Environmental Sciences Europe

RESEARCH Open Access

Development of the software tool *Sample Size for Arbitrary Distributions* and exemplarily applying it for calculating minimum numbers of moss samples used as accumulation indicators for atmospheric deposition

Werner Wosniok[1], Stefan Nickel[2*] and Winfried Schröder[2]

Abstract

Background: Do we measure enough to calculate statistically valid characteristic values from random sample measurements, or do we measure too much—without any further increase in knowledge? This question is actually one of the key issues of every empirical measurement design, but is rarely investigated in environmental monitoring.

Results: In this study, the methodology used for the design of the German Moss Survey 2015 network to determine statistically valid minimum sample numbers (MSN) for the calculation of the arithmetic mean value in compliance with certain accuracy requirements was further developed for data that are neither normally nor lognormally distributed. The core element of the procedure for estimating MSN without prerequisite to the distribution of data is an iterative Monte Carlo simulation. The methodological principle consists of using reference data (values measured in Moss Surveys preceding that in 2015) for a series of MSN candidate values to determine what accuracy would be achieved with these, and then calculating the MSN with which the specified accuracy requirement is met from a quadratic function between MSN candidates and their accuracy. The program *Sample Size for Arbitrary Distributions* (*SSAD*) was developed for the calculation of the MSN in the open programming language R.

Conclusions: The SSAD procedure closes a gap in the existing methodology for calculating statistically valid minimum sample numbers.

Keywords: Atmospheric deposition, German Moss Survey, Minimum sample size, Monte Carlo method, Scientific software, Spatial sampling design

Background

With the exception of 2010, Germany participated in the European Moss Survey conducted every 5 years between 1990 and 2015 [4]. As a methodological basis for all national contributions to this international environmental monitoring programme, a guideline is used in which, on the initiative of the authors of this article, compliance with statistically justified minimum sample numbers (MSN = number of moss sampling sites) is recommended for compliance with certain error tolerances when calculating the arithmetic mean for different spatial categories (e.g., administrative units, ecological spatial classes) [1]. The calculation formula of the International Cooperative Programme on Effects of Air Pollution on Natural Vegetation and Crops [1] ("method A") assumes that the standard deviations from many measurements are well known and the element concentrations in the mosses

*Correspondence: stefan.nickel@uni-vechta.de
[2] Chair of Landscape Ecology, University of Vechta, POB 1553, 49364 Vechta, Germany
Full list of author information is available at the end of the article

are normally distributed. In the frequent case that the substance concentrations in the mosses are lognormally distributed (41% of the German data set in 2005), Wosniok ([11], cited in [8]) extended the MSN methodology based on the calculation formula proposed by Cox for the determination of the confidence interval for the mean value in lognormally distributed data (mentioned as "personal communication" in Land ([2], cited in [5]) ("method B"). The purpose of this study is to further develop the methodology for calculating minimum sample numbers for data that are neither normal nor lognormally distributed ("method C"). To verify the methodology, the newly developed method is applied to data from the Moss Survey 2015 and compared with MSN calculations based on the previous methodology.

Method development

Theory

The core element of the procedure for estimating MSN without prerequisite to the distribution of data (method C) is an iterative Monte Carlo simulation [7]. The methodological principle consists of using reference data (previous measured values) for a series of MSN candidate values to determine what accuracy would be achieved with these, and then calculating the minimum MSN with which the specified accuracy requirement is met from a quadratic function between MSN candidates and their accuracy. This is based on the same accuracy criteria as the formula for normally distributed data (method A) or the extension based on the formula of Cox (method B) for lognormally distributed data [3, 9].

The starting point for determining MSN candidate values for method C is the same number of n_1 which is determined under the assumption of a normal distribution according to method A [1]. If $n_1 < 20$, $n_1 = 20$ will be set to numerically stabilize the further procedure. In addition to n_1 further candidate values $n_i'' i = 2, \ldots, I$ for the MSN you are looking for in the range of $\left(\sqrt{n_1}, 2 \cdot n_1\right)$ above and below the first estimate. The typical number of candidate values is $I = 11$.

For each candidate value, n_i is then determined by Monte Carlo simulation (stochastic simulation) how accurately a mean value from the distribution of the available reference data on the basis of a sample of size n_i would be determined. The maximum difference between the mean value and the limits of the 95% confidence interval serves as a measure of accuracy. For the simulation, the density of the available data is calculated once as a kernel density estimate and from this the estimate of the associated distribution function is calculated. From this distribution, random Monte Carlo samples are then taken. $x_{\mathrm{sb}}, m = 1, \ldots, n_i$ the size n_i (inversion method). For each sample, its arithmetic mean $\overline{x}_j$ is determined. Each L successive mean values form a block $B_b = \left\{\overline{x}_{(b-1)L+1}, \overline{x}_{(b-1)L+2}, \ldots, \overline{x}_{bL}\right\}$ of averages. For each block, the absolute local quality criterion

$$\delta_{b,\mathrm{abs}} = \max\left(\left|\overline{x}_{\mathrm{ref}} - Q_{2,5}(B_b))\right|, \left|Q_{97,5}(B_b) - \overline{x}_{\mathrm{ref}}\right|\right) \tag{1}$$

and the relative local criterion

$$\delta_{b,\mathrm{rel}} = \frac{100\delta_{b,\mathrm{abs}}}{\overline{x}_{\mathrm{ref}}} \tag{2}$$

is calculated. Here, $\overline{x}_{\mathrm{ref}}$ is the arithmetic mean of the available reference data, b the index of the current block and $Q_p(B_b)$ the p% quantile of those $\overline{x}_j$ which are assigned to block B_b belong to. With a block size of $L = 120$, the following results are obtained $Q_{2,5}(B_b)$ and $Q_{97,5}(B_b)$ by counting in the ascending sorted sequence of the $\overline{x}_j$ in block B_b. Both criteria are called local because they refer to only one block. The searched global quality criterion $\Delta_{\mathrm{rel}}(n_i, N)$ for a sample of size n_i is calculated by averaging the local contributions over all N drawn blocks:

$$\Delta_{\mathrm{rel}}(n_i, N) = \frac{1}{N}\sum_b \delta_{b,\mathrm{rel}}(n_i). \tag{3}$$

With growing N, the sequence of the $\Delta_{\mathrm{rel}}(n_i, N)$ converges to the true value $\Delta_{\mathrm{rel}}(n_i)$. The practical decision whether enough blocks have already been considered, because the current $\Delta_{\mathrm{rel}}(n_i, N)$ is close enough to $\Delta_{\mathrm{rel}}(n_i)$, is based on the behavior of the last 5 calculated values of $\Delta_{\mathrm{rel}}(n_i, N)$. If all last 4 calculated values of $\Delta_{\mathrm{rel}}(n_i, N)$, i.e.. $\Delta_{\mathrm{rel}}(n_i, N-3), \Delta_{\mathrm{rel}}(n_i, N-2), \Delta_{\mathrm{rel}}(n_i, N-1), \Delta_{\mathrm{rel}}(n_i, N)$, differ less than the specified convergence criterion ε (typical: $\varepsilon = 0.1$) from $\Delta_{\mathrm{rel}}(n_i, N-4)$, then $\Delta_{\mathrm{rel}}(n_i, N)$ is regarded as a sufficiently accurate determination of precision $\Delta_{\mathrm{rel}}(n_i)$. The simulation procedure for the current candidate value n_i is terminated.

After passing through the loop described above over all candidate values n_i, I support points for describing the relationship between n_i and $\Delta_{\mathrm{rel}}(n_i)$ are available. Previous experience has shown that this relationship can be approximated well by the relationship

$$\ln\left(\Delta_{\mathrm{rel}}(n_i)\right) = \beta_0 + \beta_1 n_i + \beta_2 n_i^2. \tag{4}$$

A quadratic parabola can always considered as a local second order Taylor expansion of the unknown true relationship between $\ln\left(\Delta_{\mathrm{rel}}(n_i)\right)$ and n_i. The error of a Taylor expansion in the approximated range could be calculated theoretically, if the true relation was formally known. For the present data, the quality of a quadratic approximation was checked with simulated test data (normally and log normally distributed and overlays of these) and with concentration data from the 2005 moss survey. Presumably a quadratic parabola will provide a

sufficient approximation to the simulated data in many cases; however, we advocate the future user of the SSAD approach to check the validity of this approximation for the reference data at hand. This can be done easily when using the SSAD R script described in "Implementation" section, which provides the relevant figure. If needed, Eq. (4) can be modified to either a higher-order polynomial or a spline function.

The β-coefficients in Eq. (4) can be determined by linear regression. From the estimated coefficients $\hat{\beta}_i$ Eq. (4) and the accuracy requirement $\Delta_{\text{rel,goal}}$ the MSN results from the resolution of the quadratic equation

$$\ln\left(\Delta_{\text{rel, goal}}\right) = \hat{\beta}_0 + \hat{\beta}_1 \text{MSN} + \hat{\beta}_2 (\text{MSN})^2, \tag{5}$$

when

$$a = \frac{\hat{\beta}_1}{\hat{\beta}_2}, b = \hat{\beta}_0 - \frac{\ln \Delta_{\text{rel, goal}}}{\hat{\beta}_2}, d = \sqrt{\frac{a^2}{4} - b} \tag{6}$$

$$\text{MSN}_{1,2} = -\frac{a}{2} \pm d. \tag{7}$$

Equation (5) formally has two solutions of which only MSN_1 in Eq. (7) is a solution for the problem at hand.

The SSAD approach relies on the estimated distribution function of the reference data. The accuracy of this estimate and consequently also of values derived from the estimate depends (among others) on the size of the reference data. For a typical log normal distribution with 2.5% quantile = 10 and 97.5% quantile = 50, a sample size of $n = 25$ allows estimating the mean on the linear scale with a precision of 20% (half-width of the confidence interval). This precision should not be exceeded by a reference sample, therefore the minimum requirement of $n = 25$ for the SSAD approach.

Implementation

The program *Sample Size for Arbitrary Distributions* (*SSAD*) in the open programming language R [6] was developed for the practical determination of the MSN. The SSAD program requires a file with reference data as well as further inputs to control the calculation. Reference data must be available as a column in a csv file .csv files can be created with any editor. Furthermore, most software products for data storage or evaluation allow the export of data in this format. The first line of the file must contain unique variable names, even if there is only one variable. It is recommended to use variable names with a maximum of 32 characters, letters, numbers, "." and "_" only. A distinction is made between upper and lower case. Since the csv format allows different separators between fields and also does not specify the decimal character, the character for separating columns as well as the decimal character must be given to the SSAD program, as described below. Entries for controlling the invoice are made directly in *SSAD_V9.R* at correspondingly commented places. These are:

- A description of the current analysis to identify the results (free text),
- the path and file name of the file containing the reference data,
- the character used to separate fields in the file (typically: semicolon),
- the decimal point used (typically: period or comma),
- the accuracy requirement ($\Delta_{\text{rel, goal}}$ in the previous description). The accuracy requirement is directed at the size of the 95% confidence interval with which the arithmetic mean of future samples is to be determined. The (relative) accuracy is calculated as a percentage according to Eqs. (1) and (2). Alternatively, it is possible to specify the accuracy as an absolute value, not as a relative percentage of the mean value. Further control options, which were not required for this application, are described in the program. The execution of *SSAD_V9.R* creates a table containing the input parameters and the calculated MSN. This table appears on the R console and is also saved as a text file in the *txt/*directory under the specified evaluation name. In the *Fig/*directory, images with corresponding names are stored which document the course of the simulation and calculation.

Calculation and method verification

To verify the SSAD method, the MSN was calculated using the data of the Moos Survey 2015 with the concentrations of 12 heavy metals (Al, As, Cd, Cr, Cu, Fe, Hg, Ni, Pb, Sb, V, Zn) and nitrogen [3] measured at 400 moss sampling sites in Germany. The calculation was carried out throughout Germany using the SSAD method (method C)[1] and comparing the calculation formula of the Moss Manual (method A) with identical data, i.e.. with data from the Moss Survey 2015. On the other hand, element-specific MSN for different spatial categories (Federal Republic of Germany, federal states, ecological spatial classes) were calculated with the data from the Moss Survey 2015 using the SSAD method and compared with the results of the measurement network planning on the basis of the data from the Moss Survey 2005 (Tables 1 and 2). In each case, a uniform error factor (tol) of 0.2 (= 20%) was used and a significance level

[1] R version 3.4.1 and SSAD version 9 were used to calculate the MSN [10].

Table 1 Element-specific minimum sample numbers (MSN) and actual sample sizes (*n*) for ecoregions of the Ecological Land Classification (ELCE40) [10] in Germany, calculated using Moss Survey 2005 data [3, 9]

ELCE 40	As			Cd			Cr			Cu			Fe			Hg		
	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M
B_1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
B_2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
C_0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
D_13	97	7	B	43	7	A	101	7	B	7	*7*	B	32	7	B	6	*7*	B
D_14	–	1	–	–	1	–	–	1	–	–	1	–	–	1	–	–	1	–
F1_1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
F1_2	33	30	B	10	*30*	A	61	30	A	6	*30*	B	13	*30*	A	8	*30*	B
F2_6	80	42	A	25	*42*	B	444	42	A	14	*42*	A	107	42	A	18	*42*	A
F3_1	113	75	A	39	*75*	A	152	75	A	7	*75*	B	35	*75*	A	13	*75*	A
F3_2	25	*93*	A	29	*93*	A	115	93	A	11	*93*	A	30	*93*	A	15	*93*	B
F4_1	20	19	B	13	*19*	B	82	19	A	15	*19*	B	29	19	B	19	*19*	A
F4_2	101	73	A	36	*72*	A	61	*73*	A	31	*73*	A	36	*73*	A	17	*73*	B
G1_0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
S_0	10	4	B	29	4	B	106	4	A	12	4	B	1	*4*	A	14	4	B
U_1	70	27	A	33	27	A	73	27	A	18	*27*	B	82	27	A	21	*27*	A
U_2	–	1	–	–	1	–	–	1	–	–	1	–	–	1	–	–	1	–
Other	14	10	A	13	10	A	140	10	A	12	10	B	32	10	B	13	10	A
Sum	563	382		270	381		1335	382		133	382		397	382		144	382	
Count		1			6			1			8			5			8	
%		6			36			6			47			29			47	

ELCE 40	N			Ni			Pb			Sb			V			Zn		
	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M
B_1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
B_2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
C_0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
D_13	2	*7*	B	9	7	B	44	7	B	7	*7*	A	10	7	A	4	*7*	A
D_14	–	1	–	–	1	–	–	1	–	–	1	–	–	1	–	–	1	–
F1_1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
F1_2	5	*30*	A	11	*30*	B	22	*30*	B	77	30	A	9	*30*	A	10	*30*	A
F2_6	8	*42*	B	34	*42*	A	34	*42*	A	22	*42*	B	59	42	A	16	*42*	A
F3_1	8	*75*	B	25	*74*	A	41	*75*	A	19	*75*	B	22	*75*	B	18	*75*	A
F3_2	8	*93*	B	23	*93*	A	45	*93*	A	31	*93*	B	18	*93*	B	20	*93*	B
F4_1	17	*19*	A	18	*19*	B	22	19	B	16	*19*	B	16	*19*	B	10	*19*	A
F4_2	10	*73*	B	45	*73*	A	72	*73*	A	24	*72*	B	26	*73*	A	35	*72*	A
G1_0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
S_0	9	4	A	24	4	A	78	4	A	11	4	A	9	4	B	26	4	B
U_1	11	*27*	B	37	27	B	45	27	A	42	27	B	48	27	B	103	27	A
U_2	–	1	–	–	1	–	–	1	–	–	1	–	–	1	–	–	1	–
Other	16	10	A	29	10	B	10	*10*	B	9	*10*	A	14	10	B	31	10	B
Sum	94	382		255	381		413	382		258	381		231	382		273	381	
Count		8			6			6			7			5			7	
%		47			36			36			41			29			41	

Italic numbers = MSN met or exceeded; Count = number of ELCE40 classes where MSN met or exceeded; % = percentage of ELCE40 classes where MSN met or exceeded; M = method; A = method A according to ICP Vegetation [1]; B = method B according to Wosniok ([11], cited in [8])

Table 2 Element-specific minimum sample numbers (MSN) and actual sample sizes (*n*) b calculated for the German federal states using Moss Survey 2005 data [3, 9]

Federal state	As			Cd			Cr			Cu			Fe			Hg		
	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M
BW	42	26	B	37	26	B	*26*	*26*	A	8	*26*	A	37	26	A	16	*26*	B
BY	37	*50*	A	16	*50*	A	93	50	A	8	*50*	B	18	*50*	A	14	*50*	B
BB	18	*26*	A	4	*26*	A	62	26	A	6	*26*	B	16	*26*	B	5	*26*	B
HH	5	3	A	3	*3*	A	9	3	B	9	3	A	1	*3*	A	2	*3*	A
HE	28	*30*	A	9	*30*	A	77	30	A	10	*30*	B	41	30	A	13	*30*	A
MV	15	*22*	B	11	*22*	B	60	22	A	10	*22*	B	19	*22*	A	17	*22*	A
NI	16	*39*	B	26	*39*	B	137	39	A	6	*39*	B	17	*39*	A	9	*39*	A
NW	25	*52*	B	25	*51*	A	35	52	B	14	*52*	B	23	*52*	B	9	*52*	A
RP	26	21	A	36	21	A	30	21	A	10	*21*	B	12	*21*	A	4	*21*	A
SL	29	7	A	11	7	B	128	7	B	5	*7*	B	74	7	A	6	*7*	A
SN	65	34	A	20	*34*	A	432	34	A	7	*34*	A	75	34	A	13	*34*	A
ST	113	30	A	82	30	A	60	30	A	54	30	A	69	30	A	17	*30*	A
SH	38	20	B	28	20	B	87	20	B	10	*20*	A	33	20	A	12	*20*	B
TH	95	22	A	22	*22*	B	47	22	A	6	*22*	B	62	22	A	13	*22*	A
Sum	552	382		330	381		1283	382		163	382		497	382		150	382	
Count		6			9			1			9			7			14	
%		42			64			7			64			50			100	
Federal state	**N**			**Ni**			**Pb**			**Sb**			**V**			**Zn**		
	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M
BW	7	*26*	B	17	*26*	B	45	26	B	21	*26*	A	32	26	B	19	*26*	B
BY	6	*50*	A	21	*49*	A	40	*50*	A	17	*50*	B	19	*50*	B	11	*50*	A
BB	6	*26*	B	11	*26*	B	15	*26*	B	15	*26*	B	8	*26*	B	10	*26*	B
HH	1	*3*	A	2	*3*	B	5	3	A	11	3	B	3	*3*	B	12	3	A
HE	8	*30*	B	50	30	A	44	30	A	24	*30*	B	*27*	*30*	A	13	*30*	B
MV	15	*22*	A	13	*22*	B	14	*22*	B	12	*22*	A	10	*22*	B	8	*22*	B
NI	5	*39*	A	17	*39*	B	28	*39*	A	14	*39*	A	14	*39*	B	18	*39*	B
NW	6	*52*	A	18	*52*	B	42	*52*	A	20	*51*	B	14	*52*	B	15	*51*	A
RP	3	*21*	A	19	*21*	A	32	21	A	15	*21*	B	12	*21*	B	7	*21*	A
SL	3	*7*	A	14	7	A	15	7	A	17	7	B	27	7	B	4	*7*	A
SN	7	*34*	B	27	*34*	A	36	34	B	86	34	A	44	34	A	18	*34*	B
ST	8	*30*	B	23	*30*	A	131	30	A	23	*30*	A	47	30	A	34	30	A
SH	7	*20*	A	25	20	B	29	20	B	22	20	B	23	20	B	113	20	A
TH	6	*22*	A	32	22	B	28	22	B	8	*22*	A	31	22	A	10	*22*	A
Sum	88	382		289	381		504	382		305	381		311	382		292	381	
Count		14			10			5			10			8			11	
%		100			71			36			71			57			79	

Italic numbers = MSN met or exceeded; Count = number of federal states where MSN met or exceeded; % = percentage of federal states where MSN met or exceeded; M = method; A = method A according to ICP Vegetation [1]; B = method B according to Wosniok ([11], cited in [8])

of $\alpha = 0.05$ was selected to ensure comparability with earlier surveys [1]. For validation the SSAD procedure was applied to all subsamples regardless of their distribution form. Only in the case of sample sizes ≤ 25 (method C needs sample sizes ≥ 25), SSAD was supplemented by the formula of the manual of the European Moss Survey [1] (method A) in the case of normally distributed variables or the extension proposed by Wosniok ([11], cited in [8]) in the case of differently distributed variables (method B). As in Nickel and Schröder [3] and Schröder et al. [9] for the four elements of the Convention on Long-Range Transboundary Air Pollution (CLRTAP)—i.e., Cd, Hg, Pb and N—the results of the MSN calculations were cartographically illustrated for comparisons of the spatial

Table 3 Element-specific minimum sample numbers (MSN) and actual sample sizes (*n*) for 12 elements in Germany, calculated using two methods applied to data of the Moss Survey 2015

	As	Cd	Cr	Cu	Fe	Hg	N	Ni	Pb	Sb	V	Zn
Method A	92	58	45	15	56	27	8	59	65	31	48	13
Method C	110	117	79	55	18	70	10	73	75	36	56	17
n	*400*	*398*	*398*	*399*	*400*	*400*	*400*	*400*	*400*	*397*	*400*	*400*

Italic numbers = MSN met or exceeded; method A = manual formula according to ICP Vegetation [1]; method C = SSAD method (Sample Size for Arbitrary Distributions)

distributions of deviances from the minimum sample numbers.

Results

In the nationwide data set of Moss Survey 2015, the substance concentrations with the exception of Zn (log-normally distributed) are all neither normally nor lognormally distributed ($\alpha < 0.05$). The MSN calculated on this basis for Germany on an element-specific basis are within a range of 8–92 for the Manual formula (method A) and between 10 and 117 for the SSAD method (method C) (Table 3). The results of the SSAD procedure are thus on average 38% higher than those of the Manual formula. In extreme cases, the MSN of the SSAD process are 267% above the MSN of the Manual formula for Cu and −68% below the MSN of the Manual formula for Fe. The lower limit, at which the empirically determined mean value for all elements does not differ more than 20% from the true mean value with a 95% certainty, results from the maximum MSN and is 92 (As) for the Manual formula and 117 (Cd) for the SSAD procedure.

Of the total of 312 ELCE and state-specific partial data sets (12 heavy metals and nitrogen) investigated in the Moss Survey 2015 monitoring network, 35% follow the normal distribution and 37% the log normal distribution, the remainder was distributed differently. Using the SSAD method, for the 144 ELCE-specific partial data sets, the proportion of ecological area classes in Germany where the minimum number of samples was met is between 12% (Al) and 53% (N), depending on the element (average 27%) (Table 4). With regard to the 168 country-specific data sets, the proportions vary between 21% (Al) and 100% (Cu, N) and average 58% (Table 5). In order to fully guarantee the MSN for all 13 elements, the German moss monitoring network would have to be expanded to 906 sites with regard to the ELCE and to 701 sites for the federal states of Germany. When methods A and B are applied, the proportions of ELCE classes in Germany where the minimum number of samples is reached are between 6% (Al, Cr) and 47% (Hg, N) and the average amounts to 33% (Table 1). In the federal states of Germany, the percentages vary between 7% (Cr) and 100% (Hg, N) and average 62% (Table 2). On the basis of this MSN calculation, the German moss monitoring network would have to comprise 1335 sites in relation to the ELCE40 [10] and 1283 sites in relation to the federal states in order to fully guarantee the MSN.

For 72 of the 144 ELCE40-specific partial data sets, the sample sizes are above $n = 25$, for the state-specific data sets 96 out of 168. When comparing the case numbers calculated with the two method combinations (only data sets with $n > 25$), the deviations of the MSN calculated with the methods A, B, and C with data for the year 2015 (Table 4) from those calculated with the methods A and B for the year 2005 (Table 1) for the ELCE40-related analysis range from −321 to 176 (mean value: 4.8; standard deviation: 58.8). For the federal states (only data sets with $n \geq 25$), a comparison of the MSN for 2015 (Table 5) with those for 2005 (Table 2) reveals deviations between −402 and 149 (mean value: 4.9; standard deviation: 56.6). The maximum deviations with significantly lower MSN estimates when using the SSAD method (method C) on the basis of the data for 2015 are shown in Cr for the ecoregion F2_6 (−321) and in Saxony (−402).

Figure 1 shows the spatial distribution of MSN statistics as calculated with SSAD for different spatial categories (ELCE40, federal states) using the Moss Survey data 2015 for the four elements of CLRTAP. Compared with the MSN calculations in the same measuring network using the values measured in 2005 and without the SSAD method (Figs. 2, 3, 4, 5, 6, 7 and 8), the spatial proportions of the spatial units matching the MSN are correspondingly smaller. The area shares are on average 44% lower for Cd, 23% lower for Hg, 18% lower for Pb and 14% lower for N in the calculations with the extended methodology (Fig. 1) than for the MSN calculated with the method combination A and B (Figs. 2, 3, 4, 5, 6, 7 and 8; Table 6).

Discussion and conclusions

The SSAD methodology should be regarded as a tool for spatially designing monitoring networks. A comparison of the MSN calculations using different reference data (here: 2005, 2015) clearly shows that the MSN estimates can only be transferred to future measured value variants to a limited extent depending from available

Table 4 Element-specific minimum sample numbers (MSN) and actual sample sizes (*n*) for ecoregions of the Ecological Land Classification (ELCE40) [10] using Moss Survey 2015 data

ELCE 40	As			Cd			Cr			Cu			Fe			Hg		
	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M
B_1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
B_2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
C_0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
D_13	23	6	B	8	6	A	18	6	B	13	6	A	8	6	A	12	6	A
D_14	–	1	–	–	1	–	–	1	–	–	1	–	–	1	–	–	1	–
F1_1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
F1_2	120	38	C	33	*38*	C	88	38	C	50	37	C	10	*38*	C	28	*38*	C
F2_6	181	39	C	152	39	C	36	*38*	C	123	39	C	16	*39*	C	153	39	C
F3_1	81	76	C	90	*76*	C	133	*76*	C	29	*76*	C	15	*76*	C	42	*76*	C
F3_2	75	*98*	C	201	97	C	37	*98*	C	37	*98*	C	17	*98*	C	31	*98*	C
F4_1	39	18	A	20	18	A	12	*18*	B	38	18	A	18	*18*	A	27	18	B
F4_2	102	79	C	78	*78*	C	42	*78*	C	55	*79*	C	25	79	C	88	79	C
G1_0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
S_0	31	4	A	78	4	B	21	4	A	16	4	A	7	4	A	49	4	A
U_1	25	*28*	C	27	*28*	C	33	28	C	36	28	C	18	*28*	C	34	28	C
U_2	–	1	–	–	1	–	–	1	–	–	1	–	–	1	–	–	1	–
Other	24	12	A	23	12	A	32	12	A	14	12	A	12	*12*	A	13	12	A
Sum	701	400		710	398		452	398		411	399		146	400		477	400	
Count		2			3			4			3			7			3	
%		12			18			24			18			41			18	

ELCE 40	N			Ni			Pb			Sb			V			Zn		
	MSN	*n*	M	MSN	*n*	v	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M
B_1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
B_2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
C_0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
D_13	21	6	A	4	6	A	8	6	A	14	6	A	66	6	B	18	6	B
D_14	–	1	–	–	1	–	–	1	–	–	1	–	–	1	–	–	1	–
F1_1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
F1_2	21	*38*	C	10	*38*	C	183	38	C	33	*38*	C	41	38	C	29	*38*	C
F2_6	33	*38*	C	10	*39*	C	65	39	C	55	39	C	46	39	C	138	39	C
F3_1	23	*76*	C	10	*76*	C	34	*76*	C	76	76	C	33	*75*	C	38	*76*	C
F3_2	19	*97*	C	10	*98*	C	43	*98*	C	95	98	C	32	*97*	C	42	*98*	C
F4_1	21	18	B	10	*18*	A	42	18	B	68	18	B	81	18	B	36	18	A
F4_2	21	*78*	C	11	*79*	C	75	*79*	C	72	*79*	C	28	78	C	55	*79*	C
G1_0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
S_0	64	4	A	3	*4*	A	12	4	A	44	4	A	7	4	A	45	4	A
U_1	62	28	C	12	*28*	C	40	28	C	18	*28*	C	19	*28*	C	22	*28*	C
U_2	–	1	–	–	1	–	–	1	–	–	1	–	–	1	–	–	1	–
Other	67	12	B	8	*12*	A	16	12	A	125	12	B	25	12	A	20	12	A
Sum	352	397		88	400		518	400		600	400		378	397		443	400	
Count		5			9			3			3			3			5	
%		29			53			18			18			18			29	

Italic numbers = MSN met or exceeded; Count = number of ELCE40 classes where MSN met or exceeded; % = percentage of ELCE40 classes where MSN met or exceeded; M = method; A = method A according to ICP Vegetation [1]; B = method B according to Wosniok ([11], cited in [8]); C = method C = SSAD method (Sample Size for Arbitrary Distributions)

Table 5 Element-specific minimum sample numbers (MSN) and actual sample sizes (*n*) calculated for Germany's federal states using Moss Survey 2015 data

Federal state	As			Cd			Cr			Cu			Fe			Hg		
	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M
BW	149	30	C	156	30	C	45	29	C	80	30	C	17	*30*	C	186	30	C
BY	87	60	C	50	*60*	C	81	59	C	56	*60*	C	10	*60*	C	68	60	C
BB	11	*28*	C	11	*28*	C	10	*28*	C	19	*28*	C	10	*28*	C	10	*28*	C
HH	33	3	A	12	3	A	3	*3*	A	5	3	A	2	*3*	A	4	3	A
HE	81	28	C	74	28	C	23	*28*	C	34	28	C	20	*28*	C	31	28	C
MV	48	24	B	23	*24*	A	5	*24*	A	43	24	B	9	*24*	A	28	24	B
NI	50	*57*	C	40	*57*	C	35	*57*	C	35	*57*	C	13	*57*	C	38	*57*	C
NW	49	48	C	24	*46*	C	31	*48*	C	43	*47*	C	13	*48*	C	21	*48*	C
RP	163	19	B	86	19	B	25	19	B	15	*19*	B	6	*19*	A	29	19	B
SL	95	6	B	228	6	B	53	6	A	18	6	A	3	*6*	A	37	6	B
SN	41	29	C	93	29	C	165	29	C	30	29	C	10	*29*	C	38	*29*	C
ST	33	29	C	36	29	C	51	29	C	21	*29*	C	21	*29*	C	21	*29*	C
SH	23	20	B	50	20	B	34	20	B	33	20	B	19	*20*	B	22	20	B
TH	14	*19*	A	23	19	A	21	19	A	10	*19*	A	2	*19*	A	9	*19*	A
Sum	877	400		906	398		582	398		442	399		155	400		542	400	
Count		3			5			6			7			14			5	
%		21			36			43			50			100			36	
Federal state	**N**			**Ni**			**Pb**			**Sb**			**V**			**Zn**		
	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M	MSN	*n*	M
BW	23	*29*	C	10	*30*	C	37	30	C	156	*30*	C	20	*30*	C	117	30	C
BY	26	*60*	C	10	*60*	C	96	60	C	58	*60*	C	31	*58*	C	52	*60*	C
BB	10	*28*	C	10	*28*	C	14	*28*	C	12	*28*	C	49	28	C	17	*28*	C
HH	8	3	B	1	*3*	A	1	*3*	A	12	3	A	3	*3*	A	9	3	A
HE	13	*28*	C	10	*28*	C	85	28	C	50	28	C	23	*27*	C	36	28	C
MV	23	*24*	A	6	*24*	A	36	24	B	10	*24*	A	81	24	B	51	24	B
NI	80	57	C	10	*57*	C	17	*57*	C	69	*57*	C	28	*57*	C	35	*57*	C
NW	19	*47*	C	10	*48*	C	26	*48*	C	43	*48*	C	*26*	*48*	C	25	*48*	C
RP	13	*18*	B	2	*19*	A	21	*19*	B	38	19	B	11	*19*	A	31	19	B
SL	5	*6*	B	3	*6*	B	21	6	A	39	6	B	3	*6*	A	25	6	B
SN	29	*29*	C	10	*29*	C	42	29	C	63	29	C	46	*29*	C	35	*29*	C
ST	20	*29*	C	10	*29*	C	22	*29*	C	70	29	C	27	*29*	C	24	*29*	C
SH	25	20	B	7	*20*	B	43	20	B	63	20	B	52	20	B	16	*20*	A
TH	16	*19*	B	5	*19*	A	23	19	A	88	19	B	7	*19*	A	12	*19*	A
Sum	310	397		103	400		484	400		771	400		407	397		485	400	
Count		11			14			6			6			10			7	
%		79			100			43			43			71			50	

Italic numbers = MSN met or exceeded; Count = number of federal states where MSN met or exceeded; % = percentage of federal states where MSN met or exceeded; M = method; A = method A according to ICP Vegetation [1]; B = method B according to Wosniok ([11], cited in [8]); C = method C = SSAD method (Sample Size for Arbitrary Distributions)

data. Uncertainties arise not only from different statistical distributions of the reference data, but also from the MSN methods chosen due to these differences. But even using the same data, the results of the three procedures sometimes differ greatly from each other. In addition, the SSAD method as a stochastic method produces different results even with identical data, although the extent of these differences can be reduced by tightening the accuracy requirement in the application of the method. The entire package of methods is particularly suitable for locating conspicuous non-compliances with minimum sample numbers for different spatial categories

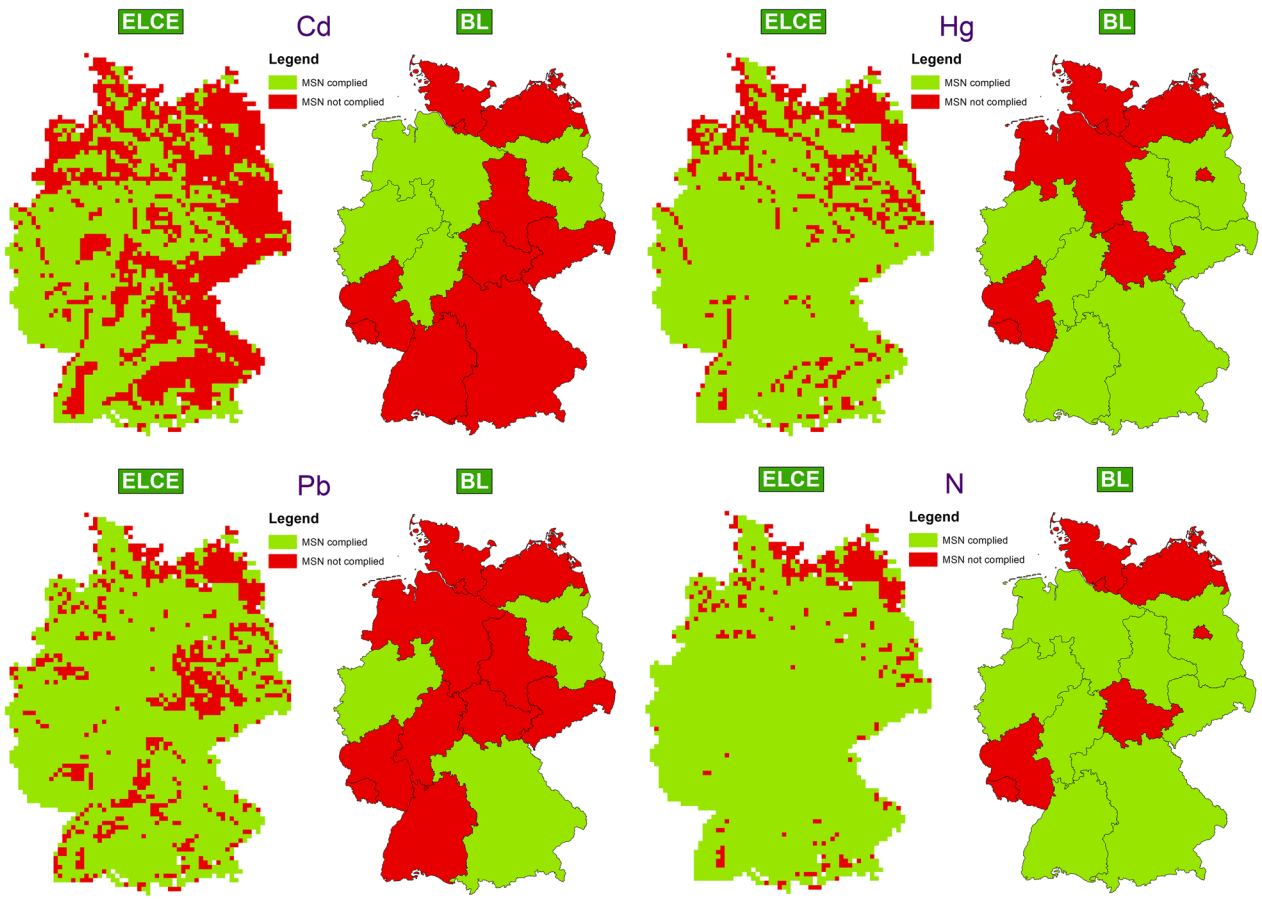

Fig. 1 Comparisons of the sample sizes in the 2015 Moss Survey network with the minimum sample numbers determined for Cd, Hg, Pb and N (reference planes): ELCE40 [10], BL (= federal states)

(e.g., administrative units, ecological spatial classes) and for correcting them in measurement network planning within the individual participating states and also across states. For a more precise quantification of the differences between the three partial methods, these would have to be applied to reference data from the same survey (2015) in addition to the data from the 2005 and 2015 moss surveys used here. In addition, mean deviations between different simulation runs with identical data should be quantified.

Due to strong deviations of the three partial methods even with identical data, the SSAD procedure cannot be recommended as the only procedure. For the application of the methodology it is recommended to still estimate minimum sample numbers from normally distributed reference data using the calculation formula of the moss manual [2; 8] (method A). For lognormally distributed reference data, the MSN formula according to Wosniok ([11], quoted in [8]) is recommended (method B). Both methods A and B are based on parametric statistics, which—assuming that the assumptions about the statistical distribution of data are correct—generally allow more accurate and precise estimates than the SSAD method (method C). The MSN formula according to Wosniok ([11], cited in [8]) is also recommended for all non-normally distributed data series with $n < 25$, since the lognormal distribution represents the more frequent case compared to the normal distribution and the SSAD procedure requires sample sizes with $n \geq 25$. For data with $n \geq 25$ that is neither normally nor lognormally distributed, the use of the SSAD procedure is recommended, since this does not impose any preconditions on the distribution of the data.

Conclusions

The SSAD procedure developed for the European Moss Survey closes a gap in the previous methodology for calculating MSN with regard to compliance with certain error tolerances in the calculation of the arithmetic mean of, e.g., element concentrations in mosses measures across Germany. Thus, for the first time a method package is available which does not impose any preconditions

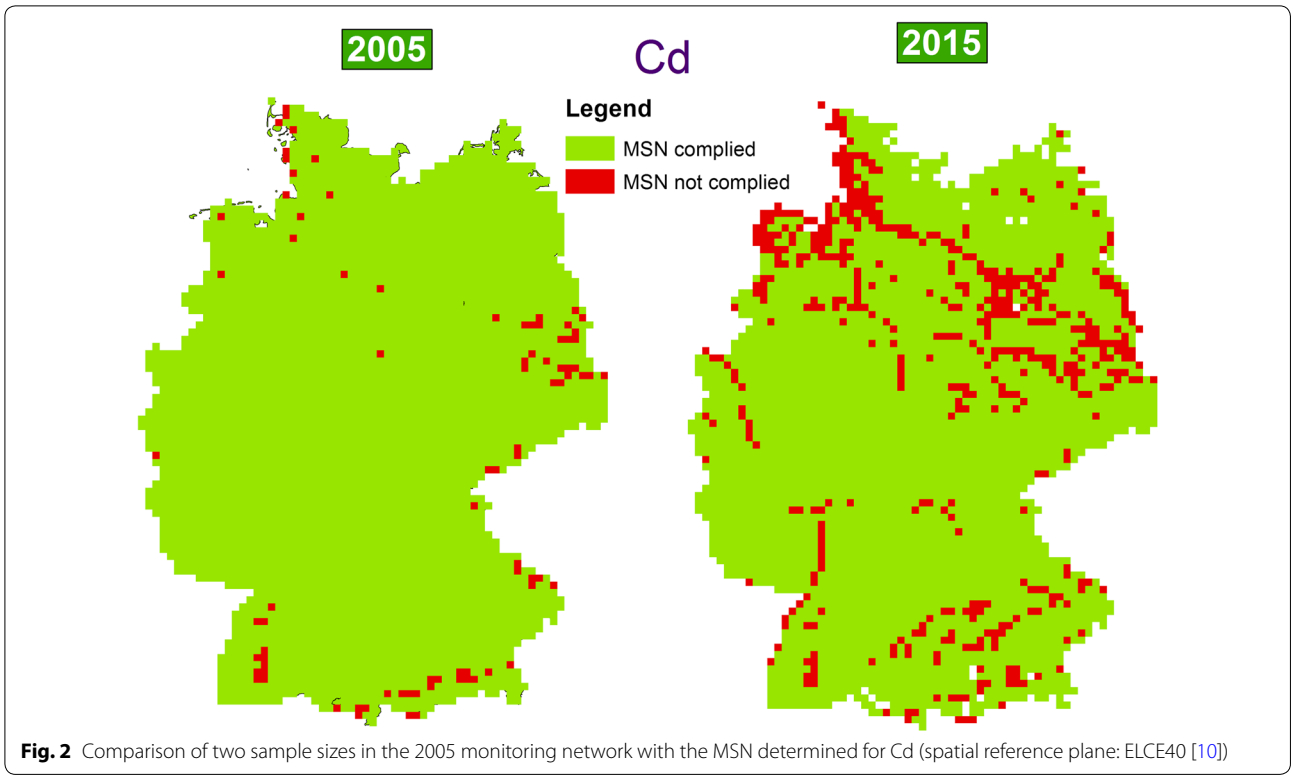

Fig. 2 Comparison of two sample sizes in the 2005 monitoring network with the MSN determined for Cd (spatial reference plane: ELCE40 [10])

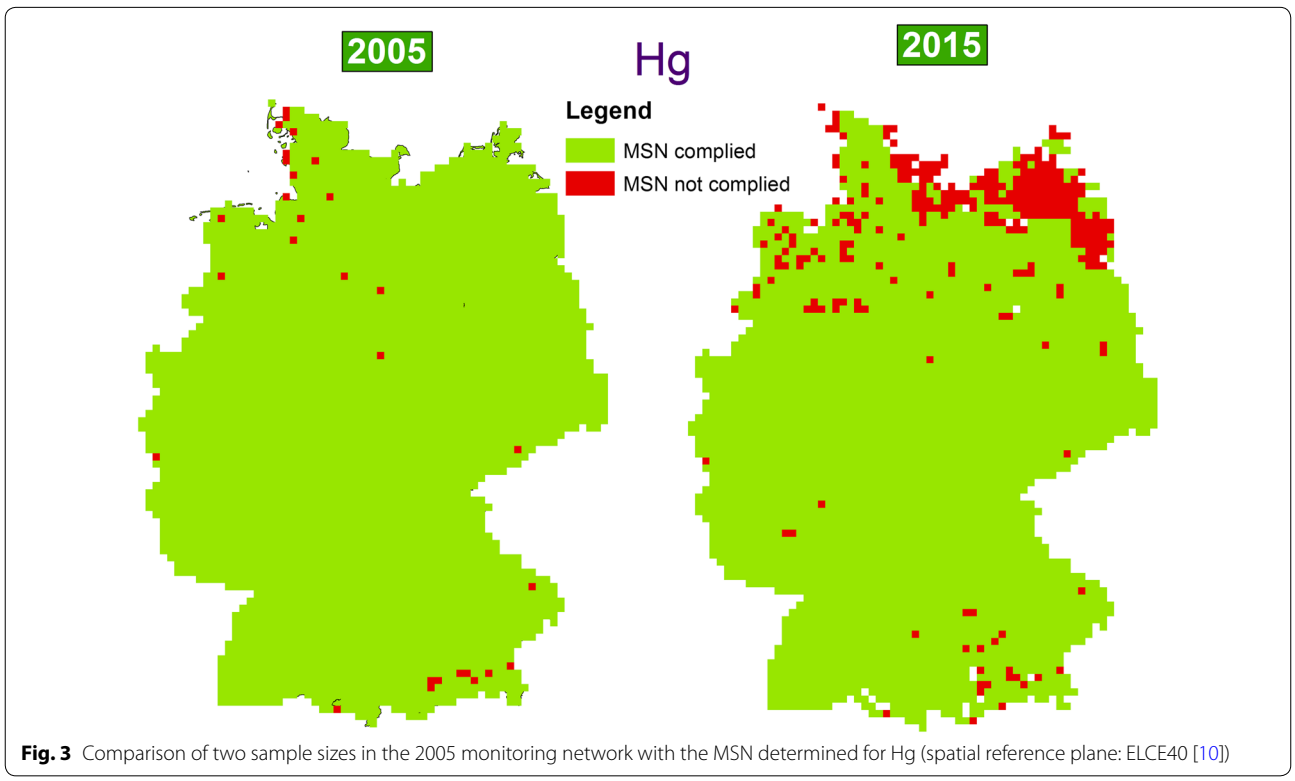

Fig. 3 Comparison of two sample sizes in the 2005 monitoring network with the MSN determined for Hg (spatial reference plane: ELCE40 [10])

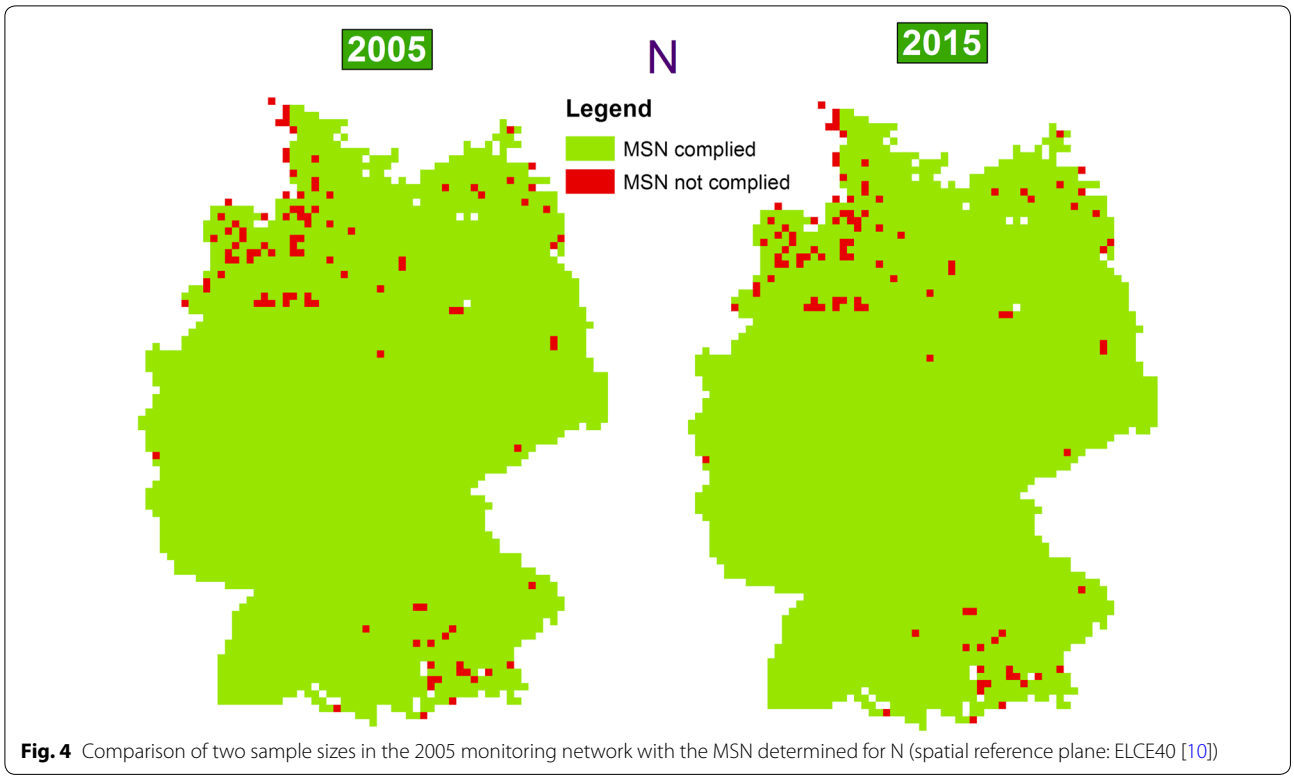

Fig. 4 Comparison of two sample sizes in the 2005 monitoring network with the MSN determined for N (spatial reference plane: ELCE40 [10])

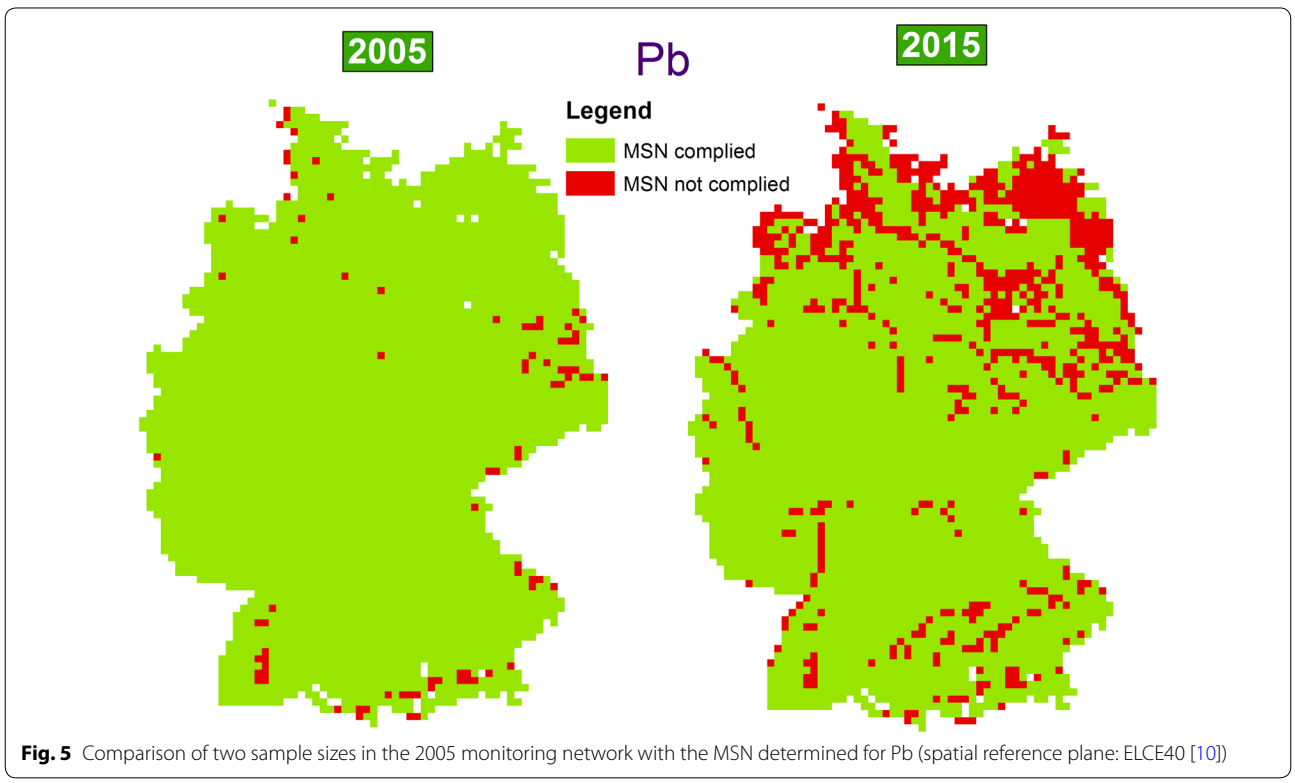

Fig. 5 Comparison of two sample sizes in the 2005 monitoring network with the MSN determined for Pb (spatial reference plane: ELCE40 [10])

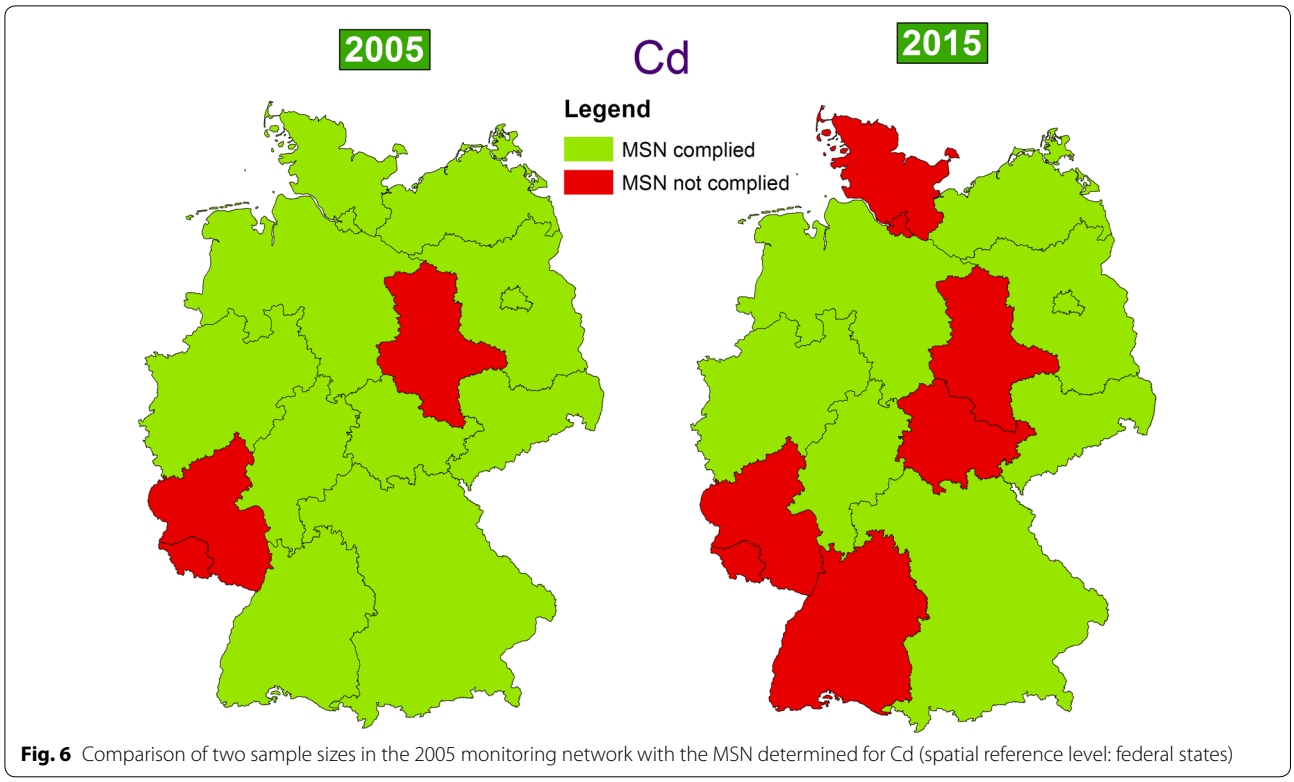

Fig. 6 Comparison of two sample sizes in the 2005 monitoring network with the MSN determined for Cd (spatial reference level: federal states)

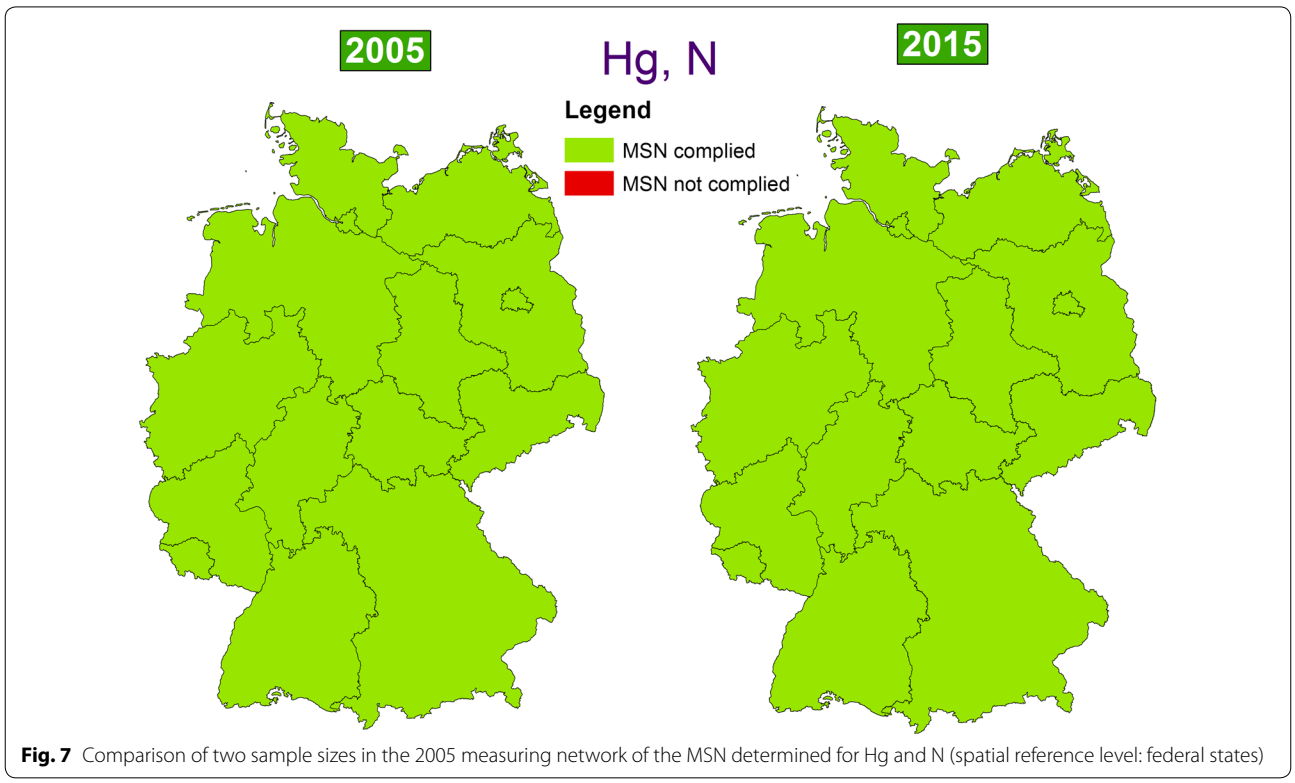

Fig. 7 Comparison of two sample sizes in the 2005 measuring network of the MSN determined for Hg and N (spatial reference level: federal states)

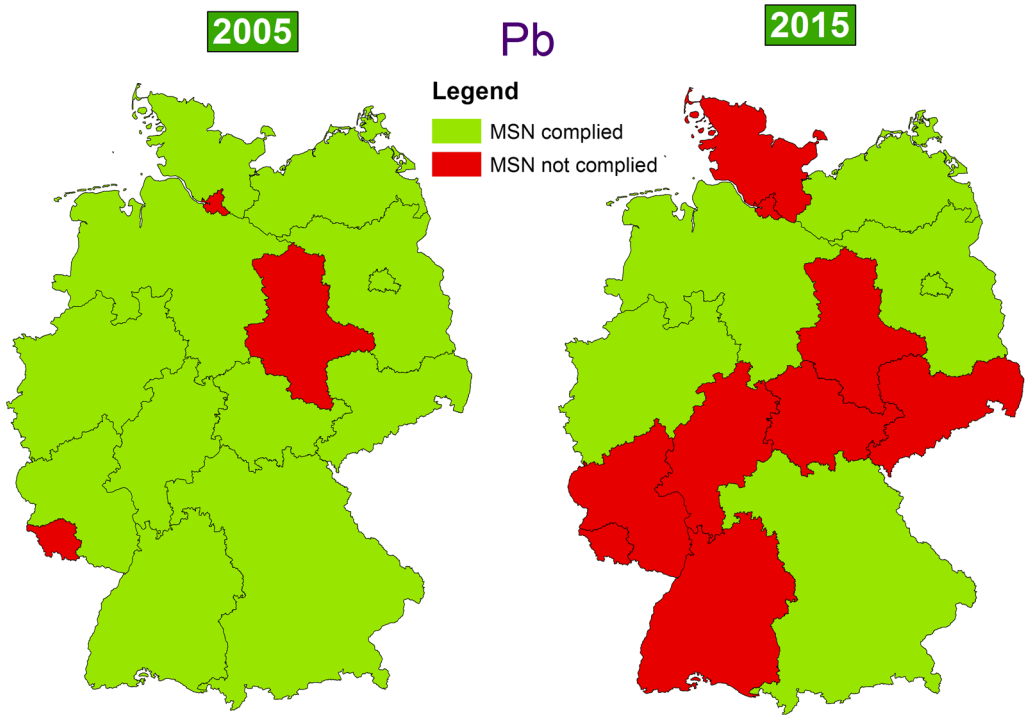

Fig. 8 Comparison of two sample sizes in the 2005 monitoring network with the MSN determined for Pb (spatial reference level: federal states)

Table 6 Area percentages of the ecological area classes (ELCE40, [10]) and federal states (BL) matching the MSN calculated based on the Moss Survey network 2015 ($n = 400$, data collected in 2005 (previous methodology) and 2015 (extended methodology)

	Cd		Hg		Pb		N	
	ELCE (%)	BL (%)	ELCE (%)	BL (%)	ELCE (%)	BL (%)	ELCE (%)	BL (%)
Moss Survey 2005	86	69	92	100	81	58	97	100
Moss Survey 2015	50	37	81	65	79	38	90	78

on the distribution of the data for the MSN calculation. The procedure is directly transferable for the planning of many other environmental monitoring networks.

Abbreviations

Al: aluminium; As: arsenic; B_1: western and northern Scandinavia, northwest Russia; B_2: The Alps, Iceland, northwest Russia; C_0: The Alps, Iceland, western and northern Scandinavia, Kola Peninsula, northwest Russia, Caucasus; BB: Brandenburg; BE: Berlin; BL: Federal state of Germany; BW: Baden-Wuerttemberg; BY: Bavaria; Cd: cadmium; Cr: chromium; Cu: copper; CSV: comma separated value; D_13: The Alps, dispersed small areas in eastern and southeast Europe; D_14: Baltic States, Belarus, western Russia; ELCE: ecological land classes of Europe; F1_1: Poland, northwest Ukraine; F1_2: Ireland, Great Britain, western and central Europe; F2_6: Central Europe, eastern and southeast Europe; F3_1: Germany, northwest Poland, Czech Republic, northern Austria, Slovenia, the Balkans; F3_2: Western Europe (including northern Spain, France, Benelux countries, western Germany), Denmark; F4_1: Southeast Great Britain, southeast Denmark, northeast Germany, northwest Poland; F4_2: Western/central and southern Europe (including southern Great Britain, eastern France, southern Belgium, Luxembourg, the Alps, Italy), eastern and southeast Europe (including the Carpathian Mountains, the Balkans); Fe: iron; G1_0: Italy, southeast Europe; HE: Hesse; Hg: mercury; HH: Hamburg; ICP: International Cooperative Programme; M: method; MSN: minimum sample numbers; MV: Mecklenburg Western-Pomerania; N: nitrogen; *n*: sample size; NI: Lower Saxony; Ni: nickel; NW: North Rhine-Westphalia; Pb: lead; RP: Rhineland Palatinate; S_0: Northern parts of Europe (including parts of Iceland, Ireland, Great Britain, Scandinavia, northwest Russia, the Baltic states and Belarus); Sb: antimony; SSAD: Sample Size for Arbitrary Distributions; SH: Schleswig-Holstein; SL: Saarland; SN: Saxony; ST: Saxony-Anhalt; TH: Thuringia; U_1: dispersed small areas within a stripe reaching form Ireland via central Europe and the Byelorussian–Ukrainian borderline to Russia; U_2: dispersed small areas in southern Europe reaching form the Iberian Peninsula via southeast Europe including, e.g., the Balkans, the Carpathians, Greece and northern Turkey to southwest Russia; V: vanadium; Zn: zinc.

Acknowledgements

We would like to thank the German Environment Agency (Dessau-Roßlau, Germany) for financial support and professional advice.

Authors' contributions
Werner Wosniok developed the methodology and wrote the R script. Winfried Schröder drafted the article and headed the computations executed by Stefan Nickel. All authors read and approved the final manuscript.

Funding
German Environment Agency, Dessau-Roßlau, Germany (Grant no. 3715 63 212 0).

Availability of data and materials
The software tool Sample Size for Arbitrary Distributions (SSAD) developed has been made freely accessible for future applications via the research data repository ZENODO® [12].

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Institute of Statistics, University of Bremen, POB 330 440, 28334 Bremen, Germany. [2] Chair of Landscape Ecology, University of Vechta, POB 1553, 49364 Vechta, Germany.

Received: 19 September 2019 Accepted: 13 January 2020
Published online: 29 January 2020

References

1. ICP Vegetation (International Cooperative Programme on Effects of Air Pollution on Natural Vegetation and Crops) (2014) Monitoring of atmospheric deposition of heavy metals, nitrogen and POPs in Europe using bryophytes. Monitoring manual 2015 survey. United Nations Economic Commission for Europe Convention on Long-Range Transboundary Air Pollution. ICP Vegetation Moss Survey Coordination Centre, Dubna, Russian Federation, and Programme Coordination Centre. Bangor, Wales, UK. https://icpvegetation.ceh.ac.uk/sites/default/files/MossmonitoringMANUAL-2015-17.07.14.pdf. Accessed 04 Apr 2019
2. Land CE (1971) Confidence intervals for linear functions of the normal mean and variance. Ann Math Stat 42:1187–1205
3. Nickel S, Schröder W (2017) Umstrukturierung des deutschen Moos-Monitoring-Messnetzes für eine regionalisierende Abschätzung atmosphärischer Deposition in terrestrische Ökosysteme. In: Schröder W, Fränzle O, Müller F (Hg) Handbuch der Umweltwissenschaften. Grundlagen und Anwendungen der Ökosystemforschung. 24. Erg.Lfg., Kap. VI-1.8:1–48
4. Nickel S, Schröder W (2018) Schwermetall- und Stickstoffkonzentrationen in Moosen deutscher Waldgebiete zwischen 1990 und 2015 – Ein Bund-Länder-Vergleich. Gefahrstoffe - Reinhaltung der Luft, vol 3. Springer, VDI, Berlin, pp 1–14
5. Olsson U (2005) Confidence intervals for the mean of a log-normal distribution. J Stat Educ 13(1). http://www.amstat.org/publications/jse/v13n1/olsson.html. Accessed 04 Apr 2019
6. R Core Team (2019) R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/. Accessed 04 Apr 2019
7. Rubinstein RY, Kroese DP (2017) Simulation and the Monte Carlo method, 3rd edn. Wiley, New York
8. Schröder W, Nickel S, Schönrock S, Meyer M, Wosniok W, Harmens H, Frontasyeva MV, Alber R, Aleksiayenak J, Barandovski L, Danielsson H, de Temmermann L, Fernández Escribano A, Godzik B, Jeran Z, Pihl Karlsson G, Lazo P, Leblond S, Lindroos A-J, Liiv S, Magnússon SH, Mankovska B, Martínez-Abaigar J, Piispanen J, Poikolainen J, Popescu IV, Qarri F, Santamaria JM, Skudnik M, Špirić Z, Stafilov T, Steinnes E, Stihi C, Thöni L, Uggerud HT, Zechmeister HG (2016) Spatially valid data of atmospheric deposition of heavy metals and nitrogen derived by moss surveys for pollution risk assessments of ecosystems. Environ Sci Pollut Res 23:10457–10476
9. Schröder W, Nickel S, Völksen B, Dreyer A, Wosniok W (2019) Nutzung von Bioindikationsmethoden zur Bestimmung und Regionalisierung von Schadstoffeinträgen für eine Abschätzung des atmosphärischen Beitrags zu aktuellen Belastungen von Ökosystemen. Ressortforschungsplan des Bundesministeriums für Umwelt, Naturschutz, Bau und nukleare Sicherheit. Forschungskennzahl 3715 63 212 0. Dessau: 1–188 Bericht, 1–288 Anhänge
10. Schröder W, Schmidt G, Hornsmann I (2006) Landschaftsökologische Raumgliederung Deutschlands. In: Handbuch der Umweltwissenschaften. Grundlagen und Anwendungen der Ökosystemforschung. Landsberg am Lech, München, Zürich, Kap V-1.9, ErgLfg 17:-100, 2006
11. Wosniok W (2015) Fallzahlen für das Moosmonitoring - Ergänzungsvorschläge für das Monitoring manual 2015 survey (ICP Vegetation 2014). Arbeitspapier vom 04.09.2015, Universität Bremen, Bremen
12. Wosniok W, Nickel S, Schröder W (2019) R Software tool for calculating Minimum Sample Sizes for Arbitrary Distributions (SSAD), link to scientific software (Version v1). Zenodo. https://doi.org/10.5281/zenodo.2583010

Publisher's Note