Ethnic-specific and Ethnic-nonspecific Factors for Ethnicity Non-identification in Bulgaria

KALOYAN HARALAMPIEV (k_haralampiev@phls.uni-sofia.bg)
DIMITAR BLAGOEV (dblagoev@phls.uni-sofia.bg)
Sofia University "St. Kliment Ohridski", Bulgaria

In this paper we present analysis of ethnic-specific and ethnic-nonspecific factors for ethnicity non-identification in 2011 Bulgarian Census. Using cluster analysis and analysis of associations we study the levels of dependencies between the share of ethnicity non-identification and the following socio-demographic and socio-economic factors on regional level: age; household size; juvenile maternity; in-country migration; cross-border migration; education; and the structure of economic activity, employment and unemployment. Our analyses show that ethnicity non-identification is, at large, not a question of individual decision based on the right of non-self-identification and its practicing, but is dependent – although at different extent in different cases – to regionally-specific sets of socio-demographic and socio-economic factors, which turn out to be ethnic-specific and ethnic-nonspecific.

Keywords: ethnicity non-identification, ethnic-specific factors, Census, cluster analysis, analysis of associations

Introduction

The issue of ethnicity non-identification in Bulgaria occurred after 2011 Census, when, for the first time in its 130-years' history, 9.3 percent of population did not answer the voluntary question on ethnicity. In our previous papers on that topic we discussed several factors for the lack of ethnic self-identification related to the specificity of ethno-cultural composition on regional and local level (Blagoev and Haralampiev, 2014; Haralampiev and Blagoev, 2014). When developing our argument we faced new research questions which framed the formulation of new hypotheses to be verified.

In this paper we present the further development of our analysis focusing on ethnic-specific and ethnic-nonspecific factors for ethnicity non-identification in Bulgaria. Using cluster analysis and analysis of associations we study the levels of dependencies between the share of the lack of ethnic identification and the following socio-demographic and socio-economic characteristics (potential factors[1]) on regional (district) level: age; education; household size; juvenile maternity (12–19 yrs); internal migration; cross-border migration; economic activity, and employment and unemployment. The choice to focus on theses characteristics is backed by the following prerequisites.

[1] We understand quite clearly that data and methods we use here do not allow us to prove causality in our analysis. A discovered strong correlation does not necessary mean it is a causal one. On the other hand, however, the lack of relationship represents the lack of causal relationship. Therefore we use the term "potential/ presumable factors" in order to denote we have revealed a relation, be it causal or not.

First, our already achieved knowledge about the lack of ethnic identification has been obtained only through analysis of publicly available Census data and has reached its limits. This situation has necessitated proceeding further with a deeper investigation within Bulgarian Census database, which however appeared to be an impossible endeavor because of the restrictions over public access to it. Therefore we undertook an intermediate research strategy for carrying on a data mining, in which we used two types of empirical data – data derived from Census publications of the Bulgarian National Statistical Institute, and data received against payment. In the first case it became necessary to overcome obstacles related to the user-not-friendly type of data publication[2], while in the second case the financial limitations restricted the data range (that is, the number of variables) we chose to buy.

Second, part of these characteristics (presumably factors for ethnicity non-identification) analyzed in this paper meet the necessity to answer new questions evolving during the course of our previous analyses. And hence our assumptions arise for possible correlation between the lack of ethnic self-identification and socio-demographic characteristics as age and education, as well as socio-economic characteristics as employment and unemployment.

Third, the rest of factors about lack of ethnic identification studied in this paper (for example, household size, juvenile maternity and migration) are included in the analysis as a result of heterogeneous reasons: Casual findings of particular correlations in some local cases (household size and juvenile maternity relations to ethnicity non-identification) or findings from our previous research experience when dealing with other empirical data, indicating a relation between migration and ethnic identification (Kabakchieva et al, 2011).

Rationale and Hypotheses

The essential reasons to opt for factors for ethnicity non-identification are as follows:

Age is chosen because in our preliminary research analyses and suppositions we related ethnicity non-identification in a greater extent with youngsters living in families with Bulgarian ethnicity (youth identity uncertainty; bias towards globalized identifications at the expense of ethnic ones) as well as in families with Roma ethnicity (with regard to the aggravating problems related to negative attitudes towards Roma during the last decades when socialization of Roma youth took place). Since a multi-component variable representing the age structure would dominate the pool of one-component variables and therefore would possibly lead to some kind of results' distortion, we will use "average age" here although we render a good account of its limitations and of the loss of the structural differentiation, which a multi-component variable of age holds.

Education is chosen within the frame of the following preliminary suppositions: First, that lower educated people would more frequently have difficulties with their ethnic identity in an increasingly globalizing and dynamising society, in which the coping with its increasingly entangling issues (including ethnic self-identification) necessitates a wider range of cultural and educational resources. Second, that people who previously identified themselves as Roma and whose educational capital is apparently weaker than that of all other ethnic groups, would possibly be more susceptible to escaping ethnic self-identification. Since education is a qualitative variable, its

[2] These data are presented as table distributions in PDF files, not amenable to any software elaboration.

usage in the form of educational structure would disrupt the methodological precision of our analysis. Therefore we decided to explore the well established measure of average schooling years (Barro & Lee, 1996).

The group of variables derivative of migration (migration during the period 2001-2011, recent migration from abroad one year prior to Census, recent internal migration one year prior to Census, and living abroad for more than one year during the period 1980-2011) is chosen in our investigation since we supposed that they are more related to the lack of ethnic identification on the following grounds. First, they imply personal dynamics and shift between different social spaces and such a social experience would possibly lead to a weakening of ethnic identity importance (which commonly stems from local communities' durability and stability) at the expense of other personal and social identifications. Second, the existing option to cope online with Census card[3] (that is, to avoid personal contact with the NSI interviewers and thus – verification of individual presence in the country) implied the possibility people living temporarily or permanently abroad to be registered online by their relatives who possibly omitted answering the ethnicity question for some ethical reasons.

The group of economic variables (economic activity, employment, and unemployment) is chosen on the grounds pertaining to education. On one hand, one can suppose that people from ethnic minority belonging (Turks, Roma, Pomaks) who are much more excluded from economic processes would possibly have ethnic identity difficulties by that same reason, since there is a possibility to perceive it as a main precondition for their economic exclusion. On the other hand, their economic troubles would possibly dominate their self-identification stronger than its ethnical dimension. We use economic activity rate, employment rate, and unemployment rate in this paper as measures of economic activity, employment, and unemployment respectively.

And, finally, the two socio-demographic variables of household size and juvenile maternity (12–19 yrs) are included in our analysis as presumable factors for the lack of ethnic identification for a common single supposition – that they both are strongly ethnically biased, particularly with respect to Roma minority, in which identity difficulties would possibly lead to a bigger share of ethnicity non-identification compared to other ethnic groups.

Following the above discussion we built several research hypotheses to be verified in our analysis.

Hypothesis #1. Socio-demographic characteristics "age" and "education" are ethnically specific factors for the lack of ethnic identification.

Hypothesis #2. Migration in its broader meaning is ethnically non-specific factor for the lack of ethnic identification.

Hypothesis #3. Socio-economic characteristics "economic activity", "employment", and "unemployment" are ethnically specific factors for the lack of ethnic identification.

Hypothesis #4. Socio-demographic characteristics "household size" and "juvenile maternity" are ethnically specific factors for the lack of ethnic identification.

[3] About 40 percent of Bulgarian population used this option.

Methodological Framework

We use an original methodological derivative of the standard cluster analysis (Haralampiev et al., 2015: 9). "Cluster analysis" is a generic name for a number of particular procedures, which are applied for obtaining homogeneous groups based on several quantitative variables. The point is to group the most similar observations (in this case – districts) in one cluster, while keeping dissimilar observations in other clusters. Thus we transform the social typological similarity or difference into a special proximity or distance, respectively. Since all variables are quantitative, each observation could be represented as a point in the multidimensional space[4], therefore the proximity and remoteness between them could be measured in a systematic way. In IBM SPSS Statistics there are seven specific algorithms for hierarchical clustering and two algorithms for non-hierarchical clustering which we have used at the beginning.

The seven hierarchical clustering methods are: Average linkage (Between groups), Average linkage (Within groups), Nearest neighbor (Single linkage), Furthest neighbor (Complete linkage), Centroid method, Median method and Ward's method. The difference between them is the way of measuring of the distances between the clusters (SPSS Inc., 2003: 6-5).

The two non-hierarchical clustering methods are: K-means clustering and Two-step clustering. The specific feature of their algorithms is that we have to define *a priory* the number of the clusters. For that reason we start our analysis with the hierarchical clustering methods, then we use the so called dendrogram to identify the number of the clusters and then we perform the non-hierarchical clustering methods with the identified number of the clusters.

Next we have to test the sustainability of the analytical results beyond the variations in the procedures, thus proving that the particular empirical result represents the quality of the studied object but not an instrumental artifact. For this purpose we follow the analytical approach of Vicente and Reis (2007: 4-6). They have devised a testing procedure – first, they apply cluster analysis by four clustering procedures. Then, they compare the obtained results by the contingency coefficient. As noted above, instead of four we compare the results of nine clustering procedures.

The best matches are taken as "best results" following the assumption that, if there is an objective reality approximately equally described by two – or more – different methods, these methods must be the right ones to provide realistic results.

Cluster Analysis

As a first step we conducted a cluster analysis and formed clusters of the 28 Bulgarian administrative regions (districts) according to the following variables on district level:

- Average age of every ethnic group and of those who have not answered the question about their ethnic identification;

- Average schooling years of every ethnic group and of those who have not answered the question about their ethnic identification;

[4] Since our "cluster variables represent entirely different scales" we made standardization before clustering (SPSS Inc., 2003: 6-8).

- Economic activity rates of every ethnic group and of those who have not answered the question about their ethnic identification;

- Employment rates of every ethnic group and of those who have not answered the question about their ethnic identification;

- Unemployment rates of every ethnic group and of those who have not answered the question about their ethnic identification;

- Share of migration (internal and from abroad) during the period 2001-2011 of every ethnic group and of those who have not answered the question about their ethnic identification;

- Share of recent internal (within-country) migration one year prior to Census of every ethnic group and of those who have not answered the question about their ethnic identification;

- Share of recent migration from abroad one year prior to Census of every ethnic group and of those who have not answered the question about their ethnic identification;

- Share of living abroad for more than one year during the period 1980-2011 of every ethnic group and of those who have not answered the question about their ethnic identification;

- Average size of household of every ethnic group and of those who have not answered the question about their ethnic identification;

- Share of juvenile maternity (12–19 yrs) of every ethnic group and of those who have not answered the question about their ethnic identification.

Using the Average linkage (Between groups) method of hierarchical clustering we reveal two clusters plus three separate districts. Average linkage (Within groups) also shows two clusters but there is only one separate district.

The hierarchical method of Nearest neighbor (Single linkage) groups the most of the districts in one single mega cluster plus two separate districts. Thus we consider that this method fails to reveal clusters because it treats almost all districts as typologically homogeneous. The Centroid and Median hierarchical methods also fail here. The hierarchical method of Furthest neighbor (Complete linkage) and Ward's hierarchical method both reveal four clusters.

Thus we have two hierarchical methods which show two clusters and two ones which show four clusters. For that reason we perform the K-means and Two-step non-hierarchical clustering twice – first, with two clusters, and secondly, with four clusters. After that we compare the obtained results.

Table 1. Comparison of the cluster membership between methods which reveal two clusters (contingency coefficients)

	Average Linkage (Between Groups)	Average Linkage (Within Group)	K-means Cluster	TwoStep Cluster
Average Linkage (Between Groups)		0,733	0,245	0,707
Average Linkage (Within Group)	0,733		0,650	0,575
K-means Cluster	0,245	0,650		0,188
TwoStep Cluster	0,707	0,575	0,188	

Table 2. Comparison of the cluster membership between methods which reveal four clusters (contingency coefficients)

	Complete Linkage	Ward Method	K-means Cluster	TwoStep Cluster
Complete Linkage		0,820	0,734	0,749
Ward Method	0,820		0,774	0,807
K-means Cluster	0,734	0,774		0,785
TwoStep Cluster	0,749	0,807	0,785	

The best match is between the method of Furthest neighbor (Complete linkage) and Ward's method. For that reason we considered these two methods as the best ones for implementation of our analysis. On the other hand the Ward's method showed bigger similarity with the other two non-hierarchical methods compared to Complete linkage. Thus we chose the Ward's method as the method describing the clusters better, which allowed us to distinguish four separate clusters of all 28 Bulgarian administrative districts, as follows:

First cluster comprises of the following eight districts: *Blagoevgrad, Dobrich, Pazardzhik, Razgrad, Silistra, Sliven, Shumen, and Targovishte*.

Second cluster comprises of the following 16 districts: *Burgas, Gabrovo, Haskovo, Kyustendil, Lovech, Montana, Pleven, Plovdiv, Russe, Sofia district, Stara Zagora, Varna, Veliko Tarnovo, Vidin, Vratza, and Yambol*.

Third cluster comprises of the following two districts: *Kardzhali and Smolyan*.

Fourth cluster comprises of the following two districts: *Pernik and Sofia city*.

The next step of our analysis is to range the variables according to their ability to distinguish the clusters. Therefore we use the coefficient of determination (eta-squared) which shows the variance between the clusters as a share of the total variance. The other part of the total variance is the variance within the clusters. Since we have chosen the Ward's method an additional advantage is that the "Ward's method creates clusters that yield the smallest possible within cluster variance" (SPSS Inc., 2003: 6-5). In that way the coefficients of determination will be the largest compared to the other methods of clustering.

The variables used to implement the cluster analysis are an outcome of the intersection between the six options of the question about ethnicity[5], on one hand, and 12 other parameters, on the other hand (the 11 socio-demographic and socio-economic characteristics specified above as well as the percentage of the ethnic groups). The resultant 72 composite variables distinguish to a different extent the four clusters of the 28 districts.

In the Table 3 below we range only those variables, which values of eta-squared exceed 0.5, which means that differences between clusters are bigger than within them. Resulting data show that there are 16 composite variables meeting that condition and they, although with a different strength, better differentiate the four clusters.

Table 3. Ranging of composite variables, which differentiate the four clusters

	Eta Squared
1. Share of migrated persons (internal and from abroad) during the period 2001-2011 of Turkish ethnicity	,771
2. Share of those who have lived abroad for more than one year during the period 1980-2011 of Turkish ethnicity	,732
3. Average schooling years of those who did not answer the ethnicity question	,701
4. Average schooling years of those who are of other ethnicity	,685
5. Average household size of those who did not answer the ethnicity question	,670
6. Average household size of those who are of other ethnicity	,631
7. Percentage of Bulgarian ethnicity	,597
8. Employment rate of those who did not answer the ethnicity question	,590
9. Percentage of those who have not self-identified themselves	,554
10. Share of those who have lived abroad for more than one year during the period 1980-2011 of other ethnicity	,532
11. Percentage of those who have not answered the ethnicity question	,531
12. Unemployment coefficient of those who have not answered the ethnicity question	,529
13. Average household size of Turkish ethnicity	,518
14. Share of those who have lived abroad for more than one year during the period 1980-2011 and have not self-identified themselves	,516
15. Economic activity rate of those who have not answered the ethnicity question	,512
16. Average age of those who have not answered the ethnicity question	,502

At first sight these results show a quite variegated picture, however a more focused analysis could make it clearer. Let us investigate the difference between the *observed* frequency of appearance of the input six (ethnic) variables and input 12 (economic and demographic) variables in the above set of composite parameters, which highly differentiate the clusters, on one hand, and their *expected* ("average") frequency, on the other hand. We can hold that those variables, which appear more than it is expected, contribute more than others to the substantive differentiation between clusters, that is, they have a greater weight in formation of essentially different groups of districts (Tables 4 and 5 below).

[5] That is, the four varieties of ethnic belonging – Bulgarian, Turkish, Roma, Other, as well as the two excluding options "I do not identify myself" and "No answer" (we deal with the latter in this paper).

Table 4. Observed and expected frequencies of the input ethnic variables (counts)

	Observed frequency	Expected frequency	Difference
"No answer"	7	2,67	4,33
Turkish ethnicity	3	2,67	0,33
Other ethnicity	3	2,67	0,33
Have not identified ethnically	2	2,67	-0,67
Bulgarian ethnicity	1	2,67	-1,67
Roma ethnicity	0	2,67	-2,67

We can assert that there is only one out of six input ethnic variables, which appearance is more than expected and this is the lack of answer on the question about ethnicity identification. That is, according to the cluster analysis' outcome, but not to our biased presumption, this is exactly the lack of ethnic identification that appears to be the main variable distinguishing the four clusters. This outcome is a strong argument for focusing our further analysis on these four clusters differentiated on the basis of ethnicity non-identification.

Table 5. Observed and expected frequencies of the input socio-economic and socio-demographic variables (counts)

	Observed frequency	Expected frequency	Difference
Share of those who have lived abroad for more than one year during the period 1980-2011	3	1,33	1,67
Average household size	3	1,33	1,67
Percentage of the ethnic group	3	1,33	1,67
Average schooling years	2	1,33	0,67
Share of migrated persons (internal and from abroad) during the period 2001-2011	1	1,33	-0,33
Average age	1	1,33	-0,33
Economic activity rate	1	1,33	-0,33
Employment rate	1	1,33	-0,33
Unemployment rate	1	1,33	-0,33
Share of recent migrated persons from abroad one year prior to Census	0	1,33	-1,33
Share of recent internal migrated persons one year prior to Census	0	1,33	-1,33
Share of juvenile maternity (12–19 yrs)	0	1,33	-1,33

With regard to the other group of 12 input variables, three of them appear with a bit larger than expected frequency, however only two of them are meaningful[6] – the share of those who have lived abroad, and the average household size. These two socio-demographic variables have slightly bigger influence on clusters' differentiation; this interim result needs additional attention further in our analysis.

[6] We exclude the third variable "Percentage of the ethnic group" from our analysis, since it does not hold any substantive meaning and therefore could not be interpreted in terms of our hypotheses.

Regression analysis

As a second step we conducted linear regression analysis with a dependent variable the rate of those who have not identified themselves ethnically, and the following independent variables:

- The difference between average ages of those with ethnic identification and those without ethnic identification; for every district
- The difference between average schooling years of those with ethnic identification and those without ethnic identification; for every district
- The difference between economic activity rates of those with ethnic identification and those without ethnic identification; for every district
- The difference between employment rates of those with ethnic identification and those without ethnic identification; for every district
- The difference between unemployment rates of those with ethnic identification and those without ethnic identification; for every district
- The difference between shares of migrated (internal and from abroad) people during the period 2001-2011 with ethnic identification and those without ethnic identification; for every district
- The difference between shares of people with ethnic identification who had lived abroad for more than one year during the period 1980-2011 and those without ethnic identification; for every district
- The difference between shares of recently internally migrated people one year prior to Census with ethnic identification and those without ethnic identification; for every district
- The difference between shares of recently migrated people from abroad one year prior to Census with ethnic identification and those without ethnic identification; for every district
- The difference between average size of household of those with ethnic identification and those without ethnic identification; for every district
- The difference between shares of young mothers (12–19 yrs) with ethnic identification and those without ethnic identification; for every district

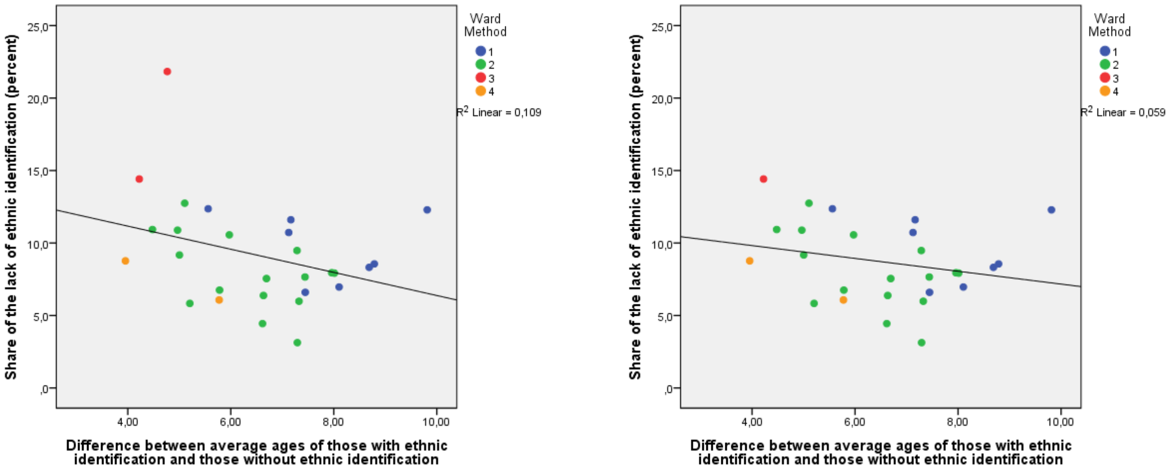
Here in this stage of the analysis we focus on the differences between every single socio-economic and socio-demographic characteristic of two groups of people – those who identify ethnically themselves and those without ethnic identification. Possible existence of such differences between these two groups will signify that ethnicity non-identification is not a question of individual decision based on the right to skip the question on one’s ethnic identity, but is dependent on a set of aforementioned factors.

We integrated the results of cluster and regression analyses and presented on a scatter plot each district as a separate item according to the values of the dependent and independent variable, and then we marked all different points of different colors according to their cluster membership. The results are presented on the *first* figures on every row below. Graphs with distribution of clusters according to the difference for every district between the shares of young mothers (12–19 yrs) with ethnic identification and those without ethnic identification are excluded because this appeared to be the only independent variable, which lacks correlation.

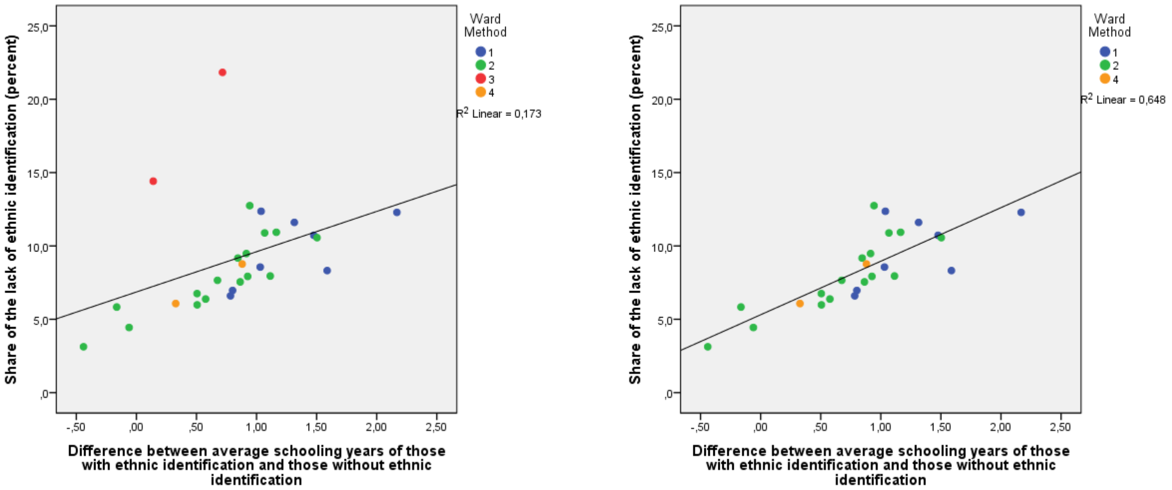
Our analysis shows that the lack of ethnic identification increases in parallel with the rise of age, unemployment, and household size and with the decrease of education, economic activity, employment, and migration in its diversity (i.e., living abroad for more than one year during the period 1980-2011, recent migration from abroad one year prior to Census, migration – internal and from abroad – during the period 2001-2011, and recent internal migration one year prior to Census). Since it is an interim outcome, we can make only some preliminary considerations at this point, that the less active, less mobile, less educated, and less autonomous a person is, the more likely he/she lacks ethnic self-identification. This proposition could be further detailed saying that this inverse ratio is dependent on the district he/she lives. The latter however is not entirely valid since a deeper consideration of clusters' distributions and cluster memberships' distributions visualized on all first graphs on every row below reveals two apparent exceptions, namely, the districts of Smolyan and Kardzhali.

We assumed that their "extreme behavior" of outliers influences both the line's slope and the strength of the relationship, and therefore we decided to repeat the procedure of combining the cluster and regression analysis without taking these two regions into account. For identification of the outliers we use the standardized residuals. When the residuals are normally distributed the probability a single standardized residual to be outside the interval -2/+2 is 4.6%. In the case of 28 districts the expected number of the standardized residuals outside the interval -2/+2 is 1, i.e. if there are two or more districts with standardized residual outside this interval we consider them as outliers. The strongest criterion is to choose the interval -3/+3. Then the probability a single standardized residual to be outside the interval -3/+3 is 0.3%. In the case of 28 districts the expected number of the standardized residuals outside the interval -3/+3 is 0, i.e. if there is even one district with standardized residual outside this interval we consider it as outlier. (If we choose the interval -3/+3 then only district of Smolyan is outlier.)

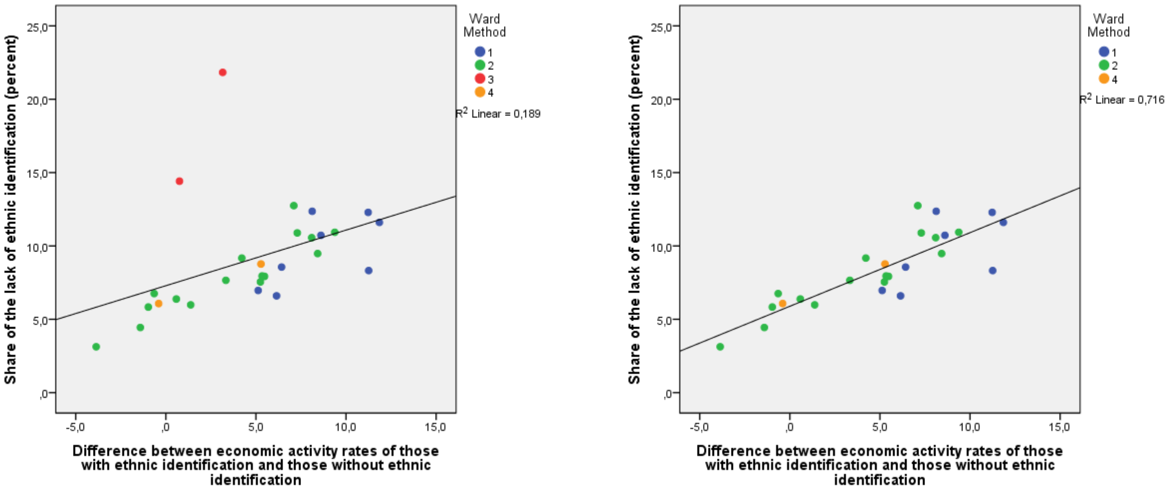
The results obtained are presented on all *second* graphs on every row below.



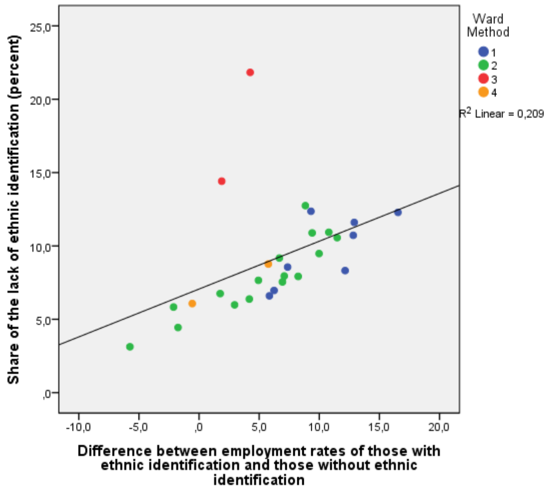
The district of Smolyan (outlier) is excluded from the analysis on the second graph above.

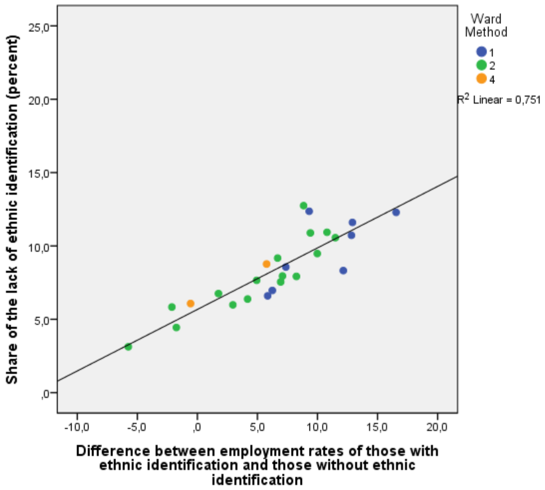


The districts of Smolyan and Kardzhali (outliers) are excluded from the analysis on the second graph above.

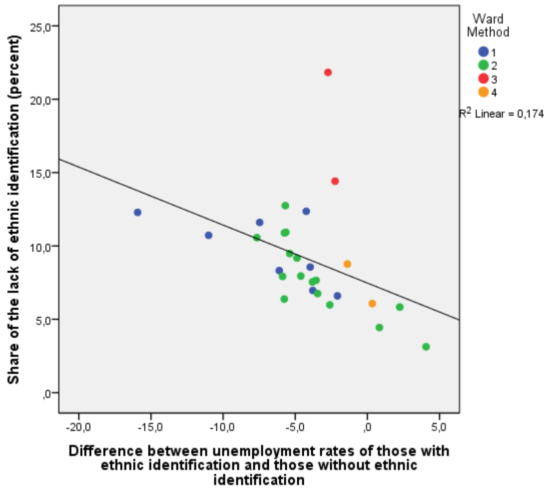


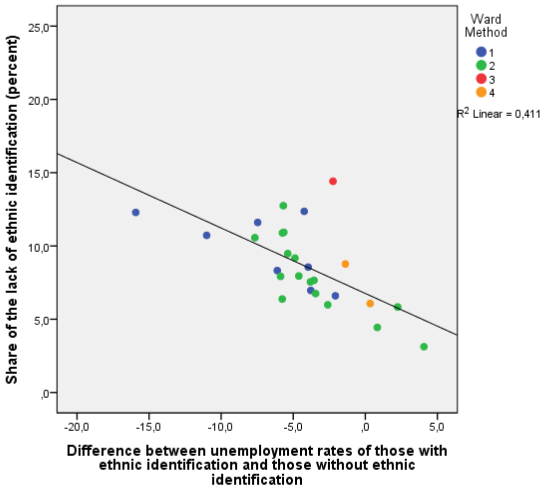
The districts of Smolyan and Kardzhali (outliers) are excluded from the analysis on the second graph above.



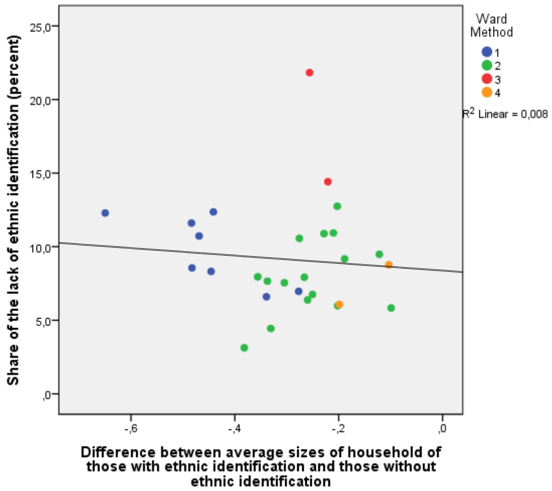


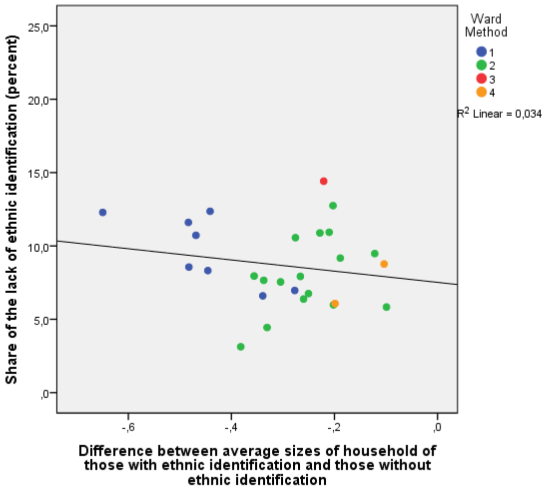
The districts of Smolyan and Kardzhali (outliers) are excluded from the analysis on the second graph above.



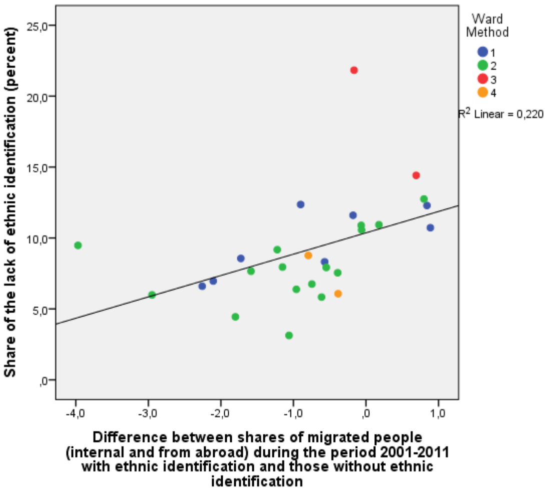


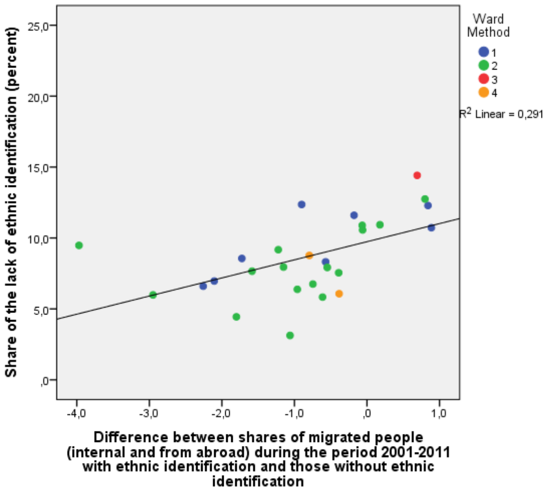
The district of Smolyan (outlier) is excluded from the analysis on the second graph above.



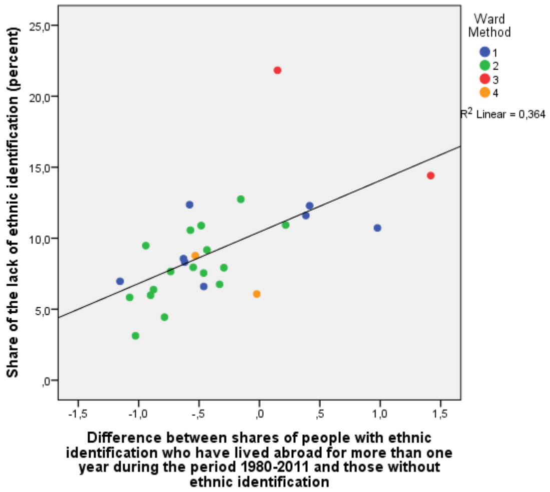


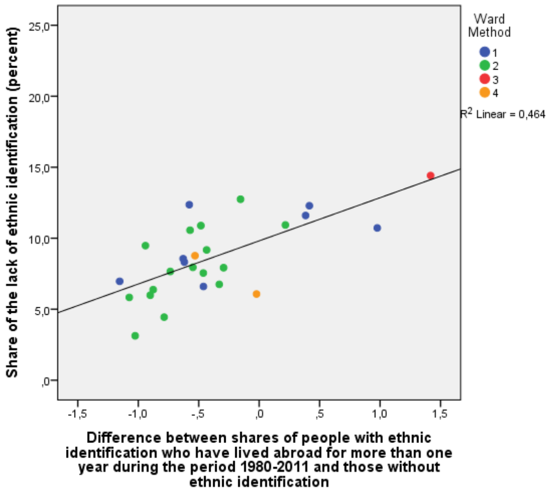
The district of Smolyan (outlier) is excluded from the analysis on the second graph above.



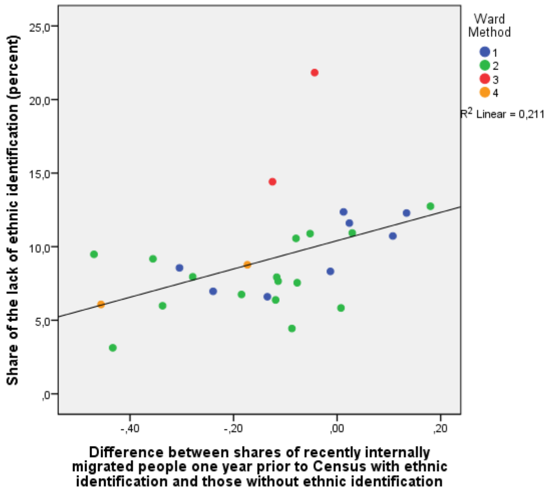


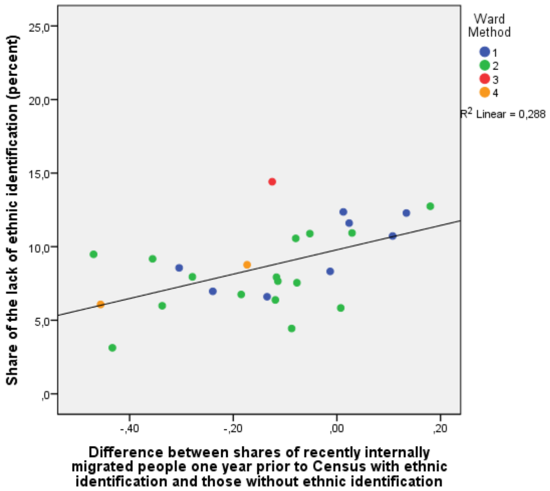
The district of Smolyan (outlier) is excluded from the analysis on the second graph above.



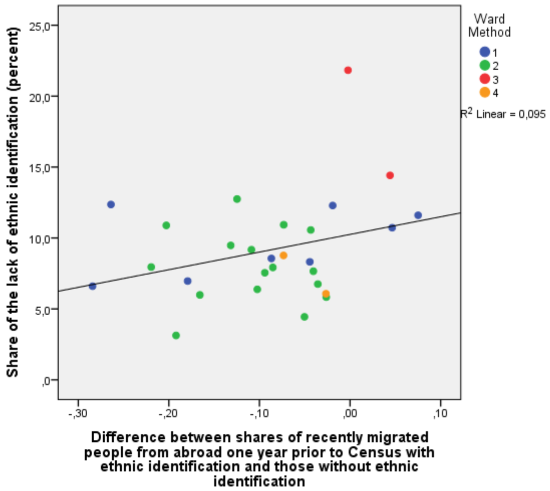


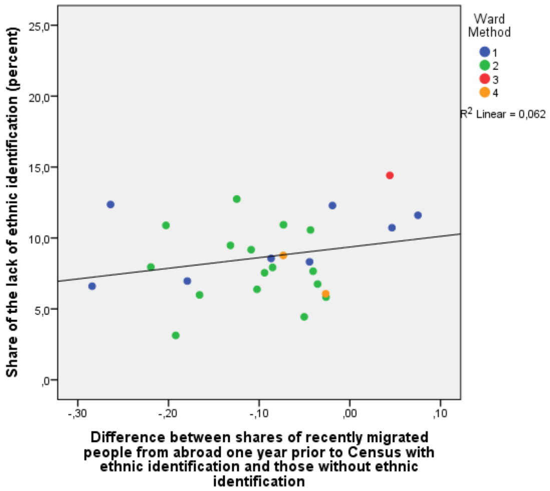
The district of Smolyan (outlier) is excluded from the analysis on the second graph above.





The district of Smolyan (outlier) is excluded from the analysis on the second graph above.





The district of Smolyan (outlier) is excluded from the analysis on the second graph above.

There is a change in the strength of relation between the dependent variable „lack of ethnic identification” and the set of independent socio-demographic and socio-economic characteristics for every district (visualized on the *second* graphs at every row above). The strength increases in the cases of economic activity, employment, and (to a lesser extent) the education, and decreases in the cases of age and (to a lesser extent) the migration from abroad one year prior to Census, and, finally, has no particular change in the cases of unemployment, household size, migration (internal and from abroad) during the period 2001-2011, recent internal migration one year prior to Census, and the living abroad for more than one year during the period 1980-2011. If we focus, however, only on the more noticeable increase of relationship's strength in the background of exclusion of the outlying districts of Smolyan and Kardzhali, then it is valid only in the cases of substantively interrelated variables of economic activity and employment. Put it differently, these two exceptional cases stray from the clear "economic relationship of non-identification" pattern, that is, economic activity and employment are not factors for the lack of ethnic identification for them.

Moreover, Smolyan analyzed alone as an outlier is the only district standing apart from other 27 districts on all studied socio-demographic and socio-economic factors. More importantly, it is among the fewer districts with the smallest differences between ethnically identified and ethnically not identified in terms of age, schooling years, economic activity, employment and unemployment. What is more, there are negligible differences between the two groups on all migration variables (migration 2001-2011, living abroad 1980-2011, recent internal migration, and recent migration from abroad). Therefore, the lack of answer on ethnicity question, which share is the biggest one exactly in Smolyan district, is conditioned by factor(s) different from the dozen of socio-demographic and socio-economic factors studied in this paper. (The latter outcome is in congruence with a supposition made in our previous papers (Blagoev and Haralampiev, 2014; Haralampiev and Blagoev, 2014) that the lack of answer on ethnicity question in Smolyan district is mostly a result of particular self-identification problem of an entire ethno-demographic group – the so called "Pomaks" – who lives there. The argument for this precondition, however, lies outside our task here.)

The discussion above takes us to the interim conclusion that some of correlations are weaker than others, which raises the question how the correlations between ethnicity non-identification and the set of socio-demographic and socio-economic characteristics are arranged.

For the purpose of ranging the correlations according to their strength we use the values of the coefficient of determination R².

Table 6. Ranging of the correlations between the dependent variable "lack of ethnicity identification" and 11 independent socio-demographic and socio-economic variables

	R²
Difference between employment rates of those with ethnic identification and those without ethnic identification	0,751
Difference between economic activity rates of those with ethnic identification and those without ethnic identification	0,716
Difference between average schooling years of those with ethnic identification and those without ethnic identification	0,648
Difference between shares of people with ethnic identification who had lived abroad for more than one year during the period 1980-2011 and those without ethnic identification	0,464
Difference between unemployment rates of those with ethnic identification and those without ethnic identification	0,411
Difference between shares of migrated people (internal and from abroad) during the period 2001-2011 with ethnic identification and those without ethnic identification	0,291
Difference between shares of recently internally migrated people one year prior to Census with ethnic identification and those without ethnic identification	0,288
Difference between shares of recently migrated people from abroad one year prior to Census with ethnic identification and those without ethnic identification	0,062
Difference between average age of those with ethnic identification and those without ethnic identification	0,059
Difference between shares of young mothers (12–19 yrs) with ethnic identification and those without ethnic identification	0,048
Difference between average size of household of those with ethnic identification and those without ethnic identification	0,034

As a result of the implementation of the analytical steps so far, we found that the employment rate, economic activity rate, and average schooling years are strongly correlated with the lack of ethnic identification. There is a moderate correlation between the latter and two other variables – the living abroad for more than one year during the period 1980-2011 and the unemployment rate. The correlation is weaker between migration (internal and from abroad) during the period 2001-2011 and recent internal migration one year prior to Census, on one hand, and the lack of ethnicity identification, on the other hand. And, finally, the other four variables (recent migration from abroad one year prior to Census, average age, juvenile maternity, and average household size) show negligible correlation with ethnicity non-identification.

This interim analytical result about correlations between the dependent variable "lack of ethnicity identification" and the 11 independent socio-demographic and socio-economic variables brings us nearer to the verification of our hypotheses. The next step is to evaluate the extent to which these factors for non-identification are ethnically specific. For that purpose we use Pearson coefficient in order to analyze the correlation between the shares of main ethnic groups (Bulgarian, Turkish and Roma) within population, on one hand, and the 11 socio-demographic and socio-economic variables, on the other hand (Table 7 below).

Table 7. Correlation between the shares of main ethnic groups (base: all) and the 11 socio-demographic and socio-economic variables

	Pearson correlation Bulgarian ethnicity	*Level of BUL ethnic specificity (Rank)*	Pearson correlation Turkish ethnicity	*Level of TUR ethnic specificity (Rank)*	Pearson correlation Roma ethnicity	*Level of ROM ethnic specificity (Rank)*	**Cumulative level of ethnic specificity**
Average household size	-,738	*High (I)*	,644	*High (I)*	,126	*Low (IX)*	**High ***
Average schooling years	,706	*High (II)*	-,605	*High (II)*	-,417	*Medium (IV)*	**High to medium**
Economic activity rate	,604	*High (III)*	-,575	*High (III)*	-,320	*Medium (VI)*	**High to medium**
Employment rate	,587	*High (IV)*	-,498	*Medium (IV)*	-,427	*Medium (III)*	**High to medium**
Unemployment rate	-,475	*Medium (V)*	,335	*Medium (VI)*	,478	*Medium (II)*	**Medium**
Average age	,469	*Medium (VI)*	-,374	*Medium (V)*	,059	*No (XI)*	**Medium ****
Share of people living abroad >1y. (1980-2011)	-,243	*Low (VII)*	,332	*Medium (VII)*	-,327	*Medium (V)*	**Low to medium**
Share of young mothers (12–19 yrs)	,113	*Low (VIII)*	-,226	*Low (VIII)*	,745	*High (I)*	**n.a.**
Share of migrated people from abroad 1y. prior to Census	-,059	*No (IX)*	,061	*No (IX)*	,176	*Low (VIII)*	**No ***
Share of internally migrated people 1y. prior to Census	,027	*No (X)*	-,040	*No (XI)*	,118	*Low (X)*	**No ***
Share of migration – internal & from abroad (2001-11)	,013	*No (XI)*	,044	*No (X)*	-,186	*Low (VII)*	**No ***

* With the exception of Roma (Low)

** With the exception of Roma (No)

Analysis of data in the above table reveals two clear mirror-patterns of ethnic (Bulgarian and Turkish) specificity of the studied variables, having a common ranking of the latter ones according to their correlation to ethnic identification. A distinctive feature of both patterns is that those variables, which are positively correlated with Bulgarian ethnicity, are negatively correlated with Turkish ethnicity, and vice versa. There is one (Roma) exception from that principle. Although it differs from the mirror-patterns on three out of 11 variables and could be considered a mixture of both, its overall correlation structure resembles their composition. These results make it possible to form three consistent groups of variables and three exceptions outside them.

The first group comprises of variables with a very strong or strong correlation with ethnic identification (that is, *they are ethnic-specific variables*): Average schooling years, economic activity rate, and (to a lesser extent) employment rate. Education and economic participation are highly positively correlated with Bulgarian self-identification, and negatively correlated with Turkish and Roma ethnicity (i.e. the lower the education and economic participation, the higher the possibility of belonging to the latter two ethnic groups), which confirms the well-established knowledge about the wider social exclusion of ethnic minorities in Bulgaria.

The second group comprises of variables with a moderate correlation with main ethnic groups (that is, *they are partly ethnic-specific variables*): Unemployment rate and (to a lesser extent) the share of people who have lived abroad for more than one year during the period 1980-2011. These two variables are quite heterogeneous and it is difficult to find a common social logic behind them. However, if we examine the first variable (unemployment rate) as an intermediate case between the first and third group (see below), then it appears as a representative of socio-economic variables but having a smaller correlation with ethnicity. By the same token we may consider the second variable here (longer living abroad 1980-2011) as a representative of the group of migration variables (see below) but with a higher correlation with ethnicity. Therefore we may conclude that the second group is an intermediate quasi-group between the ethnic-specific social and economic variables and ethnic-nonspecific migration variables.

The third group comprises of variables with very low or missing correlation with ethnic identification (that is, *they are ethnic-nonspecific variables*) and is dominated by "migration" quality: Share of recently migrated people from abroad one year prior to Census, share of recently internally migrated people one year prior to Census, and the share of migrated people (internal and from abroad) during the period 2001-2011.

And the three exceptions from the above correlation structure are the *demographic variables* household size, average age, and the share of young (12–19 yrs) mothers.

Household size is a variable highly negatively correlated with Bulgarian ethnicity, and highly positively correlated with Turkish ethnicity, with the remarkable exception of the low positive correlation with Roma ethnicity. If taken alone, this exception would raise questions about the empirical data validity and reliability, since the average Roma household size is greater than that of Bulgarian households – 4.3 members and 2.5 members respectively (NSI, 2011). Since, however, we use a relative measurement on the background of interrelation and mutual dependence between the positions within the entire ethnic structure, table data implies that the Roma household size falls approximately in the middle between Bulgarian and Turkish household sizes. With some reservation we ascribe that variable a high cumulative level of ethnic specificity taking into account the Roma exception.

Average age is moderately and positively correlated with Bulgarian ethnicity, and moderately and negatively with Turkish ethnicity, with the exception of the lack of correlation with Roma self-identification. The above explanation about the relative nature of our measurements could be also applied to this exception, with the additional argument about the weakness of the variable itself, since it levels out the inner composition of Roma age structure, which is biased to the lower age groups. And again, with some reservation, we ascribe that variable a medium cumulative level of ethnic specificity taking into account the Roma exception.

The third exceptional case is the opposite of the two already described. Juvenile maternity is weakly correlated with Bulgarian and Turkish ethnicity but is very strongly correlated to Roma ethnic group. However, the variable as a whole is difficult to evaluate in terms of ethnic specificity due to its heterogeneous composition.

Since the analysis so far revealed that only a part of these variables behave as factors for ethnic non-identification, then the final step towards verification of our hypotheses will be to study the linkage between the strength of their correlation with this non-identification and the level of their ethnic specificity. For that purpose we integrate the outcomes of both analyses in order to make our final conclusions (see Table 8 below).

Table 8. Relation between the strength of correlation of the 11 socio-demographic and socio-economic factors with non-identification, and factors' level of ethnic specificity

	Strength of correlation with non-identification	**Level of ethnic specificity**
Average schooling years	Strong	High to medium
Employment rate	Strong	High to medium
Economic activity rate	Strong	High to medium
Unemployment rate	Moderate	Medium
Living abroad > 1 year (1980-2011)	Moderate	Low to medium
Internal & external migration (2001-2011)	Weak	No
Internal migration one year prior to Census	Weak	No
Migration from abroad 1 year prior to Census	Negligible	No
Average age	Negligible	Medium*
Juvenile maternity (12–19 yrs)	Negligible	n.a.
Average household size	Negligible	High*

* With the exception of Roma (No)

We can conclude that in the cases of three out of 11 characteristics, namely, economic activity, employment and education, there is a concurrence between their strong correlation with non-identification and their relatively high level of ethnic specificity. There is a concurrence between moderate correlation of unemployment with non-identification and its medium level of ethnic specificity. These analytical outcomes almost entirely confirm our third hypothesis and the "education" part of our first hypothesis as well.

There is a partial concurrence between moderate correlation of the living abroad for more than one year during the period 1980-2011 with non-identification and its low to medium level of ethnic specificity – an outcome, which is insufficient to reject the "migration" hypotheses of ethnic-nonspecific character of non-identification. There is also concurrence between weak correlation of recent internal migration one year prior to Census and migration (internal and from abroad) during the period 2001-2011 with non-identification and their lack of ethnic specificity. Both outcomes confirm our second hypothesis as a whole.

And, finally, the fourth hypothesis about the ethnic specificity of the socio-demographic variables of household size and juvenile maternity is impossible to be verified, since the analytical outcomes are contradictory. The same is true also for the "age" part of our first hypothesis.

Conclusion

Our analyses in this paper show that ethnicity non-identification is, at large, not a question of individual decision based on the right of non-self-identification and its practicing, but is dependent – although at different extent in different cases – to regionally-specific *sets* of socio-demographic and socio-economic factors, which turn out to be ethnic-specific. These factors are economic activity, employment and unemployment, average schooling years, and – partially – the living abroad for more than one year during the period 1980-2011.

Since our analyses resulted in new and somewhat unexpected conclusions with regard to the factors for considerable share of the lack of ethnic self-identification on regional level in Bulgarian 2011 Census, both the hypotheses built and verified, and the questions raised in the course of analyses implementation need to be tested and answered in future analyses on the local level (that is, on municipality level) using methodology applied here.

Acknowledgements

We extend our special thanks to Mrs. Magdalena Kostova, Director of Demographic and Social Statistics Directorate of Bulgarian National Statistical Institute.

References

Barro, R. J., & Lee, J. W. (1996). International Measures of Schooling Years and Schooling Quality. *The American Economic Review, 86(2),* 218-223.

Blagoev, D. & Haralampiev, K. (2014). In Search of Lost Identity: Research Problems and Solutions. *Vanguard Scientific Instruments in Management,* Vol. 1(8): 16-50 (in Bulgarian)

Haralampiev, K. & Blagoev, D. (2014). Ethnicity Non-identification in 2011 Census in Bulgaria. *European Population Conference*, Budapest, 2014, http://epc2014.princeton.edu/abstracts/140592

Haralampiev, K., Dimitrov, G & Stoychev, S. (2015). 'Measuring Sociopolitical Distances between EU Member States and Candidates: A New Path', *MAXCAP Working Paper Series*, No. 15: 1-34

Kabakchieva, P., K. Haralampiev, D. Blagoev, & E. Stoykova-Doganova. (2011). *Report on Social Survey along the Route of Nabucco Gas Pipeline Project* (unpublished).

NSI (2011). *Census 2011, Vol. 1. Population, Book 4. Households*. Sofia (in Bulgarian)

SPSS Inc. (2003). *Advanced Statistical Analysis Using SPSS*.

Vicente, P. & Reis, E. (2007). Segmenting Households According to Recycling Attitudes in a Portuguese Urban Area, *Resources, Conservation and Recycling* 52 (1): 1-12.