Acta Ethnographica Hungarica, 45 (1–2), pp. 171–179 (2000)

THE METHODS AND PROBLEMS OF THE DEFINITION OF THE CULTURAL REGIONS OF HUNGARIAN-SPEAKING AREAS BY COMPUTER[1]

Balázs BORSOS

Hungarian Institute of Ethnology
H-1014 Budapest, Országház utca 30, Hungary

Abstract: First the author summarizes the attempts of defining the regions of Hungarian folk culture and he concludes that the next step in this kind of investigation must be a certain definition based on as many cultural elements as possible. He intends to do it by using the database of the Atlas of Hungarian Folk Culture and computer methods. He investigates the opportunity and the problems of the transformation of the data of the Atlas into a computer database, he presents the problems to be solved and some possible solutions. He concludes that for various reasons only a limited number of the cultural phenomena mapped in the atlas are suitable for computer analysis. He summarizes the methodological background of the correlation and cluster analysis to be used. He emphasizes that due to the special character of mapping the inconsistencies and the mixing of different points of view in the Atlas, and to the character of the computer analysis that is mechanical and not elaborate enough, this kind of definition of the cultural regions of the Hungarian-speaking areas cannot replace the previous definition of regions but it can offer a good frame of reference to define the regions more precisely.

Keywords: computer methods in ethnography, cultural regions, ethnographical atlas, ethnographical cartography

ATTEMPTS TO DEFINE THE REGIONS OF HUNGARIAN FOLK CULTURE

Attempts to define regions of the folk culture of a certain nation are nearly as old as the science of ethnography itself. Investigations both in Hungary and in other parts of the world have focused on two mutually connected and interdependent areas: in one of them the folk culture of a certain region of the investigated nation was surveyed and thoroughly described, and in the other one some special cultural phenomena were looked for that made it possible to divide the land inhabited by the investigated nation into smaller regions. At the end of the 19th and at the beginning of the 20th century the regional division was based on geographical phenomena. In Hungary this period was the time of huge descriptional monographs, some of which

[1] The article is the revised version of a lecture held at the Conference "Culture in Space" in Cieszyn, Poland, October 1998 that intended to summarize the methodological and theoretical basis of a four year long project helped by the OTKA-program F 017986 and the Bolyai János Research Scholarship aimed at finding new possibilities in the field of ethnoregional studies by the elaboration of the data of the Atlas of Hungarian Folk Culture by computer.

1216–9803/2000/$ 5.00

took ethnographical data into consideration as well.[2] But even these works dealt with those regions that were well separated from others and whose distinct status was well-known and they neglected regions not so well defined. Between the two World Wars these regional monographies based on geographical division were continued but they applied a historical point of view, too.

The first attempts by István Györffy and Károly Visky to define the regions of the whole Hungarian folk culture were also published in the 1920s and 1930s. The basis of division was still regional geography complemented by some historical, linguistic and ethnographical data. The investigations after the Second World War followed the path signed by the previous researchers: on the one hand, the classical geographical-historical division of the territory of the Hungarian nation was accepted, which was based on geographical, historical, ethnical and cultural elements alike, and, on the other hand, special cultural phenomena were looked for that could be used for differentiating among ethnographical regions. The studies of István Tálasi pointed out the weaknesses of this approach: he emphasized that the characteristics of those special cultural phenomena which were used for dividing certain cultural regions can also change owing to the effects of the history and the cultural evolution of the region, so they cannot be used for dividing regions without considering the historical element as well. He also underlined that the character of a certain cultural region depended not primarily on the absence or presence of a special cultural phenomenon but by far on the special pattern of cultural phenomena presented there as well as outside the region. Hungarian ethnographical research has not yet solved this problem: the cultural subdivision of the areas of the Hungarian nation is still based mainly on certain cultural elements (i. e. liguistic phenomena, dialects of folk musik and folk dance, etc.) and there is still a common agreement on the use of the old geographical-historical division as a basis of comparison.[3]

DEFINING ETHNOCULTURAL REGIONS ON THE BASIS OF SEVERAL CULTURAL ELEMENTS

Regarding all these, the next step in the investigation of the regional division of Hungarian folk culture is obvious. In order to define more correct cultural regions, we have to use not only some particular elements of culture but as many cultural phenomena as possible, and to avoid the influence of the historical changes and cultural evolution we have to choose a certain horizontal cross section of time. If in this particular period of time we could investigate the distribution of a long row of cultural phenomena, we can define synchronous regions of the Hungarian folk cul-

[2] Among them are the volumes of the series: Austro-Hungarian Monarchy in Text and in Pictures, and the Monograph of the Counties of Hungary. Concerning ethnography we must mention the Description of the Land of the Székely people by Balázs Orbán and the works of János Jankó, who concentrated in his volumes (for example about Kalotaszeg and the Balaton-Highland) most consciously on ethnographical data.

[3] KÓSA 1975: 30–39.

ture. This distribution does not define either ethnical or regional groups, as it has nothing to do either with the "us and them" awareness, or with the regional view of the members of the group. This kind of investigation divides ethnocultural groups, as they are defined by the ethnographical research itself.[4] The opportunity of applying this kind of investigation can be attributed to the progress of science. Primarily the progress of informatics enables us to use, to treat and to elaborate a great number of cultural data, and secondly the Atlas of Hungarian Folk Culture completed at the beginning of the 1990s provides us with this great variation of cultural phenomena in a certain period of time, as the Atlas, though intends to perform some changes in cultural elements as well, mainly lays down the ethnographical picture of the Hungarian nation around 1900.

If we want to use as many ethnographical data as possible to define the regions of the Hungarian folk culture, we have to employ computers not only for storing and in any case of need for calling forth data, but for elaborating them, too. Using computers in processing data means applying mathematical methods for this and applying mathematical methods in human sciences has its own problems, "thus above all" most of us are not educated in high mathematics well enough to understand the scientific background of the employed methods. Anyhow, for the sake of fast processing a huge amount of data, we cannot abandon computer methods and, I think, for the correct use of them we do not have to be familiar with the whole background of a chosen one, only with its principles and with the opportunities and limitations of its applicability.

To complete the computer analyzation of the regions of the Hungarian folk culture we have to fulfil three tasks.

1. to collect a wide variety of ethnographical data in the Hungarian-speaking territories,

2. to create their database on computer, and

3. to accomplish their analysis according to the regional distribution of the collected cultural elements. It means clusterization and a correlation analysis of the clusters worked out.

1. COLLECTING DATA

The collection of a wide variety of ethnographical data in the Hungarian-speaking areas was done by the authors and the editors of the Atlas of Hungarian Folk Culture, and although the opportunity of a following computer elaboration was not taken into account and therefore in the digitalization and processing we have to face numerous problems (see later), the Atlas can be used as a database on which we can carry out the analysis of cultural regions. Nevertheless, the Atlas itself is not the most extensive representative of the Hungarian folk culture.

[4] KÓSA 1975: 40–51.

1.1. The Atlas contains only 634 sheets. It means that no more than around 634 cultural phenomena can be used for the database. I say around, as some phenomena are drawn in two maps[5] and in a few maps two or more phenomena are represented.[6] The Atlas illustrates these phenomena only in 417 settlements, which are less than 10% of the number of the settlements of the Hungarian-speaking area.

1.2. The Atlas does not represent every part of the Hungarian folk culture in the same depth, though it would be crucial to define the cultural regions most precisely. The editors of the Atlas intended to collect the most typical phenomena, but while the field of material culture is more or less thoroughly represented, the folklore is not, as in an atlas phenomena must be put in maps, and folklore data are much less suitable for that than those of material culture.[7] Consequently, the cultural regions defined by the data of the Atlas will also be unilateral to a certain extent.

1.3. About 10% of the maps deal with the changes of a certain cultural phenomenon in time, so they cannot be used for drawing the cultural regions around 1900.

1.4. Theoretically every map, namely every cultural phenomenon must be equal in the later analysis, but in fact they must not be handled like that, as some maps are only a more detailed further illustration of one part of a previous map. For example map Nr 89 represents the different types of hoes, and maps Nr 90–92 are about the subtypes of the three main types. Some phenomena are in separate maps, some phenomena of the same type are contracted in one. For example the calling terms for enticing and driving away certain animals (foals, pigs, hens, ducks, geese, dogs) are in separate maps (Nr 140–143, 145–149, 151–153), while those of the sheep are concentrated in one (Nr 144).

1.5. Certain fields of culture are overrepresented in the Atlas, some are underrepresented, although in many cases the overrepresented ones are much less important parts of the culture. For example the words for guiding and ordering animals get 20 sheets (Nr 134–153), the occurrence of certain first names among the population gets 25 sheets (Nr 483–507), and all the phenomena about settlements are illustrated only in 5 sheets (Nr 1–5).

1.6. Some groups of cultural phenomena are found only in a certain part of the Hungarian-speaking area (for example the milking of sheep and the processing of sheep's milk, Nr 169–190). When these phenomena are detailed in further maps, the settlements where the phenomena occur are overrepresented and they are characterized in more detail in the sample than the settlements wihout these phenomena.

1.7. The data of a great number of cultural phenomena are absent in at least one third of the settlements which is due to the problems of collecting data.

[5] E. g. map Nr 390–391: Water gruels made from maize: accompaniments 1–2., Nr 547–548: Ways of making fun of unmarried ladies 1–2., Nr 576–577: Forewarnings, coming from animals, of death or ill luck 1–2., Nr 587–588: Elements of the wedding during the burial of a girl 1–2.

[6] E. g. map Nr 216: (1) Terms for a bundle in a linen cloth carried on the back; (2) Distribution of the double bag carried on the shoulder. Other examples are the maps about first names (Nr. 483–507) where the occurrence of the first names is shown in the case of the four main religions alike. It means that in these maps two different types of information are mixed for the benefit of more detailed presentation.

[7] KÓSA 1975: 45.

Consequently, there are much less cultural phenomena than 634 that are suitable for further analysis and even the phenomena available represent the Hungarian folk culture unilaterally to a certain degree. We may take the law of great numbers into account, which means that above a certain number of variables the result of an analysis is not modified essentially by the increasing number of variables. But in this way we can only partly remove or reduce the problem, as in the database we will have a deficiency of certain groups of cultural phenomena that are of a character different from the represented ones, and therefore they might cause vital changes in the result.

2. CREATING A DATABASE

Despite the fact that only a limited number of the sheets of the Atlas can be used for computer analysis, we decided to put every map and every datum of the Atlas on computer. This way we have got the computerized version of the Atlas of Hungarian Folk Culture and it has some advantages over the original one. Instead of dealing with huge piles of maps (the nine volumes of the Atlas are altogether 20 cm thick!) we have the database with the installer program under Windows and the application program on only three discs. With the application program we are able to arrange the database from other points of view, than only from that of the Atlas. We can collect for example all data of a certain settlement, all data of a certain region; all variants of a cultural phenomenon, the presence of only one variant of a cultural phenomenon, the simultaneous occurrence of two or more phenomena, every data of a certain field of culture (e. g. transport or housing) etc.

So there was a database defined for 417 objects (settlements) and 634 variables (maps). The co-ordinates and names of the settlements were fixed and to digitalize the different variants of the cultural phenomena, the different values of every single variable were defined. The minimum number of the values was 2,[8] the maximum was 56.[9] To each value (it means to each variant of a certain cultural phenomenon) a numerical value was connected. The program has been written so that the number of the objects can be multiplied, as in many cases at some settlements two or more variants of a certain cultural phenomenon may occur. Consequently, the database contains around 400 000 places. It took nearly three years to fill the database and to control the correctness of its values.[10]

[8] E. g. map Nr 85: The meaning of *szuszék* (wooden container) in the first half the 20th century, or Nr 580: Prophecy of the falling star.

[9] Map Nr 484: Leading female first names (1900–1910).

[10] This control had to be done thoroughly, as it was very easy to make a mistake during digitalization because the background and the frame of reference of the maps were printed too dim and in many cases the print of the signs symbolizing a certain value of a variable was not correctly drawn to their place so there was some uncertainty about which sign belongs to which co-ordinate. Although it was advised during the elaboration of the Atlas to use symbols of very different characters, in some maps the signs were far too similar and so subject to confusion.

3. THE ANALYSIS OF THE DATABASE

In order to draw the regions of Hungarian folk culture, we have to define some groups of settlements where the typical variants of cultural phenomena are more similar to each other than to other settlements that also form groups characterized by similar variants of phenomena. In each group the pattern of the similar variants differs from the pattern of other groups. This type group formation is called cluster analysis.

Cluster analysis is used frequently in human sciences (history, economics, archeology) but less often in ethnography. The author used this analysis in his PhD thesis to find out the regional differences in land use in a certain part of Hungary in the 19th century. He was also able to characterize the transformation of land use due to river canalization and to determine the most decisive factors of this process.[11]

The principle of cluster analysis says that every object has to be put in an *n*-dimensional virtual space, according to the values of the *n* variables characterizing them and the distance of these virtual points has to be investigated. The arrangement and the investigation is done by computer. Points, closest to each other get into the same cluster or group. But clusterization can be viewed as a process where at one end of the clusterization every point forms a single cluster, and at the other end all points are in the same cluster. The smaller the number of the clusters is, the farther (it means less similar) points get into the same group. The question is, after how many existing clusters we should put an end to the process. The ideal number of the clusters can be determined by investigating the value of variance at every step of the clusterization. The dramatic increase of the value means that the next point characteristically differs from the previous ones. The number of clusters at this step of the process is the ideal one.

In the process of clusterization we have to consider some factors and requirements to get the most correct result. Among the different types of clusterization methods it is advisable to use the so called Ward method, because it encourages the formation of smaller but more characteristic clusters. Before clusterization the so called standardization of the values of variables must be done. This allows the elaboration of the different variables in the same proportion, otherwise variables with many values would outbalance the result at the expense of variables with only two or three values. The Atlas has maps with two values as well as maps with 40–50 values. After standardization all variables are taken into consideration to the same measure.

To use the cultural phenomena and their variants of the Atlas as variables and values in cluster analysis we have to solve a long row of problems.

3.1. During the digitalization of the variables we gave every single value a numerical value. But the difference among these values in reality is not quantitative, but qualitative. In the normal way cluster analysis deals with numerical data, it means for example that it estimates 1 and 10 farther from each other, than 1 and 2.

[11] BORSOS 1995.

In our case, however, it is not true, as 1,2 and 10 are theoretically at the same distance from each other (We shall see later that it is not completely true). To solve this problem we have two possibilities. First we have to define the numerical values of the values so that greater difference in values means greater difference in the variants of phenomena, too. While defining the values we tried to follow this method. Nevertheless, there is another solution, which is better. It means that in the analyzer program values are not treated as numbers but as characters, so different numbers do not mean real numbers, they are only symbols to show that the values *differ* from each other. This results of course in a more difficult task for the programer.

The fact that the authors of the Atlas did not think about the opportunity of digitalization of the data causes many serious problems in the process of elaboration.

3.2. Empty space at a certain co-ordinate can mean three different things:

a) at that place the cultural phenomenon considered was not collected;

b) at that place the cultural phenomenon considered does not occur;[12]

c) at that place the most common variant of the cultural phenomenon occurs and only other variants of the cultural phenomenon are shown.

The last problem can be solved in 2 ways: either one supposes that the common variant occurs at every other place (for example in map Nr. 344, where no colour of knee-length boots is specified, means that at that place the colour is black) or one compares the data of the particular variable with the data of other variables, as this type of lack of data happens only in maps that deal with the subvariants of an important cultural phenomenon. (For example in sheet Nr 279 the name "gémeskút" for sweep-pole wells is not shown, but comparing this sheet with Nr 276 (The type of wells) we can define where sweep-pole wells occur at all.[13] We also have to write a comparing program of course.)

The problem mentioned in 3.2a and 3.2b can be solved in three different ways:

a) we can compare the particular sheet with other sheets showing similar cultural phenomena with the help of the program mentioned above. If it results in the lack of any data in other sheets as well, we can conclude that the cultural phenomenon does not occur at the given place.

b) we can decide whether 3.2a or 3.2b is true by analyzing the collecting books again or by making a re-collection of data at certain settlements. The first idea is not really promising, since if there had been any evidence of data there, the constructor of the map would have taken it into account. Re-collection after 30 years of the project seems useless, not mentioning the Sisyphus-like character of this work, as the collection was done by 159 experts and the number of the people working on the computer elaboration of the Atlas is two: a programer and the author.

[12] Even this is not a general characteristic of the Atlas as in some maps the fact that the illustrated cultural phenomenon does not occur in a certain settlement is shown with a separate symbol (i. e. map Nr 350: Material of which butter is made and terms of the remaining liquid substance).

[13] Other examples are maps Nr 49: Terms of the topmost sheaf in the pile, where the most common variant "*pap*" is not shown, or Nr 123: Terms for the elements of the frame yoke, where only some variants are shown.

c) the third solution is that after reducing the number of the cultural phenomena we are concerned with, we reduce the number of the considered settlements as well. It means that we neglect those places where we can find data for example in less than 50% of the cases. Its effect of course is that the drawing of the cultural regions will be less correct.

3.3. The editing of the Atlas and the definition of the principles of the mapping of the different cultural phenomena were not systematic. It does not mean only that instead of a comfortable number of values of different variables (5–10) the number fluctuates in different maps between 2 and 56, which means that one map-constructor concentrates only on one important aspect of the cultural phenomenon shown, while the other wants to pile up as much information about the topical cultural phenomenon as possible. More serious problems are caused by the facts that

a) if two or more variants of a cultural phenomenon occurred in certain places, the different map-constructors had different ideas to show them. In one map every possible constellation gets a separate symbol and so a separate value in the digitalization (for example map Nr 9: Animal power used for ploughing by farmers c. 1900). In another map the constellation is shown by the presence of all the symbols and so all the values of the basic variants which occur at the places considered, and the constellation itself does not get a separate symbol (in the majority of the cases). These maps are incomparable, so before cluster and correlation analysis the separate symbols of constellations must be turned into the basic variants. (Now we think that this is the easiest way.)

b) although the differences among the values of certain variables must be qualitatively equal, in many cases it is not so. In some cases in the definition of the variants of the cultural phenomenon shown in a map two points of view have been mixed. For example in the map which is about the way how the horse brakes the cart (Nr 117) three symbols show the different ways of this activity, but the fourth shows the occurrence of the first type by the changing in time. In other cases very similar variants of a certain cultural phenomenon get separate symbols (for example if only one letter differs in the name of a tool) so they seem to be qualitatively as different from each other as from other variants (which have for example a totally different word for the same tool).[14] In further cases totally different names are drawn together under only one symbol.[15]

These problems are mainly caused by the lack of unified editing principles of the Atlas, and cannot be cured easily. They of course weaken the correctness of the drawing of the cultural regions even more.

[14] A good example appears in the map Nr 298: The terms for looms used by farmers c. 1900. Here variants Nr 10–17 are the variants of the term *osztováta* (*szováta, eszváta, esztováta, esztaváta, eszteváta, isztováta, asztaváta*), Nr 1–3 are the variants of the term *szövőszék* (*szüvőszék, szüszék*), Nr. 4–6 are the variants of the term *szövőfa* (*szövő, szüöfa*), Nr 8–9 are the variants of the term *szátva* (*szátyva*) and Nr. 7 is a totally different term: *fajz*.

[15] I. e. map Nr 151: Call terms for driving off geese.

CONCLUSIONS

Although computer elaboration is thought to be precise and exact, we have seen that because of the special character of mapping, the inconsistencies, and the mixing of different points of view in the Atlas of Hungarian Folk Culture the correctness of our attempt to draw the cultural regions of the Hungarians based on the data of the Atlas leaves much room for improvements. These can be summarized as follows:

1. We could improve our correctness by taking all the characteristic cultural phenomena into account, as those shown in the Atlas are a bit unilateral.

2. We could draw the border of the different cultural regions more precisely if we took samples more often, which means that we should make investigations in all settlements of the Hungarian-speaking area, which is of course impossible. The correctness of cultural regions is inevitably influenced by the geographical-historical point of view, too, no matter how we try to concentrate on cultural phenomena, as the settlements appearing in the Atlas were chosen so that they represent all the smaller regions of the whole Hungarian-speaking area, and those smaller regions were partly defined by the historical-geographical approach.

3. The correctness of the analysis decreases because the aspects of the Atlas and the computer elaboration cannot be reconciled without difficulties. (The problematic points are the following: the analysis of numerical values or characters, the collation of the inconsistent definition of the values of certain variables, the problem of missing data and multiple data.)

The analysis based on the data of the Atlas is mechanical and not elaborate enough to determine small, regional distinctions of the culture; it can determine only rough territorial differences. Because of all these problems, this kind of determination of cultural regions cannot replace the previous "mutually agreed" cultural-historical-geographical determination of regions, but it can offer a good frame of reference to define the regions more precisely.

LITERATURE

BARABÁS, Jenő (ed.)
1987–92: Magyar Néprajzi Atlasz. Atlas of Hungarian Folk Culture. Budapest, Akadémiai.

BORSOS, Balázs
1995: Ecosystem – Geographic Area – Economic Region: A Computer Analysis of the Environmental and Economic Changes in the Bodrogköz, NE Hungary in the Second Part of the 19th Century. Acta Ethnographica Hungarica 40 (1–2): 131–184.

KÓSA, László
1975: Néprajzi csoportok és tájak a magyar népismeretben. In: KÓSA, László–FILEP, Antal: A magyar nép táji történeti tagolódása. Budapest, Akadémiai. 7–51.