Considerations for Zero-Value Data Recorded in Student Surveys

Teruo ARASE, Hiroaki SHIRASAWA, Hajime KOBAYASHI,
Wataru KINOSHITA, Yukio NOMIZO and Toshinobu SAKAI

Education and Research Center of Alpine Field Science,
Faculty of Agriculture, Shinshu University

学生実習における調査票の 0 値について

荒瀬輝夫，大塚　大，小林　元，木下　渉，野溝幸雄，酒井敏信

信州大学農学部附属アルプス圏フィールド科学教育研究センター

要旨： 生態学的な植物調査では，植物の枯死や消失が生じうるため，0 値や欠測値は珍しくない。2019 年に行われた信州大学農学部の学生実習での立木調査において，多くの 0 値が記録された。これらの 0 値について，本研究では，調査票の表現を精査して生育状態を推定した。さらに，不明確な 0 値の生じた原因について，仕事量（調査本数）と調査の複雑さまたは単調さ（立木の生育状態の順序パターン）に注目して分析を行った。立木調査では計 243 本のうち 82 本が枯死または消失しており，さらにそのうち 13 本についての調査票の記録が不明確な 0 値であった。こうした 0 値は 7 つの学生グループのうち 3 グループで記録され，様々な表現が用いられていた。調査本数，立木の生育状態の順序パターン（連の数と長さ，i 番目と i+1 番目の生育状態の関連性）と，不明確な 0 値との相関はほとんど認められなかった。よって，不明確な 0 値の発生には立木とは関わりのない要因が影響したと推測された。その要因としてグループ内の学生間のコミュニケーション不足が挙げられ，グループ分けの方法の改善や，0 値のときの明確な表現についての指示が必要と考えられた。

キーワード：学生実習，立木調査，0 値
Key words: Practical training for students, Tree survey, Zero-value data

Introduction

Missing data potentially cause an imbalance in randomized block design studies and make many statistical analyses and tests invalid (Scheiner and Gurevitch, 2001). Zero values also produce mathematical problems in compositional data such as petrochemical compositions and faunal compositions, since logarithm and geometric mean cannot be taken on zero values (Arai and Ohta, 2006). In particular, mortality and loss in field animals and plants are common enough to make missing and zero values not so rare in field surveys.

To date, many of the methods suggested for

replacing missing values range from simple methods, such as using marginal means, to complicated methods, such as multiple regression analysis involving the other variables. Deletion of data sets with missing data is also used in order to avoid bias caused by estimation based on incomplete data sets in the case that random sampling can be postulated (Arai and Ohta, 2006; Abe, 2016). These methods are often utilized by researchers and engineers for statistical analysis of data.

In contrast, replacement methods for zero values are not generally known. Zero values can be classified as being 'rounded zero' for values that are below the detection limit or as being 'essential zero' (Aichison, 1986; Martín-Fernández et al., 2003). In compositional data, the replacement of zero value data is necessarily accompanied by changes in the rest of the data set due to constraints of the constant sum. For rounded zero values, several replacement methods have been suggested, which can be classified as additive replacement and multiplicative replacement methods (Martín-Fernández et al., 2003; Arai and Ohta, 2006). For essential zero values, distinguishing the data as either having zero value or not having zero value, or merging the parameter having zero values with another parameter of a similar category have been suggested (Martín-Fernández et al., 2003). Further, missing data would be confused with being zero values, but these categories are intrinsically different from each other (Arai and Ohta, 2006).

It is inevitable that problems arising due to missing and zero values are barely realized by students before they learn and apply statistics. In the application of using a tree survey for the practical training of students, methods for dividing the survey site into zones of equal areas and allotting them to student groups results in differences in the number and size of trees among zones, but complete data can be collected from every tree (Arase et al., 2017a,b). In contrast, following the method of a MECE (mutually exclusively and collectively exhaustively) search for trees based on each student group's own selections considerably reduces the differences in number and size of trees among groups, but enhances the probability of missing data (Arase et al., 2018a,b).

Many zero-value data which need further treatment in data sets were obtained in a tree survey conducted by our students in 2019. The survey was conducted in a young broad-leaved forest three years after planting, where 13% of the trees were dead but remaining at the site and 21% of the trees had been lost; in other words, nearly one-third of the trees at this site need to be planted again. The dead trees would be regarded as 'rounded zero' and the lost trees would be regarded as 'essential zero', but students recorded these states with various confusing expressions. Although both states are the same from the viewpoint of supplementary planting in forest management, they are intrinsically different from each other in respect to assessing growth states.

In the present study, we examined and evaluated zero-value data obtained by students in the 2019 tree survey. As causative factors of unclear zero-value data, we assessed the tiredness or boredom of students during the survey; workload (the number of surveyed trees); and the complexity or monotony of the work (the sequences of growth states encountered in the survey).

Method

Tree survey

The survey site was located in Terasawayama Research Forest at Shinshu University (Ina City, Nagano Prefecture, central Japan). Amur cork (*Phellodendron amurense*) trees had been planted at this site in 2017, in 18 rows (from A to R) with a density of 0.30 trees per square meter in an area of 800 m^2. At this site, tree markers (thin pole with a height of about 0.7 m) remained close to the point where each tree had been planted. The slope direction was SW, at an elevation of 1,170 to 1,195 m above sea level (Arase et al., 2020).

A tree survey was conducted for practical student training as part of the program 'Training for field science of agriculture and forestry' at the Faculty of Agriculture of Shinshu University on June 5, 2019. Twenty-eight students, almost all beginners at

fieldwork, participated in the survey. Five members of the educational staff (two teachers, two engineers and an assistant senior student) instructed the students in attendance.

After arriving at the survey site, the students (15 male, 13 female) observed Amur cork trees and learned the purpose and significance of the survey. Then, the students were divided into seven groups (groups 1 to 7) each having four people with two male and two female students in groups 1, 2, 3, 4, 5, and 7 and three male and one female students in group 6.

Tree height and growth state were surveyed. For tree height, the height of the living woody part (i.e., the position of highest leaf axil or buds) was measured. To assess growth state, five stages (a combination of 'presence or absence of leaves', 'position of leaves' and 'presence or absence of root tension') were defined as follows.

I: Highest leaves come out on the top,

II: Highest leaves come out on a lateral branch,

III: Leaves are limited on the base of the tree,

IV: No living leaves are found but root tension remains,

V: No living leaves are found and no root tension remains (i.e., the tree is 'dead').

To determine growth states IV and V, root tension was checked by pulling on the tree by hand for a second.

The students were instructed to record tree height as zero for growth states VI and V. In addition, if no tree was found where there should be a tree (i.e., the tree has been 'lost'), the students were instructed to check neither the columns of growth state nor to record tree height (the columns of growth stage and tree height remain blank) but to record comments in the remarks column about the condition of the spot.

One group was assigned to each of the first seven rows of afforested young trees on the lowest part of the slope. Then, the remaining rows were surveyed by the MECE search. In this method, after completing the survey of the first assigned row, each group was instructed to seek out and survey the next adjacent unsurveyed row.

Estimation of unclear zero-value data

After the survey was completed, the number of trees in each row (A to R) and growth states (I to V and lost trees) were compiled for each student group (1 to 7). The number of trees in each height range (in 10-cm intervals) was also compiled but is omitted in this paper.

The zero-value data for tree height recorded as '0' with a check mark in the growth state column is reliable. Zero-value data with no record for tree height was allowed if the tree had been lost, but the expressions written by students in remarks column in the survey form were so confusing that the state of the tree was unclear; thus, the expressions in remarks column were examined carefully to estimate the growth condition of zero-value data.

Factors coursing unclear zero-value data

As in the cause of many unclear zero-value data, we suspected tiredness or boredom during the survey. Two perspectives, load and variation of work, were considered. As for the workload, the number of foliaged trees (I, II and III for which students have to measure tree height) and defoliated trees (IV, V and lost trees for which students do not have to measure) were counted. As for the complexity or monotony of work, the sequence of growth states was analyzed: the number and the maximum length of runs in the survey by each student group were examined, and the relation between tree state and that of the next tree was analyzed.

The number and the maximum length of runs were as follows: at first, the data of growth state was divided into two categories, foliaged trees and defoliated trees as mentioned above. Then, the number of runs and the maximum run length for each state were determined; for example, if the sequential order of the growth state was 'FFDDDFDFFFFFF' (F and D mean foliaged and defoliated, respectively), the number of runs is 5 and the maximum length is 6 for foliaged trees and 3 for defoliated trees.

The relationship between the state of a tree and that of the next tree was expressed by the φ-coefficient given by a χ^2-value. A foliaged tree was recorded as 1 and a defoliated tree was recorded as 0

Table 1 Number of trees classified by growth states in each afforestation row

Growth state	Student group 1			2			3		4			5		6		7			Total
Row	A	K	O	E	H	Q	G	N	B	I	R	D	M	C	J	F	L	P	
I	3	0	1	2	0	0	4	0	4	0	0	5	0	5	0	3	0	0	27
II	1	0	0	2	1	2	2	0	2	2	2	0	0	2	1	1	1	1	20
III	2	0	0	4	0	0	5	0	4	1	0	3	2	2	1	1	1	1	27
IV	6	2	2	10	4	3	7	6	5	4	2	8	3	8	4	8	3	2	87
V (Dead)	0	0	1	5	2	0	7	1	1	0	0	0	0	10	1	4	0	0	32
Lost	10	4	2	0	0	0	3	0	8	1	1	7	2	0	0	9	2	1	50
Total	22	6	6	23	7	5	28	7	24	8	5	23	7	27	7	26	7	5	
			34			35		35			37		30		34			38	243

"I' to 'V' indicate the growth state of individual trees (described in text).

Table 2 Recorded entries of zero-value data and the presumed growth states in the present study

Student group	Expression	Presumed growth state	Reason for uncertainty in the record	Number of trees	
				Certain	Uncertain
1	No tree	'Lost'		16	
2	Dead	'Dead'	Students could not judge 'Dead' without the tree. (There is a possibility of this being 'Lost' by misinterpretation.)		7
3	(Omitted)	'Lost'	The students could not correctly interpret the meaning of blank columns. (There is a possibility that the tree was alive but omitted.)		3
4	Nothing	'Lost'		10	
5	Absent, Dead	'Lost'	Based on the expression of 'Absent'. (Adding the word 'Dead' is inconsistent.)		2
	Absent	'Lost'		4	
	None	'Lost'		2	
	(Blank)	'Lost'	Not necessary to record anything in this case. (There is a possibility that the tree was alive but omitted.)		1
7	Nothing	'Lost'		12	
			Total	44	13

in order to simply encode the sequential order of the tree survey. Then, the i-th and i+1-th values were compared, and each tree was classified into one of four categories, comprised of (1,1), (1,0), (0,1), (0,0), in a two-way table by i-th and i+1-th values. If we assign the number of trees in (1,1), (1,0), (0,1), and (0,0) categories as a, b, c, and d, we can take the total as N (= a+b+c+d), and calculate

$\chi^2 = N (|ad-bc|-N/2)^2 / [(a+b) (b+d) (a+c) (b+d)]$

where the value of ad-bc gives the sign of φ.

The influence of the number of trees and the pattern of their arrangement on unclear zero-value data were analyzed using the correlation coefficient by the least square method. Except for φ coefficients, data were counting data; each counting data was transformed into a square root before conducting analysis to approximate a normal distribution.

Results

Estimation of zero-value data

At the survey site, a total of 243 Amur cork trees were surveyed, including 161 living trees and 82 dead or lost trees. Each student group measured 30 to 38 trees during their survey (Table 1). In addition, each group spent approximately 1 h, and all groups completed the survey almost simultaneously.

Table 2 shows the student records of zero-value data and estimated growth conditions in the present study. Six out of seven student groups (except group 6) recorded zero-value data (57 trees in total) as required. Five of the student groups used only one expression for zero-value data, but group 5 used as many as four different expressions. After examining

Table 3 Sequence of growth states of trees in each afforestation row

Row		Lower part of slope ←																				→ Upper part of slope					
SE	A	IV	IV	II	L	L	I	L	L	IV	L	L	L	III	III	L	IV	IV	IV	L	I	L	I				
↑	B	L	IV	L	III	III	I	V	IV	III	L	II	I	L	L	L	L	I	IV	II	L	IV	I	IV	III			
	C	IV	IV	IV	IV	V	V	I	IV	V	IV	I	IV	V	IV	V	I	I	V	V	V	III	I	V	II	II	V	III
	D	IV	III	IV	III	[L]	I	[L]	I	L	L	IV	IV	L	L	IV	[L]	I	III	I	I	IV	IV	IV				
	E	IV	[V]	IV	IV	IV	[V]	[V]	III	II	I	IV	I	II	III	[V]	IV	IV	IV	III	[V]	IV	III	IV				
	F	L	L	IV	IV	IV	IV	L	V	L	L	I	IV	I	L	III	II	IV	IV	V	IV	L	L	V	V	I	L
	G	IV	IV	[L]	[L]	[L]	II	V	V	V	IV	V	V	IV	V	II	III	I	I	III	I	IV	IV	III	III	V	I
	H	[V]	IV	IV	IV	IV	[V]	II																			
	I	III	II	II	IV	IB	IV	L	IV																		
	J	III	V	II	IV	IV	IV	IV																			
	K	L	L	IV	L	L	IV																				
	L	L	L	IV	III	II	IV	IV																			
	M	IV	III	L	IV	IV	L	III																			
	N	IV	IV	IV	IV	IV	IV	V																			
	O	IV	IV	L	I	V	L																				
	P	IV	II	IV	L	III																					
↓	Q	IV	IV	II	IV	II																					
NW	R	IV	IV	L	II	II																					

'I' to 'V' indicate the growth state of trees (described in text), and 'L' indicates a lost tree.
Entries enclosed in boxes are unclear zero-value data.

Table 4 Number of trees surveyed and complexity of sequential order of growth states and their relationships to unclear zero-value data

Item	Student group							Correlation coefficient with unclear zero-value data	
	1	2	3	4	5	6	7	Number of trees	Number of expressions
Number of surveyed trees									
Foliaged	7	11	11	15	10	10	10	0.10	0.00
Defoliated	11	24	21	12	11	24	16	0.36	-0.63
Lost	16	0	3	10	9	0	12	-0.47	0.53
Complexity of sequential order of growth states									
Number of runs	13	14	11	16	15	16	14	-0.36	-0.15
Maximum length of run of foliaged trees	2	3	6	4	4	2	3	0.47	0.46
Maximum length of run of defoliated trees	9	7	8	8	8	6	10	-0.28	0.48
φ-coefficient between i-th and i+1-th states	-0.09	0.11	0.33	0.13	-0.02	0.00	0.08	0.38	-0.03
Unclear zero-value data									
Number of trees	0	7	3	0	3	0	0	1	0.44
Number of expressions	1	1	1	1	4	0	1		1

Correlation coefficient significant at $p < 0.05$ ($n = 7$, F-test) is 0.755.

the expressions, unclear zero-value data were detected in three groups (group 2, 3 and 5). Seven trees were judged as ‘dead’ and six trees as ‘lost’ out of 13 trees with unclear zero-value data (Table 2).

Factors coursing unclear zero-value data

Causative factors for unclear zero-value data

Table 3 shows the sequence of growth states of trees in each afforestation row. Table 4 shows the number of surveyed trees and the complexity of the sequence order of growth state, and their relationships to unclear zero-value data.

The numbers of living trees, dead trees and lost trees ranged from 17 to 28, from 0 to 11 and from 0 to 16, respectively, among the student groups (Table 4). The number of runs, maximum length of runs of foliaged trees, and defoliated trees ranged from 11 to 16, 2 to 6 and 6 to 10, respectively, among student

groups. The φ-coefficient ranged roughly between -0.10 and +0.10 among the six student groups, but was larger for one student group (group 3, φ= 0.33; p= 0.0558, χ^2-test) (Table 4). Although the number of trees and the complexity of the sequence of growth states seemed to vary among student groups, those cannot be designed experimentally (i.e., not caused by the student groups) due to the random occurrence of dead trees.

Most of the correlation coefficients between number of surveyed trees and complexity of the sequence order of growth state and unclear zero-value data were not significant, and they were roughly within the range of -0.50 and +0.50 (Table 4). Only the number of dead trees slightly correlated to the number of unclear zero-value data expressions (r= -0.63, p= 0.129, F-test) (Table 4).

Discussion

Compared to our previous studies (Arase et al. 2017ab, 2018ab), there were many unclear zero-value data were detected in this student survey. The occurrence of many unclear zero-value data itself is due to the fact that the mortality and loss of trees were frequent among the afforested Amur cork trees in the survey site.

However, our data do not show any relationship between workload (the number of surveyed trees) and occurrence of unclear zero-value data (Table 4). A significant negative correlation was detected only between the number of dead trees (i.e., growth state IV) and the number of unclear expressions of zero-value data; that is, students might improve their skills for examining the growth state through the repeated practice of pulling defoliated trees.

Further, our results also do not indicate complexity or monotony of work (sequence of growth states) as being related to the occurrence of unclear zero-value data (Table 4). These results imply that factors other than the trees being surveyed cause the occurrence of unclear zero-value data. We suspect lack of communication among students to be one of the factors.

Lack of communication might be a problem of personality or morale among the students. In group 5, in particular, as many as four different expressions were used, and two out of the four expressions were treated as unclear zero-value data. The students could have easily noticed that the novel expressions were different from previously used expressions when they made subsequent entries, but they disregarded the discrepancy. Although some members of the group might have noticed a change in terminology, this information was not shared or discussed within the group.

The composition of members of groups is reported to influence the results of cooperative work. Groups formed by consensus rather than by lottery tend to be more active and efficient (Matsumoto et al., 2008). The grouping of students was not made by consensus at the start of the present survey, which may have decreased the communication and morale during the survey. To divide the students into groups formed by consensus or to encourage communication within each group, additional educational staff may be needed. Further, it should be most effective for educational staff to give explicit directions for how to express zero-value data to the students.

Conclusions

In the present study, we examined and estimated unclear zero-value data detected in a tree survey conducted by students in 2019. As the causative factors of unclear zero-value data, we considered workload (number of trees surveyed) and the complexity or monotony of the work (sequence of growth states encountered in the survey). The results were as follows:

1. In the survey, 161 living trees and 82 dead or lost trees were examined, and the latter group included 13 unclear zero-value data. Three student groups out of seven groups recorded zero-value data but used various expressions.
2. The workload (the number of trees surveyed) and the complexity or monotony of the work (sequence order of growth state) showed no significant relationship with unclear zero-value data.
3. These results imply that the factors other than the

state of the trees to be surveyed are likely cause the occurrence of unclear zero-value data; we suspect lack of communication among students as a factor. Improvement of methods for dividing students into groups and providing explicit directions for expressing zero-value data are needed.

References

Abe, T. (2016) Statistical Analysis of missing data. Asakura Publishing, Tokyo. 190 pp. (in Japanese)

Aichison, J. ed (1986) The statistical analysis of compositional data. Chapman & Hall, London. 416 pp.

Arai, H. and Ohta, T. (2006) Note on zero and missing values in compositional data. Journal of the Geological Society of Japan, 112 (7): 439-451 (in Japanese with English summary)

Arase, T., Otsuka, D., H., Kobayashi, H., Kinoshita, W., Nomizo, Y. and Sakai, T. (2020) Survey of an afforested area with broad-leaved trees conducted as practical training for students: the case of an Amur cork forest established in 2017. Bulletin Shinshu University Alpine Field Center, 18: in press (in Japanese with English summary)

Arase, T., Shirasawa, H., Kobayashi, H., Kinoshita, W., Nomizo, Y. and Sakai, T. (2017a) Surveys and thinning of animal-damaged trees in a Shinshu University research forest as practical training for students. Bulletin Shinshu University Alpine Field Center, 15: 61-65 (in Japanese with English summary)

Arase, T., Shirasawa, H., Kobayashi, H., Kinoshita, W., Nomizo, Y. and Sakai, T. (2017b) Issues in student surveys of animal-damaged trees in a research forest. The Annals of Environmental Science Shinshu University, 39: 68-73

Arase, T., Shirasawa, H., Kobayashi, H., Kinoshita, W., Nomizo, Y. and Sakai, T. (2018a) Survey of a permanent experimental stand in a research forest of Shinshu University conducted as practical training for students: the case of a Japanese yew forest established in 1976. Bulletin Shinshu University Alpine Field Center, 16: 55-60 (in Japanese with English summary)

Arase, T., Shirasawa, H., Kobayashi, H., Kinoshita, W., Nomizo, Y. and Sakai, T. (2018b) Issues in student surveys of a permanent experimental stand in a research forest. The Annals of Environmental Science Shinshu University, 40: 57-63

Martín-Fernández, J., Balceló-Vidai, C. and Paulowski-Glahn, V. (2003) Dealing with zeros and missing values in compositional data sets using nonparametric imputation. Mathematical Geology, 35 (3): 253-278

Scheiner, S.M. and Gurevitch, J. eds. (2001) Design and analysis of ecological experiments. Oxford University Press, New York. 415 pp.

（原稿受付　2020. 3. 19）