**Сравнение методов предварительной обработки временных рядов**

**Макеев Д.А.1**

1студент

Московский государственный университет имени М.В.Ломоносова,физический факультет, Москва, РоссияE–mail: [*makeev.da21@physics.msu.ru*](mailto:makeev.da21@physics.msu.ru)

В настоящее время остро стоит вопрос контроля процесса изменения климата, порождающий множество вторичных задач: сбор данных, проверка их на валидность, дальнейшая обработка входного набора параметров [1]. Так как точное измерение физической величины невозможно в силу конструкционных особенностей измерительной аппаратуры, получаемые в результате данные помимо полезной информации содержат в себе шумовые компоненты. В данной работе рассматриваются методы очистки данных.

Сравнительный анализ проводился на искусственно сгенерированном временном ряде, описываемом уравнением (1):

(1)

Здесь 0.00001×(t+20)0.89 – медленно изменяющийся тренд, следующие два слагаемых соответствуют двум циклическим компонентам с периодами p1 = 3 дня, p2 = 365 дней. Rand{−1,1} – шумовая компонента, которая добавляется к ряду отдельно. Таким образом, получены две зависимости: полезные данные, а также шум. Далее суммарная зависимость подается на вход алгоритмам. На выходе получаются две компоненты ряда: очищенные данные и остатки ряда. Производится сравнение очищенных данных с исходно сгенерированными рядами визуально, а также при помощи метрик.

Первый метод, который был использован в данном исследовании – Анализ сингулярного спектра (англ.: Singular Spectrum Analysis, сокр.: SSA) [2]. Он основан на представлении временного ряда в виде многомерного объекта – траекторной матрицы. Далее, находя собственные векторы и собственные значений траекторной матрицы, а также оставляя только наиболее значимые собственные векторы, можно выделить наиболее информативные частоты спектра. Второй рассматриваемый метод – алгоритм скользящего среднего. Основное преимущество данного метода перед первым – меньшее число вычислительных операций по сравнению с построением сингулярного разложения траекторной матрицы. Оба рассматриваемых метода реализованы в стандартных пакетах и имеют всего один управляющий параметр, однако при использовании алгоритма скользящего среднего теряются некоторые значения ряда, это связано с особенностями метода.

На рисунках 1 и 2 показаны результаты применения алгоритмов к сгенерированному ряду. На графиках представлены компоненты, на которые алгоритмы разделили исходный зашумлённый сигнал. Вместе с тем, приведены значения исходных данных для информации, признанной полезной (сверху) и для шума (снизу). На графиках нанесены значения коэффициента детерминации и среднеквадратичная ошибка при сравнении сигнала без шума с восстановленным.

|  |  |  |
| --- | --- | --- |
|  | |  |
| ***Рис. 1.*** Результаты обработки при помощи метода SSA.  Из графиков и метрик видно, что метод SSA более точно передаёт полезную информацию, сохраняя при этом все наборы значений временного ряда, и, несмотря на свою низкую по сравнению со скользящим средним эффективностью, может быть рекомендован к применению, так как результаты вычислений могут быть использованы при дальнейшем частотном анализе или восстановлении. | | | |
|  | |
| ***Рис. 2.*** Результаты обработки при помощи скользящего среднего. | | | |

**Литература**

1. Kurbatova J., Tatarinov F., Molchanov A. et al. Partitioning of ecosystem respiration in a paludified shallow-peat spruce forest in the southern taiga of European Russia. // Environ. Res. Lett., 2013. T. 8. №. 4. C. 045028.
2. N. Golyandina, V. Nekrutkin, A. Zhigljavsky. Analysis of time series structure: SSA and related techniques. Monographs on statistics and applied probability, 2001.
3. Wikipedia.org https://en.wikipedia.org/wiki/Moving\_average
4. Box George E.P., Jenkins Gwilym M. Time series analysis, forecasting and control. Holden-day, Inc., 1976. С.300