681.518

ПОМЕХОУСТОЙЧИВАЯ М-ОЦЕНКА ВЫБОРОК МНОГОКРАТНЫХ НАБЛЮДЕНИЙ

Н. С. ИВАНОВ

Институт геологии и геохимии УрО РАН, Екатеринбург, Россия,
e-mail: ivanovks@igg.uran.ru

Описан способ определения М-оценок, позволяющий при обработке выборок многократных наблюдений, засоренных промахами с неизвестным законом распределения, значительно повысить помехоустойчивость обработки. Предлагаемый подход применен к вычислению М-оценки с весовой функцией Тьюки.

Ключевые слова*: М-оценка, двусторонний доверительный интервал, многократные наблюдения, помехоустойчивость.*

The M-estimation determination method allowing to merease the noise immunity of processing of samplings of multiple observation clogged by blunders with unknown distribution law is described. The suggested approach is applied to computation of M-estimation with Tuki weight function.

Key words: *M-estimation, two-sided confidence interval, multiple observations, noise immunity.*

В ряде областей науки и практики необходима обработка малых наборов повторяющихся отсчетов (выборок из нескольких десятков наблюдений), где доля аномально сильно искаженных наблюдений (в дальнейшем «промахов» или «выбросов») велика и составляет до половины и более половины всех наблюдений. Например, при геофизических измерениях в промышленно развитых районах, при акустических измерениях, осложненных шумами и т.п.

Количественной характеристикой помехоустойчивости методов обработки многократных наблюдений (далее методов) является точка срыва, становящаяся пределом доли промахов, с которой может справиться обработка [1]. Вторая характеристика методов – эффективность, т. е. отношение минимально возможной дисперсии результатов измерений к ее фактическому значению для данного метода при распределении наблюдений по нормальному закону.

Эти характеристики связаны, и одновременное улучшение их обычно невозможно, чем выше устойчивость, тем меньше эффективность и наоборот. Крайние известные решения – среднее арифметическое (эффективность 1, устойчивость 0) и выборочная медиана (эффективность около 0,6, устойчивость немного меньше 0,5). Многие методы, называемые М-оценками, ранговыми R-оценками [1, 2], имеют эффективность 0,8 – 0,9, а точки их срыва лежат между 0,2 и 0,3. Все известные универсальные методы работоспособны, если время сильного искажения сигнала относительно невелико. Даже наиболее помехоустойчивые из них, в том числе выборочная медиана (в последующем медиана), отказывают, когда это время приближается к половине времени измерения.

В работе описан новый подход к вычислению М-оценок. Его применение к М-оценке Тьюки повысило ее помехоустойчивость более чем в 1,5 раза. Полученная оценка по помехоустойчивости превосходит все известные универсальные оценки, включая медиану. По эффективности же она почти не отличается от М-оценки Тьюки, определяемой общепринятым способом. Это особенно актуально при подавлении однонаправленных промахов, наиболее «опасных» для существующих методов, поскольку каждый промах отклоняет оценку в одну и ту же сторону. Такие промахи вызываются, например, помехами от ускорения или торможения электропоездов, когда период полезного сигнала много меньше длительности переходного режима поезда.

М-оценки обязаны своим названием обобщенному методу максимального правдоподобия [1]. По нему для определения оценки a решается экстремальная задача на минимум вида

$$\min Q(\Psi) = \sum_{1}^{n} \Psi[(x_i - a)/S];$$

или уравнение

$$\sum_{1}^{n} \varphi[(x_i - a)/S] = 0,$$

где S – оценка масштаба, т. е. среднее квадратическое отклонение (СКО) флуктуационной помехи; $\Psi(x)$ – выбранная весовая функция, $\varphi(x)=\Psi'(x)$.

Уравнение можно записать в эквивалентном виде как

$$\sum_{1}^{n} (\omega_i (x_i - a)) = 0\,, \tag{1}$$

где $\omega_i = \dfrac{\varphi[(x_i - a)/S]}{(x_i - a)/S}$.

Тогда оценка приводится к взвешенному среднему $a = \sum_{1}^{n} \omega_i x_i \Big/ \sum_{1}^{n} \omega_i$, где ω_i – весовые коэффициенты, зависящие от выбранного вида весовой функции $\Psi(x)$ и выборки $x_1, x_2, \ldots, x_i, \ldots, x_n$ из n наблюдений.

М-оценки не выражаются в явном виде, и их определяют, решая (1) с помощью итерационных методов. При этом за начальное приближение оценки a берется помехоустойчивая оценка – медиана, а оценка масштаба S при очередной итерации k вычисляется через медиану модулей отклонений наблюдений x_i от результатов предыдущей итерации [2]:

$$S_k = 1{,}483\,\mathrm{med}\left\{\left|x_i - a_{k-1}\right|\right\}. \tag{2}$$

Для повышения помехоустойчивости М-оценок автором предложено изменить описанное выше их определение по четырем пунктам. Так, для предварительной отбраковки части промахов, во-первых, и для более помехоустойчивого, чем по медиане, определения начального приближения, во-вторых, нужно использовать предложенный ранее метод достоверного сравнения [5, 6], первоначальный вариант которого назывался методом сравнения [3, 4]. Его точка срыва близка к 0,6 и помехоустойчивость выше, чем у любой универсальной оценки.

Метод подавляет сильно искаженные наблюдения–выбросы–промахи, даже если они составляют большинство в выборке. Абсолютного правила, подавляющего большинство наблюдений, быть не может, если искаженные наблюдения выглядят подделкой неискаженной выборки. К счастью, промахи имеют важное отличие от наблюдений, искаженных только флуктуационной помехой: плотность распределения первых ниже, чем вторых. Поэтому их подавление при большинстве в выборке вполне возможно. Нужно прямо или опосредованно использовать указанную разность плотности распределения.

Сравним воздействие на оценки по методам медианы и среднего арифметического от добавления наблюдения, искаженного на величину x, к выборке наблюдений с нормальным распределением $N(0, \sigma^2)$. Это возможно с помощью функции влияния, впервые введенной Хампелем [7]. Функция влияния $IF(x)$ оценивает воздействие на оценку, оказываемое загрязнением в точке x, нормированным массой этого загрязнения. В [7] получено, что для медианы $IF(x)=\sigma(\pi/2)^{1/2}\mathrm{sign}(x)$, а для среднего арифметического $IF(x)=x$. Следовательно, наблюдение, искаженное более чем на $\pm\sigma(\pi/2)^{1/2}$, отклоняет оценку по среднему арифметическому в соответствующую сторону больше, чем оценку по медиане.

Используем это для отбраковки промахов. Найдем медиану и среднее арифметическое значение выборки. Если среднее арифметическое больше медианы, отбракуем максимальное наблюдение, а если меньше – минимальное. По оставшимся наблюдениям вычислим их повторно и повторим отбраковку. Процедура выполняется до тех пор, пока разность очередных среднего арифметического и медианы не станет меньше некоторой величины Δ, или не изменит свой знак, или пока не отбракуются 2/3 наблюдений выборки. Если разность меняет знак, то за оценку по методу достоверного сравнения примем полусумму двух последних медиан, в других случаях – последнюю медиану. Если метод применяется как начальное приближение при вычислении М-оценок, то $\Delta = 14{,}2R/n$, где R – размах четверти наблюдений с наиболее плотным распределением в ранжированной по величине наблюдений выборке. Если метод используется самостоятельно, то $\Delta=(14{,}2R/n)((n-1{,}5j)/n)^2$, где j – количество отбракованных наблюдений. В этом случае Δ уменьшается в процессе отбраковки.

Метод проверен моделированием и на практике [4]. Например, он реализован в программно-управляемом геофизическом вольтметре ЭВП-802 [8]. Прибор прошел государственные испытания и был выпущен малой серией, изготовленной Свердловским опытно-экспериментальным заводом скважинной геофизической аппаратуры

(ОЭЗСГА). Имеются акты и отзывы геологоразведочных организаций, утверждающие, что метод применялся при геофизических измерениях на участках с очень высоким уровнем помех от ЛЭП, газопроводов, электропоездов и т.п., где ранее вести измерения было невозможно.

Определение S по (2) имеет точку срыва менее 0,5. Для повышения помехоустойчивости автором предложено при очередной итерации k вычислять значение S_k не по (2), а через 33 %-й квантиль $Q(0,33)$ ранжированных модулей отклонений наблюдений x_i от результата предыдущей итерации как $S_k=2,32Q\{|x_i - a_{k-1}|\}$. Причем вычислять $Q(0,33)$ можно только по наблюдениям x_i, оставшимся после предварительной отбраковки части наблюдений методом достоверного сравнения. Определяемая таким образом величина S имеет точку срыва около 0,6.

Из известных М-оценок автору представляется наилучшей М-оценка Тьюки, весовая функция $\Psi(x)$ которой не имеет резких изломов и которая немного помехоустойчивее других М-оценок, что установлено моделированием на ЭВМ [4]. Поэтому предложенная процедура и была применена к данной оценке, весовые коэффициенты которой при k-й итерации вычисляются на основе бивеса Тьюки:

$$\omega(u_i) = \begin{cases} (1 - u_i^{\,2})^2, & |u_i| \le 1; \\ 0, & |u_i| > 1, \end{cases}$$

где $u_i=(x_i - a_{k-1})/(4S_{k-1})$.

При обработке выборок размером 16 – 800 наблюдений помехоустойчивость полученной оценки оказалась выше, чем у любой ранее известной универсальной оценки, а ее эффективность почти не отличалась от эффективности исходной.

Оценка проверена моделированием на ЭВМ на выборках наблюдений, искажения которых распределены по закону:

$$p(y;\varepsilon) = \frac{1-\varepsilon}{\sqrt{2\pi}\ \sigma}\exp(-\frac{y^2}{2\sigma^2}) + \varepsilon h(y). \qquad (3)$$

Указанная модель искажений предполагает, что принятый сигнал содержит небольшую флуктуационную помеху с приблизительно нормальным законом распределения $N(0;\sigma^2)$. Иногда с вероятностью ε сигнал зашумлен аномальной, обычно импульсной помехой, имеющей неизвестное, часто несимметричное распределение $h(y)$ и амплитуду, превышающую или соизмеримую с амплитудой измеряемого сигнала. Моделирование проводили согласно рекомендации [9], но более широко и более жестко.

Полезный сигнал m выражали каким-либо постоянным числом и суммировали с искажениями, распределенными по (3). Для моделирования $h(y)$ применяли не только три рекомендованных в [9] симметричных распределения, но и «трудные» для всех оценок односторонние распределения; двумодальное распределение – «трудное» для оценок, основанных на медиане, так как ожидаемая оценка соответствует минимуму плотности распределения; распределение Лоренца–Коши, имеющее очень «тяжелые хвосты» – всего 7 различных распределений. Типичные результаты моделирования с однополярными выбросами представлены на рис. 1. В этом случае выбросы моделировали абсолютными величинами случайных чисел с нормальным распределением $N(0;2)$, остальные наблюдения – числами с нормальным распределением $N(0;\ 0{,}1)$. Кривые *1* – *4* представляют собой известные методы, наиболее сильные из применяемых на практике: среднего арифметического, медианы, медианы Ходжеса–Лемана, М-оценки Тьюки. Кривые *5*, *6* – метод достоверного сравнения и предлагаемую М-оценку. Рис. 1 получен для 32 наблюдений в выборке, но моделирование, в том числе и при исследовании доверительных интервалов, проводили с выборками из 4, 8, 12, 16, 24, 32, 96, 320 и 800 наблюдений со всеми упомянутыми выше распределениями.

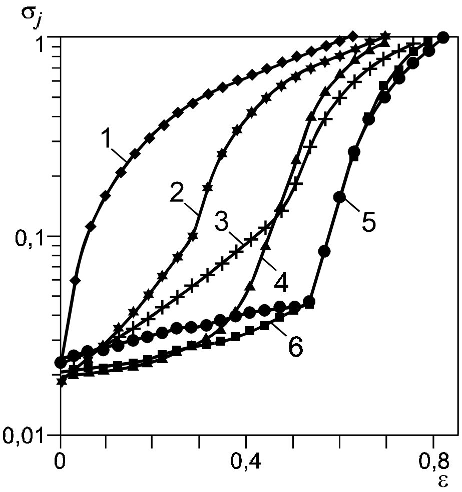

Рис. 1. Зависимость СКО результатов измерения σ_j от доли ε однонаправленных промахов для методов:

1 – среднего арифметического; *2* – медианы Ходжеса–Лемана; *3* – медианы; *4* – М-оценки Тьюки; *5* – достоверного сравнения; *6* – предложенной М-оценки

На рис. 1 видно преимущество предлагаемой М-оценки по сравнению с М-оценкой

Тьюки. По эффективности они близки, до ε=0,3 кривые *4* и *6* идут почти сливаясь. М-оценка Тьюки работоспособна, пока аномально искажено менее 1/3 наблюдений выборки, после ε=0,3 кривая *4* идет резко вверх, а при предлагаемом подходе допустимо аномальное искажение до половины наблюдений выборки, а иногда даже немного более половины. Более чем полуторократный выигрыш в помехоустойчивости позволяет утверждать, что разработан новый вид М-оценки. При его сопоставлении с методом медианы видно, что предлагаемая М-оценка и помехоустойчивее медианы, и значительно превышает ее по эффективности и точности – кривая *6* везде идет значительно ниже кривой *3*. Преимущество оценки по медиане лишь в простоте и возможности обрабатывать выборки размером в 3 и более наблюдений, а предлагаемую М-оценку целесообразно применять только с 16 и более наблюдений.

При симметричном распределении выбросов выигрыш предлагаемой М-оценки перед известными оценками заметно меньше. Кроме того, в этом случае частично взаимно компенсируется влияние разнополярных выбросов и все оценки искажаются намного меньше, чем от однонаправленных выбросов. В [4] для случая рис. 1 показано, что если доля ε однополярных выбросов меньше 60 % от всех наблюдений выборки, то почти все они отбраковываются методом достоверного сравнения. При симметричном распределении выбросов отбраковывается достоверным сравнением их малая доля, но количество выбросов на обоих концах ранжированной выборки при этом выравнивается, и взаимная компенсация разнонаправленных выбросов усиливается.

Сопоставив метод достоверного сравнения и предлагаемую М-оценку (кривые *5*, *6*), отметим большую эффективность М-оценки – 0,47 и 0,77, соответственно. За большую эффективность приходится платить большей сложностью вычислений и несколько меньшей относительно метода достоверного сравнения точкой срыва при некоторых распределениях промахов: 0,5 против 0,6.

Можно ожидать, что действия, выполненные с М-оценкой Тьюки, могут оказаться успешными также и с любой М-оценкой.

При обработке многократных наблюдений недостаточно получить оценку результата измерения. Не менее важно знать ее доверительные интервалы. Как известно [2], для нормально распределенных оценок их симметричные двусторонние доверительные интервалы (в дальнейшем «доверительные интервалы») $U(P)=\pm t_{n-1}(P)\sigma_j$, где $t_{n-1}(P)$ –

значение коэффициента Стьюдента для доверительной вероятности P и выборки из n наблюдений: σ_j – СКО соответствующих оценок. По [10] дисперсия оценок по медиане $\sigma^2_{med}=\pi\sigma^2/(2(n-1)+\pi)\approx\pi\sigma^2/2n$, где σ – СКО наблюдений в выборке из выражения (3). Таким образом, доверительные интервалы оценок по медиане определяются как

$$U(P) \approx \pm \frac{t_{n-1}(P)\sigma}{\sqrt{n}}\sqrt{\pi/2}\,. \tag{4}$$

Помехоустойчивость указанного определения $U(P)$ зависит от помехоустойчивости определения σ. Если σ вычислять через сумму квадратов отклонений, то точка срыва такого определения σ равна нулю. Используем из [2] более помехоустойчивое определение σ через медиану модулей отклонений наблюдений x_i от результата измерения T: $\sigma = (\mathrm{med}\{|x_i - T|\})/0{,}675$. Подставив в (4) определенную таким образом σ, получим доверительный интервал $U(P)$ для оценок по медиане. Моделирование с $U(0{,}95)$ показало необходимость введения дополнительного множителя 1,05. В итоге для медианы получили:

$$U(0{,}95) = \pm \frac{1{,}857 \cdot 1{,}05 t_{n-1}(0{,}95)\mathrm{med}\{|x_i - T|\}}{\sqrt{n}}. \tag{5}$$

Для определения 95 %-го доверительного интервала $U(0{,}95)$ предлагаемой М-оценки из эвристических соображений воспользуемся выражением, аналогичным (5). Вычисление предлагаемой М-оценки включает отбраковку j наблюдений методом достоверного сравнения. Поэтому для определения ее доверительного интервала используем только наблюдения x_i, оставшиеся после отбраковки. В (5) заменим n — первоначальное количество наблюдений на $(n-j)$, а коэффициенты Стьюдента t_{n-1} на t_{n-j-1}. Выполнив умножение, получим:

$$U(0{,}95) = \pm \frac{1{,}95 t_{n-j-1}(0{,}95)\mathrm{med}\{|x_i - T|\}}{\sqrt{n-j}}. \tag{6}$$

Правильность (5) и (6) для доверительной вероятности P=0,95 проверяли моделированием по описанной ранее методике. Вычисляли процент d попаданий исходного сигнала m в доверительные интервалы $U(0{,}95)_l$ результатов измерений T_l. Также определяли среднее квадратическое отношение h фактических ошибок результатов измерений T_l к их доверительным интервалам $U(0{,}95)_l$:

$$h=\sqrt{\frac{1}{L}\sum_{l=1}^{L}(T_l-m)^2/U(0{,}95)_l^2}\,, \qquad (7)$$

где L – число повторений опыта с тем же количеством и распределением промахов, но с другими случайными числами и другими пораженными наблюдениями.

Типичные результаты моделирования представлены на рис. 2 – 4 для 32 наблюдений сигнала в одном измерении. На рис. 2, 4 приведены воздействия однополярных выбросов, получаемых так же, как для рис. 1, а на рис. 3 – разнонаправленных выбросов, созданных с помощью случайных чисел равномерно распределенных на интервале от –3 до 3. На рис. 2, 4 метод достоверного сравнения по формуле [5, 6] представлен кривой *5*, предложенная по (6) М-оценка – кривой *4*, метод медианы по (5) – кривой *3*, метод медианы по [11] – кривой *2* и метод среднего арифметического (для примера, если неправомерно получение доверительного интервала по [12]) - кривой *1*. Прямой *6* показан 95 %-й уровень.

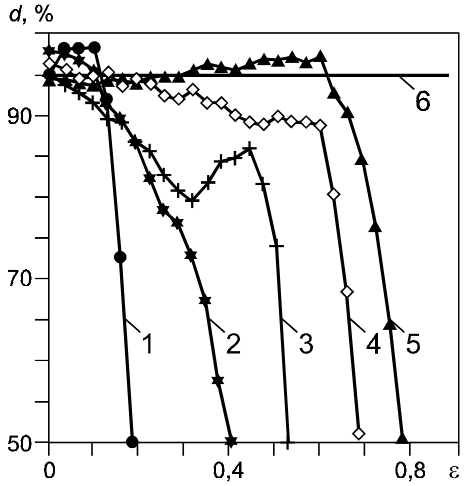

Рис. 2. Зависимость процента попаданий d исходного сигнала m в доверительные интервалы $U(0{,}95)_l$ результатов измерений T_l от доли ε однонаправленных промахов для методов:

1 – среднего арифметического по [12]; *2* – медианы по [11]; *3* – медианы по [5, 6]; *4* – предложенной М-оценки; *5* – достоверного сравнения по (8);
прямая *6* – 95 %-й уровень

Из рис. 2, 4 следует, что предложенные М-оценка и метод достоверного сравнения (кривые *4*, *5*) при доле ε однополярных выбросов меньше 0,60 и 0,65 имеют процент попаданий d, не сильно отличающийся от 95 %, а их среднее квадратическое отношение h ошибки к доверительному интервалу сравнительно стабильно и близко к 0,55. Определение доверительных интервалов оценок помехоустойчивее определения самих оценок: соответствующие точки срыва предлагаемой М-оценки 0,6 и 0,5, метода досто-

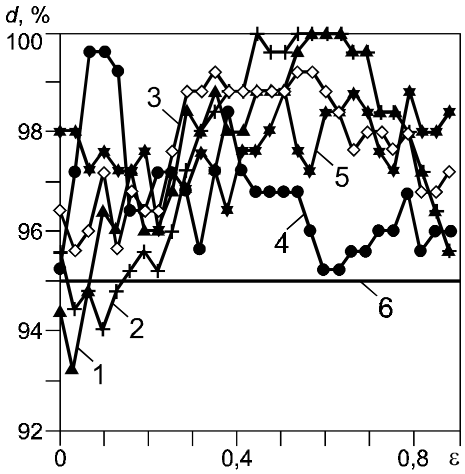

Рис. 3. Зависимость процента попаданий d исходного сигнала m в доверительные интервалы $U(0{,}95)_l$ результатов измерений T_l от доли ε симметричных промахов для методов:

1 – достоверного сравнения по [5, 6]; *2* – медианы по (5); *3* – предложенной М-оценки; *4* – среднего арифметического по [12]; *5* – медианы по [11]; прямая *6* – 95 %-й уровень

верного сравнения – 0,65 и 0,6. Это позволяет уверенно определять точность и достоверность данных оценок и их применимость при существующем в момент измерения уровне помех.

Таким образом, доверительные интервалы оценки по медиане получать по (5) в данном случае предпочтительнее, чем по [11] (кривые *3* и *2*, соответственно). Последнее наблюдается в подавляющем большинстве случаев, но все же не всегда. Поэтому для оценки по медиане можно рекомендовать применять оба упомянутых определения ее доверительных интервалов. При этом для среднего арифметического (кривая *1*) процент попаданий d при ε>15 % очень сильно отличатся от 95 %. Заметим, что при ε>25 % все исходные значения сигнала m лежат вне доверительных интервалов этой оценки: d=0, а h>1, если ε>0,25.

При симметричном распределении выбросов (рис. 3) процент попаданий d для всех методов незначительно отличается от 95 %.

Таким образом, предложен модернизированный способ определения М-оценок, значительно повышающий их помехоустойчивость. Способ опирается на предложенный ранее метод достоверного сравнения [5, 6]. Рассмотренный способ, примененный к М-оценке с весовой функцией Тьюки, увеличивает ее помехоустойчивость более чем 1,5 раза практически без снижения эффективности. Эффективности полученной М-оценки и метода достоверного сравнения равны соответственно 0,77 и 0,47, точки срыва – 0,5 и 0,6, а точки срыва определения их доверительных интервалов – 0,6 и 0,65. Предложенные М-оценку, а также метод достоверного сравнения рекомендуется применять для обработки любых выборок многократных наблюдений в любых приложениях, если выборки содержат соответственно более 15 или 6 – 7 наблюдений. Использование подобных методов особенно полезно, когда доля промахов велика, т. е. сильно искажены до половины (и даже до 60 %) всех наблюдений. Выигрыш велик весом при подавлении промахов от наиболее «опасных» для существующих методов однополярных помех.

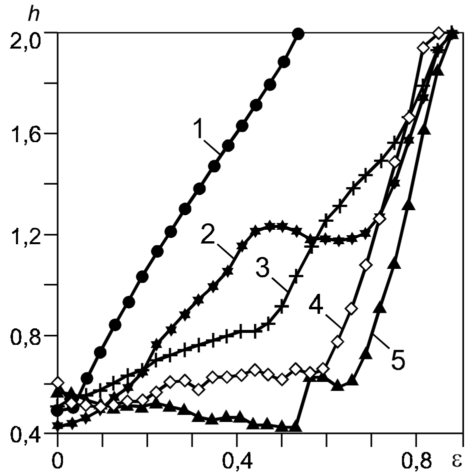

Рис. 4. Зависимость среднего квадратического отношения *h*, определяемого по (7), фактических ошибок результатов измерений к их доверительным интервалам от доли ε однонаправленных промахов:

обозначения те же, что и на рис. 2

Кроме того, предложено новое выражение для определения симметричного двустороннего 95 %-го доверительного интервала медианной оценки многократных наблюдений, которое рекомендуется применять одновременно с определением по [11], особенно при антисимметричном распределении промахов, доля которых велика.

Часть работы выполнена при поддержке интеграционного проекта СО-УрО РАН № 12-С-5-1028.

Л И Т Е Р А Т У Р А

1. **Хьюбер П.** Робастность в статистике. М.: Мир, 1984.

2. **Грановский В. А., Сирая Т. Н.** Методы обработки экспериментальных данных при измерениях. Л.: Энергоатомиздат, 1990.

3. **А. с. 1800927 РФ.** Способ преобразования периодического электрического сигнала в код и устройство для его осуществления /Н. С. Иванов, А. И. Человечков, С. В. Байдиков // Изобретения. 1995. № 13.

4. **Иванов Н. С.** Новые методы цифровой нелинейной фильтрации аномальных помех с неизвестным законом распределения. Екатеринбург: УрО РАН, 2002.

5. **МУ 88-16360-10–2010.** Робастный метод обработки результатов измерений засоренных до 60 % выбросами с неизвестным законом распределения (метод достоверного сравнения).

6. **Иванов Н. С.** Помехоустойчивый и достоверный метод обработки многократных наблюдений // Ежегодник-2010. Екатеринбург: Труды ИГГ УрО РАН. 2011. Вып. 158. С. 209 – 212.

7. **Хампель Ф., Рончетти Э., Рауссеу П., Штаэль В.** Робастность в статистике. Подход на основе функции влияния. М.: Мир, 1989.

8. **Иванов Н. С. и др.** Цифровой вольтметр для метода вызванной поляризации с программно-задаваемыми алгоритмами функционирования // Геофизическая аппаратура. 1993. Вып. 97. С. 61 – 67.

9. **МИ 2174–91.** ГСИ. Аттестация алгоритмов и программ обработки данных при измерениях. Основные положения.

10. **Гильбо Е. П., Челпанов И. Б.** Обработка сигналов на основе упорядоченного выбора. М.: Советское радио, 1975.

11. **ГОСТ Р ИСО 16269-7–2004**. Статистическое представление данных. Медиана. Определение точечной оценки и доверительных интервалов.

12. **ГОСТ Р 50779.21–2004.** Статистические методы. Правила определения и методы расчета статистических характеристик по выборочным данным. Часть 1. Нормальное распределение.

Дата принятия 05.07.2012 г.