*Вестник Томского государственного университета. 2013. № 375. С. 169–173* DOI: 10.17223/15617793/375/36

# НАУКИ О ЗЕМЛЕ

УДК 504.3.054

*Г.Г. Журавлев, Э.В. Иванова, А.И. Кусков*

## ПРИМЕНЕНИЕ РОБАСТНЫХ ПРОЦЕДУР ДЛЯ ОЦЕНКИ УРОВНЯ ЗАГРЯЗНЕНИЯ АТМОСФЕРЫ г. ТОМСКА

В работе рассмотрен один из путей оценки параметров распределения полей загрязнителей – использование процедур, нечувствительных к структуре данных, а именно робастных процедур оценивания. Применены две наиболее часто встречающиеся робастные оценки: винзоризованная и усеченная. Использование робастных оценок позволяет не только корректировать средние значения выборки с учетом «выбросов», но и определять аномальные, в статистическом смысле, значения концентрации загрязнителей.

**Ключевые слова:** робастные процедуры; уровень загрязнения атмосферы; винзоризованная и усеченная оценка; годовой ход.

Для решения проблемы загрязнения окружающей среды первостепенное значение имеет тщательное изучение уровней загрязнения в пространстве и во времени. При измерении концентраций загрязнителей атмосферы, особенно при массовых измерениях и значительных вариациях полей концентраций, трудно избежать ошибок. Поэтому для принятия оптимальных природоохранных решений необходима корректная оценка уровня загрязнения [1].

Как правило, в выборке встречаются грубые погрешности измерений или аномальные значения, которые сильно искажают величину среднего значения и, особенно, дисперсии. В работе [2] приводится ссылка на пример, когда 10% измерений, представляющие собой аномальные значения, увеличивают оценку дисперсии в 2 раза.

Многие исследователи исключают аномальные значения из дальнейшей обработки как не относящееся к данному распределению. Другие, после удаления выпадающих наблюдений, исследуют их отдельно, потому что аномалии могут представлять больший интерес, чем сама выборка. Можно встретить большое количество рекомендаций по выявлению и отсеву аномальных значений [2].

Одним из путей оценки параметров распределения полей загрязнителей является использование процедур, нечувствительных к структуре данных. Такие процедуры оценивания называют робастными [3–5]. Среди робастных оценок наиболее часто применяют две: винзоризованные и усеченные [5].

В данной работе рассмотрены возможность применения робастных процедур для оценки уровня загрязнения воздушного бассейна г. Томска, а также основные статистические характеристики ряда концентраций двуокиси азота. Материалом для исследования послужили результаты измерений двуокиси азота, проводимых в течение четырех лет на одном из наблюдательных постов города.

В работе применены два варианта расчета статистических характеристик уровней загрязнения. Для первого варианта расчет статистических характеристик проводился с использованием всего массива, в который входили и случаи, когда уровень загрязнения был равен нулю. Для второго варианта применялись только те случаи, когда отмечался уровень загрязнения, отличный от нуля. Результаты расчетов по двум вариантам приведены в табл. 1.

Сравнение средних значений и средних квадратических отклонений обоих вариантов позволил установить, что величина среднего уровня загрязнения при расчете по второму варианту в 2–3 раза выше по сравнению с первым. Величины стандартных отклонений, полученные в разных вариантах, практически не отличаются. В структуре годового хода стандартных отклонений четко прослеживаются два максимума (в мае и декабре), которые отличаются в 2–3 раза от значений других месяцев. Для выявления причины возникновения двух максимальных пиков в годовом ходе среднего квадратического отклонения была проведена классификация уровней загрязнения по месяцам.

В табл. 2 приведено число случаев по месяцам с различными уровнями загрязнения. Уровни загрязнения выражены в долях ПДК (разовая ПДК для двуокиси азота составляет 0,085 мг/м). Из табл. 2 следует, что для всех месяцев года, кроме мая и декабря, концентрация двуокиси азота не превышает трех ПДК. В мае отмечаются уровни от 3,5 до 5 ПДК, в декабре – от ПДК и выше, причем таких случаев в мае и декабре отмечалось всего по два. Данные случаи можно отнести к разряду аномальных. Для проверки этого использовались робастные процедуры, а именно винзоризованные и усеченные оценки.

Винзоризованные оценки применяют при оценивании среднего значения и дисперсий, построении доверительных интервалов, а также при проверке гипотез относительно генерального среднего в ситуациях, когда можно предположить присутствие в выборке аномальных значений (выбросов). В этой процедуре крайние значения в упорядоченном возрастанию или убыванию ряду не отбрасываются, а изменяются.

Если выборка состоит из $n$ наблюдений, тогда $g$-винзоризованные наблюдения получаются путем замены первых g наблюдений на значение $(g + 1)$ наблюдения. Последние $(g)$ наблюдений также заменяются значением $(n – g – 1)$ наблюдения.

Таблица 1

**Статистические характеристики загрязнения атмосферы двуокисью азота для различных вариантов расчета**

| Месяц | Статистические характеристики | | | | | | |
|---|---|---|---|---|---|---|---|
| | Среднее | | Ср. кв. отклонение | | Число случаев | | Вероятность |
| | п | о | п | о | п | о | % |
| Январь | 0,021 | 0,047 | 0,031 | 0,03 | 273 | 124 | 45,4 |
| Февраль | 0,026 | 0,044 | 0,028 | 0,022 | 263 | 153 | 58,2 |
| Март | 0,025 | 0,051 | 0,032 | 0,028 | 275 | 137 | 49,8 |
| Апрель | 0,018 | 0,042 | 0,025 | 0,022 | 298 | 126 | 42,3 |
| Май | 0,023 | 0,055 | 0,046 | 0,057 | 281 | 118 | 42 |
| Июнь | 0,02 | 0,042 | 0,027 | 0,024 | 304 | 142 | 46,7 |
| Июль | 0,018 | 0,04 | 0,025 | 0,022 | 312 | 145 | 46,5 |
| Август | 0,018 | 0,042 | 0,025 | 0,021 | 305 | 133 | 43,6 |
| Сентябрь | 0,018 | 0,042 | 0,025 | 0,021 | 305 | 133 | 43,6 |
| Октябрь | 0,012 | 0,037 | 0,021 | 0,021 | 299 | 100 | 33,4 |
| Ноябрь | 0,025 | 0,045 | 0,032 | 0,031 | 273 | 150 | 54,9 |
| Декабрь | 0,028 | 0,047 | 0,062 | 0,074 | 301 | 180 | 59,8 |

*Примечание.* о – оптимальная выборка; п – полная выборка.

Таблица 2

**Распределение числа случаев с концентрациями загрязнения в долях ПДК для двуокиси азота по месяцам**

| Загрязнения в ПДК | Месяц | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | янв. | фев. | март | апр. | май | июнь | июль | авг. | сент. | окт. | ноябрь | дек. |
| Нет | 149 | 110 | 137 | 172 | 163 | 162 | 167 | 204 | 172 | 199 | 123 | 121 |
| 0,0–0,5 | 67 | 98 | 74 | 87 | 68 | 100 | 106 | 91 | 96 | 79 | 97 | 117 |
| 0,5–1,0 | 49 | 44 | 47 | 34 | 38 | 33 | 35 | 19 | 30 | 17 | 42 | 55 |
| 1,0–1,5 | 5 | 10 | 13 | 4 | 2 | 8 | 2 | 6 | 6 | 3 | 6 | 5 |
| 1,5–2,0 | 2 | 1 | 4 | 1 | 2 | 1 | 1 | 3 | 1 | 1 | 3 | 2 |
| 2,0–2,5 | 0 | 0 | 0 | 0 | 5 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| 2,5–3,0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 3,0–3,5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3,5–4,0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4,0–4,5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4,5–5,0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5,0–5,5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5,5–6,0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6,0–6,5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6,5–7,0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7,0–7,5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7,5–8,0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8,0–8,5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| > 8,5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Таким образом, по определению

$$z_1 = z_2 = ... = z_g = x_{g+1}$$

$$z_{g+i} = x_{g+i}\ ,\ 2 < i < n - g - 1$$

$$z_n = z_{n-1} = ... = z_{n-g-1} = x_{n-g}.$$

При этом выборочное среднее ($\overline{z}_g$) и среднее квадратическое отклонение ($s_g$) оценивают как

$$\overline{z}_g = \frac{1}{n}\sum_{i=1}^{n} z_i\ ,\ i = 1, 2, ..., n$$

$$s_g^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(z_i - \overline{z}_g\right)^2.$$

Приближенный (1–α)100%-ный $g$-винзоризованный доверительный интервал для среднего (μ) определяется как

$$\overline{z}_g - t_{(1-\alpha/2),h-1}\left[\frac{n-1}{h-1}\right]\left[\frac{s_g}{n}\right] <$$

$$< \mu < \overline{z}_g + t_{(1-\alpha/2),h-1}\left[\frac{n-1}{h-1}\right]\left[\frac{s_g}{n}\right],$$

где $h = n-2g$, $t_{(1-\alpha/2),\,h-1}$ – (1–α)100% – точка в $t$-распределении Стьюдента с ($h$–1) степенями свободы.

Другая робастная процедура – усеченные оценки среднего получаются путем отбрасывания $g$ крайних

значений выборки. Тогда *a*-усеченная оценка среднего равна

$$\overline{z}_u(a) = \frac{1}{h}\sum_{i=g+1}^{n-g} z_i,$$

где (*a*) выбирается так, чтобы (*g* = *na*), величина 100(1–2a) показывает долю наблюдений, расположенных в середине упорядоченного ряда, по которым производился расчет среднего. Можно провести переход от *a*-усеченной оценки к *g*-усеченной, которая в использовании является более удобной. При этом $\overline{z}_u = \overline{z}_g$, если выполняется условие *g* = *na*.

Стандартное отклонение от среднего вычисляется по формуле

$$s_u(g) = \frac{ss\ (g)}{h\ (h-1)},$$

где *ss* (*g*) обозначает винзоризованную сумму квадратов:

$$ss(g) = (g+1)\ [z_{g+1} - \overline{z}_u(g)]^2 +$$
$$[z_{g+2} - \overline{z}_u(g)]^2 + \ldots + [z_{n-g-1} - \overline{z}_u(g)]^2 +$$
$$(g+1)\ [z_{n-g} - \overline{z}_u(g)]^2.$$

Приближенный (1–α)100%-ный a-усеченный интервал для среднего равен

$$\overline{z}_u(a) - t_{(1-\alpha/2),\,h-1}\ s_u\ (g) < \mu < \overline{z}_u(a) + t_{(1-\alpha/2),\,h-1}\ s_u\ (g).$$

Применение робастных процедур, как правило, приводит к изменению (уменьшению) оценки среднего значения, стандартного отклонения и длины доверительного интервала, которые могут отмечаться на всем выбранном диапазоне.

Уменьшение длины доверительного интервала, как известно, увеличивает точность оценки среднего. Поэтому при использовании робастных оценок представляется возможность выбора между получением более точной оценки и изменением слишком большого числа наблюдений. При выборе робастной оценки следует руководствоваться не только длиной доверительного интервала, но и изменением его границ. Известно, что распределение концентраций подчиняется логарифмически нормальному закону [6]. При этом область возможных значений ограничена нулем слева и возможность изменения концентраций загрязнителя задана в одну сторону. Особенно это справедливо, если стандартное отклонение велико по сравнению со средним значением [7], что имеет место в распределении двуокиси азота в г. Томске. Применение робастных оценок для определения статистических характеристик концентраций загрязняющих веществ оправдано тем, что в условиях распределения концентраций, близких к логарифмически нормальному закону, формируется «хвост», состоящий из редко наблюдаемых больших или даже очень больших значений концентраций.

По исходной выборке были получены *g*-винзоризованные оценки среднего и дисперсии для всех месяцев года. Величина *g* изменялась от 0 до 6. Результаты расчетов для мая и декабря приведены в табл. 3. Из анализа табл. 3 следует, что в мае при *g* = 2 величина 95%-ного доверительного интервала стабилизируется и дальнейшее увеличение практически не отражается на величине доверительного интервала (рис. 1, 2).

Для декабря при *g* = 1 величина доверительного интервала и среднее квадратическое отклонение уменьшаются более чем в 3 раза, а оценка среднего значения уменьшилась почти на 10% против среднего, полученного без применения робастной процедуры. Следует отметить, что по мере увеличения *g* изменялся только максимум выборки. Таким образом, *g*-винзоризованные оценки получаются путем корректировки максимального значения выборки.

Таблица 3

**Винзоризованные оценки среднего, среднего квадратического отклонения, границы интервалов, длина интервала, максимальное и минимальное значение, попавшее в расчет**

| *g* | Статистические характеристики | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Среднее | Ср. кв. | Нижняя граница | Верхняя граница | Интервал | Max | Min | Доля, % |
| Май | | | | | | | | |
| 0 | 0,0549 | 0,0572 | 0,0445 | 0,0653 | 0,0209 | 0,02 | 0,39 | 1,00 |
| 1 | 0,0543 | 0,054 | 0,0443 | 0,0643 | 0,02 | 0,02 | 0,32 | 0,94 |
| 2 | 0,0526 | 0,0466 | 0,0438 | 0,0614 | 0,0176 | 0,02 | 0,22 | 0,81 |
| 3 | 0,0524 | 0,0457 | 0,0436 | 0,0612 | 0,0176 | 0,02 | 0,21 | 0,80 |
| 4 | 0,052 | 0,0445 | 0,0433 | 0,0608 | 0,0174 | 0,02 | 0,2 | 0,78 |
| 5 | 0,0516 | 0,0431 | 0,043 | 0,0602 | 0,0172 | 0,02 | 0,19 | 0,75 |
| 6 | 0,0516 | 0,0431 | 0,0428 | 0,0604 | 0,0176 | 0,02 | 0,19 | 0,75 |
| Декабрь | | | | | | | | |
| 0 | 0,0474 | 0,0744 | 0,0365 | 0,0584 | 0,0219 | 0,02 | 1 | 1,00 |
| 1 | 0,0427 | 0,0223 | 0,0394 | 0,046 | 0,0066 | 0,02 | 0,15 | 0,30 |
| 2 | 0,0425 | 0,0213 | 0,0393 | 0,0457 | 0,0064 | 0,02 | 0,13 | 0,29 |
| 3 | 0,0425 | 0,0213 | 0,0393 | 0,0457 | 0,0064 | 0,02 | 0,13 | 0,29 |
| 4 | 0,0421 | 0,0199 | 0,039 | 0,0452 | 0,0061 | 0,02 | 0,11 | 0,27 |
| 5 | 0,0416 | 0,0182 | 0,0387 | 0,0444 | 0,0057 | 0,02 | 0,09 | 0,24 |
| 6 | 0,0416 | 0,0182 | 0,0387 | 0,0444 | 0,0057 | 0,02 | 0,09 | 0,24 |

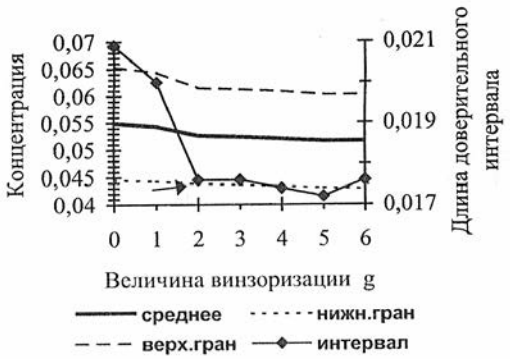


Рис. 1. Влияние процесса винзоризации на величину среднего значения концентрации $NO_2$ и длину 95%-ного доверительного интервала (май)

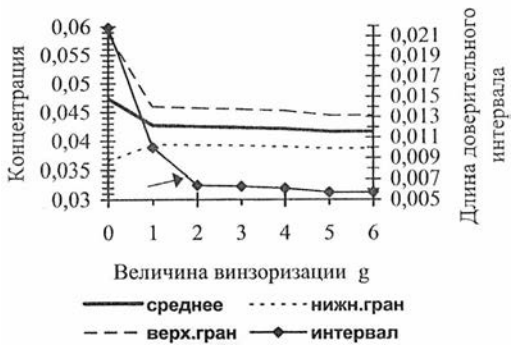


Рис. 2. Влияние процесса винзоризации на величину среднего значения концентрации $NO_2$ и длину 95%-ного доверительного интервала (декабрь)

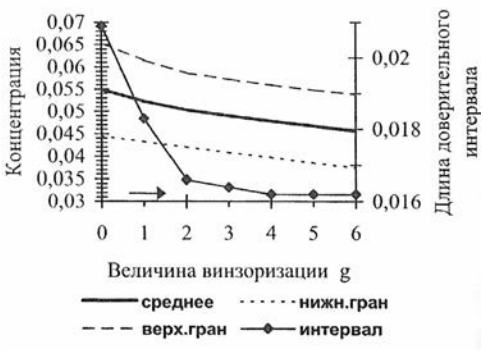


Рис. 3. Влияние процесса усечения на величину среднего значения концентрации $NO_2$ и длину 95%-ного доверительного интервала (май)

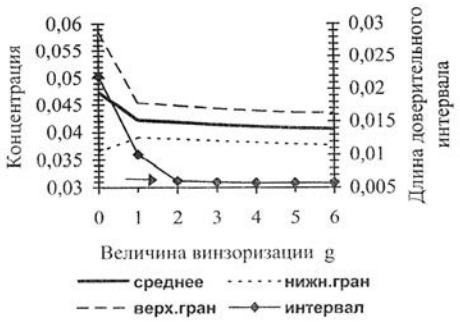


Рис. 4. Влияние процесса усечения на величину среднего значения концентрации $NO_2$ и длину 95%-ного доверительного интервала (декабрь)

Таблица 4

**Усеченные оценки среднего, среднего квадратического отклонения, границы интервалов, длина интервала, максимальное и минимальное значение, попавшее в расчет**

| $g$ | Статистические характеристики | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Среднее | Ср. кв. | Нижняя граница | Верхняя граница | Интервал | Max | Min | Доля, % |
| | | | Май | | | | | |
| 0 | 0,0549 | 0,0572 | 0,0445 | 0,0653 | 0,0209 | 0,02 | 0,39 | 1,00 |
| 1 | 0,0523 | 0,0496 | 0,0432 | 0,0615 | 0,0183 | 0,02 | 0,32 | 0,87 |
| 2 | 0,0503 | 0,0447 | 0,042 | 0,0586 | 0,0166 | 0,02 | 0,22 | 0,78 |
| 3 | 0,049 | 0,0437 | 0,0408 | 0,0572 | 0,0164 | 0,02 | 0,21 | 0,76 |
| 4 | 0,0478 | 0,043 | 0,0397 | 0,0559 | 0,0162 | 0,02 | 0,2 | 0,75 |
| 5 | 0,0467 | 0,0426 | 0,0385 | 0,0548 | 0,0162 | 0,02 | 0,19 | 0,74 |
| 6 | 0,0456 | 0,0425 | 0,0374 | 0,0538 | 0,0164 | 0,02 | 0,19 | 0,74 |
| | | | Декабрь | | | | | |
| 0 | 0,0474 | 0,0744 | 0,0365 | 0,0584 | 0,0219 | 0,02 | 1 | 1,00 |
| 1 | 0,0422 | 0,0213 | 0,0391 | 0,0454 | 0,01 | 0,02 | 0,15 | 0,29 |
| 2 | 0,0418 | 0,0201 | 0,0388 | 0,0448 | 0,006 | 0,02 | 0,13 | 0,27 |
| 3 | 0,0414 | 0,0194 | 0,0385 | 0,0443 | 0,0058 | 0,02 | 0,12 | 0,26 |
| 4 | 0,041 | 0,019 | 0,0382 | 0,0439 | 0,0057 | 0,02 | 0,11 | 0,26 |
| 5 | 0,0408 | 0,0187 | 0,0379 | 0,0436 | 0,0057 | 0,02 | 0,09 | 0,25 |
| 6 | 0,0406 | 0,0188 | 0,0377 | 0,0435 | 0,0057 | 0,02 | 0,09 | 0,25 |

Усеченные оценки концентрации двуокиси азота были получены также для мая и декабря (табл. 4). Сравнение $g$-усеченных оценок с соответствующими винзоризованными для мая показывает, что стабилизация длины 95%-ного доверительного интервала в обеих процедурах начинается с $g = 2$.

Среднее значение, среднее квадратическое отклонение и величина доверительного интервала при всех использованных значениях $g$-усеченных оценок меньше, чем винзоризованных, или равны им. Аналогичное соотношение оценок отмечается при обработке концентраций двуокиси азота, наблюдавшихся в декабре (рис. 3, 4).

При использовании робастных оценок представляется возможность выбора между получением более точной оценки и изменением слишком большого числа наблюдений.

Применение робастных процедур, как правило, приводит к изменению оценки среднего значения и уменьшению длины доверительного интервала, которое может отмечаться на всем выбранном диапазоне $g$ .

Поэтому при выборе робастной оценки следует руководствоваться не только длиной доверительного интервала, но и изменением его границ. Выбор будет оправдан, если среднее значение, рассчитанное стандартным образом, не попадает внутрь доверительного интервала робастной оценки.

На примере табл. 4 можно показать, что для мая среднее значение концентрации двуокиси азота, рассчитанное обычным образом, попадает внутрь доверительного интервала винзоризованной оценки при $g = 6$.

Следовательно, изменение первоначального среднего нецелесообразно. Иная ситуация отмечается при расчете среднего значения для декабря. Уже при $g = 1$ среднеквадратическое отклонение уменьшается в 3,5 раза, что приводит к уменьшению длины доверительного интервала более чем в 2 раза. При этом первоначальное среднее ($g = 0$) не попадает в 95%-ный доверительный интервал среднего при $g = 1$.

Отсюда можно сделать вывод, что концентрация 1 мг/м$^3$, присутствующая в первоначальной выборке, сильно влияет на ошибку расчета среднего значения, и эту величину можно считатать аномальной для данной выборки. В этом случае изменение оценки представляется целесообразным.

Таким образом, использование робастных оценок позволяет не только корректировать средние значения выборки с учетом «выбросов», но и определять аномальные, в статистическом смысле, значения концентрации загрязнителей.

## ЛИТЕРАТУРА


1. *Безуглая Э.Ю., Расторгуева Г.П., Смирнова И.В.* Чем дышит промышленный город. Л. : Гидрометеоиздат, 1991. 255 с.
2. *Львовский Е.Н.* Статистические методы построения эмпирических формул. М. : Высш. шк., 1988. 239 с.
3. *Афифи А., Эйсен С.* Статистический анализ: Подход с применением ЭВМ : пер. с англ. М. : Мир, 1982. 488 с.
4. *Уланова Е.С., Забелин Б.Н.* Методы корреляционного и регрессионного анализа в агрометеорологии. Л. : Гидрометеоиздат, 1990. 207 с.
5. *Хьюбер П.* Робастность в статистике. М. : Мир, 1984. 303 с.
6. *Безуглая Э.Ю.* Использование статистических методов для обработки данных наблюдений за загрязнением воздуха // Труды ГГО. 1969. Вып. 238. С. 42–47.
7. *Петрович М.Л., Давидович М.И.* Статистическое оценивание и проверка гипотез на ЭВМ. М. : Финансы и статистика, 1989. 191 с.


Статья представлена научной редакцией «Науки о Земле» 18 июня 2013 г.