

ВАРИАЦИОННЫЙ КОНТРОЛЬ КАЧЕСТВА ДАННЫХ В ЗАДАЧЕ ВАРИАЦИОННОГО УСВОЕНИЯ ДАННЫХ ИЗМЕРЕНИЙ ПРИЗЕМНОЙ ТЕМПЕРАТУРЫ

В.С. Рогутов

*Гидрометеорологический научно-исследовательский центр Российской Федерации
rogutovv@mail.ru*

Введение

Классические методы усвоения данных неустойчивы к большим ошибкам измерений и требуют предварительной обработки данных. Обычно такая обработка включает в себя несколько предварительных шагов: контроль по первому приближению, контроль по соседям, и выполняется до использования данных в схеме усвоения. Вариационный контроль качества измерений, предложенный в [3, 5], встроен в схему вариационного усвоения и учитывает наличие больших ошибок благодаря использованию негауссовой статистики измерений.

Вариационный контроль качества измерений работает как часть решения вариационной задачи. Достоинством этого метода является одновременное использование всех доступных измерений и первого приближения для решения задачи контроля качества данных. В вариационном контроле качества данных классический штрафной функционал искривляется, чтобы учесть наличие больших ошибок, не подчиняющихся Гауссову распределению. Это приводит к тому, что данные измерений учитываются при усвоении с весами, уменьшающимися при росте невязки. Данные с большими невязками по отношению к текущему состоянию не отбрасываются полностью и продолжают влиять на усвоение при дальнейших итерациях минимизации.

Вариационный контроль качества данных.

Опишем кратко алгоритм трехмерного вариационного усвоения данных (3D-var). Предположим, нам известны первое приближение вектора состояния атмосферы x_b и вектор измерений y . Тогда задача вариационного усвоения данных сводится к минимизации функционала

$$J = \frac{1}{2}(x - x_b, B^{-1}(x - x_b)) + \frac{1}{2}(R^{-1}(y - Hx), y - Hx), \quad (1)$$

где B – матрица ковариации ошибок первого приближения; R – матрица ковариации погрешностей измерений; x – вектор состояния атмосферы; H – оператор наблюдений. Если оператор наблюдений H – линейный, то решением задачи является оценка метода оптимальной интерполяции

$$\hat{x} = BH^* (HBH^* + R)^{-1} \quad (2)$$

В случае независимых наблюдений их ошибки не скоррелированы и матрица R становится диагональной

$$J = \frac{1}{2}(x - x_b, B^{-1}(x - x_b)) + \frac{1}{2} \sum_i (y^i - H^i x)^2 / r^i. \quad (3)$$

Вторую часть штрафного функционала можно представить в виде:

$$J_o^N = \frac{1}{2} \sum_i (y^i - H^i x)^2 / r^i = \sum_i J_{oi}^N = - \sum_i \ln p_i + c, \quad (4)$$

где $p_i = \frac{1}{r^i \sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{y^i - H^i x}{r^i}\right)^2\right]$ – плотность вероятности Гауссова распределения.

Идея вариационного контроля данных, предложенная в [3, 5] – заменить в штрафном функционале J_o^N на $J_o^{qc} = - \sum_i \ln p_i^{qc} + c$, где p^{qc} – плотность вероятности ошибки измерений с учетом наличия больших ошибок. Предположим, что все измерения делятся на две группы: измерения, ошибки которых имеют нормальное гауссово распределение p^n и измерения с большими ошибками с равномерным распределением p^g . Пусть априорная вероятность отсутствия большой ошибки равна A , вероятность большой ошибки $F = 1 - A$.

Тогда плотность вероятности ошибок наблюдений с учетом наличия больших ошибок равна

$$p^{qc} = Ap^n + (1 - A)p^g, \quad (5)$$

где $p^g = 1/D = 1/2dr^i$ – плотность вероятности больших ошибок, D – ширина распределения. Тогда штрафной функционал можно переписать в виде

$$J_o^{qc} = - \ln p_i^{qc} = - \sum_i \ln \left(\frac{\gamma + \exp(-J_{oi}^N)}{\gamma + 1} \right), \quad (6)$$

где $J_{oi}^N = \frac{1}{2}(y^i - H^i x) / r^i$ — второе слагаемое «классического» штрафного функционала вариационного усвоения. Градиент штрафного функционала вариационного контроля качества равен

$$\nabla_{\hat{y}} J_o^{qc} = \sum_i \nabla_{\hat{y}} J_{oi}^N \left[1 - \frac{\gamma}{\gamma + \exp(-J_{oi}^N)} \right], \quad (7)$$

где $\gamma = \frac{A\sqrt{2\pi}}{2d(1-A)}$.

$p_{ap}^g = \frac{\gamma}{\gamma + \exp(-J_{oi}^N)}$ — это апостериорная вероятность большой ошибки [5]. Тогда

выражение для градиента функционала можно переписать в виде

$$\nabla_{\hat{y}} J_o^{qc} = \sum_i \nabla_{\hat{y}} J_{oi}^N (1 - p_{ap}^g).$$

Таким образом, отдельные измерения используются в вариационном контроле качества с весами, равными апостериорной вероятности большой ошибки, и никогда не отбрасываются полностью.

Теоретически вариационный контроль качества не требует предварительной обработки данных.

Однако, если начальное приближение для минимизации функционала (6) было грубым, то на практике $\nabla_{\hat{y}} J_o^{qc}$ может оказаться сопоставимым с машинным нулем, и метод сопряженных градиентов не будет работать корректно. Кроме того, функционал (6) может иметь несколько локальных минимумов.

Таким образом, для минимизации функционала вариационного контроля качества требуется настолько качественное начальное приближение, насколько это возможно. На практике качественное начальное приближение можно получить, выполнив сначала несколько итераций метода сопряженных градиентов с классическим штрафным функционалом с предварительной обработкой больших ошибок (background check). Затем полученное решение можно использовать как начальное приближение при минимизации функционала J_o^{qc} .

Численные эксперименты

Используем метод вариационного усвоения данных с вариационным контролем качества для оценки приземной температуры по измерениям приземной температуры на

метеостанциях (SYNOP) и кораблях (SHIP). В качестве первого приближения используем прогнозы глобальной модели атмосферы ПЛАВ [1] на географической сетке $0,72^\circ \times 0,9^\circ$ (251×400 узлов).

Матрицу ошибок первого приближения будем задавать с помощью предложенной в [4] формулы для функции корреляции с компактным носителем, заданной на сфере:

$$B_{ij} = f(x) = \begin{cases} x^2 \left(-\frac{5}{3} + x \left(\frac{5}{8} + x(0.5 - 0.25x) \right) \right) + 1 & 0 \leq |x| \leq 1/2 \\ x \left(-5 + x \left(-\frac{5}{3} + x \left(\frac{5}{8} + x(-0.5 + x/12) \right) \right) \right) + 4 - \frac{2}{3*x}, & 1/2 \leq |x| \leq 1, \quad x = r/c, \quad (8) \\ 0 & |x| > 1 \end{cases}$$

где c – радиус корреляции; r – расстояние между двумя точками.

Для понижения размерности задачу усвоения приземной температуры можно разбить на независимое усвоение температуры в отдельных регионах с радиусом, равным (или превышающим) радиус корреляции. Однако при радиусе корреляции 800 км в приполярных регионах размерность поля температуры достигает 3600, и матрица ковариации ошибок первого приближения становится плохо обусловленной. Этих недостатков можно избежать при использовании в вариационном усвоении редуцированной широтно-долготной сетки [2].

Сравним результаты усвоения температуры методом оптимальной интерполяции на регулярной широтно-долготной сетке с вариационным методом без вариационного контроля качества и вариационным методом с вариационным контролем качества на редуцированной сетке. При усвоении приземной температуры будем использовать случайно выбранные 90 % измерений, остальные 10 % используем для оценки погрешности $e = \langle y^i - H^i x \rangle$.

На рис. 1 приведены примеры разностей полученной оценки и первого приближения (инкрементов анализа). В табл. 1 приведены полученные оценки погрешности объективного анализа с 01.04 по 31.05.2011 года.

Как видно из рисунка и таблицы, результаты почти совпадают. Поскольку для стационарных измерений приземной температуры оператор наблюдений H линейный, то 3D-var и метод оптимальной интерполяции эквивалентны при условии задания одинаковых матриц B и R .

**Среднеквадратичные погрешности анализа приземной температуры (в градусах)
в апреле–мае 2011 года**

	Мир	Евразия	Европа	Россия
Оптимальная интерполяция	2,08	2,09	1,86	1,79
3D-var	2,13	2,14	1,90	1,84
3D-var с вариационным контролем качества	2,14	2,15	1,90	1,85
3D-var без предварительной проверки	2,14	2,15	1,90	1,85
3D-var с предварительной проверкой, с вариационным контролем качества	2,15	2,16	1,91	1,85

Небольшие отличия объясняются тем, что в методе оптимальной интерполяции вместо матрицы ковариации B задаются матрицы BH^* и $NBH^* + R$ по формуле (7). Так же небольшие отличия в приполярных областях объясняются использованием редуцированной сетки.

Побочным продуктом вариационного контроля качества является апостериорная вероятность большой ошибки. Разделим измерения на 4 группы: измерения с вероятностью большой ошибки меньше 25 %, с вероятностью от 25 до 50 %, от 50 до 75 % и больше 75 %. Построим гистограмму распределения измерений в каждой из групп в зависимости от инкремента анализа $y - Hx_a$ (рис. 2) и инкремента первого приближения $y - Hx_b$ (рис. 3).

Как видно из табл. 1 и 2, при объективном анализе стационарных измерений приземной температуры наличие или отсутствие вариационного контроля качества почти не влияет на результаты усвоения, так как больших ошибок в исходных данных немного, и они незначительно влияют на качество объективного анализа. Так же почти не влияет на результаты анализа использование данных без предконтроля.

Однако, как видно из рис. 3, можно ожидать существенных отличий результатов объективного анализа с использованием вариационного контроля качества и контроля по первому приближению при усвоении менее качественных данных.

Результаты применения вариационного контроля качества

	Большие ошибки, найденные контролем по первому приближению	$P_{ap}^g, \%$			
		больше 75	от 50 до 75	от 25 до 50	меньше 25
Весь мир, данные с предконтролем.	436	6400	2195	3777	1636032
Россия, данные с предконтролем.	42	1004	364	616	348600
Весь мир, данные без предконтроля.	596	8667	2446	4039	1638815
Россия, данные без предконтроля.	65	1445	418	651	349090

Выводы

Результаты усвоение станционных измерений приземной температуры с использованием метода оптимальной интерполяции и вариационного усвоения с вариационным контролем качества и без него весьма близки. Необходимо проверить работу вариационного контроля качества при усвоении менее качественных данных.

Список использованных источников

1. Толстых М.А. Глобальная полулагранжева модель численного прогноза погоды. – М.; Обнинск: ОАО ФОР, 2010 – 111 стр.
2. Фадеев Р.Ю. Построение редуцированной широтно-долготной сетки для задач глобального численного прогноза погоды // Метеорология и гидрология. – 2006. – № 9. – С. 5–21.
3. Andersson E., Jarvinen H. Variational quality control // Quart.J.Roy.Meteorol.Soc. – 1999. – No. 125. – P. 697–722.
4. Gaspari G., Cohn S. Construction of the correlation functions in two and three dimensions. // Quart.J.Roy.Meteorol.Soc. – 1999. – No. 125. – P. 723–757.
5. Ingeby, N.B., Lorenc A.C. Bayesian quality control using multivariate normal distributions. // Quart.J.Roy.Meteorol.Soc. – 1993. – No. 119. – P. 1195–1225.

Поступила в редакцию 27.09.2012 г.

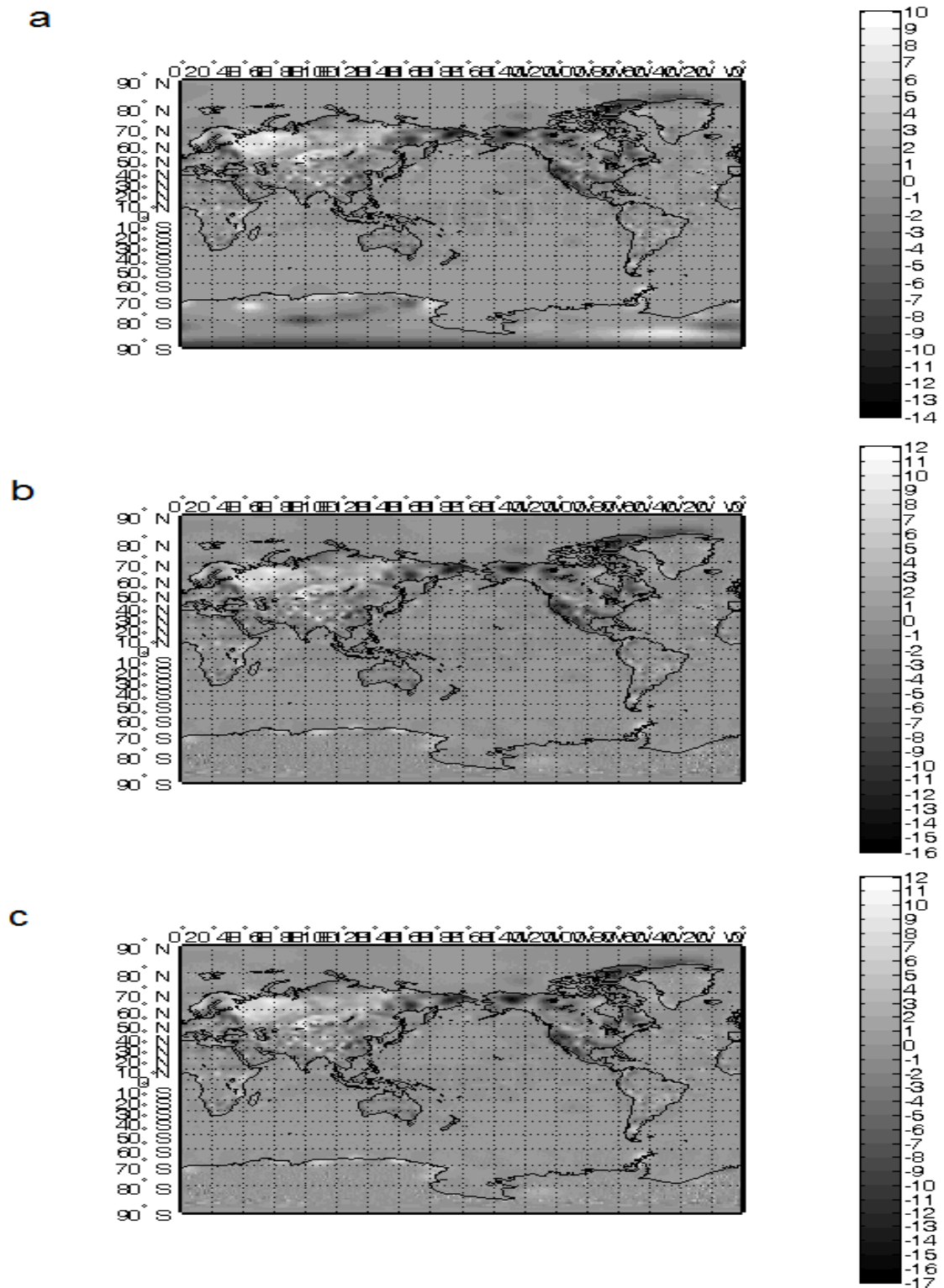


Рис. 1. Пример разности оценок приземной температуры 01.04.2011 г. и температуры первого приближения, полученной методом оптимальной интерполяции (а), вариационным усвоением данных без вариационного контроля качества (б) и с контролем качества (с).

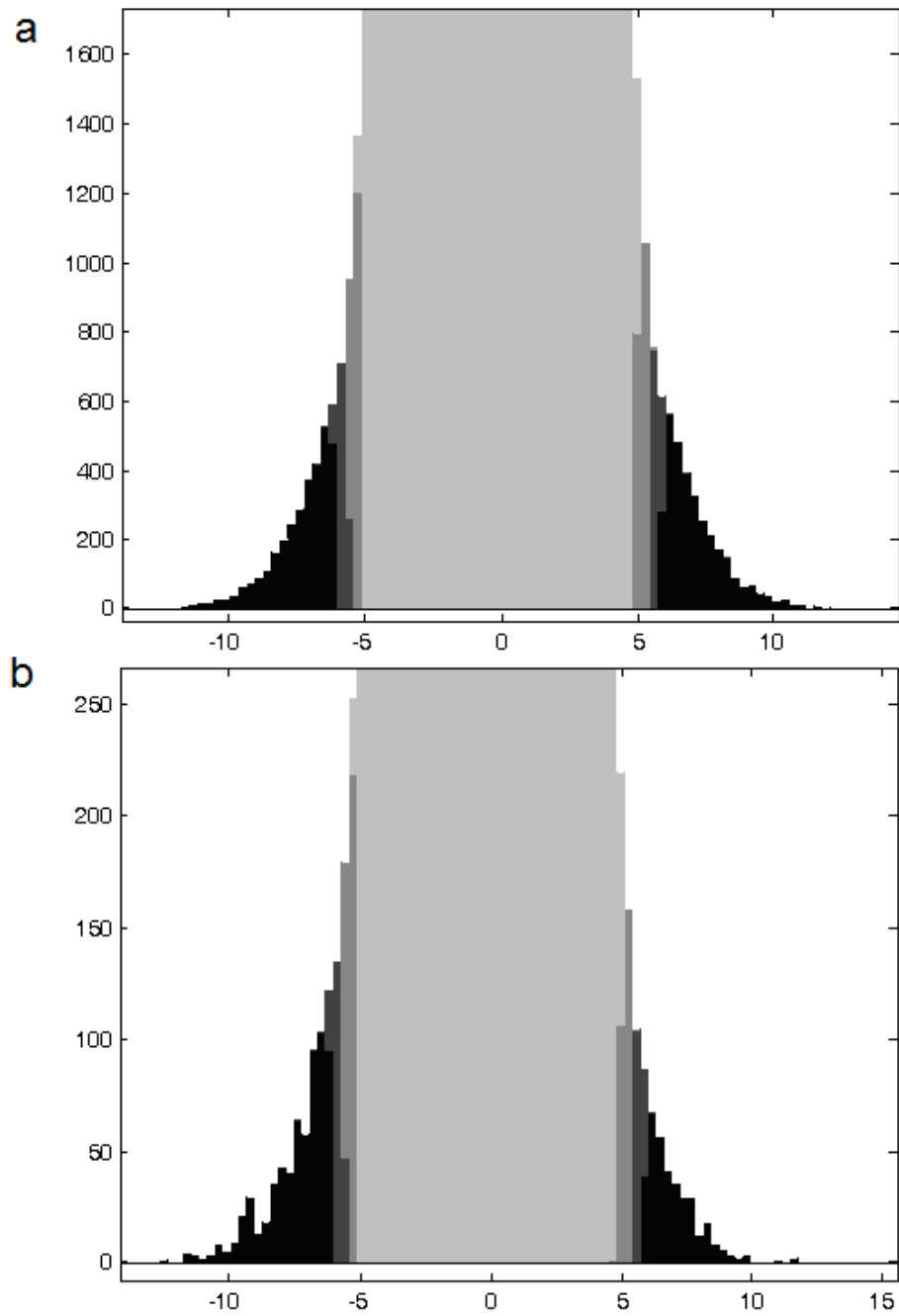


Рис. 2. Гистограмма измерений минус анализ в мире (а) и в России (б).

Измерения разбиты на 4 группы: с апостериорной вероятностью большой ошибки p_{ap}^g меньше 25 %, от 25 до 50 %, от 50 до 75 % и больше 75% (показаны, соответственно, оттенками от светлого к черному).

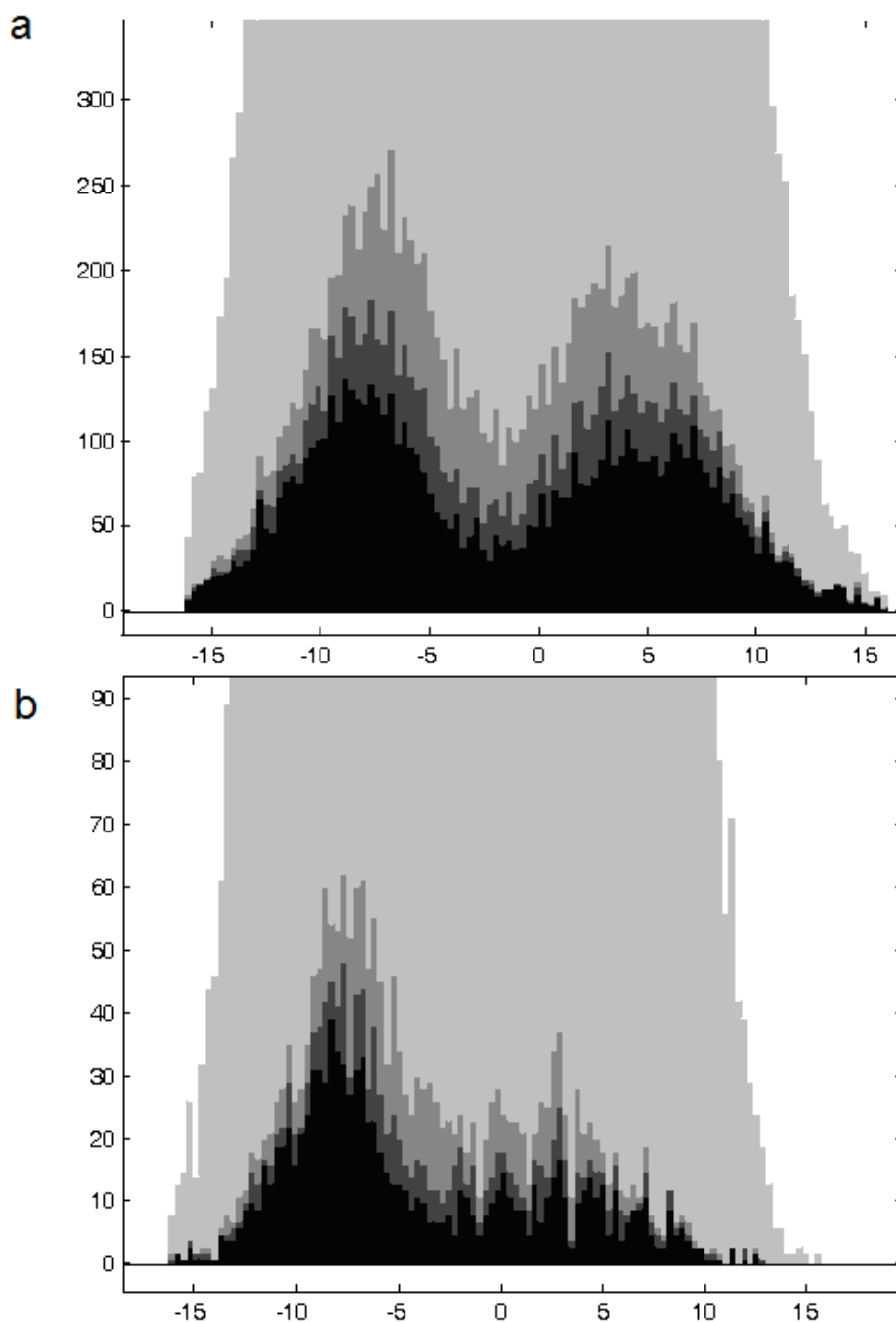


Рис. 3. Гистограмма измерений минус первое приближение в мире (а) и в России (б).
 Измерения разбиты на 4 группы: с апостериорной вероятностью большой ошибки p_{ap}^g меньше 25 %, от 25 до 50 %, от 50 до 75 % и больше 75% (показаны, соответственно, оттенками от светлого к черному).