

Эффективный алгоритм устранения выбросов из данных измерений глобальных навигационных спутниковых систем

И. В. БЕЗМЕНОВ, А. В. НАУМОВ, С. Л. ПАСЫНОК

Всероссийский научно-исследовательский институт физико-технических и радиотехнических измерений, г. п. Менделеево, Московская обл., Россия, e-mail: bezmenov@vniiftri.ru

Разработан эффективный алгоритм детектирования и устранения выбросов, применяемый при обработке данных измерений в глобальных навигационных спутниковых системах. Алгоритм основан на поиске оптимального решения, для которого среднее квадратическое отклонение и максимальное абсолютное отклонение измеренных данных от средних значений не превышают заданных пороговых значений, а количество детектированных выбросов минимально. Проведено тестирование алгоритма для линейной комбинации Мельбурна–Вуббена, образованной из данных ГНСС-измерений на двух несущих частотах.

Ключевые слова: ГНСС-измерения, выбросы, чистка данных, комбинация Мельбурна–Вуббена, оптимальное решение.

An effective outlier detecting and elimination algorithm used in data processing of observations in the Global Navigation Satellite Systems is developed. The algorithm is based on search of optimal solution for which both RMS and absolute deviation of the observed data do not exceed the preset threshold values, with the number of the detected outliers being minimal. The algorithm test was carried out for the Melbourne–Wübbena linear combination formed of GNSS-measurements on two carrier frequencies.

Key words: GNSS-observations, outliers, data screening, Melbourne–Wübbena linear combination, optimal solution.

Исходными данными в современных методах решения прикладных задач в области геодинамики, координатно-временного и навигационного обеспечения, а также сопутствующего ему метрологического обеспечения средств и систем измерений часто являются данные наблюдений, получаемые от радиоприёмных устройств глобальных навигационных спутниковых систем (ГНСС) – ГНСС-приёмников. Ряд станций, снабжённых ГНСС-приёмниками, входят в международную сеть IGS (International GNSS Service) и передают данные наблюдений в международные базы данных в режиме реального времени [1, 2]. Данные наблюдений формируются на основе измерений фаз и времени распространения радиосигналов, излучаемых спутниками навигационных систем [3]. В зависимости от решаемой прикладной задачи данные ГНСС-измерений подвергаются соответствующей обработке, одним из важнейших элементов которой является детектирование грубых измерений или выбросов в линейной комбинации Мельбурна–Вуббена [4–6]. Понятие грубых измерений является ключевым в теории обработки измерений [4], и общего определения для него не существует. Во многих случаях грубыми называют измерения, которые отличаются от среднего значения на величину, превышающую предустановленное пороговое значение. Устранение выбросов из данных измерений, называемое чисткой данных от выбросов, часто приводит к неоправданной отбраковке нужной измерительной информации. Уменьшение количества наблюдений, возникающее из-за неоправданной отбраковки данных при обработке ГНСС-измерений, может существенно повлиять на точность оцениваемых величин и привести к ухудшению их метрологических характеристик.

В настоящей работе авторы разработали алгоритм чистки от выбросов данных, полученных при ГНСС-измерениях. Предложенный алгоритм основан на поиске так называемого оптимального решения с минимальным количеством отбракованных данных.

Постановка задачи. Часто данные y_j измерений, выполненных в моменты времени j , можно представить в виде

$$y_j = z + \xi_j; j=1, \dots, N, \quad (1)$$

где z – неизвестная постоянная, не зависящая от j ; ξ_j – шумовая составляющая.

Так, например, в обработке данных ГНСС-измерений для выявления потерь захвата фазы несущих частот на предварительной стадии обработки используется комбинация Мельбурна–Вуббена, которую можно представить в виде суммы трёх слагаемых [7]. Одно из слагаемых включает целочисленные фазовые неоднозначности для двух несущих частот [3, 4], второе – учитывает смещения, вызванные инструментальными задержками для каждой пары приёмник – спутник, а третье слагаемое обусловлено измерительным шумом. Таким образом, на временном интервале, где фазовые неоднозначности не изменяются, комбинацию Мельбурна–Вуббена можно записать в виде формулы (1). В этом случае в задачу предварительной обработки входит отбраковка данных грубых измерений или выбросов. Иными словами, нужно найти такой набор $Y_P = \{y_{j_1}, \dots, y_{j_P}\}$ из P элементов, т. е. длины P , для которого выполняются условия

$$\sigma_{Y_P} = \sqrt{(P-1)^{-1} \sum_{j \in \{j_1, \dots, j_P\}} (y_j - z)^2} \leq \sigma_{\max}; \quad (2)$$

$$|y_j - z| \leq 3\sigma_{\max}; y_j \in \{y_{j_1}, \dots, y_{j_P}\}; \quad (3)$$

$$P \geq \text{MINOBS}, \quad (4)$$

где σ_{Y_P} , $\sigma_{\max}$ – среднее квадратическое отклонение (СКО) и его заданное пороговое значение, соответственно; MINOBS – параметр, ограничивающий снизу длину искомого набора измеренных значений (например, 10); далее будем считать, что $\text{MINOBS} < N$.

Значения y_j , не входящие в множество Y_P , рассматриваются как выбросы и удаляются из дальнейшей обработки. Часто выражения (2)–(4) являются единственными условиями, учитываемыми при удалении из исходного набора грубых измерений или выбросов в программе сглаживания данных [4]. Сформулируем новую задачу поиска решения, дополнив условия (2)–(4) двумя условиями экстремума.

1. Потребуем, чтобы длина искомого набора была максимальной (количество отбракованных значений минимально):

$$P \rightarrow \max. \quad (5)$$

При заданных значениях y_j решения u задачи (2)–(4) может и не существовать (например, если y_j – арифметическая прогрессия с шагом, превышающим $\sigma_{\max}$). В случае, когда решение существует, значение P , при котором достигается максимум (5), обозначим $P_{\max}$. Заметим, что условие (5) не обеспечивает единственности решения, так как может быть найдено несколько наборов длиной $P_{\max}$, удовлетворяющих условиям (2)–(4).

2. Из всех возможных наборов, удовлетворяющих условиям (2)–(5), выберем тот, для которого величина σ_{Y_P} принимает наименьшее значение

$$\sigma_{Y_P} \xrightarrow{Y_P = \{y_{j_1}, \dots, y_{j_P}\}; P = P_{\max}} \min. \quad (6)$$

Набор значений

$$Y_{\text{opt}} = \{y_{j_1}, \dots, y_{j_{P_{\max}}}\}, \quad (7)$$

удовлетворяющий условиям (2)–(6), назовём оптимальным решением задачи (2)–(6). Соответствующее значение СКО обозначим σ_{opt} . Таким образом, задача состоит в построении алгоритма поиска оптимального решения задачи (2)–(6). Сформулируем утверждение, являющееся ключом к построению эффективного алгоритма поиска оптимального решения.

Утверждение. Пусть набор (7) является оптимальным для заданной последовательности значений $\{y_j\}$ и

$$y_{\min} = \min \{y_{j_1}, \dots, y_{j_{P_{\max}}}\}; y_{\max} = \max \{y_{j_1}, \dots, y_{j_{P_{\max}}}\},$$

тогда в интервале $(y_{\min}; y_{\max})$ не содержится значений y_j , не входящих в множество Y_{opt}

Действительно, предположим противное: пусть $y_j \notin Y_{\text{opt}}$, $y_{\min} < y_j < y_{\max}$ и y_k, y_l – значения из множества Y_{opt} , для которых $y_k = y_{\min}$, $y_l = y_{\max}$. Возможен один из случаев:

1. $z < y_j$ и

$$0 < (y_j - z) < (y_l - z) \Rightarrow (y_j - z)^2 < (y_l - z)^2;$$

2. $z \geq y_j$ и

$$0 \leq (z - y_j) < (z - y_k) \Rightarrow (y_j - z)^2 < (y_k - z)^2.$$

В первом случае заменим в наборе Y_{opt} значение y_l на y_j , а во втором – y_k на y_j . В любом из рассмотренных случаев будем иметь другой набор той же длины $P_{\max}$, для которого выполнены условия (2), (3), но СКО, как следует из (2), меньше σ_{opt} . Следовательно, набор Y_{opt} не является оптимальным, так как для него не выполнено требование (6). Полученное противоречие доказывает справедливость сформулированного утверждения.

В реальной ситуации точное значение z , входящее в условия (2), (3), неизвестно. Определим его приближённо по формуле

$$z = P^{-1} \sum_{j \in \{j_1, \dots, j_P\}} y_j. \quad (8)$$

Таким образом, величина z зависит от искомого решения, что затрудняет его поиск.

Рассмотрим алгоритм Спрингера, реализованный в программе сглаживания данных [4] и предназначенный для поиска набора Y , удовлетворяющего условиям (2)–(4). Данный алгоритм основан на итерациях (номер итерации обозначен верхним индексом в скобках).

Шаг 1. Инициализация:

$$Y^{(0)} = \{y_1, \dots, y_N\}; \text{Level}^{(0)} = 10^{20}; k=0.$$

Шаг 2. Проверка длины набора $Y^{(k)}$; если она меньше MINOBS, то процесс заканчивается, решение не найдено.

Шаг 3. Вычисление значений $z^{(k)}$, $\sigma^{(k)}$ по имеющемуся набору $Y^{(k)}$ с помощью формул (8), (2).

Шаг 4. Проверка выполнения неравенства $\sigma^{(k)} \leq \sigma_{\max}$. Если оно выполнено, то набор $Y^{(k)}$ принимается за решение и процесс поиска заканчивается. В противном случае осуществляется переход к Шагу 5.

Шаг 5. Определение уровня $\text{Level}^{(k+1)}$ для детектирования выбросов:

$$\text{Level}^{(k+1)} = 3\sigma^{(k)}.$$

Чтобы избежать бесконечного повторения итераций, осуществляется проверка

$$\text{Level}^{(k+1)} < \text{Level}^{(k)}.$$

Если это неравенство не выполнено, то полагается:

$$\text{Level}^{(k+1)} = \text{Level}^{(k)} / 2.$$

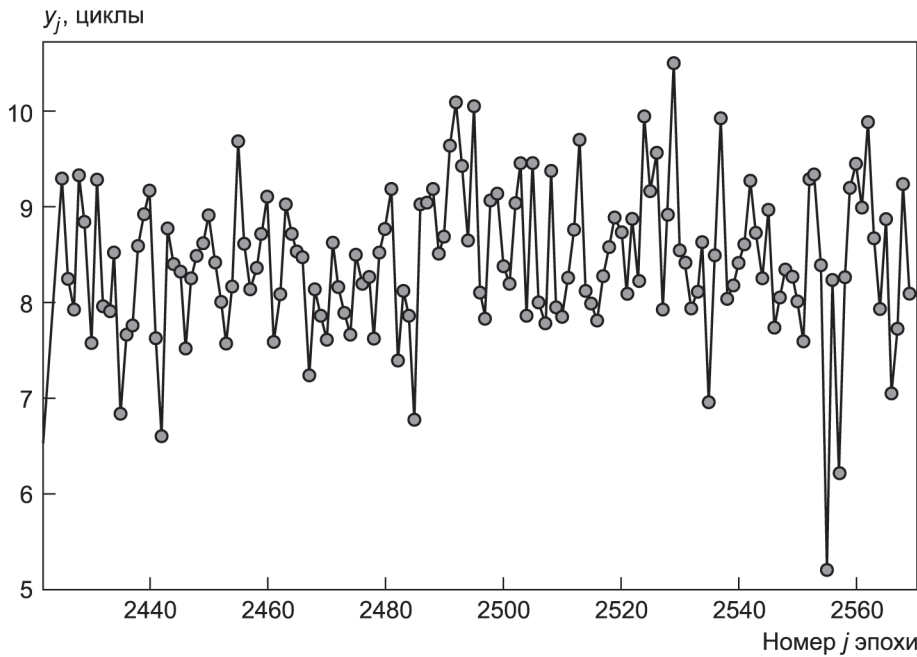


Рис. 1. Комбинация Мельбурна–Вуббена для станции ONSA (PRN=26 за 2010 г., день 207)

Шаг 6. Определение нового набора $Y^{(k+1)}$, включающего те и только те y_j , для которых

$$|y_j - z^{(k)}| \leq \text{Level}^{(k+1)}.$$

Шаг 7. $k++$. Переход к Шагу 2.

Если предположить, что в описанном выше процессе значения $\sigma^{(k)}$ монотонно убывают по k и сходятся к некоторому значению $\sigma^{(\infty)}$, так что $\text{Level}^{(\infty)} = 3\sigma^{(\infty)}$, то для решения, полученного на Шаге 4, будут выполнены условия

$$\sigma^{(\infty)} \leq \sigma_{\max};$$

$$|y_j - z^{(\infty)}| \leq 3\sigma^{(\infty)} \leq 3\sigma_{\max}.$$

Таким образом, решение задачи (2)–(4) получено.

В реальности убывание $\sigma^{(k)}$ может не быть монотонным, что приводит на Шаге 5 к двукратному изменению уровня детектирования выбросов и, как следствие, к неоправданной отбраковке данных. Заметим, что оптимальное решение таким способом можно и не найти, что подтверждают численные расчёты.

Алгоритм решения задачи (2)–(6). Для поиска оптимального решения рассмотренной выше задачи (2)–(6) авторы разработали алгоритм.

Предварительно расположим значения $\{y_j\}$ в порядке возрастания. Чтобы избежать сложностей в обозначениях, будем считать, что $y_1 \leq y_2 \leq \dots \leq y_N$. Более того, для простоты рассуждений будем считать, что все y_j различны, т. е. $y_1 < y_2 < \dots < y_N$. Заметим, что в силу сформулированного утверждения, если Y_{opt} – оптимальный набор и y_k, y_l – его наименьшее и наибольшее значения, соответственно, то все значения y_j из интервала (y_k, y_l) принадлежат Y_{opt} .

Следовательно, $Y_{\text{opt}} = \{y_k, \dots, y_l\}$. Таким образом, вместо перебора всевозможных наборов различной длины с общим количеством 2^N для поиска решения задачи (2)–(6) достаточно варьировать только два параметра k, l , связанных условием:

$$k + \text{MINOBS} < l. \quad (9)$$

Количество всех пар индексов k, l , подчинённых условию (9), равно

$$(N - \text{MINOBS})(N - \text{MINOBS} + 1) / 2.$$

Введём обозначения:

$$z(k; l) = (l - k + 1)^{-1} \sum_{j=k}^l y_j; \quad (10)$$

$$\sigma^2(k; l) = (l - k)^{-1} \sum_{j=k}^l (y_j - z(k; l))^2. \quad (11)$$

Имеют место следующие рекуррентные соотношения, позволяющие находить $z(k; l)$, $\sigma^2(k; l)$ через вычисленные значения $z(k; l+1)$, $\sigma^2(k; l+1)$ за семь арифметических действий:

$$z(k; l) = z(k; l+1) - A_{k, l+1} [y_{l+1} - z(k; l+1)]; \quad (12)$$

$$\sigma^2(k; l) = B_{k, l+1} \sigma^2(k; l+1) - C_{k, l+1} [y_{l+1} - z(k; l+1)]^2; \quad (13)$$

$$A_{kl} = (l - k)^{-1}, \quad B_{kl} = \frac{l - k}{l - k - 1}, \quad C_{kl} = \frac{l - k + 1}{(l - k)(l - k - 1)}. \quad (14)$$

Значения дробей можно вычислить заранее как элементы двумерного массива.

Аналогичные формулы позволяют выразить $z(k; l)$, $\sigma^2(k; l)$ через $z(k-1; l)$, $\sigma^2(k-1; l)$:

$$z(k; l) = z(k-1; l) - A_{k-1, l} [y_{k-1} - z(k-1; l)], \quad (15)$$

$$\sigma^2(k; l) = B_{k-1, l} \sigma^2(k-1; l) - C_{k-1, l} [y_{k-1} - z(k-1; l)]^2. \quad (16)$$

Условия (2), (3) запишем в новых обозначениях:

$$\sigma^2(k; l) \leq \sigma_{\max}^2, \quad (17)$$

$$\left. \begin{aligned} y_l - z(k; l) &\leq 3\sigma_{\max}, \\ z(k; l) - y_k &\leq 3\sigma_{\max}. \end{aligned} \right\} \quad (18)$$

Разработанный авторами настоящей статьи алгоритм основан на переборе всевозможных пар (k, l) , удовлетворяющих условию (9). Организуем этот перебор следующим

образом. Разобьём алгоритм на шаги, на каждом из которых рассмотрим наборы заданной длины, начиная с наибольшего значения N и уменьшая его на единицу на каждом следующем шаге, пока не достигнем значения MINOBS–1 или не найдём решение. На первом шаге имеется один набор $\{y_1, \dots, y_N\}$ длины N . Проверим его на выполнение условий (17), (18). Если они выполнены, то данный набор является решением, и дальнейший поиск прекращается. В противном случае переходим к рассмотрению наборов длиной $N-1$.

В общем случае, если решение не найдено среди наборов длиной большей L , то переходим к наборам длиной L . При этом если $L = \text{MINOBS} - 1$, то процесс поиска заканчивается – решение не найдено. При $L \geq \text{MINOBS}$ рассмотрим $N-L+1$ наборов длиной L :

$$\{y_1, \dots, y_L\}, \{y_2, \dots, y_{L+1}\}, \dots, \{y_{N-L+1}, \dots, y_N\}. \quad (19)$$

Каждый из этих наборов проверим на выполнение условий (17), (18). Если ни один из них не удовлетворяет данным условиям, то переходим к рассмотрению наборов длины $L-1$. В противном случае возможны два варианта:

если найден только один набор из (19), удовлетворяющий условиям (17), (18), то он и будет решением поставленной задачи;

если условиям (17), (18) удовлетворяют сразу несколько наборов, то в качестве решения выберем тот из них, для которого значение $\sigma^2(k; l)$ окажется наименьшим. Процесс поиска на этом прекращается, при этом $P_{\max} = L$.

Для вычисления значений $z(k; l)$, $\sigma^2(k; l)$ используем рекуррентные формулы (12)–(16) в соответствии со схемой, на которой показаны только $z(k; l)$:

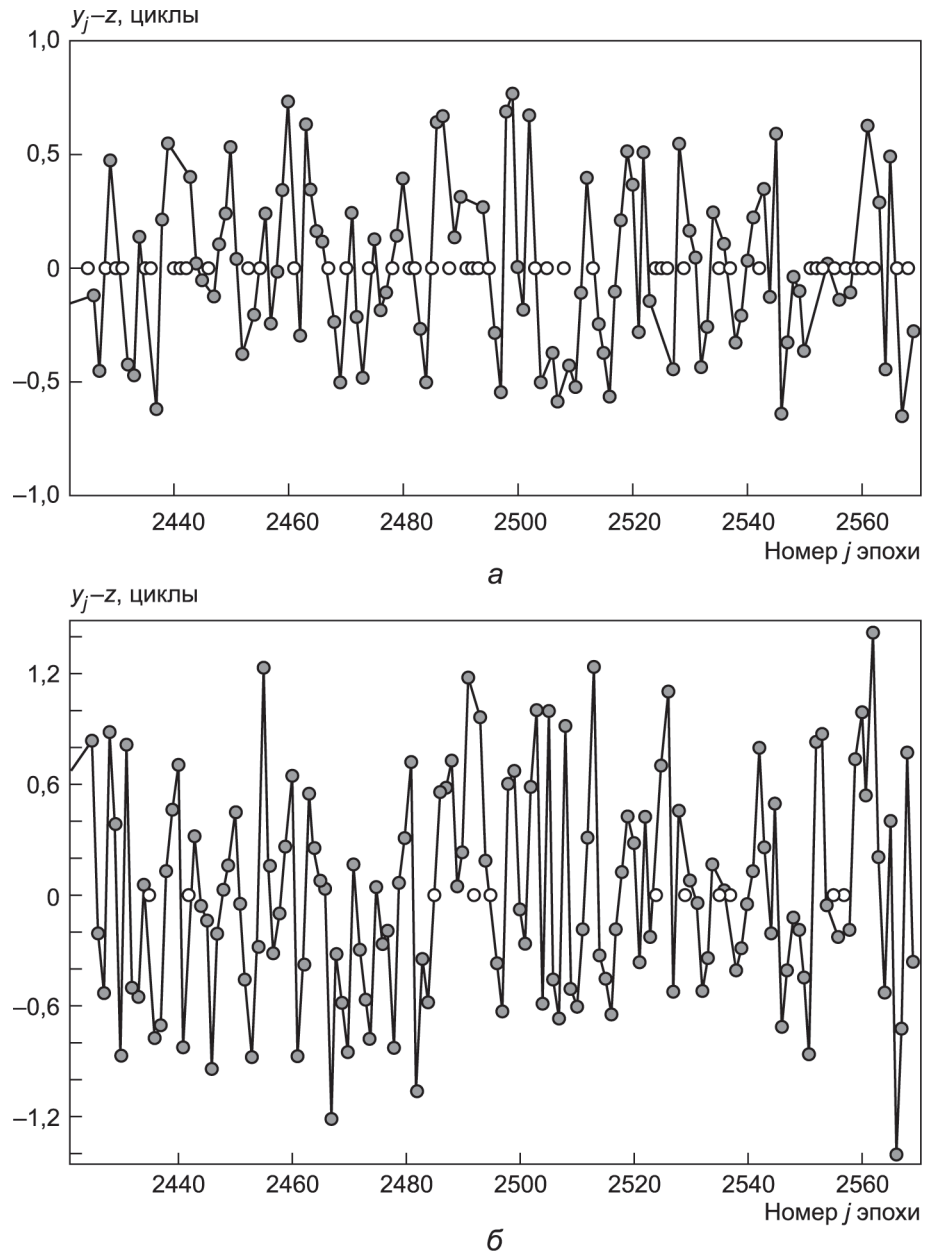
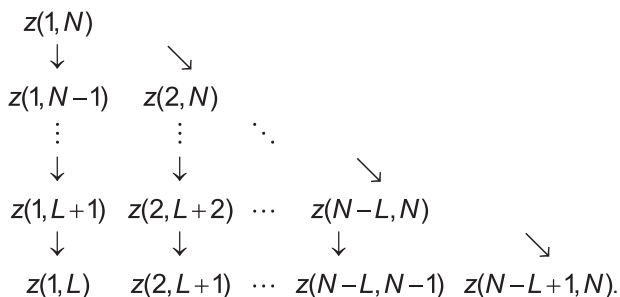


Рис. 2. Отклонения значений комбинации Мельбурна–Вуббена от среднего значения после чистки данных от выбросов по алгоритму, применяемому в [4] (а), и по разработанному алгоритму (б)

Согласно представленной схеме на первом шаге алгоритма по формулам (10), (11) вычислим значения $z(1; N)$, $\sigma^2(1; N)$, для чего потребуется $4N$ арифметических операций. Для нахождения $z(k; l)$, $\sigma^2(k; l)$ на всех последующих шагах применим рекуррентные формулы (12)–(14), позволяющие рассчитывать значения указанных пар величин за семь арифметических операций каждая, основываясь на результатах вычислений предыдущего шага. Так, например, на втором шаге, исходя из известных значений $z(1; N)$, $\sigma^2(1; N)$, по формулам (12)–(14) найдём значения $z(1; N-1)$, $\sigma^2(1; N-1)$ (вертикальная стрелка на схеме), а с помощью формул (15),

(16) – значения $z(2; N)$, $\sigma^2(2; N)$ (косая стрелка на схеме). В общем случае переход на следующий шаг в направлении вертикальных стрелок (см. схему) осуществляется согласно формулам (12)–(14), а в направлении косых – согласно формулам (15), (16). Количество арифметических действий, необходимых для нахождения решения, не должно превышать

$$4N+9[(N-P_{\max}+2)(N-P_{\max}+1)/2-1].$$

Во внимание приняты затраты на проверки выполнения неравенств (17), (18), которые составляют от 0 до 2 арифметических действий.

Результаты расчётов. Для проверки работоспособности предложенного авторами алгоритма использованы данные измерений станции ONSA [2] для спутника GPS с системным номером PRN=26 за 2010 г., день года 207. Эти данные поставляются вместе с дистрибутивом программного пакета [4]. На рис. 1 представлены значения комбинации Мельбурна–Вуббена на временном интервале, составляющем около 73 минут ($N=145$). По горизонтальной оси отложены номера j эпох, отсчитываемых от начала суток с 30-секундным интервалом. По вертикальной оси отложены значения y_j комбинации, выраженные в циклах с длиной волны $\lambda_5 \approx 0,86$ м [4]. На рис. 2 показаны значения отклонений от среднего значения данных, очищенных от выбросов с помощью применяемого в [2] и разработанного авторами алгоритмов. В обоих случаях $\sigma_{\max}=0,6$; MINOBS=10. По вертикальным осям отложены значения $y_j - z$ в циклах с длиной волны λ_5 , по горизонтальной оси – номера j эпох. Эпохи, в которых информационные данные были забракованы, обозначены светлыми кружками. В первом случае (см. рис. 2, а) было отбраковано 46 данных проведённых измерений, что составляет 31,7 % от общего числа данных. Во втором случае (см. рис. 2, б) было отбраковано 11 данных измерений (7,5 %), что почти на 25 % меньше, чем в предыдущем расчёте.

Заключение. Разработан устойчивый алгоритм обработки данных, полученных от измерительных систем, которые применяются в геодинимике, навигации и других направлениях научно-технической деятельности. Алгоритм позволяет эффективно чистить от выбросов данные наблюдений, состоящие из неизвестной составляющей и шума. После проработки алгоритма выполнены проверочные расчёты по обработке реальных данных, полученных в навигационных спутниковых системах. Характерной чертой предложенного алгоритма является то, что выборка представляется не просто набором случайных чисел (результатов измерений), а характеризуется также моментами времени получения этих данных. Оптимизация метода обработки для заданных предустановленных значений параметров $\sigma_{\max}$ и MINOBS обеспечивает минимальное количество

отбракованных данных измерений, при этом для поиска оптимального решения требуется не более $N - \text{MINOBS} + 1$ шагов, за которые решение будет гарантированно найдено, если только оно существует.

Отличие рассмотренной процедуры от классической статистической схемы исключения промахов и применения методов непараметрической статистики заключается в привязке данных к временной оси. Оптимизация метода обработки предполагает исключение возможности неоправданной отбраковки какой-то части данных, т. е. минимизацию исключённых результатов при высокой надёжности выводов.

Л и т е р а т у р а

1. **Перов А. И., Харисов В. Н.** ГЛОНАСС: принципы построения и функционирования. Изд. 4-е, перераб. и доп. М.: Радиотехника, 2010.
2. **Сайт** Международной службы GNSS. [Электрон. ресурс]: <http://www.igs.org/network> (дата обращения: 05.04.2018).
3. **Перов А. И., Харисов В. Н.** ГЛОНАСС: принципы построения и функционирования. Изд. 4-е, перераб. и доп. М.: Радиотехника, 2010.
4. **Dach R., Andritsch F., Arnold D., Bertone S., Fridez P., Jäggi A., Jean Y., Maier A., Mervart L., Meyer U., Orliac E., Ortiz-Geist E., Prange L., Scaramuzza S., Schaer S., Sidorov D., Sušnik A., Villiger A., Walser P., Baumann C., Beutler G., Bock H., Gäde A., Lutz S., Meindl M., Ostini L., Sošnica K., Steinbach A., Thaller D.** Bernese GPS Software Version 5.2 / edited by R. Dach, S. Lutz, P. Walser, P. Fridez. Astronomical Institute, University of Bern, 2015.
5. **Melbourne W. G.** The case for ranging in GPS based geodetic systems. In Goad, C., editor, Proceedings of the 1st International Symposium on Precise Positioning with the Global Positioning System, US Department of Commerce, Rockville, Maryland, 1985. P. 373–386.
6. **Wubben G.** Software developments for geodetic positioning with GPS using TI 4100 code and carrier measurements. In Goad, C., editor, Proceedings 1st International Symposium on Precise Positioning with the Global Positioning System. US Department of Commerce, Rockville, Maryland, 1985. P. 403–412.
7. **Subirana J. Sanz, Zornoza J. M. Juan., Hernández-Pajares M.** Detector based in code and carrier phase data: The Melbourne–Wubben combination. ESA navipedia. [Электрон. ресурс]: http://www.navipedia.net/index.php/Main_Page (дата обращения: 05.04.2018).

Дата принятия: 09.06.2018.