

Метод “Гусеница”-SSA: анализ временных рядов

Ссылка на полную версию работы находится на странице

<http://www.gistatgroup.com/gus/books.html>

9 Приложение В: пример анализа временного ряда методом “Гусеница”-SSA

Продемонстрируем методику применения метода “Гусеница”-SSA на примере анализа временного ряда FORT (объемы месячных продаж крепленых вин в Австралии с января 1984 года по июнь 1994 года). Так как мы будем рассматривать как сам ряд длины $N = 174$, так и его подряд, состоящий из первых 120 точек, то там, где это необходимо, будем использовать обозначения FORT174 для всего ряда и FORT120 для укороченного ряда. Приведенные графики будут одновременно являться иллюстрацией понятий, возникающих при изложении теоретических аспектов метода.

Визуальный анализ ряда, изображенного на рис. 3, говорит о том, что ряд имеет тренд, который должен хорошо описываться либо линейной функцией либо убывающей экспонентой, а также сезонное поведение довольно сложной меняющейся формы. Периодограмма центрированного ряда (для увеличения резкости изображения она построена по первым 168 точкам, так как 168 кратно двенадцати, числу месяцев в году) подтверждает такое предположение (рис. 4).

Поставим задачу поиска разложения данного ряда на три компоненты — тренд, сезонную компоненту и шум.

Выбор длины окна. Посчитаем, какая может быть размерность у сигнала (состоящего из тренда и периодической компоненты). Тренд ряда FORT скорее всего описывается 1–2 собственными тройками, сезонная компонента раскладывается на гармонические составляющие с частотами $1/12$ (годовая периодичность), $1/6=2/12$ (полугодовая), $1/4=3/12$, $1/3=4/12$, $1/2.4=5/12$, $1/2=6/12$. Все гармоники с частотами, меньшими, чем $1/2$, имеют размерность 2, гармоника с периодом два имеет размерность 1 (мы предполагаем, что амплитуды каждой их гармоник меняются по экспоненте, возможно с разными показателями для разных частот). Таким образом, предполагаемая размерность ряда не должна превосходить 13. Если бы ряд не содержал шума и компоненты ряда были строго разделимы, то для получения разложения ряда на тренд и сезонную компоненту достаточно было бы взять длину окна, равной 13. Однако для реальных данных точная разделимость практически не встречается, поэтому необходимо воспользоваться результатами о приближенной (асимптотической) разделимости медленно-меняющегося тренда и гармоники. Для достижения лучшей точности мы должны выбрать длину окна большой, близкой к половине длины ряда (так как скорость сходимости ошибки разделимости к нулю имеет порядок $1/\min(L, K)$, где $K = N - L + 1$, N — длина ряда). Так как кратность L и K периоду должна улучшать отделимость периодической компоненты, причем более важна кратность меньшего из этих чисел, то выберем длину окна, равной $L = 84$ ($N = 174$, следовательно, $K = 91$).

Подход к идентификации собственных троек. Рассмотрим результат сингулярного разложения траекторной матрицы ряда при таком выборе длины окна. На рис. 5 приведено изображение собственных векторов из первых шести собственных троек сингулярного разложения. Заметим, что вид факторных векторов, или правых сингулярных векторов, практически не отличается от левых, так как длина окна близка к половине ряда. Если бы мы взяли длину окна меньше, то собственные вектора имели бы более регулярный вид по сравнению с факторными, которые содержали бы в себе, например, изменение амплитуд для гармонических компонент.

Для идентификации собственных троек воспользуемся результатами о виде собственных векторов, соответствующих тренду и гармоникам при условии из приближенной разделимости.

Идентификация тренда. Начнем с идентификации тренда. Мы знаем, что сингулярные (в частности, собственные) вектора имеют в целом такой же вид, как и компонента исходного ряда, которой они соответствуют. Поэтому на одномерных диаграммах собственных векторов нужно найти медленно-меняющиеся собственные вектора. В данном случае только один, а именно, первый собственный вектор имеет требуемый вид. Так как получилось, что в данном случае тренд

описывается единственной собственной тройкой, то это означает, что тренд аппроксимируется экспонентой. Чем сложнее форма тренда, тем больше его (приближенная) размерность и тем большее число собственных троек ему соответствует.

Идентификация гармоник. Займемся теперь идентификацией гармонических (возможно, с меняющейся амплитудой) компонент, порожденных сезонной компонентой исходного ряда. На рис. 5 видно, что собственные тройки с номерами 2—6 возможно соответствуют каким-либо гармоникам, так как имеют регулярное периодическое поведение. Однако, поскольку гармоника с периодом большим, чем 2, порождает две собственные тройки (косинус с периодом 2 порождает только одну собственную тройку, сингулярные вектора которой имеют пилообразный вид, а таких на одномерных диаграммах в данном случае не обнаружено), то искать относящиеся к гармоникам пары собственных троек удобнее на двумерных диаграммах. Так как при достаточно большой длине ряда соответствующая пара собственных чисел имеет близкие значения, то достаточно рассматривать двумерные диаграммы собственных векторов из соседних, упорядоченных по собственным значениям, собственных троек. На рис. 6 можно различить регулярные двумерные изображения, образующие двумерные траектории с вершинами, лежащими на кривой, имеющей спиралеобразную форму. Это означает, что соответствующая пара собственных векторов порождена модулированной гармонической компонентой исходного ряда. Таким образом, получаем, что собственные тройки (используем аббревиатуру ET — eigentriples) ET_{2,3} соответствуют периоду 12, ET_{4,5} — периоду 4, ET_{6,7} — периоду 6, ET_{8,9} — периоду 2.4 (говоря о дробном периоде, мы имеем в виду гармонику с частотой, обратной этому периоду, в данном случае — $5/12$), ET_{10,11} — периоду 3.

Вспомогательные характеристики. Рассмотрим, какая дополнительная информация может помочь для идентификации собственных троек (или подтвердить то, что получено). Рис. 7, на котором изображены логарифмы собственных значений, подтверждает найденные пары собственных троек (каждой паре соответствует “ступенька”).

Другим полезным средством идентификации является матрица w -корреляций между восстановленными компонентами ряда. На рис. 8 изображены первые 6 формально восстановленных, по каждой собственной тройке отдельно, компонент исходного ряда (будем называть такие восстановленные компоненты *элементарными*). Напомним, что w -корреляция — это взвешенная корреляция между восстановленными рядами и равенство ее нулю является необходимым условием разделимости компонент ряда. Рис. 9 подтверждает проведенную идентификацию тем, что внутри пары, порождаемой гармоникой, w -корреляция высока, а между парами и трендом — близка к нулю (это показывает белый цвет соответствующих ячеек матрицы корреляций).

Отделение сигнала от шума. Отдельно остановимся на вопросе отделения компонент, относящихся к сигналу, от шумовых компонент. Во-первых, нерегулярное поведение сингулярных векторов может говорить о принадлежности их к набору, порожденному шумовой компонентой (эти случаи нужно не путать с перемешиванием компонент, порожденным отсутствием сильной разделимости рядов). Во-вторых, медленное, почти без скачков, убывание собственных значений с некоторого номера также говорит об этом. В-третьих, большой набор собственных троек, порождающих коррелирующие друг с другом восстановленные компоненты, скорее всего относится к шуму. Рис. 9 показывает, что собственные тройки с номерами 14—84 как раз образуют такой блок. Вопрос вызывает пара ET_{12,13}. С одной стороны, она довольно хорошо отделяется от остатка. С другой стороны, периодограмма ряда показывает, что период у восстановленной на основе пары ET_{12,13} компоненты равен примерно 2.33. Такой период, необъяснимый наличием сезонности, может быть порожден шумом, однако в принципе, причиной может быть также сильная модуляция гармоники с периодом 2.4. На данном конкретном ряде невозможно сделать достоверный вывод, поэтому мы все же отнесем ET_{12,13} к шуму.

И последним подтверждением правильности разделения (естественно, приближенного) сигнала и шума является проверка на принадлежность к шуму компоненты ряда, полученной с помощью восстановления по так сказать “шумовым” собственным тройкам, стандартными статистическими методами. На рис. 10 изображено разложение исходного ряда на три компоненты — тренд (ET₁; на фоне исходного ряда), периодику (ET₂₋₁₁) и шум (ET₁₂₋₈₄). Статистические критерии подтверждают (не отвергают) то, что третья компонента является реализацией белого шума (вероятностные уровни трех разных критериев независимости наблюдений больше, чем 0.4).

Разложение исходного ряда на компоненты. Таким образом, рис. 10, полученный на основе группировки элементов сингулярного разложения и последующего диагонального усреднения, дает решение поставленной в начале раздела задачи построения разложения исходного ряда на интерпретируемые аддитивные компоненты. Таблица 1 содержит значения w -корреляций между изображенными на рис. 10 компонентами.

Кроме общего разложения может представлять интерес более подробное исследование ряда. Для

Таблица 1: w -Корреляции для разложения ряда на тренд, сезонную компоненту и шум

	ET1	ET2-11	ET12-84
ET1	1	0	0
ET2-11	0	1	0.016
ET12-84	0	0.016	1

нашего ряда представляет интерес более подробное исследование нерегулярного на вид поведения сезонной компоненты. На рис. 11 представлены первые три компоненты разложения сезонной компоненты на гармоники. Обращает на себя внимание уменьшение амплитуды годовой гармоник, более-менее постоянное поведение полугодовой и увеличение амплитуды 4-месячной гармоник. Заметим, что стандартные статистические методы обычно предполагают либо неизменность амплитуды (аддитивные модели), либо ее изменение, одинаковое для всех гармоник, пропорциональное тренду (мультипликативные модели). Видно, что в данном случае эти оба варианта не годятся.

Проблема отсутствия сильной разделимости. При анализе исходного временного ряда нам не пришлось столкнуться с проблемой перемешивания компонент ряда, т. е., с проблемой отсутствия сильной разделимости, вызванной близкими собственными числами (весами), соответствующими разным компонентам. Чтобы продемонстрировать эту проблему, укоротим исходный ряд до первых 120 точек. Соответственно выберем длину окна, равной 60. На рис. 12 изображена матрица w -корреляций. Сравнивая ее с матрицей, изображенной на рис. 9, и анализом полного ряда мы видим, что по-прежнему собственные тройки, начиная с двенадцатой, можно отнести к шуму. Однако темный блок, который образуют ET8-11, показывает, что, скорее всего, две гармонические компоненты смешались. Периодограммный анализ собственных векторов подтверждает это — смешались гармоники с частотами $1/3$ и $1/2.4$. Для выделения всей сезонной компоненты целиком такое смешение не столь существенно (оно лишь вызывает неудобства при идентификации). Если же мы хотели бы выделить, например, квартальную (трехмесячную) компоненту отдельно, то отсутствие сильной разделимости не позволило бы нам это сделать. Сравнение периодограмм (почти) полного исходного ряда (рис. 4) и его первых 120 точек (рис. 13) объясняет, почему проблемы отсутствия сильной разделимости не было для исходного ряда: вклады частот $1/3$ и $1/2.4$ в первом случае немного различаются, в то время как во втором случае они практически совпадают.

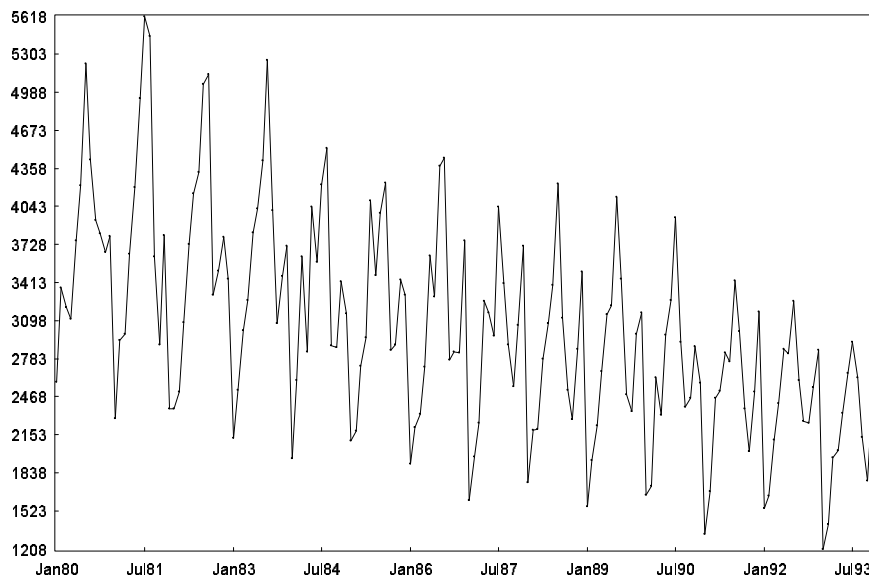


Рис. 3: FORT174: график исходного ряда месячные объемы продаж крепленых вин в Австралии

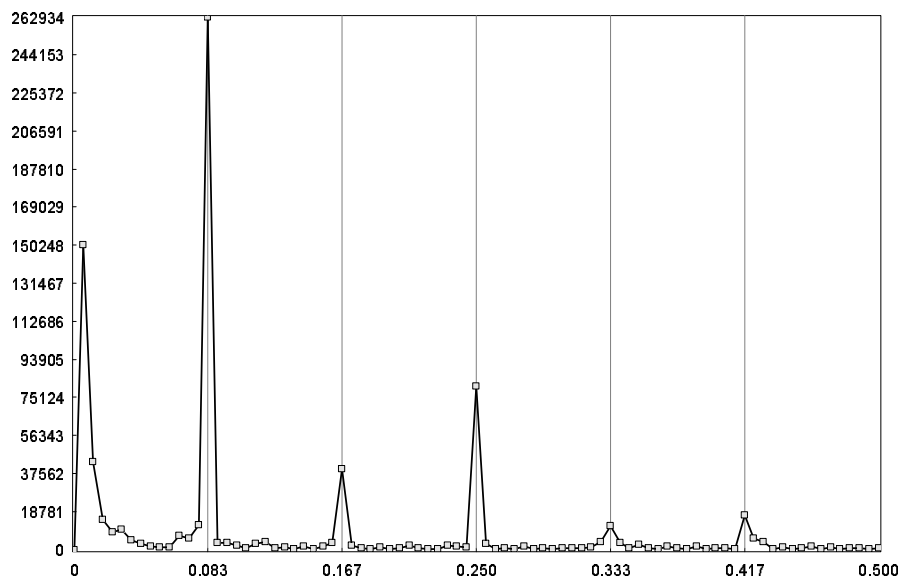


Рис. 4: FORT174: периодограмма центрированного ряда (первых 168 точек)

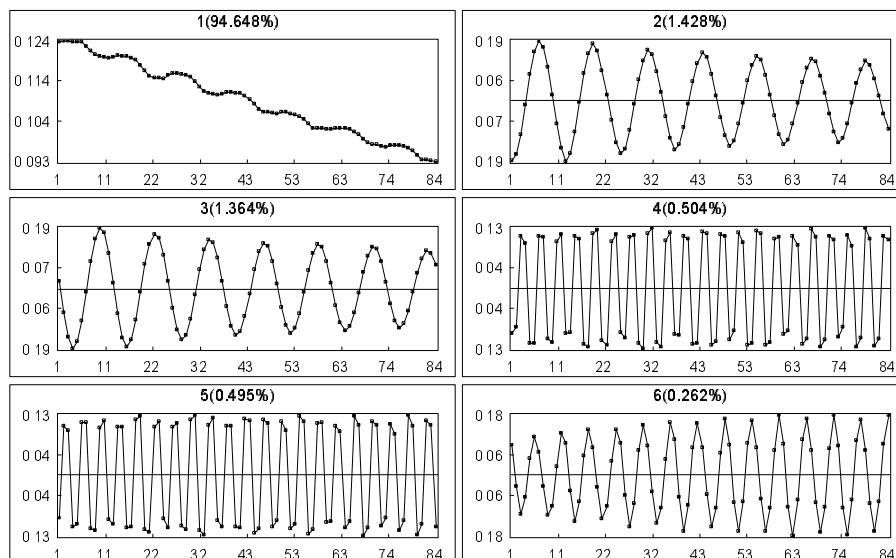


Рис. 5: FORT174: первые шесть собственных векторов сингулярного разложения траекторной матрицы ряда (одномерные диаграммы)

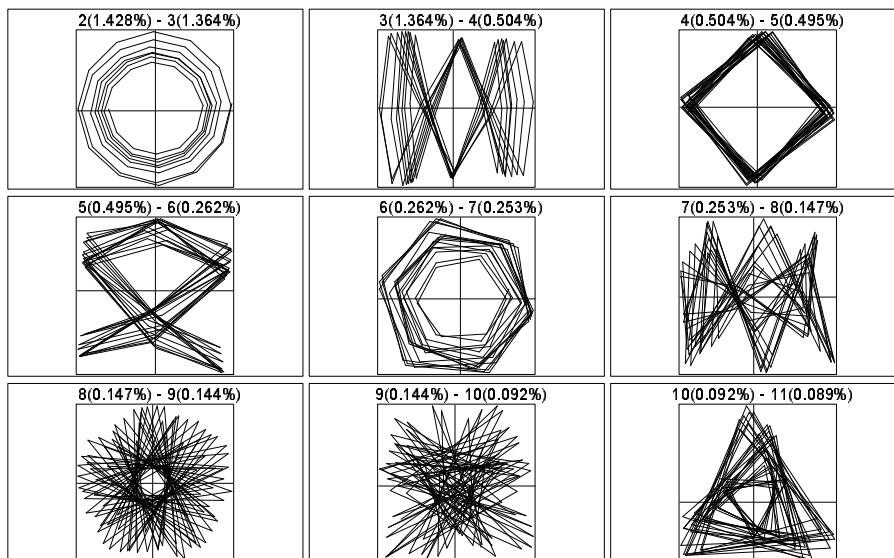


Рис. 6: FORT174: собственные вектора сингулярного разложения траекторной матрицы ряда (двумерные диаграммы)

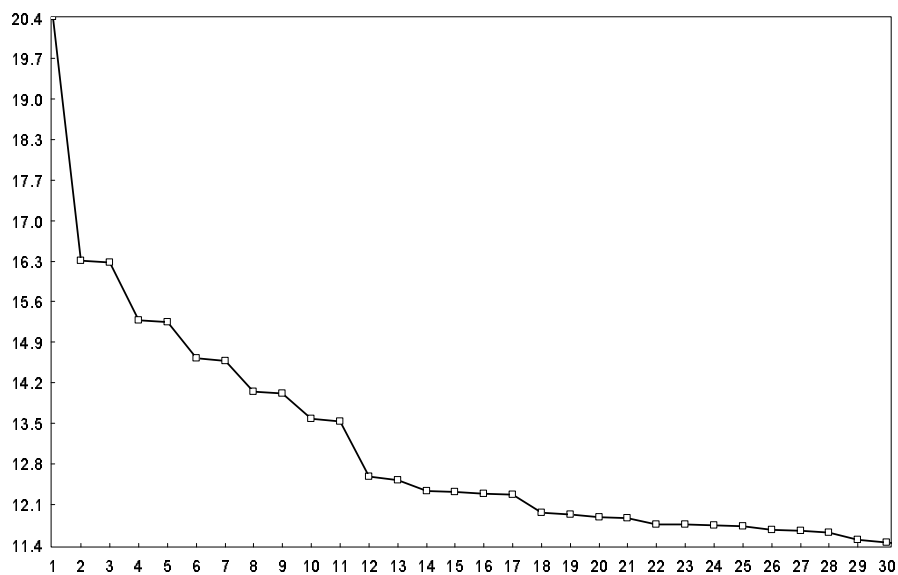


Рис. 7: FORT174: логарифмы первых 30 собственных чисел сингулярного разложения траекторной матрицы ряда

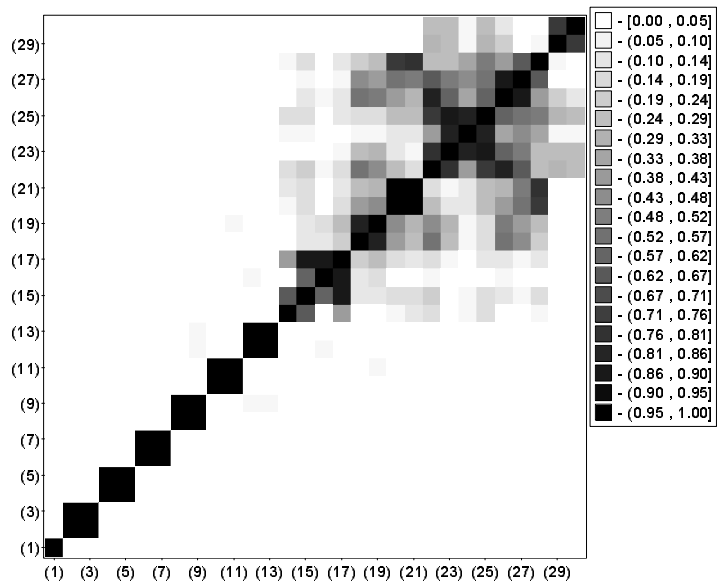


Рис. 8: FORT174: матрица w -корреляций элементарных восстановленных компонент ряда

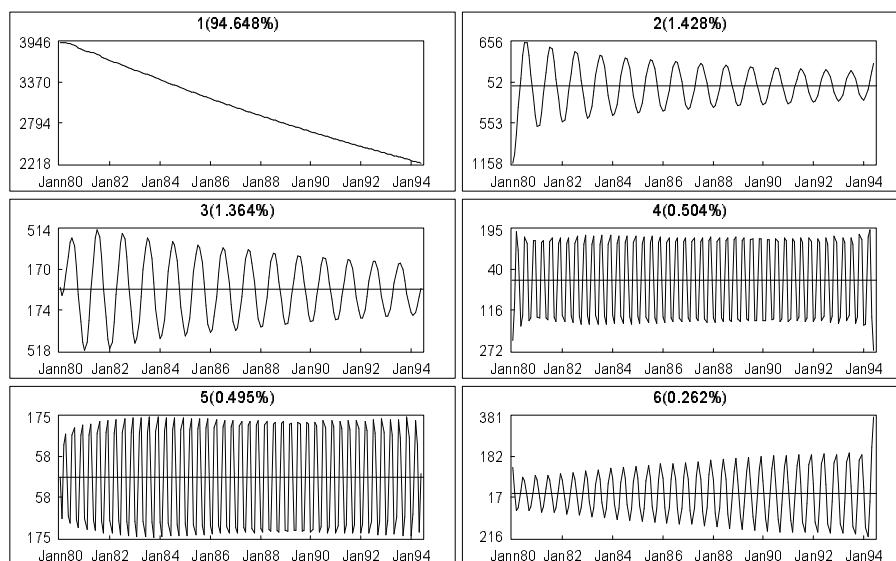


Рис. 9: FORT174: элементарные восстановленные компоненты ряда

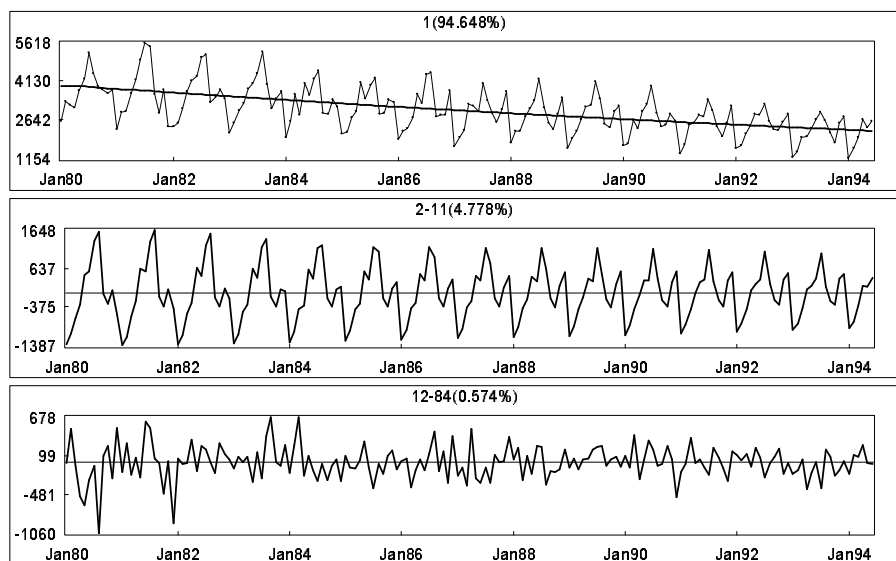


Рис. 10: FORT174: разложение ряда на тренд (вверху, на фоне исходного ряда), сезонную компоненту (в середине) и шум (внизу)

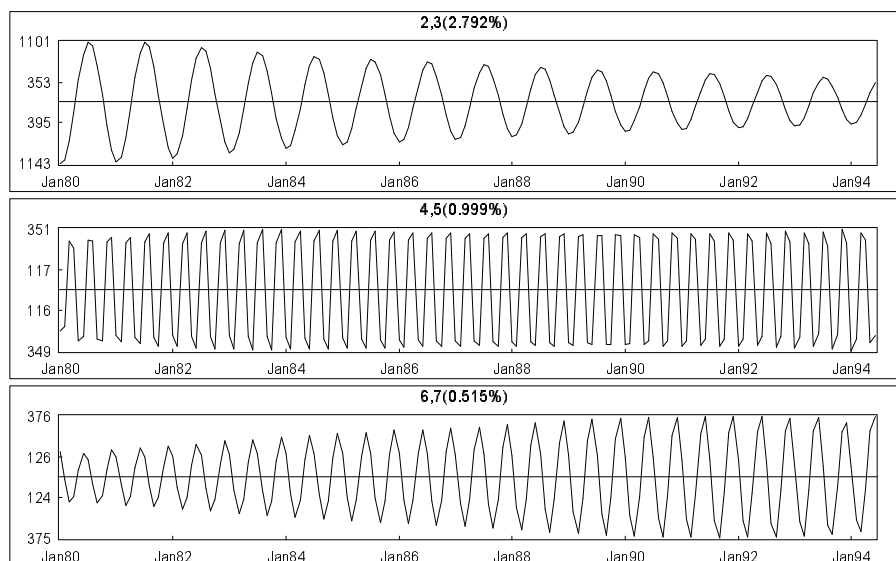


Рис. 11: FORT174: разложение сезонной компоненты ряда на (модулированные) гармоники

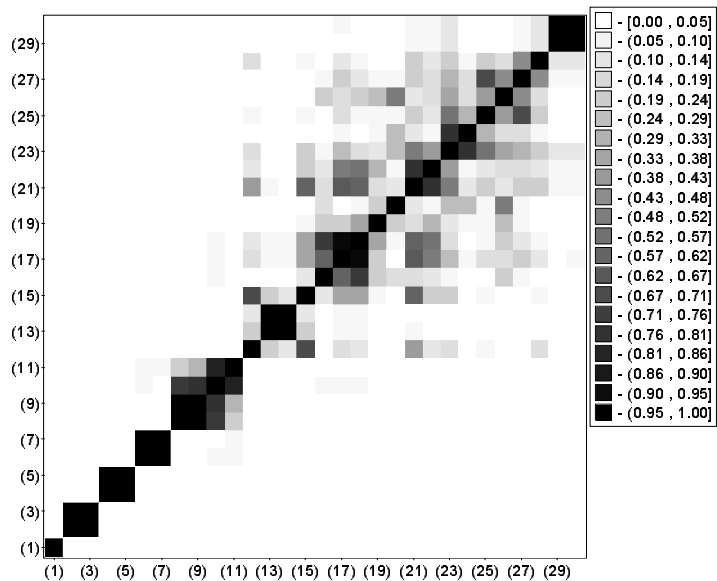


Рис. 12: FORT120: матрица w -корреляций элементарных восстановленных компонент ряда

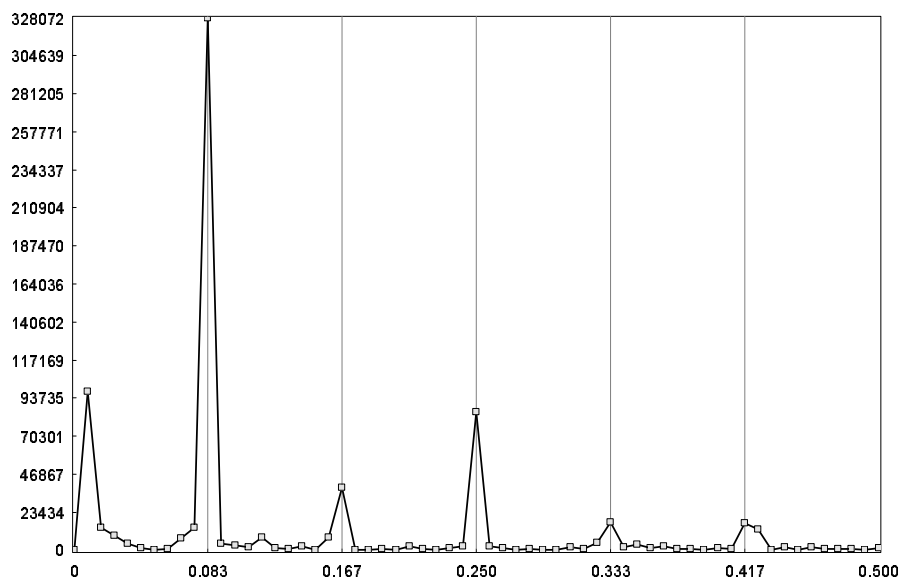


Рис. 13: FORT120: периодограмма центрированного ряда