# TYPE OF DATA AND EFFICIENCY OF CHOROPLETH MAPS

Marta Tomaszewska
Chair of Cartography, Faculty of Geography and Regional Studies
University of Warsaw
Krakowskie Przedmieście 30
00-927 Warsaw, Poland
mjablons@astercity.net

## Type of data and efficiency of choropleth maps

Choropleth technique is one of the most widely used types of visualization in thematic maps. However, cartographers still have not specified the principles of elaboration of choropleth maps to accurately represent a spatial pattern of a given dataset. One of the issues is the type of data that may be used for choropleth map design.

Most authors researching choropleth methodology point out that the most correct approach is when choropleth maps are based on *relative data applied to the entire area of enumeration units* (for example: population density). For choropleth map design cartographers also partially accept the use of other relative data: *amounts related to a part of the area of enumeration units* (e.g. share of wheat sown area in the total sown area), *amounts related to the total population of a unit* (e.g. share of population aged 15–59 in total population), *amounts related to a part of the population of a unit* (e.g. share of the unemployed with elementary education in the total number of the unemployed) and *amounts referred to other elements* (e.g. number of inhabitants per pharmacy). Cartographers have not agreed on the use of *absolute data* for choropleth map design. Some authors seem to intentionally avoid this, while others do not allow such option at all.

Validity of cartographers' theses was verified in an experiment performed at the Faculty of Geography and Regional Studies at University of Warsaw. The efficiency of choroplethic transmission of information, depending on data type, was evaluated by a test. Students (128) were divided into four groups. Each group (of 32 students) was presented with a choropleth map based on different data: group I – *relative data applied to the entire area of enumeration units*, group II – *relative data applied to elements other than entire area of enumeration units*, group IIIA – *absolute data that can be applied to the entire area of enumeration units, group IIIB – absolute data that cannot be logically applied to the entire area of enumeration units.* The efficiency of choropleth maps based on various data was tested on an elementary level of map reading. In six tasks respondents were asked to compare values of presented phenomenon in two marked enumeration units. Each task differed from the others with proportion of marked enumeration units area or/and with proportion of values of presented phenomenon in marked enumeration units. Information about three dependent variables (evidencing efficiency of maps) was collected: accuracy of the given answer, time needed to give and check the answer, and conviction of respondent that the given answer was correct.

Although the results of experiment show slight differences between groups in three dependent variables, results of Kruskal-Wallis tests made for each dependent variable in most cases are not statistically significant at the 0,05 level of significance.

Thus, experiment has not confirmed theoretical statements of cartographers. Its results do not allow saying that type of data influences the efficiency of choroplethic transmission of information about values of phenomenon on elementary level of map reading.

## Introduction

Choropleth maps are known since 19th century, but profound exploration of this method took place in the second half of 20th century. Nowadays it is one of the most common used methods of cartographic representation. Nevertheless cartographers still have not specified the principles of elaboration of choropleth maps to accurately represent a spatial pattern of a given dataset. One of the issues is the type of data that may be used for choropleth map design. We can distinguish „raw" data, not related to any other data (e.i. absolute data) or „transformed" data – related to other data (e.i. relative data). The second type of data can be written as ratio (or rate), where *entire area of enumeration units* or *other elements than whole area of enumeration units* can be denominator.

## Different data type and its use for choropleth map design

Most authors researching choropleth methodology point out that the most correct approach is when choropleth maps are based on *relative data applied to the entire area of enumeration units* (G.F. Jenks, 1976; M. J. Kraak and F. Ormeling, 1996; J. Pasławski, 2003; J. Pravda, 1983; T. A. Slocum and others 2009). One of the reason of this statement is that we can represent a choropleth map as 3D construction, where for each enumeration unit we can build a prism with a base that parallels enumeration unit and with an altitude that equals presented value (e.g.: population density, share of forest area in total area). Than the volume of block can be interpreted as absolute value (e.g. number of people, forest area), that has been used for calculating ratio presented on the map. Furthermore choropleth maps use whole area of enumeration units, so it seems reasonable to apply data related to this area (L. Ratajski 1989).

For choropleth map design cartographers also partially accept the use of other relative data: *amounts referred to a part of the area of enumeration units* (e.g. share of wheat sown area in the total sown area), *amounts related to the total population of a unit* (e.g. share of population aged 15–59 in the total population), *amounts referred to a part of the population of a unit* (e.g. share of the unemployed with elementary education in the total number of unemployed) and *amounts related to other elements* (e.g. number of inhabitants per pharmacy) (D. J. Cuff, K. R. Bieri, 1979; G. F. Jenks, 1976;

M. J. Kraak, F. Ormeling 1996; L. Ratajski 1989; A. H. Robinson and others, 1995). Interpretation of choropleth maps based on *relative data applied to elements other than whole area of enumeration units* causes more difficulties compare to interpretation of choropleth map based on *relative data referred to the entire area of enumeration units*. It happens because values of denominator influence values of the presented ratio. However, map reader is not informed about denominators' values in any way. Cartographers make an assumption that denominator should represent phenomenon with homogeneous distribution but in practice this theoretical statement is not abided.

Cartographers have not agreed on the use of *absolute data* for choropleth map design. Some authors seem to intentionally avoid this, while others do not allow such option at all. D. J. Cuff and K. R. Bieri (1979) find large enumeration unit easier to behold than small enumeration unit although in both enumeration units value of presented phenomenon is the same – thus colour in both enumeration units is also the same. On the other hand G.F. Jenks (1976) emphasizes the meaning of map readers' subconscious assumption about homogeneous distribution of phenomenon. As he says map user instinctively join large enumeration unit with large number of objects and small enumeration unit with small number of objects. Thus, mentioned authors point out that difficulties with choropleth map reading are caused by different area of the enumeration units. Cartographers do not mention any other reason not to use absolute data for choropleth maps design. They seem to forget that different areas of enumeration units also impede interpretation of choropleth maps based on *relative data*.

## Design of experiment

I supposed that assumption of cartographers about possibilities of interpretation of choropleth map based on different type of data could be verified in experiment. If theoretical statements of cartographers are valid, the choropleth maps designed with use of *absolute data* are less efficient in transmission of information than the choropleth maps based on *relative data*.

To evaluate data type influence on efficiency of choropleth maps I divided sample of population of choropleth map users (128 students of 1st semester at Faculty of Geography and Regional Studies at University of Warsaw, before courses of basic cartography) into four experimental groups. Students were selected to groups according to the sex proportions in the entire sample. Each experimental group was presented with a choropleth map based on different data: group I – *relative data applied to the entire area of enumeration units* (population density – number of people per square km), group II – *relative data referred to elements other than entire area of enumeration units* (share of people aged 0–18 in the total population – %), group IIIA – *absolute data that can be related to the entire area of enumeration units* (number of people), group IIIB – *absolute data that cannot be logically applied to the entire area of enumeration units* (number of people aged 0–18). To eliminate confounding variables the elaboration of choropleth maps for all experimental groups was unified and maps were given the same appearance. The only difference between maps in experimental groups was the name of the phenomenon shown in the legend at the right side of the map. In the six tasks respondents were asked to compare two marked enumeration units and define if value of phenomenon presented on the map in the unit marked with a grey outline was:

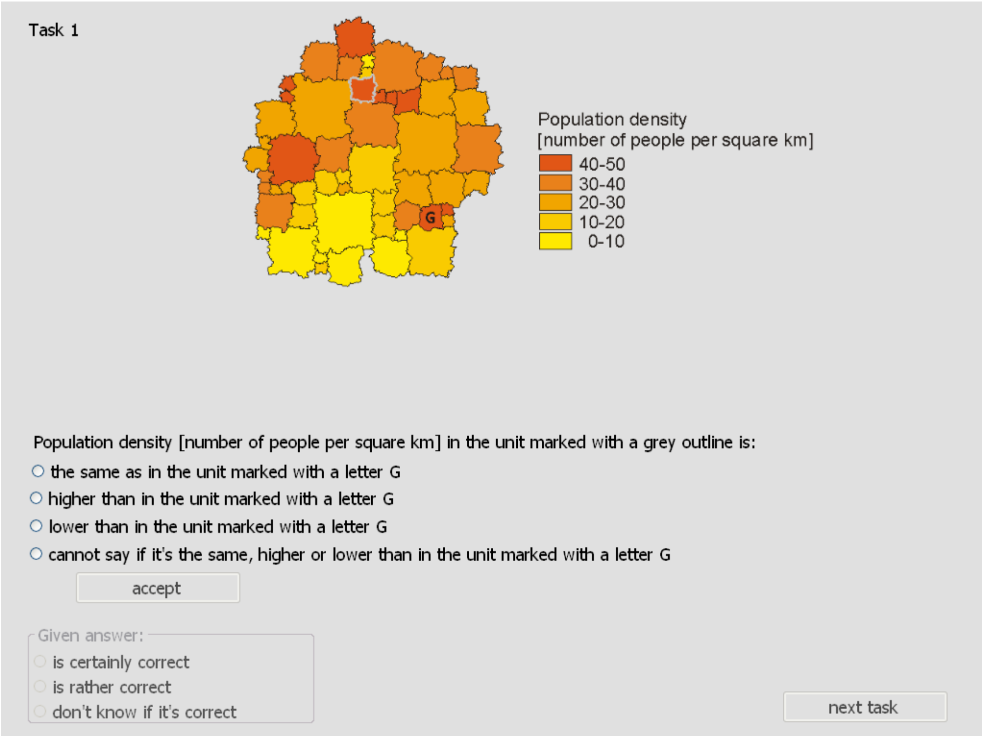


Figure 1. Task 1 in experimental group I (original in polish).

the same as value in the unit marked with a letter, higher than value in the unit marked with a letter, lower than value in the unit marked with a letter or cannot say if it was the same, higher or lower than value in the unit marked with a letter (figure 1). Therefore efficiency of maps on elementary level of reading (J. Bertin 1983) was evaluated.

Task differed with proportion of enumeration units areas and/or with values of presented phenomenon in marked enumeration units. Respondents were to compare:

– in task 1: enumeration units with the same values of phenomenon presented on the map and the same area (in figure 2 – unit no 1 and unit G). Therefore they should choose first answer in the task form (e.i. the same value).
– in task 2: enumeration units with different values but with the same area. Second answer in the task form was accurate as in unit no 2 value was higher than in unit E.
– in task 3: enumeration units with the same values but with different area (unit no 3 and unit B). Respondents should choose first answer in the task form (e.i. the same value).

– in task 4: small enumeration unit with low value (unit no 4) to large enumeration unit with high value (unit F). Thus, third answer was accurate (i.e. lower value).
– in task 5: small enumeration unit with high value (unit no 5) to large enumeration unit with low value (unit E) – second answer correct (i.e. higher value).
– in task 6: large enumeration unit with low value (unit no 6) to small enumeration unit with high value(unit G) – third answer accurate (i.e. lower value).

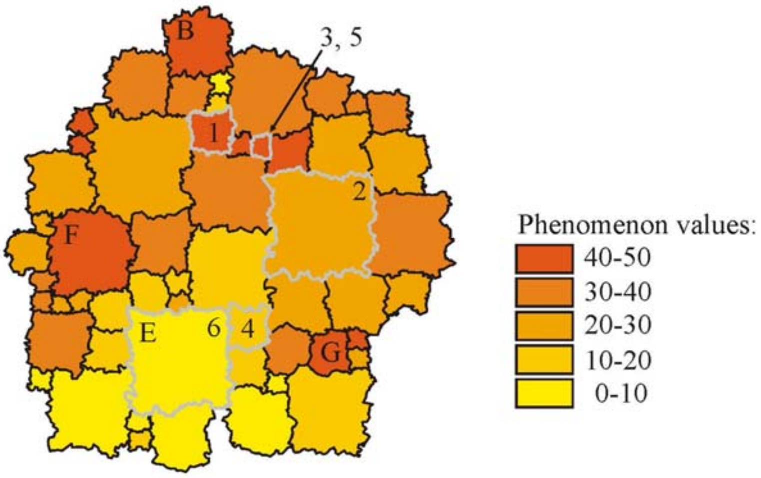


Figure. 2. Enumeration units compared in tasks of experiment.

Information about three dependent variables was collected: accuracy of the given answer, time needed to give and check the answer (time measured from map show on the monitor of computer to button 'accept' click), and conviction of respondent that the given answer was correct (in each task respondents were asked to designate if given answer: 'was certainly correct', 'was rather correct' or 'didn't know if it was correct').

If choropleth maps based on the *relative data applied to the entire area of enumeration units* are more efficient in transmission of information comparing to choropleth maps based other types of data, accuracy of the given answers should be better, time needed to give and check the answer should be shorter and the conviction of respondent that the given answer was correct should be higher in the experimental group I than in groups: II, IIIA and IIIB. Furthermore, according to theoretical statements of cartographers, results of respondent reading maps based on *absolute data* (group IIIA, IIIB) should be worse than results of respondents using maps based on *relative data* (group I, II).

**Results of experiment**

Figure 3 shows that the majority of the given answers was accurate. In tasks 1, 2, 4 and 6, maximum difference between groups in number of correct answers were very small. They varied from 1 in the task 6, to 3 in the task 1. In tasks 2 and 4 difference in number of correct answers equalled 2. High and similar accuracy of the given answers in those tasks surely cannot evidence that type of data does influence efficiency of choropleth maps. Results of Kruskal-Wallis test (table 1) seem to confirm this statement at the 0,05 level of significance (α). In task 3 maximum difference between groups in number of correct answers equalled 5. Many respondents (21) stated that values of phenomenon in two marked enumeration units were not the same. Most of them (14) chose answer 3 in the task form. Those respondent assumed that value in small enumeration unit was smaller than in large enumeration unit.

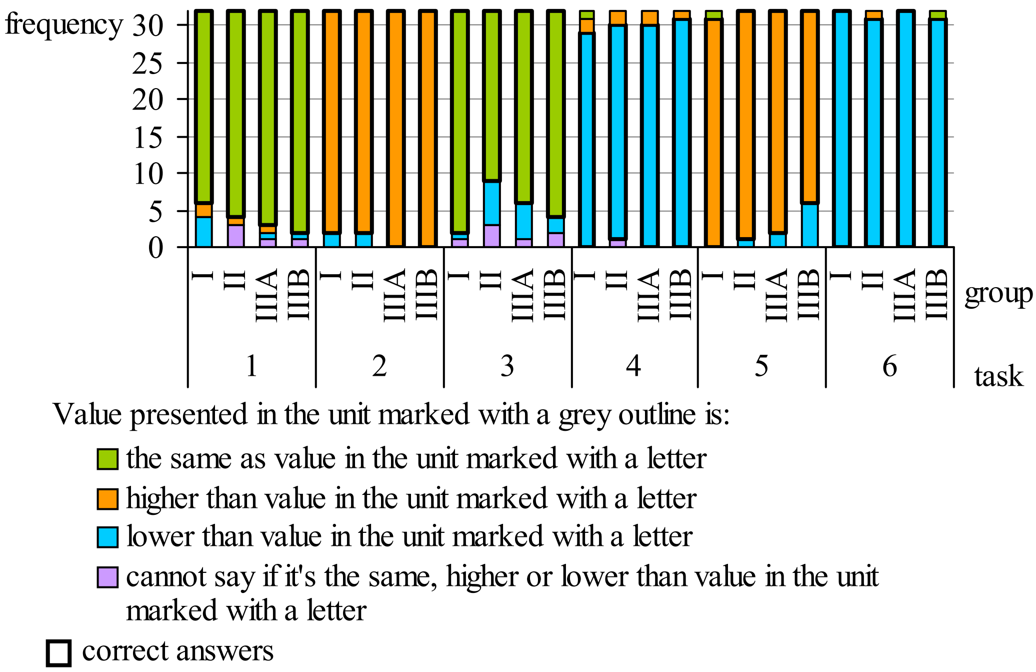


Figure 3. Answers given in the tasks 1–6 and their accuracy.

I suppose that they subconsciously matched value of represented phenomenon with area of enumeration unit. This effect was especially significant in group II and it seemed to be weak in group I. The influence of matching values with areas of compared enumeration units was also observed in the results of the task 5. Most of that wrong answers (6) were given in group IIIB, while in group I there was no such answer and in group II only one respondent gave such answer. I thought that differences in accuracy of given answers in task 3 and 5 could evidence that type of data influenced accuracy of given answers. However, the results of Kruskal-Wallis test, shown in the table 1, did not allow saying that at the 0,05 level of significance. Also number of respondents, who gave together 6, 5, 4 or 3 correct answers in tasks 1–6, was similar (figure 4). Moreover not significant results of Kruskal-Wallis *H* test (α=0,05) did not confirm hypothesis that type of data influenced the accuracy of given answers.

| | task 1 | task 2 | task 3 | task 4 | task 5 | task 6 | total tasks 1-6 |
|---|---|---|---|---|---|---|---|
| $\chi^2$ | 2,622 | 4,097 | 6,048 | 1,304 | 7,319 | 2,016 | 3,203 |
| df | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| p | 0,454 | 0,251 | 0,109 | 0,728 | 0,062 | 0,569 | 0,361 |

Table 1. Results of Kruskal-Wallis *H* test (accuracy of answers in tasks 1–6).

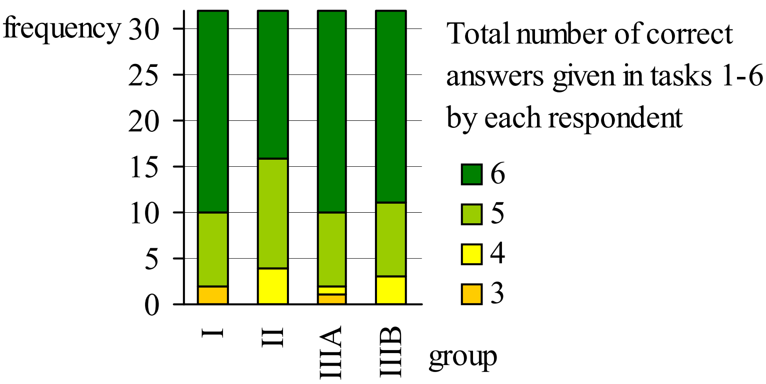


Figure 4. Total number of correct answers given in the tasks 1–6.

Figure 5 shows time that respondent needed to give and check the answer. The first thing to see is that average time needed to give and check answers reduces gradually from 24–27 sec. in task 1 to 9–12 sec. in task 6. But in almost every task average time in groups II and IIIB is longer than in groups I and IIIA. Also maximum time tends to be longer in groups II and IIIB than in two others. The distribution of the dependent variable indicated that the largest differences between groups occurred in task 2, 3 and 5, but only differences of the respondents' results in task 5 were statistically significant at the 0,05 level of significance (table 2). I also verified which groups (in pairs) differed. The Mann-Whitney *U* rank sum test with Bonferroni correction revealed significance (at the 0,05 level of significance) of difference group II and IIIA (*U*=309,5; N=32;
(*Z*=-2,719); p=0,042).

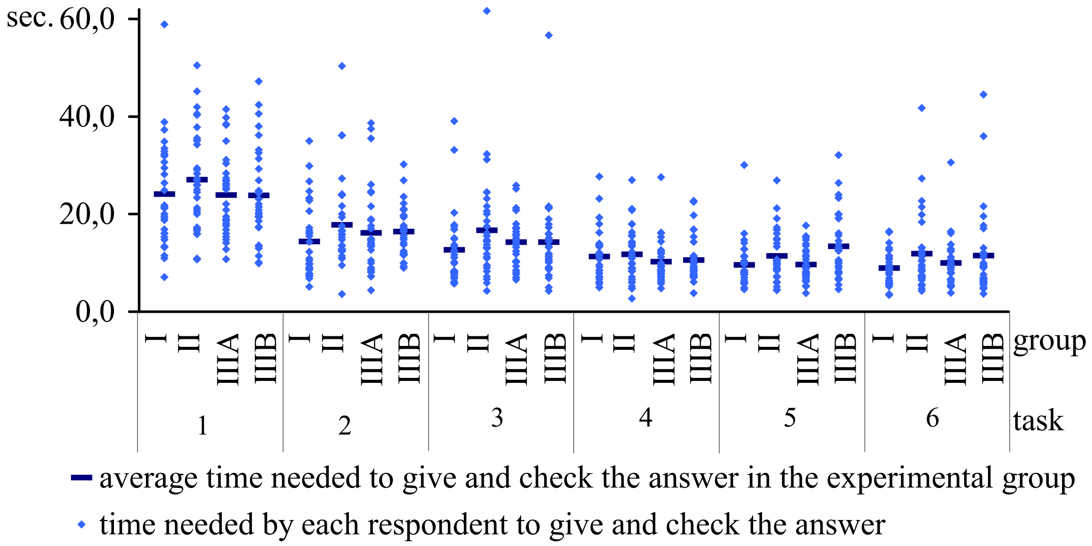


Figure 5. Time needed to give and check the answer in tasks 1–6.

| | task 1 | task 2 | task 3 | task 4 | task 5 | task 6 | total tasks 1-6 |
|---|---|---|---|---|---|---|---|
| $\chi^2$ | 2,364 | 5,844 | 5,681 | 2,011 | 9,642 | 1,724 | 6,209 |
| df | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| p | 0,500 | 0,119 | 0,128 | 0,570 | 0,022 | 0,632 | 0,102 |

Table 2. Results of Kruskal-Wallis *H* test (time needed to give and check the answer in tasks 1–6).

In the figure 6 we can see that total average and total maximum time needed to give and check the answers in tasks 1–6 was longer in groups II and IIIB compare to groups I and IIIA. However, results of Kruskall-Wallis test ($\alpha$=0,05) did not allow saying that type of data influenced total time needed to give and check the answers in task 1–6.

Figure 7 shows that respondents were mostly certain that the given answer was correct. The maximum difference between groups in number of respondents that chose the highest rank in tasks 2–6 was small (vary from 3 to 5). As rank of conviction was subjective estimation of the respondent, I did not find those results confirming hypothesis that type of data influenced conviction of respondents about accuracy of the given answers. The results of Kruskal-Wallis *H* test (table 3) showed validity of my opinion. Only results of test made for task 1 were statistically significant at the 0,05 level of significance. Here the maximum difference between groups in number of respondents that were certain about accuracy of given answer equalled 9. I also verified which groups (in pairs) differed in conviction of respondent that given answer was correct. The Mann-Whitney *U* rank sum test with Bonferroni correction revealed that none of observed differences between groups (in pairs) were statistically significant.

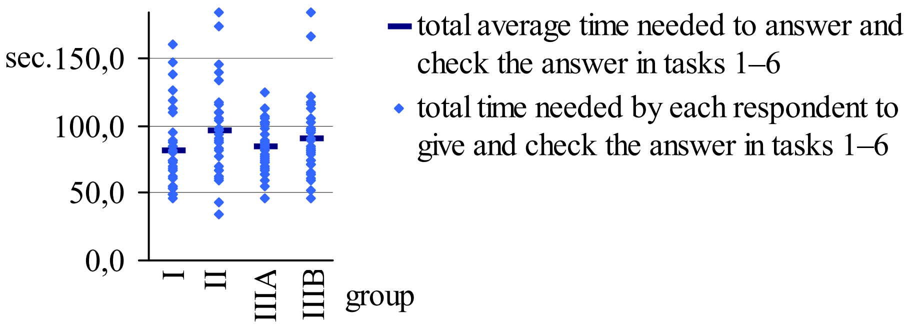


Figure 6. Total time needed to give and check the answer in tasks 1–6.

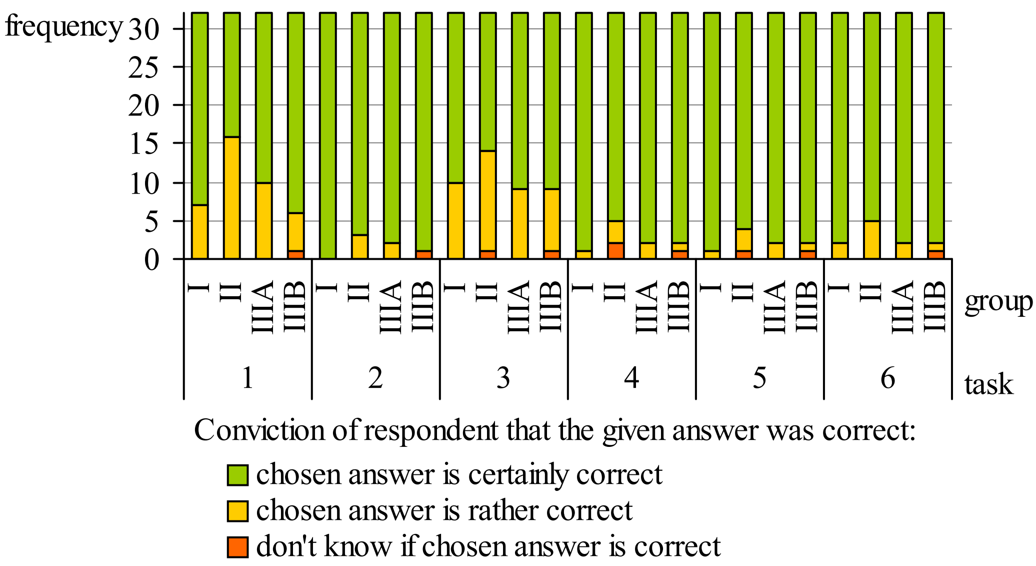


Figure 7. Conviction of respondent that the given answer was correct in tasks 1–6.

To compare total results acquired in the tasks 1–6, I summarized ranks of conviction of each respondent that the given answer was correct (answer ‘is certainly correct’ was given value 3, answer ‘is rather correct’ – value 2, and answer ‘don’t know if chosen answer is correct’ – value 1). The number of respondents who achieved respective sums of the ranks is shown in the figure 8. There was a difference between groups but it was not statistically significant (α=0,05).

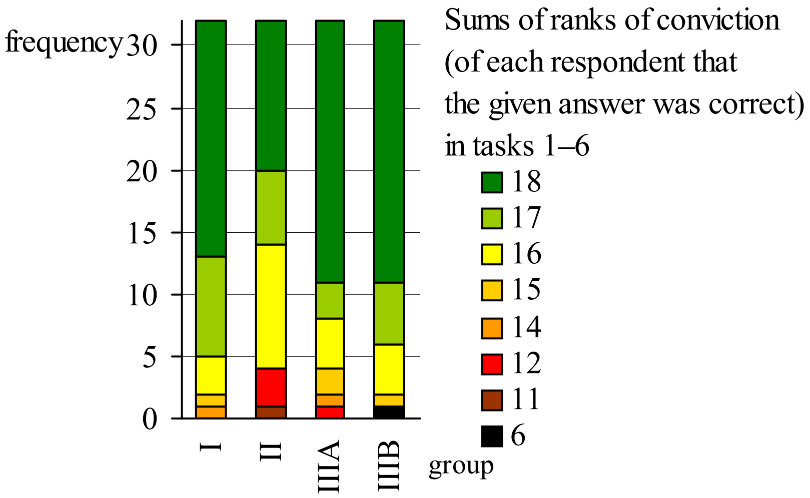


Figure 8. Sums of all ranks of conviction of each respondent that the given answer was correct in tasks 1–6.

| | task 1 | task 2 | task 3 | task 4 | task 5 | task 6 | total tasks 1-6 |
|---|---|---|---|---|---|---|---|
| $\chi^2$ | 8,403 | 3,384 | 2,496 | 3,992 | 2,282 | 2,573 | 7,618 |
| df | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| p | 0,038 | 0,336 | 0,476 | 0,262 | 0,516 | 0,462 | 0,055 |

Table 3. Results of Kruskal-Wallis *H* test
(conviction of respondent that the given answer was correct in tasks 1–6).

## Conclusions

The experiment did not confirm theoretical statements of cartographers. Its results showed that efficiency of choropleth maps based on *absolute data that can be related to the entire area of enumeration units* was similar to efficiency of choropleth maps elaborated with use of *relative data referred to the entire area of enumeration units*. Furthermore it seems that choropleth maps based on *relative data applied to elements other than entire area of enumeration units* and based on *absolute data that cannot be applied the entire area of enumeration units* were less efficient than choropleth maps designed with use of two other types of data. Nevertheless the results of statistical tests did not allow saying that type of data influence the efficiency of choropleth maps in transmission of information about values of phenomenon on elementary level of map reading.

## References


Bertin J., 1983: *Semiology of Graphics: Diagrams, Networks, Maps.* Madison. University of Wisconsin Press,

Cuff D.J., Bieri K.R., 1979: *Ratios and absolute amounts conveyed by a stepped statistical surface.* „The American Cartographer" Vol. 6, no 2, pp. 157−168.

Jenks G.F., 1976: *Contemporary statistical maps – evidence of spatial and graphic ignorance*. „The American Cartographer" Vol. 3, no 1, pp. 11−19.

Kraak M. J., Ormeling F., 1996: *Cartography. Visualization of spatial data.* Addison. Wesley Longman Limited.

Pasławski J., 2003: *Jak opracować kartogram. (How to make a choropleth map)*. Warszawa. Uniwersytet Warszawski. Wydział Geografii i Studiów Regionalnych.

Pravda J., 1983: *Zákon kartogramu a problem vyjadrovania nerozlohových charakteristik (The Law of Cartogram [choropleth map] and the Problem of Expressing Non-Areal Characteristics)* „Geogr. Casopis" R. 35, c. 2, pp.136−159.

Ratajski L., 1989: *Metodyka kartografii społeczno-gospodarcze (Handbook of socio–economic cartography).* Warszawa. PPWK – 2nd ed.

Robinson A. H., Morrison J. L., Muehrcke P. C., Kimerling A. J., Guptill S. C., 1995: *Elements of Cartography.* New York, Chichester, Brisbane, Toronto, Singapore. John Wiley & Sons Inc. 6th ed.

Slocum T.A., McMaster R.B., Kessler F.C., Howard H.H., 2009: *Thematic cartography and geographic visualization*. Upper Saddle River, NJ: Person Education, Inc.