

ОЧИСТКА ВРЕМЕННЫХ РЯДОВ ОТ ВЫБРОСОВ.

Муравьев П.А¹, Соловьев М.Н.¹, Дементьев М.В.²

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И.

Ульянова (Ленина)¹, г. Санкт-Петербург

Челябинская областная клиническая больница², г. Челябинск

Ни один файл данных, содержащий значения различного рода показателей, подлежащий дальнейшему анализу программными средствами в качестве временного ряда (серии) не может быть гарантированно лишен ошибок в данных. Их источники многообразны – это могут быть и случайные опечатки при ручном вводе данных, и ошибки регистрации при протоколировании результатов наблюдений, и многие другие причины. В частности, при автоматическом мониторинге артериального давления ошибочные данные могут регистрироваться при случайных напряжениях руки пациента в момент нагнетания воздуха в манжету, при записи электрокардиограммы – вследствие случайных нарушений проводимости сигнала от тела к датчикам прибора. Такие искажения, со статистических позиций, относятся к выбросам, поиск которых представляет крайне важную задачу, которая должна быть качественно реализована перед тем, как приступить к дальнейшему анализу и обработке временного ряда.

Практический прием очистки наблюдений от выбросов хорошо известен – исключение данных, значения которых отстоят от среднего арифметического более, чем на три стандартных отклонения (SD). При наличии трендов, в частности, периодических колебаний, при слепом применении указанного приема многие выбросы не будут учтены [1]. В то же время, вследствие наличия выраженной циркадианной ритмичности наибольшие значения показателей на высоте волны ритма и наиболее низкие на глубине ее спада могут отстоять от среднесуточного значения больше, чем на 3SD, и они при таком стандартном подходе могут оказаться ошибочно исключенными.

Для таких случаев был предложен анализ разностей последовательных значений [2], но конкретные программные реализации не были описаны.

Разработан алгоритм очистки временных рядов от выбросов независимо от наличия трендов, результаты которого проверены на материале 48 записей автоматического суточного мониторинга артериального давления (СМАД). Для демонстрации работы алгоритма была создана программа, которая позволяет наглядно продемонстрировать алгоритм в действии.

Интерфейс приложения носит демонстрационный характер и позволяет пользователю анализировать временные ряды как произвольного вида, так и ряды данных СМАД. Алгоритм основан на анализе стандартных отклонений разностей соседних измерений. Выбросы возможно выявлять автоматически с любой задаваемой пользователем статистической вероятностью. Вместе с тем, такая очистка рядов требует и экспертной оценки, поскольку отмеченное резкое отклонение может явиться следствием болезненного состояния; поэтому крайне важно предоставить пользователю ручной режим контроля.

В перспективе планируется, что программа будет работать не только с записями артериального давления, но и с любым временными рядами, независимо от их реальной физической природы (например, очистка рядов наблюдений профиля поверхностей, геомагнитных данных и др.). Конечно, и в данных случаях необходима экспертная оценка соответствующих специалистов. Данная программная разработка является только небольшой частью будущего программного комплекса, предназначенного для анализа временных рядов, а стадия очистки ряда от выбросов является наиболее важной составляющей, без которой дальнейшая работа комплекса невозможна.

Апробация программы очистки проведена при реальной автоматической регистрации систолического артериального давления (САД). Были обследованы 48 машинистов локомотивных бригад. У каждого из них проводили суточное мониторирование артериального давления (СМАД) прибором Кардиотехника-4000 (производство ЗАО "Инкарт") и очистку записей от ошибок программными средствами фирменного обеспечения. В этих файлах подсчитывали общее количество записей и среди них – количество измерений, превышающих 140 мм рт. ст. (величину, принимаемую за верхнюю границу нормы САД). После этого производили очистку от выбросов программой, представляемой в настоящем сообщении, и производили подсчеты снова. Результаты приведены в таблице.

Таблица. Количество выбросов при автоматическом измерении систолического артериального давления (САД)

Показатель	До очистки от выбросов	После очистки от выбросов	Удалено измерений	Доля удаленных измерений (%)
Общее количество наблюдений	2593	2543	50	1.93

Из них превышают 140 мм рт. ст.	243	221	22	9.05
---------------------------------------	-----	-----	----	------

Очистка показала, что она лишь незначительно (на 2%) изменяет общее количество результатов мониторинга, но количество случайных завышений САД снижает почти на 10%. Если оставлять такие результаты для принятия суждений, в некоторых случаях не исключена возможность так называемой гипердиагностики, когда диагноз болезни может быть поставлен здоровому человеку.

Предлагаемая предварительная обработка данных первичных измерений позволит при хронобиологических исследованиях более эффективно осуществлять их дальнейший анализ, в частности, частотную фильтрацию и выявление спектра колебаний.

ЛИТЕРАТУРА

1. Карп В.П., Катинас Г.С. Устранение выбросов. В кн.: Вычислительные методы анализа в хронобиологии и хрономедицине. СПб, Восточная корона, 1997, с. 68-69.
2. Cornélissen G. Automatic detection of multiple outliers in physiologic time series notably temperature. In: Annual Review of Chronopharmacology, v.1, pp 157-160.
3. Сборник научных тезисов и статей «Здоровье и образование в XXI веке» РУДН, Москва, 2010г.
4. Сборник научных тезисов и статей «Здоровье и образование в XXI веке» РУДН, Москва, 2009г.
5. Сборник научных тезисов и статей «Здоровье и образование в XXI веке» РУДН, Москва, 2008г.
6. Сборник научных тезисов и статей «Здоровье и образование в XXI веке» РУДН, Москва, 2007г.
7. Сборник научных тезисов и статей «Здоровье и образование в XXI веке» РУДН, Москва, 2006г.
8. Сборник научных тезисов и статей «Здоровье и образование в XXI веке» РУДН, Москва, 2005г.
9. Сборник научных тезисов и статей «Здоровье и образование в XXI веке» РУДН, Москва, 2004г.
10. Сборник научных тезисов и статей «Здоровье и образование в XXI веке» РУДН, Москва, 2003г.
11. Сборник научных тезисов и статей «Здоровье и образование в XXI веке» РУДН, Москва, 2002г.
12. Сборник научных тезисов и статей «Здоровье и образование в XXI веке» РУДН, Москва, 2001г.
13. Сборник научных тезисов и статей «Здоровье и образование в XXI веке» РУДН, Москва, 1999г.