Colloquium Biometricum 39
2009, 137–145

MOSAIC PLOTS HELP VISUALIZE CONTINGENCY TABLES. EXAMPLE FOR A QUESTIONNAIRE SURVEY ON KNOWLEDGE OF AND ATTITUDE TOWARDS GMO

Agnieszka Wnuk[1,2], Marcin Kozak[2*], Małgorzata Rochalska[1]

[1]Department of Physiology, [2]Department of Experimental Design and Bioinformatics
Warsaw University of Life Sciences, Nowoursynowska 159, 02-776 Warsaw, Poland
* Corresponding author: nyggus@gmail.com

Summary

Mosaic plots can help visualize contingency tables, even those complex ones, consisting of many categorical variables. This kind of display can be very helpful in understanding simple and complex associations among categorical variables, especially for three- and more way tables. That said, applications of this display are all-too-rare. This paper aims to direct the readers' attention to this useful graphical display. To present mosaic plots, they are applied to visualize associations among questions from a survey of knowledge of and attitude towards genetically modified organisms.

Key words and phrases: categorical data, contingency tables, genetically modified organisms, mosaic display, visualization.

Classification AMS 2000: 62-07, 62-09, 62P10

1. Introduction

Visualization can powerfully support statistical analysis. In Cleveland's (1994) words, "Data display is critical to data analysis. Graphs allow us to explore data to see overall patterns and to see detailed behavior; no other ap-

proach can compete in revealing the structure of data so thoroughly." Indeed, in certain situations graphical data analysis may be much more powerful than regular statistical analysis. This can be especially the case when some hidden inconsistency in data occurs or for large data sets, when it is difficult to grasp the whole data set with the standard output from any statistical method.

Mosaic plots, proposed by Hartigan and Kleiner (1981, 1984) and then further developed (e.g., Friendly 1994, 1999), may help visualize contingency tables. If simple (as is with 2×2 tables), contingency tables might (though do not have to) be easy to understand just by exploring numbers, but this does not have to be the case with more complex tables–especially when the variables are associated. In such instances mosaic plots (or a fourfold display for 2×2 tables with an additional stratum variable) can help grasp the associations between categorical variables.

Such great usefulness of the mosaics to visualize categorical data notwithstanding, they are used very infrequently. In biological literature it is quite difficult to find them. Three interesting examples are those from Laffont et al. (2007), Noser and Byrne (2007) and Love and Yoklavich (2008). Laffont et al. (2007) applied a mosaic plot to visualize the two-way partitioning of total sums of squares into genotype and genotype-by-environment components in studying genotype-by-environment interaction. Noser and Byrne (2007) applied extended mosaic display (in which the information from the log-linear modeling is incorporated by colored shading) to show observed frequencies of path segments leading to fruit (FR), seed (SE) and miscellaneous (MI) food sources during outward and inward journeys of wild chacma baboons, *Papio ursinus*. Love and Yoklavich (2008) used quite complex mosaic and extended mosaic displays to show habitat factors of juvenile cowcod, *Sebastes levis*.

This paper aims to point the readers' attention to a mosaic display. We show its application to a questionnaire study on the knowledge of and attitude to genetically modified organisms (GMO), conducted at two faculties of Warsaw University of Life Sciences.

2. Material and Methods

The survey

The questionnaire survey was conducted in 2007 among the students of the third academic year of undergraduate studies of the Faculty of Agriculture and Biology and the Faculty of Human Nutrition and Consumer Sciences of the Warsaw University of Life Sciences. Aimed to discover the knowledge and attitude towards GMO among the students, the questionnaire consisted of 41 questions, of which we have chosen four to the present work. These questions along with the possible answers are presented in Table 1. In the questionnaire the term "GMO products" was used in order to facilitate understanding the questions by the respondents. Correctly we should have used two terms: "a genetically modified product" to mean the product that is totally modified genetically, and "a product that is stamped with the term 'genetically modified'" to mean the product with some components being genetically modified.

The study was planned as a total enumeration of the population of these students, but some non-responds appeared (owing to absence at classes during which the survey was conducted). In both cases, 117 students of both faculties were surveyed; there were 115 questionnaires with answers to all the four questions considered for the Faculty of Agriculture and Biology, and 110 for the Faculty of Human Nutrition and Consumer Sciences.

Graphing

We did not consider these samples random, which is why we did not apply probability-based statistical methods to analyze the data. The mosaics (Hartigan and Kleiner 1981, 1984, Friendly 1999) for three-way contingency tables were applied to visualize the associations between the questions 1a and 1b, and between 2a and 2b. The plots were drawn with R's mosaicplot() function (R Development Core Team 2008).

Friendly (1994, 1999) described how the mosaics are constructed and interpreted. In summary, this is done in the context of conditional probabilities. Take for example two-way tables. For such a table we have cell frequencies n_{ij} and cell probabilities $p_{ij} = n_{ij}/n$ (n standing for the sum of counts from the whole table). A unit square, which represents all the counts n, is vertically divided into rectangles with widths proportional to the observed marginal frequencies $n_{i\cdot}$;

of course, at the same time they are proportional to the marginal probabilities $p_i = n_{i\cdot}/n$. Then these rectangles are subdivided horizontally proportionally to the conditional probabilities of the second variable given the first, $p_{j|i} = n_{ij}/n_{i\cdot}$. Thanks to such construction of the mosaics, the area of each tile is proportional to the observed probability p_{ij}. This is how the mosaic display should be understood and interpreted—the bigger the tile, the more counts occurred for the corresponding combination of variables. Clearly, interpretation of the mosaic is straightforward. This can be easily generalized for multi-way tables.

Table 1. Questions taken to the present study along with possible answers to them

Question/Answers
1a. Do you believe that the available information on GMO is true? Yes/No/Don't know
1b. What is the knowledge on GMO in the society? Big/Small/None
2a. Are in your opinion GMO products well stamped? Yes/No/Don't know the stamp/Didn't notice/Not interested
2b. Are in your opinion GMO products available in Poland's market? Yes/No/Don't know/Didn't notice

3. Results

Contingency tables for both pairs of questions are given in Tables 2 and 3. Studying such tables in order to find possible associations and understand them as well as to compare the two faculties is not easy. Instead, interpretation can be done with the help of mosaic plots presented in Figures 1 and 2.

Note that no one replied that the knowledge of the society on GMO is big (Table 2, Figure 1). Clearly there is no association between the answers to questions 1a and 1b from the Faculty of Agriculture and Biology. The same can be seen for the Faculty of Human Nutrition and Consumer Sciences, although among those who think that the available information on GMO is true more respondents decided that the knowledge is small than among those who do not believe in the validity of the available information. Much more respondents from the Faculty of Human Nutrition and Consumer Sciences answered that they do not know if the information on GMO is true or not than from the Faculty of Agriculture and Biology.

Table 2. Contingency table for questions 1a and 1b

What is the knowledge on GMO in the society?	Do you believe that the available information on GMO is true? Yes	No	Don't know
	Faculty of Agriculture and Biology		
Big	0	0	0
Small	27	29	36
None	6	7	10
	Faculty of Human Nutrition and Consumer Sciences		
Big	0	0	0
Small	17	8	52
None	5	7	21

Table 3. Contingency table for questions 2a and 2b

Are in your opinion GMO products available in Poland's market?	Are in your opinion GMO products well stamped? Yes	No	Don't know the stamp	Didn't notice	Not interested
	Faculty of Agriculture and Biology				
Yes	0	25	6	7	2
No	1	10	7	6	0
Don't know	0	4	4	9	0
Didn't notice	2	5	13	13	1
	Faculty of Human Nutrition and Consumer Sciences				
Yes	5	25	14	5	0
No	3	4	3	3	0
Don't know	3	3	6	2	0
Didn't notice	0	11	10	12	1

Only few respondents from both faculties claimed (rather incorrectly) that GMO products are well stamped (Table 3, Figure 2), although their fraction was slightly bigger for the Faculty of Human Nutrition and Consumer Sciences. In both cases many respondents decided that GMO products are not well stamped, the fraction of whom being similar for both faculties. Interestingly, 10 respondents from the Faculty of Agriculture and Biology and 4 from the Faculty of Human Nutrition and Consumer Sciences claimed that the products containing GMO are not well stamped even though they responded that no GMO products are available in Poland's market, which is quite an inconsistency. Quite a significant fraction of respondents replied that they were not interested in whether the GMO products are stamped or not.

4. Discussion and Conclusions

That mosaics facilitate reading contingency tables is easy to prove—it suffices to compare data from Tables 2 and 3 with the corresponding plots from Figures 1 and 2. Careful examination of tables may provide some information

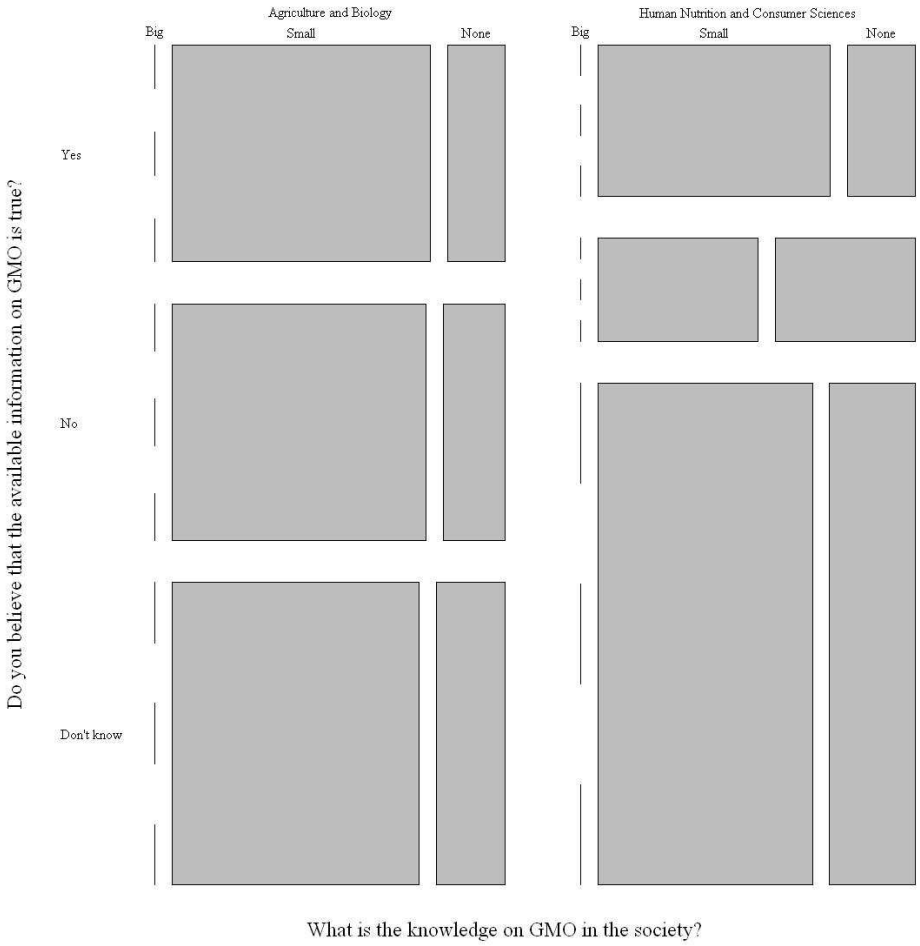

Figure 1. Mosaic plot visualizing the contingency table presented in Table 2

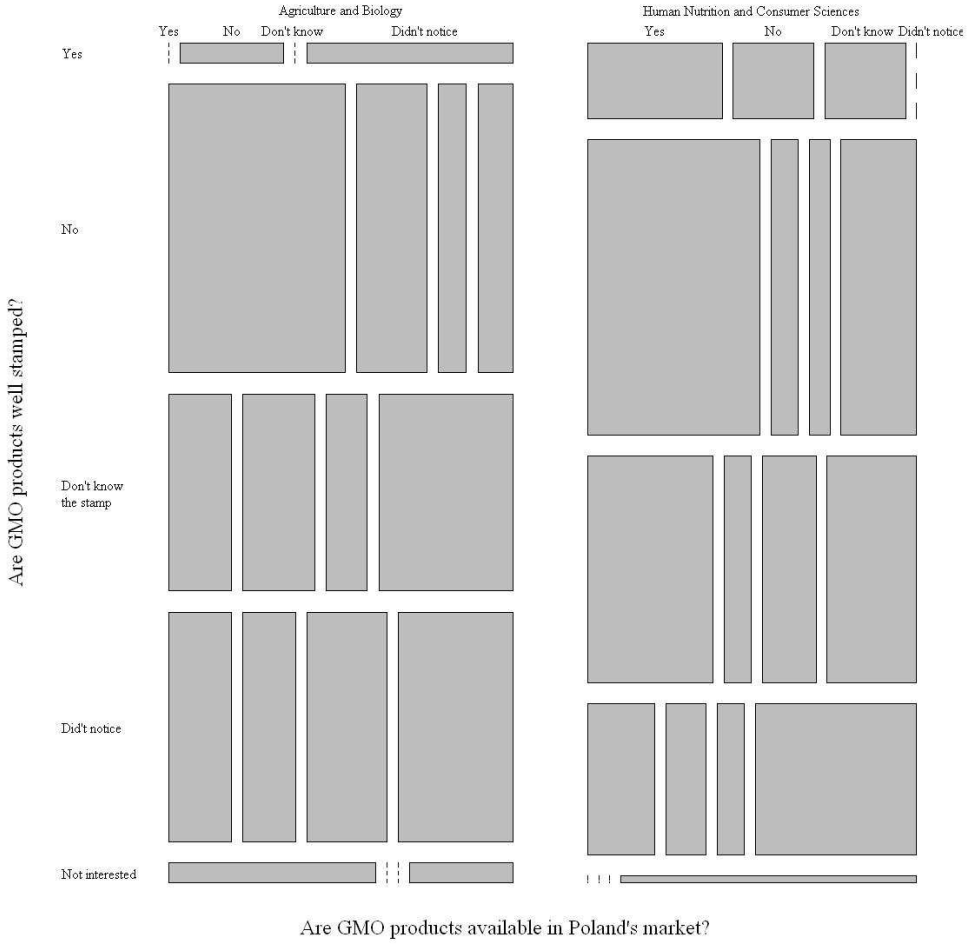

Figure 2. Mosaic plot visualizing the contingency table presented in Table 3

on the data and associations among the variables, but the information offered by the mosaic plots is incomparable. For example, if you look at Figure 1, the lack of association between the answers to the two questions actually jumps out at you. If you spend some time on Table 2 and pay much attention to the numbers, maybe you will see the pattern. Note that this was the obvious thing to see from Figure 1, and yet not so obvious to see from Table 2. Compare now Table 3 and

Figure 2 and decide yourself what gain in understanding the patterns there the mosaic display offers compared to the corresponding contingency table. We believe that after that you will agree with us that the mosaic display may be recommended for use in exploratory analysis and interpretation of categorical data. Of course, the plots do require some time and attention as well, but does them not pay off greatly?

One can go a step further with the mosaic display. Extended—or enhanced–mosaics are yet another powerful tool to visualize such kind of data (Friendly 1994, 1999). This extension comes from adding the information on residuals from log-linear models (Agresti 2002), which are used as a standard to analyze contingency tables. Hence, at only minor cost of slightly more careful examination of the graph, one is offered even more possibilities of interpretation of the associations. However, one needs to be careful to choose the appropriate model for both the particular data set and the question one aims to answer by means of the model. Still, the classical mosaic display is a very easy and powerful tool to understand even complex contingency tables, also by non-statisticians with no expertise in statistical modeling and graphics. This is, of course, a big advantage of the display for the purpose of consultancy to scientists.

References

Agresti A. (2002). *Categorical Data Analysis*. 2nd edition, John Wiley & Sons, New Jersey.

Cleveland W. S. (1994). *The elements of graphing data*. 2nd ed. Summit, NJ: Hobart, USA.

Friendly M. (1994). Mosaic Displays for Multi-way Contingency Tables. *Journal of the American Statistical Association* 89, 190–200.

Friendly M. (1999). Extending Mosaic Displays: Marginal, Conditional, and Partial Views of Categorical Data. *Journal of Computational and Graphical Statistics* 8(3), 373–395.

Hartigan J.A., Kleiner B. (1981). Mosaics for Contingency Tables. In: *Computer Science and Statistics*: *Proceedings of the 13th Symposium on the Interface*, ed. W. F. Eddy, New York: Springer, 268–273.

Hartigan J.A., Kleiner B. (1984). A Mosaic of Television Ratings. *The American Statistician* 38, 32–35.

Laffon J.L., Hanafi M., Wright K. (2007). Numerical and Graphical Measures to Facilitate the Interpretation of GGE Biplots. *Crop Science* 47, 990–996.

Love M.S., Yoklavich M. (2008). Habitat characteristics of juvenile cowcod, *Sebastes levis* (Scorpaenidae), in Southern California. *Environmental Biology of Fishes* 82, 195–202.

Noser R., Byrne R.W. (2007). Travel routes and planning of visits to out-of-sight resources in wild chacma baboons. *Papio ursinus*. *Animal Behaviour* 73, 257–266.

R Development Core Team (2008). R: *A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

WIZUALIZACJA TABEL KONTYNGENCJI PRZY POMOCY WYKRESU MOZAIKOWEGO: PRZYKŁAD DLA BADANIA NAD WIEDZĄ O I STOSUNKIEM DO GMO

Streszczenie

Wykresy mozaikowe są bardzo pomocne w wizualizacji tabel kontyngencji—choćby i bardzo złożonych, przedstawiających wiele zmiennych—pozwalając zrozumieć proste i bardziej skomplikowane związki między zmiennymi jakościowymi, zwłaszcza w przypadku tabel trzy- i więcej kierunkowych. Pomimo to zastosowania wykresu tego typu są niezwykle rzadkie. Celem tego artykułu jest zwrócenie uwagi czytelników na wykresy mozaikowe. Ich zastosowanie jest przedstawione na przykładzie danych pochodzących z badania nad wiedzą o genetycznie modyfikowanych organizmach i nastawieniem do nich.

Słowa kluczowe: dane jakościowe, tabele kontyngencji, organizmy modyfikowane genetycznie, wykres mozaikowy, wizualizacja.

Klasyfikacja AMS 2000: 62-07, 62-09, 62P10