) и* ***Худшей*** *(****rMax****) невязок между исходными данными и Моделями, полученными по МНК. Горизонтальная ось периодов* n *сокращена с* ***max*** *значения* n*=1000 (в нашем случае) до* n*=500.*

Для каждого значения периода (каждое n_L*) невязки* ***rMax*** *и* ***rMin*** *равны значениям на красной и синей кривой, соответственно. Эти* ***rMax*** *и* ***rMin*** *являются максимальным и минимальным, соответственно, значениями из массива* ***<u>всех</u>*** *невязок, насчитанных* ***для каждого*** *периода* n_L*. Каждый такой массив состоит из невязок для* ***<u>всех</u>*** *фаз (сдвигов) в диапазоне сдвига от 0 до 400 (в нашем случае) дней. Т.о., для каждого значения периода*

***n_L**, невязки **rMax** и **rMin** отобраны из массива в 400 невязок. Величины прочих, не экстремальных 398 невязок, лежат между **rMax** и **rMin** по оси **Y**.*

На полученном графике отчётливо виден **экстремум**, т.е.:

глобально лучшее (минимальное) значение невязки, являющейся разностью между исходными данными и Моделью, для диапазонов полного перебора по n_L и по k_m.

Этот экстремум (***rMin***) и соответствует разыскиваемому периоду для Лучшей модели.

Т.о., легко определяется как период n_0, так и фаза k_0 Лучшей Модели, имеющей, в нашем случае, вид единственного синуса: $\textbf{Модель}_0 = C_0 * \sin[(t_j + k_0) * (2\pi/n_0)]$.

7. Определение **периода** и **фазы** производится следующим образом.
   Период. По полученному графику определяется точное значение минимума **nMin**. Это значение равно координате (на горизонтальной оси периодов) для минимального значения невязок - синей кривой – см. Рис.2. (*В нашем случае это значение составляет 225 дней.).*

8. **Фаза**. Файл результатов счёта **REZ-1** в режиме «**ПОДРОБНО**», пересчитывается только для найденного значения **nMin.** Такой пересчёт делается в связи с тем, что выводить в файл сразу необходимые нам только теперь подробные результаты бессмысленно. Во первых: т.к. эти данные насчитываются для ВСЕХ вариантов перебора, файл приобретает огромные для **EXCEL** размеры и содержит бессмысленно подробную информацию. Из всей такой информации нам требуется только один «блок сдвигов». Этот блок содержит невязки **для всех** сдвигов для нужного значения периода.
   (*В нашем случае в файле* **REZ-1** *м.б. записано 999 таких блоков- для каждого периода в диапазоне от 2 до 1000 суток. Каждый такой блок оформлен в виде, показанном в Таблице 1.*
   ***Первая строка**: **k** – значение сдвига (фазы), **rR** – невязка для указанных в таблице значений **n** и **k**, **C** – коэффициент при синусе (Модели), **n** – напоминание периода в днях для обобщающих значений, расположенных в самом низу блока, для невязок **rMin** и **rMax**.*
   ***Вторая строка**: **n** – ещё раз значение периода в днях.*

k	rR	C[1...1]	n	rMin	rMax
n=	225				
0	1.051391	-0.09501			
1	1.036914	-0.09605			
2	1.0232	-0.09701			
3	1.010328	-0.09789			
…	…	…			
12	0.943307	-0.10177			
13	0.94216	-0.10177			
14	0.942354	-0.10168			
…	…	…			
398	1.526539	0.02696			
399	1.53343	0.024372			

400	1.539642	0.021764			
			225	0.942089	1.563739

Таблица 1. *Блок результатов для конкретного периода (**n_{225}=225** дней). Диапазон сдвигов: **k_m=0…400** дней. Цветом отмечено значение фазы (сдвига) для нашего **МИНИМУМА** невязки. Сдвиг **k_{13}= 13** дней. (см. п. 9).*

9. По первым двум столбцам выбранного блока строится график. (*В нашем случае – Рис.3)*. На графике горизонтальная ось соответствует сдвигу (фазе), а вертикальная – невязке для этой фазы. Естественно, в данной модели мы получаем синусоиду с периодом n_L и максимумом= ***rMax***, а минимумом= ***rMin***. Далее, для использования фазы Модели, выбирается минимум. Т.к. от минимума до минимума аргумент синуса изменяется на **2π**, то в принципе, можно выбирать любое значение сдвига, соответствующее любому минимуму. (*В нашем случае мы всегда выбираем самый левый минимум. Т.о., **k**= 13)*.

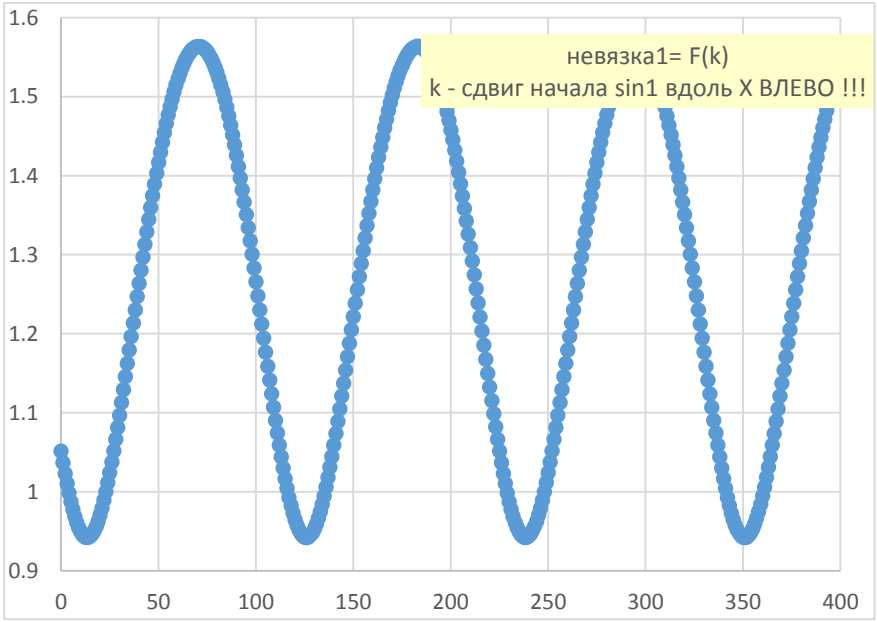

*Рис.3 Невязка1 от сдвига (фазы) Модели с периодом **n**= 225.*

10. Далее, используя найденные значения n_{225}, k_{13} и коэффициент при синусе: $C_{225,13}$ (см. Таблицу.1), для проверки **вычисляется лучшая Модель**. Модель, вычисленная для всех исходных точек t_j выводиться в файл **REZ-2** и **забирается в EXCEL,** где отрисовывается на графике вместе с исходными данными. (*В нашем случае: **n**= 225, **k**= 13 и $C_{225,13}$ = -0.10177).*

Т.о., в нашем случае, окончательно, Лучшая модель записывается в виде:

$$y = -0.10177 * \sin [(t_j + 13) * (2\pi/225)],$$

где t_j номер дня, который, в нашем случае, отсчитывается от даты 09-12-2018, соответствующей дню с **n=** 1.

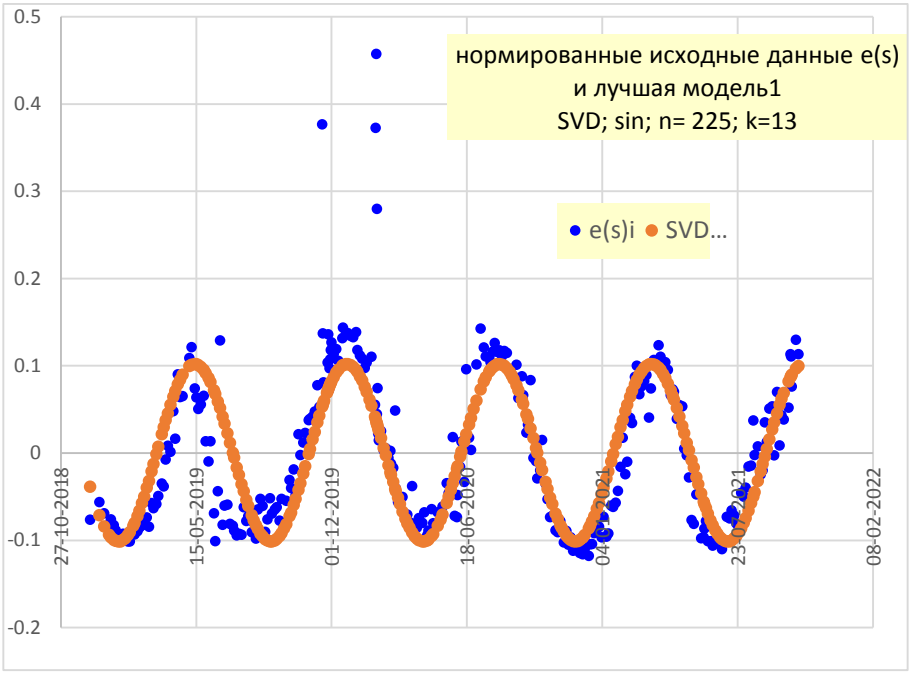

*Рис. 4 Нормированные исходные данные и **лучшая Модель1**, полученная по описанному алгоритму.*

Фаза определяется следующим образом.

Т.к., в данном случае, коэффициент при синусе **<0**, это означает, что интересующий нас минимум синусообразной кривой на графике рис.3, на самом деле является максимальным значением функции **sin(φ)** и равен 1**.** Т.о., нам надо решить простейшее уравнение:

$(t_j + 13) * (2\pi/225) = \pi/2$.

В нашем случае, мы получаем ***t= 43,25*** – номер дня, отсчитываемый от дня ***1*** (который соответствует указанной выше дате 09-12-2018). Это 26-10-2018 - дата, когда **(φ)=0.**

Напоминаем, что параметр k определяет сдвиг ВЛЕВО!

При положительном коэффициенте при синусе, интересующий нас минимум на графике, аналогичном графику рис.3, и является минимумом, т.е. **sin(φ) = -1.** Т.о., уравнение превращается (значение k - условно) в:

$(t_j + 13) * (2\pi/225) = 3*\pi/2$.

Далее, можно взять остатки/ невязки между полученной Лучшей Моделью и исходными данными и аппроксимировать уже эти остатки по описанному алгоритму.

Понятно, что процесс аппроксимации последовательно получаемых остатков можно повторять столько раз, сколько это посчитает разумным исследователь. В результате, если это необходимо, будет получен спектр, аналогичный стандартному спектру по методу Фурье. Этот будет спектр коэффициентов {$\mathbf{C_m}$} при синусах, для найденных описанным выше способом периодов {$\mathbf{n_m}$}, для каждой {**Лучшей модели$_m$**}, которая определяется в каждой последовательной {**аппроксимации$_{(m)}$**}. Естественно, что для получении спектра, в каждой последующей аппроксимации период Лучшей модели указывает самый «глубокий» из минимумов, наблюдаемых на графиках, аналогичных графику рис. 2.

Однако, необходимо помнить, что **техническая осуществимость процедуры не гарантирует её физического смысла**. Например, проведённые таким образом аппроксимации на глубину **12** остатков - В НАШЕМ СЛУЧАЕ - не показали каких-то интерпретируемых результатов, кроме **первого периода**. Все эти периоды носили произвольный характер в том смысле, что оказалось невозможно привязать их к какому-то явлению. Это отчётливо видно уже по графику **первых остатков** - рис. 5, который выглядит схоже с графиком шумов, хотя на нём и просматривается некоторая структура. Графики следующих остатков, с номерами от 2 и до 12, сохраняют, в целом, такой-же случайный вид. Невязки последующих моделей уменьшаются весьма медленно, а минимумы, аналогичные минимуму рис.2, быстро превращаются из минимума на гладкой кривой, в одну из хаотически расположенных точек. Всё это говорит, естественно, о присутствие значительной шумовой компоненты в исходных данных.

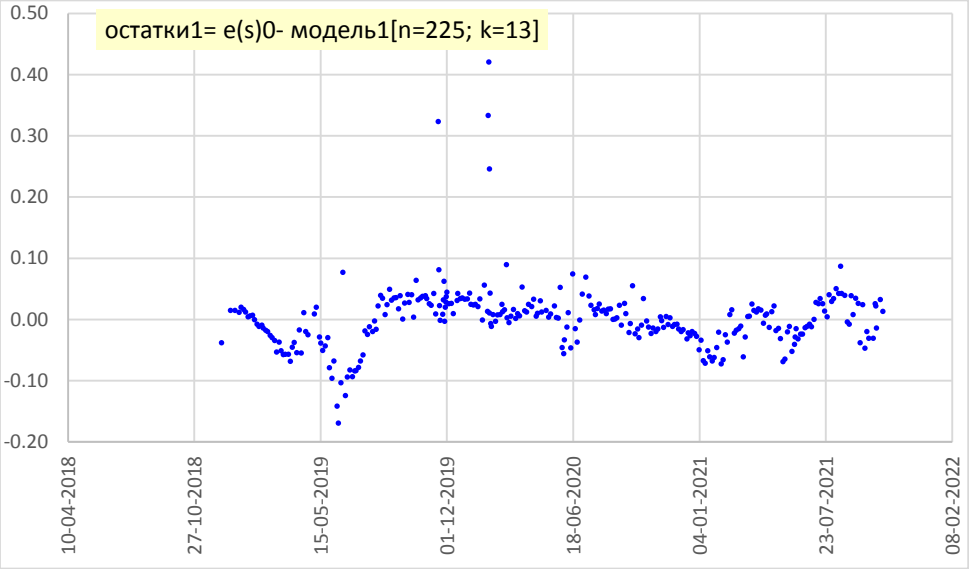

Рис. 5. Остатки1 = (нормированные исходные данные) – (Лучшая Модель1)

<u>ТОЛЬКО В КАЧЕСТВЕ ДЕМОНСТРАЦИИ</u> мощности и работоспособности алгоритма (***<u>и в качестве курьёза</u>***), на рис. 6 отрисована кривая, являющаяся суммой моделей **1…12** (все модели - синусы). Модель 1 аппроксимирует исходные синусообразные данные и показана на рис. 4. Модели 2…12 являются аппроксимациями последовательно получающихся остатков: т.е., например, Модель7 – это аппроксимация остатков 6.

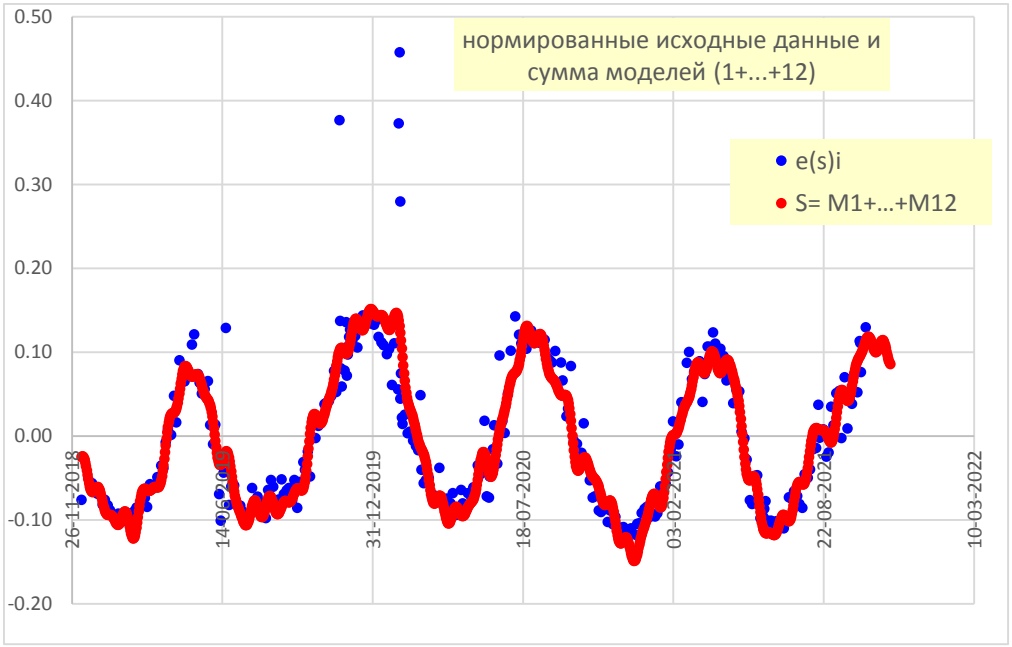

*Рис. 6. Иллюстрация **технической работоспособности** описываемого метода. Исходные данные (синие точки) и сумма моделей 1…12 (красная кривая).*

Т.о., разработанный алгоритм показал свою **мощность, прекрасную работоспособность и теоретическую прозрачность.** С нашей точки зрения, этот метод определения периодов и фаз, вполне может оказаться **лучшим вариантом** для корректного получения спектра колебаний в относительно сложных случаях. К таким случаям относятся массивы неравномерных по времени (или другой переменной) исходных данных, данные с пропусками и другие не стандартизованные, в этом смысле, данные, обработка которых с помощью метода Фурье «в лоб» – затруднительна, или невозможна.

Как уже указывалось выше, переход от дискретных значений по оси X к «непрерывным» значениям, также осуществляется без каких-либо принципиальных трудностей.

К достоинствам способа стоит отнести его неприхотливость к подготовке исходных данных. Алгоритм не требует предварительного вычитания среднего или тренда, т.к. модель МНК в виде, например, ***M= C1*1+C2*x+C3*SIN(φ)*** автоматически учитывает и средний уровень и тренд.

Литература

[1] Дж. Форсайт и др. Машинные методы математических вычислений. М. 1980.