**Использование метода главных компонент для восстановления многомерных временных рядов**

**Арефьев С. А1., *Макеев Д.А.*2**

1*студент*

*Московский государственный университет имени М.В.Ломоносова,*  
*физический факультет, Москва, Россия*  
*E–mail*: [arefev.sa23@physics.msu.ru](mailto:arefev.sa23@physics.msu.ru)2*студент*

*Московский государственный университет имени М.В.Ломоносова,*  
*физический факультет, Москва, Россия*  
*E–mail*: [*makeev.da21@physics.msu.ru*](mailto:makeev.da21@physics.msu.ru%20)

В настоящее время активно обсуждается вопрос контроля процесса изменения климата, в связи с чем возникает множество проблем: необходимо проводить сбор данных, проверять их на валидность, проводить анализ входного набора данных [1].

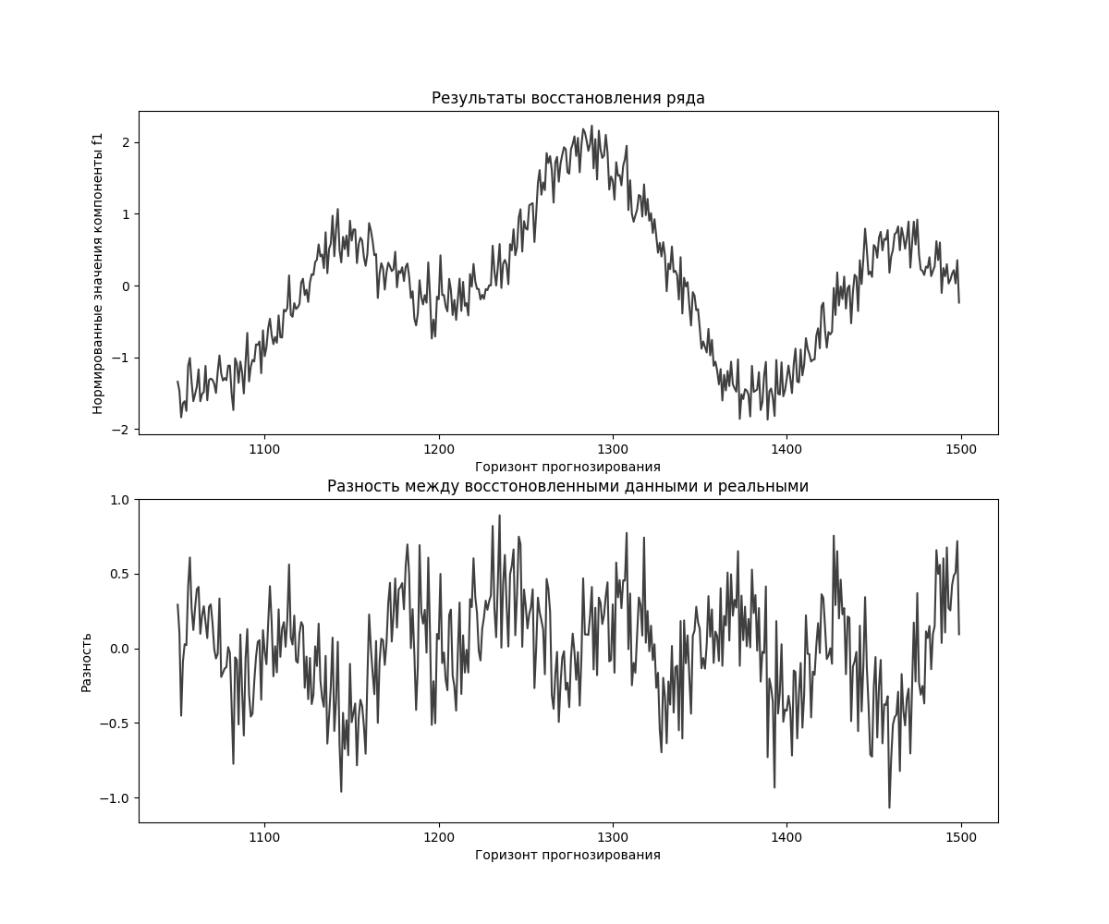
Ряды, описывающие различные геофизические процессы, можно представить в виде отображений , где *i* — индексы, нумерующие тот или иной процесс, зависящий от времени. Таким образом, эти процессы можно рассматривать как многомерный временной ряд. Это представление порождает удобный аппарат для анализа рассматриваемых явлений и работы с ними. Однако в силу несовершенства методов сбора, хранения и передачи информации могут отсутствовать некоторые из значений как одной, так и нескольких компонент ряда за различные промежутки времени, что приводит к ухудшению результатов дальнейшей обработки.

В работе был рассмотрен метод, позволяющий восстанавливать отсутствующие компоненты временных рядов по известным данным, тем самым он может быть полезен в задаче заполнения пропусков в ряде.

Данный алгоритм основан на методе главных компонент [4], который позволяет оценивать значимость компонент и эффективно понижать размерность, проецируя исходный временной ряд на гиперплоскость меньшей размерности. По построению метода в осях этой гиперплоскости дисперсия данных вокруг среднего будет наибольшей. Для восстановления исходных данных, то есть для возвращения к исходной размерности, используется псевдообратная матрица [2] оператора проектирования на выбранную гиперплоскость. Тем самым, при реализации этой процедуры в данных остаются только значимые компоненты и удаляются шумовые.

В работе исследовался синтетический временной ряд с известными компонентами, длина – 1500 точек. Первая компонента этого ряда: *,* где – это случайный вектор, имеющий многомерное нормальное распределение. Вторая компонента: сглаженная методом “locally weighted scatterplot smoothing” [3] компонента с добавлением шумовой компоненты Третья компонента: сглаженная методом “locally weighted scatterplot smoothing” [3] компонента . Четвертая компонента:

Для оценки качества модели были выбраны метрики RMSE и R2.



На графике показан результат восстановления ряда, при неизвестных последних 30% наблюдений.

Метрики для данного примера: . Они свидетельствуют о высоком качестве предложенной модели, демонстрируя отличное соответствие восстановленного ряда с исходными данными.

В результате работы, было получено, что метод восстановления данных на основе анализа главных компонент способен восстанавливать утраченные компоненты многомерных рядов. Основным преимуществом данного метода является высокая скорость работы и качество восстановления.

**Литература**

1. Kurbatova J., Tatarinov F., Molchanov A. Partitioning of ecosystem respiration in a paludified shallow-peat spruce forest in the southern taiga of European Russia. et al. – Environ. Res. Lett., 2013. – Vol. 8(4). – 045028.
2. Moore E.H. On the reciprocal of the general algebraic matrix // Bulletin of the American Mathematical Society. 1920. Vol. 26. P. 394—395
3. Cleveland, W. S. Robust Locally Weighted Regression and Smoothing Scatterplots // Journal of the American Statistical Association. 1979. Т. 74. No. 368. С. 829-836.
4. Gewers, Felipe L.; Ferreira, Gustavo R.; Arruda, Henrique F. De; Silva, Filipi N.; Comin, Cesar H.; Amancio, Diego R.; Costa, Luciano Da F. Principal Component Analysis: A Natural Approach to Data Exploration // ACM Comput. Surv. 2021. Vol. 54, No. 4. С. 70:1–70:34. doi:10.1145/3447755.