**Построение эмпирической автоковариации в реальном времени для больших временных рядов с пропусками**

***Семёнова Софья Эдуардовна***

*Студент*

*Московский государственный университет имени М.В.Ломоносова,*

*физический факультет, Москва, Россия*

*E–mail:* [*semenovasofyaeduardovna@gmail.com*](mailto:semenovasofyaeduardovna@gmail.com)

В ходе широкомасштабного сбора временных данных часто возникают пропущенные значения. Они мешают корректно проводить анализ временных рядов и использовать их в дальнейшем. При этом большинство существующих методов анализа рядов не рассчитаны на работу с пропусками. Одной из самых частых стратегий работы с пропущенными значениями является их «заполнение» [1]. К сожалению, такое заполнение обычно является эвристическим и может существенно исказить стохастические свойства временного ряда. Возникает проблема: для заполнения пропусков требуется проанализировать ряд, а для анализа стандартными методами требуется ряд без пропусков. В связи с этим появляется необходимость разработки подходов и алгоритмов, которые могут полноценно работать с рядами, имеющими пропуски [3].

В современных исследованиях ряды часто имеют огромные объёмы и непрерывно пополняются. Для эффективной обработки таких данных оказывается удобным разбить обработку данных на две фазы [2]. А именно, по мере поступления данных, в реальном времени, производить накопление некоторой информации, а окончательный результат обработки строить по мере необходимости на основании всей накопленной на данный момент информации.

В данной работе рассматривается метод построения эмпирической автоковариации для временного ряда с пропусками непосредственно в процессе поступления данных [4]. Автоковариация является одной из важнейших характеристик стационарного временного ряда. Она часто используется, например, при прогнозировании будущих значений [5] и для оценки пропущенного фрагмента ряда по имеющимся соседним. Предложенный подход использует идеи накопления информации в условиях больших данных. А именно, путем анализа поступающих данных временного ряда формируется специального вида каноническая информация, на основании которой легко вычисляется оценка автоковариационной функции. Далее, по мере поступления данных каноническая информация постоянно обновляется. В результате, актуальная оценка автоковариационной функции обновляется в режиме реального времени. При этом устраняется необходимость хранить в оперативном доступе полный набор данных.

Цель данной работы – показать, что достаточно содержательный анализ временных рядов, а именно, оценка автоковариации и ее использование, могут быть реализованы даже в случае высокой доли пропусков. В работе исследовано влияние пропусков различного типа и доли пропусков на точность построения оценки автоковариации и рассмотрено их влияние на качество заполнения пропущенных фрагментов временного ряда, рис. 1, 2.

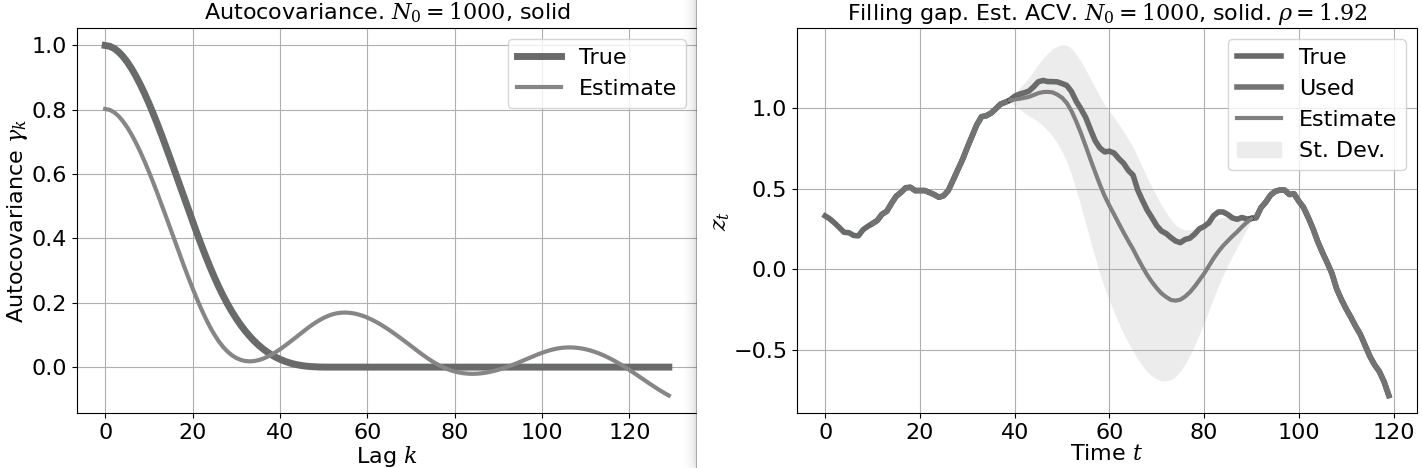


Рис. 1. Оценка автоковариации (слева) для ряда без пропусков длины 1000 и, основанная на ней, оценка пропущенного фрагмента (справа).

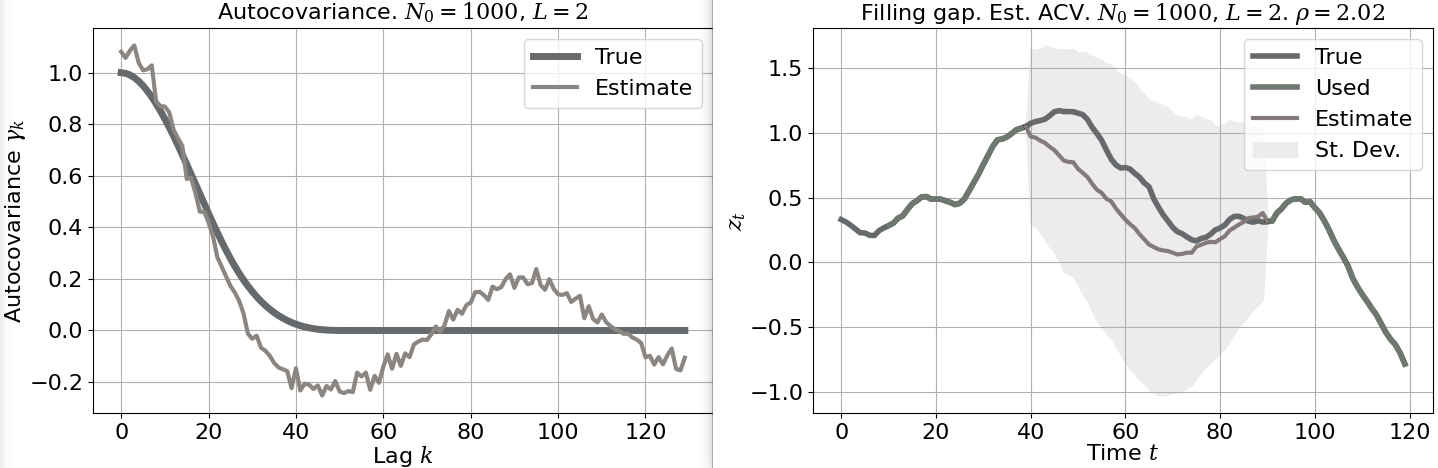


Рис. 2. Оценка автоковариации (слева) для ряда с вероятностью пропуска 0.5, содержащего 1000 валидных значений, и основанная на ней оценка пропущенного фрагмента (справа).

Показано, что оценки автоковариации, построенные на основании рядов с большой долей частых пропусков, могут утрачивать некоторые важные свойства автоковариации, такие как гладкость и положительная определенность, что негативно отражается на качестве заполнения пропусков.

**Литература**

1. Грачев А.В. К восстановлению 2004. пропусков в экспериментальных данных // Вестн. ННГУ им. Н.И. Лобачевского. Сер. Радиофизика. Вып.2. C.15–23. http://www.unn.ru/pages/elibrary/vestnik/99990216\_West\_rdfisika\_2004\_1(2)/2.pdf

3. Голубцов П.В. Понятие информации в контексте задач обработки больших данных // НТИ Сер. 2. Информационные процессы и системы, 2017. №1, С.31–36.

2. Дещеревский А.В., Журавлев В.И., Никольский А.Н., Сидорин А.Я. Проблемы анализа временных рядов с пропусками и методы их решения в программе WinABD // Геофизические процессы и биосфера, 2016. T.15, №3, С.5–34.

4. Golubtsov P. Scalability and Parallelization of Sequential Processing: Big Data Demands and Information Algebras // Advances in Intelligent Systems and Computing, Springer: 2020. Vol.1127. p.274–298. <http://dx.doi.org/10.1007/978-3-030-39216-1_25>

5. G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time Series Analysis: Forecasting and Control. 5th Edition. 2015. John Wiley and Sons Inc., Hoboken, New Jersey.