An indirect approach to map ethnic identities in post-conflict societies

Daniel Bochsler*
NCCR Democracy, Universität Zürich
Affolternstrasse 56
CH- 8050 Zurich
Switzerland

E-mail: daniel.bochsler@uzh.ch
web: www.bochsler.eu
Tel. +41 44 634 54 53
Fax +41 44 634 52 00

* *corresponding author*

Basil Schläpfer
Centre for Democracy Studies Aarau (ZDA)
Küttigerstrasse 21
5000 Aarau
Switzerland

E-mail: basil.schlaepfer@uzh.ch

Ethnicity remains one of the most salient layers of individuals' social identities, and information about the distribution of ethnic identities turns out to be crucial for many studies that investigate political or social processes in divided societies. In the aftermath of civil wars, however, censuses providing such data are controversial and often delayed. Where census data is lacking, researchers can make use of the multiplier method to infer the distribution of ethnic identities based on indirect observations from sub-samples of the population. However, due to the selective nature of their data, the sub-samples might not be representative of the population. This paper proposes a method, which corrects for such selection effects. Based on the ethnic identity of birth-giving parents, we estimate the distribution of ethnic identities in the municipalities of the Federation of Bosnia and Herzegovina for 2008-10. We correct for possible selection biases by including economic, demographic and war related variables. Multiple tests of validity show that our estimation appears to be the most accurate procedure currently available for the distribution of ethnic identities in municipalities in Bosnia and Herzegovina.

Keywords: Ethnic politics; identities; Bosnia and Herzegovina; multiplier method.

Introduction

In post-war societies, population censuses can be highly politicised issues. Whether and how ethnicity should be subject to census questions is often controversial.[1] With good reason: asking citizens about their ethnicity contributes to the consolidation of historical categories, and can help to foster political divisions along ethnic lines (Horowitz, 1985, pp. 194-195; Kertzer & Arel, 2001; Szreter et al., 2004; Aitken, 2007). In consociational settings, quotas and the political control over the territory are sometimes directly linked to the officially recognised ethnic structure of the population (McCulloch, 2014; McCrudden & O'Leary, 2013; Williams & Husk, 2013, p. 297). Therefore, political elites might even be incentivised to mobilize a sufficiently large 'support' for their ethnic category in the official census. In such settings, census questions about ethnicity, language, or religion are politically controversial (Visoka & Gjevori, 2013), and censuses are often delayed (recently in Fiji or Macedonia) or not conducted at all (Lebanon).

At the same time, the lack of systematically collected data about the distribution of ethnic identities in the intermediate period is a major obstacle for academic investigations that would require this kind of data (Mitchell et al., 2009; Buhaug & Rød, 2006; Raleigh & Hegre, 2009). For example, ethnic identities can be related to clientelism (Franck & Rainer, 2012), and studies of political behaviour and radicalisation in post-conflict societies often require information on ethnic diversity (Pugh & Cobble, 2001; Belloni, 2007, pp. 183-184). The lack of data has consequences for academics and for policy makers. The distribution of identities can affect security and inter-ethnic cooperation (Fearon & Laitin, 1996; Varshney, 2001), and it is one main predictors of refugee returns (International Crisis Group, 1999; Nalepa, 2012, pp. 349, note 369). Therefore, research occasionally resorts to indirect proxies (Caspersen, 2004).

Where official data is lacking, indirect estimators, which do not gain the same legitimacy as official censuses, are used to map the distribution of ethnic identities. There is a possibly unlimited

[1] Cf. BBC News, 31 August 2010, "Kenya defends tribal census figures", http://www.bbc.co.uk/news/world-africa-11143914.

range of sources, which are and can be used as a proxy of ethnicity. Often, studies infer identities from the distribution of names in the population (Harris, 2015; International Crisis Group, 2003, pp. 6-7,15; Elliott et al., 2008; Mateos et al., 2011). Other useful indicators can include membership in religious communities (Susewind, forthcoming), relationship to kin communities, or the consumption of goods and services, which are typical for different ethnic communities. Della Vigna and colleagues even used beer brands sold in local bars as proxies of ethnicity (DellaVigna et al., 2014, pp. 112-113).

Such indirect indicators might in many contexts come close to a perfect predictor of ethnicity at the individual level. Nevertheless, such indirect indicators relate to a sub-sample of the overall population. Data on membership of religious communities or the consumption of consumer goods relate to church members or consumers, not to the entire population. While they might help to infer the identity of individuals, the membership rate in religious communities, or the per capita consumption might vary between ethnic groups, *and* between municipalities: e.g., church membership might be more frequent for group A than B, *and* more frequent in rural than in urban municipalities. Even register data (marriage registers, death or birth registers) (Elliott et al., 2008; Mateos et al., 2011) only capture a sub-sample of the population, e.g. newly married couples, rather than the entire population, and the inclusion into this sub-sample varies between groups and municipalities. Hence, they are not representative of the overall population.

This paper takes this kind of sub-sample biases into account and proposes a multiplier approach, to infer the distribution of ethnic identities, i.e. it estimates the characteristics of the overall population from a sub-sample. We use information on the identity of parents of newly born babies in the post-war period in the Federation of Bosnia and Herzegovina, in order to estimate the distribution of identities in each of its 77 municipalities. For this period, no census data are available. In Bosnia and Herzegovina, the first post-war census was only very recently conducted (in 2013), 18 years after the peace agreement of 1995. Data from the pre-war census in Bosnia and Herzegovina is outdated, due to significant movements in the population and `ethnic cleansing'

during the war (Melander, 2007). Our approach relies on data from official birth registries (Federalni zavod za statistiku, 2003-2010), and multiple tests of validity.

Both the applied methodology and the estimated ethnic composition of the municipalities of the Federation of Bosnia and Herzegovina contribute to the research on ethnic politics, and on post-war societies: first, both quantitative and qualitative studies in the field often rely on measurements of the ethnic composition of a territory, of a town or a region of interest. The methodology proposed in this paper provides for estimates where other, particularly official data is lacking. Second, and in particular reference to Bosnia and Herzegovina, the lack of data on ethnic identities is a major challenge to academic research, whether or not the applied method is of qualitative or quantitative nature. This paper provides for the first estimate for the post-war period, although limited to the Federation. As multiple methods of validation show, our estimates are more plausible than earlier attempts by state and international agencies to estimate the distribution of ethnic identities, which have never been published. At the same time, the fact that a census has been conducted in 2013 means that we apply our approach to a country in which decision makers have decided to collect and provide such data, although the data collection has been subject to irregularities in the most contested municipalities.[2] Our estimation will provide a basis of comparison.

Addressing social science methodology more generally, the paper innovates by providing a new methodology, that corrects for uneven distributions of birth rates across ethnic groups and municipalities. The method can similarly be applied to other cases, where social scientists infer data indirect, from sub-samples of the population, and where alternative estimation methods are not feasible.

When interpreting and using such methods, one needs to keep in mind that the nature of the data generation process has an important impact on the results (Nagle & Clancy, 2012). In our case, the birth registers have been compiled by state officials. Parents register with only one nominal

[2] See, for instance, Oslobodjenje, 16 October 2013, http://www.oslobodjenje.ba/vijesti/bih/salkic-instruktori-u-bratuncu-ponistavaju-popisnice-iako-je-popis-zavrsen (last accessed on 15 February 2015).
[4]Results and replication data available from the authors' website.

identity, which implies that the registries do not include information on multiple or multi-layered identities (Chandra, 2006; Lee, 2009). As identities are highly politicised in Bosnia and Herzegovina, the registration through a state official might increase the number of persons registered as members of one of the locally dominant identities (i.e., Serbs, Croats, or Bosniaks). Similar concerns have been raised for the official census, which is underway.

We proceed as follows: After a brief review of the role of ethnicity in Bosnia and Herzegovina, the second section of this paper develops our estimator of ethnic diversity and applies it to municipalities of the Federation of Bosnia and Herzegovina.[4] The third section provides a series of tests to examine the validity of our procedure. The conclusion discusses the broader applicability of our research.

1. Ethnicity in Bosnia and Herzegovina

In Bosnia and Herzegovina, the debate about the ethnic denominations is a result of the war of 1992-5. This section provides for a historical account and an overview over earlier estimations. For nearly half a century, Bosnia and Herzegovina has been defined as a multi-ethnic republic, with three state people: Serbs, Croats, and Muslims (or `Bosniaks'). Muslims were recognised in 1968, when the country was still part of the Socialist Federal Republic of Yugoslavia, as a constituent people, on an equal basis to Serbs and Croats, which led to an increase in number of citizens identifying as Muslims (Burg & Shoup, 2000, pp. 41-42). In 1993, the ethnic Muslims were officially re-named "Bosniaks".

The Dayton peace agreement of 1995 carried on and reinforced the ethnicisation of politics in Bosnia and Herzegovina. It created a complex federal structure: one entity, which has been called 'Republika Srpska', is dominated by the Serbs. The other entity, the Federation, consists of ten cantons, eight of which are either Bosniak- or Croat-dominated. The ethnic separation is largely the product of the war, during which large-scale displacement along ethnic lines took place. The Dayton agreement would allow displaced persons to return, and to re-establish the pre-war ethnic

diversity, but the process has not been completed. However, a more nuanced picture of the ethnic distribution after the war is lacking. In this situation, conducting a census is politically controversial: while some politicians understand it as a preliminary snapshot of the achievements of the politics of returns, it can also serve to officially recognise the post-war territorial divisions. Against this backdrop, no census has been conducted that informed about the distribution of ethnic identities up until 2013. As a result, for more than two decades, data on the territorial distribution of ethnic identities is lacking.

In 2005, the Federal Office of Statistics has estimated the ethnic structure of all 79 municipalities of the Federation of Bosnia and Herzegovina, but it is possible that their estimates might partially still rely on the pre-war census (Federalni zavod za statistiku, 2005). Based on more recent data, the OSCE has estimated the distribution of ethnic identities in 65 municipalities, but their methodology has not been published. Toal and Dahlman (2011) provide estimates for five municipalities of the Federation, the International Crisis Group (2003, pp. 6-7,15) has estimated the distribution of ethnic identities in Mostar from the entries in the electoral register, Caspersen {, 2004 #5310} relies on indirect proxies from electoral results. They seem to be straightforward, as the political landscape of Bosnia and Herzegovina is highly ethnicised. We refrain from using such a strategy, however, as it would not allow the clear identification of voters of multi-ethnic parties, and because data derived from election results would be highly endogenous for the explanation of political behaviour or political cooperation.

2. A two-stage procedure for the estimation of the distribution of ethnic identities

We propose an approach that is based on a sub-sample of the population in order to estimate the distribution of ethnic identities for some of these years.

There are a few common methods to estimate populations based on incomplete data, including capture-recapture, enumeration, and multiplier techniques. The capture-recapture technique uses

two or more lists of individually identifiable subjects to estimate the total number of subjects belonging to a given population (e.g. Bishop et al., 1975, pp. 229-256; Ball & Asher, 2002). Enumeration methods are similar to census methods, but limited to pre-defined segments of the target population. Finally, multiplier methods work with two independent records of data that relate to each other and extrapolate certain characteristics from a sub-sample of the population to the overall population, assuming that the sub-sample of the population is representative of the overall population (UNAIDS/WHO 2003, pp. 13-23). Approaches based on indirect observations rely on the multiplier method, even though their sub-samples are usually not representative, and therefore, they violate the underlying assumptions.

Indirect approaches are based on observable manifestations of ethnicity, such as the presence of languages or symbols in the public space, individual (consumer or political) behaviour, which is related to certain ethnic groups in a specific context (DellaVigna et al., 2014). They can also be related to direct observations of ethnicity or closely related identities (religious affiliation, languages, names) for a sub-sample of the entire population (Abrahamse et al., 1994; International Crisis Group, 2003, pp. 6-7,15; Elliott et al., 2008; Mateos et al., 2011).

Our estimation of the distribution of ethnic identities in Bosnia and Herzegovina is based on the multiplier method, but corrects for the selectivity of the sub-sample. We exploit the fact that while no representative survey or census is available to measure the distribution of ethnic identities at the municipal level, information on ethnic identities is collected upon the registration of newly born children. Every birth registration includes the ethnic identity of both parents of newborns. From the parents' identities, we extrapolate to the overall municipal populations.

Parents of newly born children are, of course, not a random or representative sample of the full population. Birth rates vary geographically and across ethnic and social groups, so that the number of parents cannot directly inform us about the overall size of the group they belong to. In other words, parents of newborns are not only a selective sub-sample of the overall population. Variation in birth rates may also be linked to those attributes of the population, which we are particularly

interested in, namely ethnicity and geography. Therefore, our methodology relies on a multiplier method (extrapolating from a sub-group to the overall population), with a correction procedure introduced in order to account for the variance in birth rates across social groups, ethnic groups, and municipalities.[5]

The model

Data on the identities of parents is reported in the birth statistics of the Federation of Bosnia and Herzegovina, for each of the three constituent groups (Serb, Croat, Bosniaks), and several other groups. We also have data on the overall size of the population of the municipalities. Our model will thus be based on the following information (and a set of further control variables, such as numbers of refugees/returnees and GDP estimates):

- p_{Ti} total population size of municipality *i*
- b_i total number of newly born children in municipality *i*
- $b_{e,i}$ parents (mothers and fathers) of newly born children in municipality *i*, by ethnic identity *e*, where *e* distinguishes Bosniaks (B), Croats (C), Serbs (S), and others (O)

We will use this information to derive an estimation of the following variable:

$p_{e,i}$ population of ethnic group *e* in municipality *i*

($p_{S,i}$ = ethnic Serbs in municipality *i*, etc.)

Naive Birth Rate Model

A first, direct application of the multiplier method allows us to obtain an estimate of the ethnic population structure. In this case, we assume that the parents of newly born children are a non-selective sample of the overall population, and treat them thus as a representative sub-group. Hence,

[5] The reliance on birth-related baseline data from medical institutions is not uncommon when it comes to the estimation of population figures (WHO, 2003, pp. 3-5).

we estimate a single birth rate for each municipality ($\gamma_i = \frac{b_i}{p_i}$). We derive the population structure by ethnic group by applying the same single birth rate to each of the ethnic groups and each municipality, $p_{e,i} = \frac{b_{e,i}}{\gamma_i}$.

This direct application would lead to valid results under the assumption that there are no significant differences between the birth rates, age structures, or gender balances of different ethnic groups, i.e. that birth rates only co-vary with local characteristics, and that they are not connected to ethnicity. We call this the *naive birth rate model*, which assumes that those conditions hold. If, however, the birth rate varies across ethnic groups, either due to demographic or value-related differences between the ethnic groups, the estimates will be biased.

Full Birth Rate Model

The naive model underlies the assumption that birth-giving parents are a representative sample of the overall population. This is implausible: war and migration have significantly altered the structure of the population, and not only particularly affected the ethnic, but also the age structure and gender balance of the society. For this reason, a valid estimation should consider possible variation in birth rates, due to differences between the municipalities, which might affect the rate of inclusion into our sub-sample (parents). In particular, these are the distribution of age groups and the gender balance across municipalities, and further correlates of birth rates, such as economic development or urbanity.

Therefore, we allow the birth rate $\gamma_{e,i}$ to vary between municipalities i and between ethnic groups e. We define the birth rate as the quotient of the number of newly-born children $b_{e,i}$ from ethnic group e, in municipality i and the number of inhabitants declaring themselves as belonging to this ethnic group $p_{e,i}$. We do not have any information of $\gamma_{e,i}$ or of $p_{e,i}$.

$$\gamma_{e,i} = \frac{b_{e,i}}{p_{e,i}} \qquad (1)$$

We know, however, that the sum of the population of all ethnic groups $p_{e,i}$ residing in a municipality i is equal to the total size of the population, p_{Ti}.

$$p_{Ti} = p_{Bi} + p_{Ci} + p_{Si} + p_{Oi} \tag{2}$$

The two equations (1) and (2) can be re-arranged as follows:

$$p_{Ti} = \left(\frac{b_{B,i}}{\gamma_{B,i}} + \frac{b_{C,i}}{\gamma_{C,i}} + \frac{b_{S,i}}{\gamma_{S,i}} + \frac{b_{O,i}}{\gamma_{O,i}} \right)$$

We divide the overall municipal population p_{Ti} by twice the overall number of local births $2 \cdot b_i$, and this results in the inverse overall birth rate, $\frac{1}{\gamma_i} = \frac{p_{Ti}}{b_i}$. We add an arrow $\beta_x X_i$, which stands for further covariates of the local birth rate (demographic structure, economy, etc.).

$$\frac{1}{\gamma_i} = \frac{p_{Ti}}{b_i} = \left(\frac{b_{Bi}}{\gamma_B \cdot b_i} + \frac{b_{Ci}}{\gamma_C \cdot b_i} + \frac{b_{Si}}{\gamma_S \cdot b_i} + \frac{b_{Oi}}{\gamma_O \cdot b_i} \right) + \beta_x X_i \tag{3}$$

The resulting equation (3) is a linear function, where each of the five unknown variables – the birth rates γ_B, γ_C, γ_S, γ_O, and the parameter for the control variables, β_x – is part of one summand. We can estimate the function with a linear regression model, where $\frac{1}{\gamma_i}$ is our dependent variable.

The equation is not free of assumptions. We allow for birth rates to vary by ethnic groups, i.e. we estimate birth rates for each ethnic group, γ_B, γ_C, γ_S, γ_O. Further control variables X allow birth rates to vary across municipalities. However, we assume that birth rates vary similarly for all ethnic groups across municipalities. For instance, if rural municipalities have higher birth rates than urban municipalities, this will be captured by our control variable. Their effect is fixed for all groups in a municipality.

Model specification and variables

This paragraph discusses the data on which we rely in our estimation and issues which we consider to be crucial for estimating population structures in a post-war society.

The birth registration statistics of Bosnia and Herzegovina include information about the ethnic identity of both parents of newborns (Federalni zavod za statistiku, 2003-2010). As the data refers to the parents, not the children, and as there are two entries for each child, the estimates would lead to a result, which is twice higher than the birth rate; hence we divide the numbers by two. The identity should be self-reported, and registered by the state official ('matičar'). In practice, however, state officials occasionally do register parents who do not claim to belong to any of the official ethnic groups by selecting a group based on the parents' names.[6]

Further demographic and socio-economic variables were included to control for possible biases. First, due to migration and war casualties, the gender balance and the age structure might vary across municipalities, leading to varying birth rates. The available data allows us to restrict the population to those aged 18 to 64, and to test for the robustness of the results, compared to the overall population. We cannot control directly, however, for possible effects of the gender balance. Birth rates tend to vary in different socio-economic contexts. We control for the economic development (local GDP estimate), the number of refugees and returnees, and indirect measures of the degree of urbanity (population size of the municipality and population density, logarithmised).

In some municipalities, our estimates will be less precise. In particular, this will be the case for small municipalities, as we use a small sub-sample of the entire population (parents), and with a smaller size of the population, inference from a sub-sample becomes less accurate. Similarly, as refugees were usually expelled from their homes based on their ethnic identity, and their return

[6] Interview with a political analyst, Sarajevo, on 21 May 2012 (kindly provided by Elena Stavrevska). The registration card can be found in the appendix of Federalni zavod za statistiku (2009).

depends on the economic and social context of the municipalities, refugees and returnees are no homogeneous group, and thus have not a homogeneous effect on municipal birth rates. The estimation will also be biased in municipalities, where large parts of the population do not belong to the age cohorts, who might have children, or with an unbalanced gender distribution. Hence, these control variables have heteroscedasticity effects. Therefore, our models will perform less well for municipalities with high numbers of refugees and returnees, small municipalities, and those with a specific demographic structure.

Because we expect such heteroscedasticity effects, we run linear models with separate variance-predictors. Our models use two sets of predictors. On the one hand, the model estimates the average effect on the mean of our dependent variable. On the other hand, a second set of explanatory variables is expected to estimate the accuracy of the model, and therefore is related to the unexplained variance of the dependent variable, σ^2 (Davidian & Carroll, 1987; Braumoeller, 2006). In this study, the main variables, introduced in equation 3 – birth numbers by ethnic identity and a first set of control variables – are expected to affect the mean birth rate, and are therefore included in the main part of the model. A second set of control variables is related to the accuracy of the predictions (effect on variance), consisting of the share of refugees and returnees among the population, the size of the municipality, and the age structure (population <15 years). There is no information available for the gender balance of the municipalities. Data on refugees and returnees is missing for 9 out of 77 municipalities. We have imputed missing information for these municipalities. Estimation results remain stable with/without imputation.

Two municipalities have not registered any births at all (Ravno), or only very few births (Neum). While Ravno is a rather small municipality, and thus susceptible to be an outlier, Neum is surrounded by Croat territory, and is according to all sources overwhelmingly inhabited by ethnic Croats. Most parents with residence in Neum give birth in hospitals in neighbouring Croatia, where

they can receive health care as domestic citizens.[7] We decided to exclude the two municipalities from the estimation, as the estimates for the two cases are unreliable. However, the inclusion of the two municipalities does not alter the overall results of the estimation substantially (results available from the authors).

Results of the OLS estimation of the birth rates are presented in table 1. Ethnic Bosniaks are our remote category, meaning that the coefficients measure the differences between the birth rates of the Bosniaks and of other ethnic groups. The dependent variable (birth rates) is inverted, so that the positive coefficient in the main model for population density implies that birth rates are lower in municipalities with a high population density. Likewise, they tend to be lower in places with a high share of non-returned refugees. After controlling for these factors, we find that the overall birth rate lowers as the percentage of Serbian parents increases. Among Croats, birth rates tend to be higher than among Serbs, but lower than among Bosniaks. There are also significant effects in the variance part of the models: size of the municipality, the age structure, and the share of refugees and returnees affect the accuracy of the estimation.

xxx include table 1 about here xxx

In the second step, based on the models presented in table 1, we calculate predicted values for the birth rates in 77 municipalities for each ethnic group (γ_B, γ_C, γ_S, γ_O, for Bosniaks, Croats, Serbs, 'Others'), whilst considering the socio-economic context. Equation 1 allows us to predict the overall distribution of ethnic identities in these municipalities, which we display for 2008 in figures 1 and 2. Certainly, the distinction of municipalities in figure 1 by the largest group after the war reproduces well-established knowledge. More importantly however, our results also provide an estimation of the actual size of each group. The category ‘Others’ is barely visible; our highest

[7] http://wikileaks.org/cable/2003/04/03ZAGREB866.html (last accessed on 11 July 2013).

estimation is for Bosanski Grahovo (0.5% of the 2008 population), but due to the very small number of births by parents in this group, the estimation might not be reliable.

xxx include figure 1 about here xxx

xxx include figure 2 about here xxx

Reliability and validity

In what follows we provide the results of a number of tests that assess the reliability and the validity of our estimates reported above. These tests compare the results of our estimation model to earlier estimations (by the Statistical Office and by the OSCE), and to the naive model. We aim to show that our estimation model outperforms the naive estimation, and that they are also better suited than earlier estimations to measure ethnic identities in the Federation of Bosnia and Herzegovina.

Reliability test: Although the age structure – and, especially in small municipalities – the number of births can change over time, the distribution of ethnic identities should correlate strongly from one year to the next in the post-war period. Hence, if reliable, our estimation results should almost be stable over the years. As a test of the reliability of our results, we therefore correlate our estimates for the available years (2008-2010) (table 2).

Validity tests: We assess whether our population estimates correspond to the actual distribution of ethnic identities in municipalities in Bosnia and Herzegovina. Due to the absence of official data, this can only be done indirectly. We perform two well-established tests of validity.

First, we test whether the new measures reproduce the results of other estimates (criterion validity). We compare them to unofficial and not publicly available previous estimates by the

Federal Statistical Office (for all municipalities, in 2005) and the OSCE (covering only 31 of the municipalities of the Federation, in 2008, and matching 30 of our cases). We correlated them with our estimates for the year closest in time to these, namely 2008. Ideally, the correlation coefficient (according to Pearson) should amount to 1. Our estimates of the full model are all above 0.96 for the ethnic Bosniaks and the ethnic Croats, and around 0.83-0.91 for the estimates of ethnic Serbs (table 2). Our results estimate the share of Serbs to be higher than the Federal Statistical Office; the difference is particularly pronounced for the municipality of Drvar (32% versus 78%). Given these differences in the estimates of ethnic Serbs, it is crucial to test which of these measures is more precise in explaining theoretically expected patterns.

xxx include table 2 about here xxx

Second, we employ the new estimates to test for theoretically expected relationships (construct validity). In many ethnically divided countries, ethnic identities are good predictors of the vote for ethno-nationalist parties. Likewise, elections in Bosnia and Herzegovina are perceived as de-facto ethnic censuses. We use the vote share of ethno-nationalist parties in the 2006 parliamentary elections and in the 2008 local elections to assess our estimates. Ethno-nationalist parties obtained 73%-75% of the votes in the Federation in these elections.

Two aspects are tested for. First, we expect that the municipality-level voting results can inform us about the distribution of ethnic identities. Due to the contextual effects of ethnic diversity on the propensity of voters to vote for nationalists, the relationship is often curvilinear.[8] We analyse, thus, to what extent our measure explains the vote share of ethno-nationalist parties. Construct validity should be high, if our measure is accurate, and if patterns of ethnic voting are strong. Figure 3

[8] For models explaining the curvilinearity of this relationship, see Bochsler (2013) and Grofman and Handley (1995).

shows such a model exemplarily for the Croat parties, based on our full estimation model.[9] The curvilinear line is the best fit, and the graph demonstrates that our population estimates can explain the vote share of Croat parties reasonably well.

Because we expect a curvilinear pattern, we use a measure of explanatory power (the adjusted R^2), rather than a correlation coefficient. High values, that is, values close to 1, indicate a high degree of construct validity (figure 4). However, some variation unexplained by this model remains. Our second indicator of construct validity tests for the degree of underestimation. As we are less concerned with minor deviations than with larger ones, we use a measure of squared deviations (dev_e), only counting deviations, where the vote share of ethnic parties is higher than their group's share of the local population (figure 5).[11] Small squared deviations (on a scale from 0 to 10000) indicate a high degree of validity. This is illustrated in figure 3: points located below the diagonal line might indicate cases where Croats might not vote for Croat parties. Points located above the diagonal line mark municipalities indicate places where Croat parties reached a higher vote than the estimated share of ethnic Croat residents. This suggests that (assuming roughly equal turnout across all groups) our model underestimates the Croat population for these cases.

xxx include figure 3 about here xxx

[9] Party classification in appendix A.

[11] The measure is calculated separately for each ethnic group, as follows: $\frac{dev_e = \sum_{i=1}^{N} \left[\max\left(vote_{e,i} - p_{e,i}, 0\right)\right]^2}{N}$, where N is the number of municipalities, $vote_{e,i}$ is the vote share of the ethno-nationalist party of group e in municipality i.

Based on these two indicators, our results clearly outperform two previous estimates of ethnic diversity in Bosnia and Herzegovina, the estimates of the Statistical Office of the Federation of 2005, and the estimates provided by the OSCE.

Our procedure appears to provide a much more accurate estimation of the size of the Serb and the Bosniak communities than the estimates of the Statistical Office: validation models that are based on the figures of the Statistical Office result in only moderate explanatory power for the success of Serb parties, with an adjusted R^2 around 0.48 (2008 elections, figure 4, left panel, and table B1 in appendix B). Our figures allow a fairly accurate explanation (adj R^2: 0.96). The Statistical Office appears to under-estimate the size of the Serbian community (dev_e: 23.6), and of the ethnic Bosniak community (dev_e: 15.5, figure 5, left panel, and table B2 in appendix B) in certain municipalities, which our full model does not. In contrast, we under-estimate the size of the Croat community in several municipalities.[12] Our measure (dev_e=13.7) is slightly less accurate than the Statistical Office's estimation, but it deviates less the Statistical Office's measure for Bosniaks and Serbs. In part, the inaccuracies of the Statistical Office's figures for the Serb community might stem from the considerable migration between 2005 (when the estimation was done) and 2008 (when the elections were held). Therefore, we have repeated our validity tests also for the 2006 parliamentary elections. Even in this comparison, our measure considerably outperforms the population figures of the Statistical Office for two indicators (and is only moderately less accurate regarding a third indicator) (see appendix B).

To compare our model to the OSCE estimates, we restrict the validation procedure to the 30 municipalities for which OSCE estimates are available (figures 4 and 5, right panels). The OSCE appears to under-estimate the number of ethnic Bosniaks in some municipalities, and our results

[12] This is driven by Usora, we count 69% ethnic Croats and 31% ethnic Bosniaks, but 92% of the votes are cast for ethnic Croat parties. However, the number of registered births is very low, with 26 births on 7000 inhabitants in 2008.

outperform the OSCE on both indicators for this group. We also seem to have a slightly better estimation of the ethnic Serbs, judging by the adjusted R^2. Both the OSCE and our figures seem to under-estimate the number of ethnic Croats in certain municipalities, but here the OSCE slightly outperforms our measures.

Quite clearly, our *full model* improves the estimation, compared to the *naive model*, which does not correct for different birth rates across municipalities and ethnic groups. The difference is most notable for the squared deviation measure. The naive model tends to under-estimate the size of the ethnic Croat and the ethnic Serb community systematically for all municipalities. This also explains why the naive model and the full model have almost the same explanatory power (adjusted R^2).

xxx include table 3 about here xxx

3. Conclusions

In those countries, where data on the distribution of ethnic identities is most relevant for academic research, they are often also subject to political struggles. As a result, questions about ethnicity in the population census are highly salient, and thus get politicised, or the census might get delayed altogether. In the absence of (credible) census data, ethnic identities are often estimated based on indirect data, such as register data (e.g. birth, marriage, death registers), or proxies of ethnic identities, such as membership in religious communities or consumer data for goods and services, associated with particular groups.

This paper proposes a new methodology to estimate the distribution of ethnicity at the sub-national level (municipalities, districts, etc.), relying on indirect data from a sub-sample of the entire population. Such indirect approaches, also addressed as multiplier method, assume that the sub-sample is representative of the entire population.

However, sub-samples, providing for indirect observations of identities, are not only incomplete, but also selective: the likelihood that individual citizens are included into the sub-

sample will often correlate with ethnic identities and with further socio-economic variables. For instance, birth and marriage rates, life expectancy, but also the share of the population who are members of religious communities varies across groups, and due to other contextual factors across municipalities. The method proposed in this paper corrects for these effects. It allows thus to improve the estimation of ethnic identities from indirect data, but could also transfer to other estimations of populations, where some characteristics are indirectly observed on a sub-sample of the entire population.

This paper applies the procedure the to 77 municipalities in Bosnia and Herzegovina in the post-war period. We rely on data from birth registers on the ethnic identity of parents of newly born children. To our knowledge, this is the first attempt to provide estimates of ethnic identities for the post-war period in Bosnia, based on a transparent and systematic procedure that covers a large number of municipalities.

Multiple tests of robustness and validity indicate that our procedure leads to a good approximation of the distribution of the hitherto officially recognised ethnic identities in many municipalities of Bosnia and Herzegovina. This, despite the fact that the estimation is based on subset of citizens, which relates to a specific age group, and despite potential problems as the gender balance of the population might vary across municipalities. Most importantly, our numbers also seem to outperform previous measures by the OSCE and the Statistical Office of the Federation. In municipalities with very low birth figures, the measurements are less accurate, as they seem to under-estimate the size of the Croat population.

While the finding that municipalities of the Federation of Bosnia and Herzegovina are either Bosniak- or Croat-dominated might be little surprising to area specialists, the estimation also identifies the places with a genuinely multi-ethnic face. Examples are the municipalities of Glamoć or Bosanski Petrovac, where none of the ethnic groups constitutes a clear majority of the population, and over a dozen of further places, which are ethnically diverse. Apart from the many-fold applications for research, these results can serve as a basis of comparison once data on ethnic

identities from the census of Bosnia and Herzegovina will be released. Due to the inclusion of more nuanced questions on ethnicity - including proxies such as the use of language and religion - the 2013 census might show a more multi-faceted and diversified picture of ethnic identitfication (Perry, 2013, p. 12; Bieber, 2013).[13] On the other hand, the heated political debate about the census might also leave its footprint on the empirical results (Friedman, 2002; Perry, 2013, p. 4). Overall, the results of the census might therefore reveal a more diverse and civic face of the country, where non-recognised, civic, or multi-faceted identities are more visible. Our data might be used in studies that test hypotheses about identity shifts, return processes, registration politics, and other ethnicity-related processes in the years between the war, and the first post-war population census. Our method can travel to other cases, where ethnic identities are estimated indirectly, based on observations from a sub-sample of the population.

References

Abrahamse, A. F., Morrison, P. A., & Bolton, N. M. (1994) Surname analysis for estimating local concentration of Hispanics and Asians, *Population Research and Policy Review*, 13, pp. 383-398.

Aitken, R. (2007) Cementing divisions? An assessment of the impact of international interventions and peace-building policies on ethnic identities and divisions, *Policy Studies*, 28 (3), pp. 247-267.

Ball, P., & Asher, J. (2002) Statistics and Slobodan: Using Data Analysis and Statistics in the War Crimes Trial of Former President Milosevic, *Chance*, 15 (4), pp. 17-24.

Belloni, R. (2007) *State Building and International Intervention in Bosnia* (Abingdon, Routledge).

Bieber, F. (2013). When counting counts. The Bosnian Census (Vol. 2013).

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975) *Discrete Multivariate Analysis: Theory and Practice* (Cambridge (MA), MIT Press).

Bochsler, D. (2013) Radicalizing Electoral System Effects on Support for Nationalist Hardliners, *Nations and Nationalism*, 19 (1), pp. 21-43.

Braumoeller, B. F. (2006) Explaining Variance; Or, Stuck in a Moment We Can't Get Out Of, *Political Analysis*, 14, pp. 268-290.

Buhaug, H., & Rød, J. K. (2006) Local determinants of African civil wars, 1970–2001, *Political Geography*, 25 (3), pp. 315-335.

Burg, S. L., & Shoup, P. S. (2000) *The war in Bosnia-Herzegovina: ethnic conflict and international intervention* (Armonk, M.E.Sharpe).

Caspersen, N. (2004) Good Fences Make Good Neighbours? A Comparison of Conflict-Regulation Strategies in Postwar Bosnia, *Journal of Peace Research*, 41 (5), pp. 569-588.

[13] See also Lee (2009) and Chandra (2006) on multi-layered and multi-faceted identities.

Chandra, K. (2006) What is Ethnic Identity and Does It Matter?, *Annual Review of Political Science*, 9, pp. 397-424.
Davidian, M., & Carroll, R. J. (1987) Variance Function Estimation, *Journal of the American Statistical Association*, 82 (400), pp. 1079-1091.
DellaVigna, S., Enikolopov, R., Mironova, V., Petrova, M., & Zhuravskaya, E. (2014) Cross-border media and nationalism, *American Economic Journal: Applied Economics*, 6 (3), pp. 103-132.
Elliott, M. N., Fremont, A., Morrison, P. A., Pantoja, P., & Lurie, N. (2008) A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity, *Health Research and Educational Trust*, 43 (5), pp. 1722-1736.
Fearon, J. D., & Laitin, D. D. (1996) Explaining Interethnic Cooperation, *American Political Science Review*, 90 (4), pp. 715-735.
Federalni zavod za statistiku. (2003-2010). Živorođeni prema nacionalnoj - etničkoj pripadnosti majke i oca. Sarajevo: Federalni zavod za statistiku.
Federalni zavod za statistiku. (2005). *Procjene ukupnog broja prisutnih stanovnika, po nacionalnoj strukturi, stanje 30.06.2005.* Sarajevo: FZS.
Federalni zavod za statistiku. (2009). *Demografska statistika 2008*. Sarajevo: FzS.
Franck, R., & Rainer, I. (2012) Does the Leader's Ethnicity Matter? Ethnic Favoritism, Education, and Health in Sub-Saharian Africa, *American Political Science Review*, 106 (2), pp. 294-325.
Friedman, V. (2002) Observing the Observers: Language, Ethnicity, and Power in the 1994 Macedonian Census and Beyond, *New Balkan Politics*, 2002 (3/4).
Grofman, B., & Handley, L. (1995) Racial context, the 1968 Wallace vote, and southern presidential dealignment: evidence from North Carolina and elsewhere. In M. Eagles (Ed), *Spatial and Contextual Models in Political Research.*, pp. 151-162 ((London, Taylor and Francis).
Harris, J. A. (2015) What's in a Name? A Method for Extracting Information about Ethnicity from Names, *Political Analysis*, 23 (2), pp. 212-224.
Horowitz, D. L. (1985) *Ethnic Groups in Conflict* (Berkeley, Los Angeles, London, University of California Press).
International Crisis Group. (1999). *Preventing Minority Return in Bosnia and Herzegovina* (No. 73). Sarajevo: ICG.
International Crisis Group. (2003). *Building Bridges in Mostar*. Brussels: International Crisis Group.
Kertzer, D. I., & Arel, D. (Eds). (2001) *Census and Identity. The Politics of Race, Ethnicity and Language in National Censuses*. Cambridge, Cambridge University Press.
Lee, T. (2009) Between Social Theory and Social Science Practice. Toward a New Approach to the Survey Measurement of 'Race'. In R. Abdelal, Y. M. Herrera, A. I. Johnston & R. McDermott (Eds), *Measuring Identity*, pp. 113-144 ((Cambridge, Cambridge University Press).
Mateos, P., Longley, P. A., & Sullivan, D. O. (2011) Ethnicity and Population Structure in Personal Naming Networks, *PLoS one*, 6 (9).
McCrudden, C., & O'Leary, B. (2013) *Courts and Consociations. Human Rights versus Power-Sharing* (Oxford, Oxford University Press).
McCulloch, A. (2014) Consociational settlements in deeply divided societies: the liberal-corporate distinction, *Democratization*, 21 (3), pp. 501-518.
Melander, E. (2007). Ethnic Cleansing in Bosnia-Herzegovina, 1992-1995, *Conference "Disaggregating the Study of Civil War and Transnational Violence"*. Colchester.
Mitchell, P., Evans, G., & O'Leary, B. (2009) Extremist Outbidding In Ethnic Party Systems Is Not Inevitable: Tribune Parties in Northern Ireland, *Political Studies*, 57, pp. 397-421.

Nagle, J., & Clancy, M.-A. C. (2012) Constructing a shared public identity in ethno nationally divided societies: comparing consociational and transformationist perspectives, *Nations and Nationalism*, 18 (1), pp. 78-97.

Nalepa, M. (2012) Reconciliation, Refugee Returns, and the Impact of International Criminal Justice: The Case of Bosnia and Herzegovina. In M. Williams, R. Nagy & J. Elster (Eds), *NOMOS, Proceedings of the American Society for Political and Legal Philosophy*, pp. 316-360 ((New York, New York University Press).

Perry, V. (2013). *The 2013 Census in Bosnia and Herzegovina - A Basic Review*. Sarajevo: Democratization Policy Council.

Pugh, M., & Cobble, M. (2001) Non-Nationalist Voting in Bosnian Municipal Elections: Implications for Democracy and Peacebuilding, *Journal of Peace Research*, 38 (1), pp. 27-47.

Raleigh, C., & Hegre, H. (2009) Population size, concentration, and civil war. A geographically disaggregated analysis, *Political Geography*, 28 (4), pp. 224-238.

Susewind, R. (forthcoming) What's in a Name? Probabilistic Inference of Religious Community from South Asian Names, *Field Methods*.

Szreter, S., Sholkamy, H., & Dharmalingam, A. (Eds). (2004) *Categories and Contexts: Anthropological and Historical Studies in Critical Demography*. Oxford, Oxford University Press.

Toal, G., & Dahlman, C. T. (2011) *Bosnia Remade. Ethnic Cleansing and its Reversal* (Oxford, Oxford University Press).

UNAIDS/WHO Working Group on Global HIVAIDS and STI Surveillance. (2003). *Estimating the Size of Populations at Risk for HIV: Issues and Methods*: Family Health International (FHI).

Varshney, A. (2001) Ethnic Conflict and Civil Society. India and Beyond, *World Politics*, 53, pp. 362-398.

Visoka, G., & Gjevori, E. (2013) Census politics and ethnicity in the Western Balkans, *East European Politics*, 29 (4), pp. 479-498.

WHO. (2003). *HIV/AIDS Epidemiological Surveillance Update for the WHO African Region 2002*. Harare: WHO.

Williams, M., & Husk, K. (2013) Can we, should we, measure ethnicity?, *International Journal of Social Research Methodology*, 16 (4), pp. 285-300.

An indirect approach to map ethnic identities in post-conflict societies

Figures and tables

Table 1: Explanatory model for inverted birth rates ($1/\gamma_i$) in the Federation BiH, 2008-2010
Variance model (Linear regression with multiplicative heteroscedasticity)
[a] Variable imputed for 9 municipalities.

	year=2008		year=2009		year=2010	
	Coef.	Robust Std. Err.	Coef.	Robust Std. Err.	Coef.	Robust Std. Err.
Main model (mean)						
b_{Ci} / b_i	10.02 **	3.22	10.27 *	4.18	11.93 **	4.42
b_{Si} / b_i	49.98 **	7.63	76.69 **	22.34	-23.14 *	10.54
b_{Oi} / b_i	-49.16	30.97	-126.26 **	32.43	-99.83 **	23.82
% employed women	-10.99	12.24	5.53	11.53	-15.35	18.46
population density (log)	.54	.85	1.69	1.08	1.88 *	.85
GDP per capita (log)	.33	1.99	-1.69	1.99	-2.83	2.50
population (log)	-.23	.1.49	-.31	1.53	1.31	1.94
returnees (log)[a]	.09	4.54	-4.58	4.93	-1.64	5.27
refugees non-returned (log)[a]	21.77 **	2.61	47.03 **	5.41	45.31 **	3.12
constant	30.94	19.17	39.86 *	18.59	39.23 *	19.57
Variance model						
returnees (log)[a]	1.96 *	.93	1.85	1.14	2.72 **	.91
refugees non-returned (log)[a]	-8.03 **	1.26	-4.26 **	1.23	-8.08 **	1.22
population (log)	-.89 **	.27	-.78 **	.30	-1.02 **	.26
% children	-34.54 **	6.02	-32.96 **	6.59	-27.2 **	6.23
constant	31.92 **	8.84	18.34 **	3.02	20.10 **	2.78
Number of obs	77		77		77	
Model $\chi^2(12)$	83.9		121.2		72.035	
Prob > χ^2	0		0		0	
Pseudo R^2	0.1309		0.1736		0.1093	
VWLS R^2	0.975		0.5641		0.9897	

Table 2: Results of reliability and criterion validity tests

		Bosniaks		Croats		Serbs	
Type of test	*Indicator*	*naive birth-rate model*	*full model*	*naive birth-rate model*	*full model*	*naive birth-rate model*	*full model*
Reliability	mean correlation across years	0.975	0.947	0.962	0.966	0.984	0.978
Criterion validity	correlation with OSCE estimates (N=30)	0.956	0.962	0.964	0.976	0.831	0.829
	correlation with FZS estimates (N=77)	0.962	0.970	0.956	0.961	0.881	0.907

Figure 1: Population estimates by ethnic group, 2008, municipalities of the Federation of Bosnia and Herzegovina. Full model estimation.

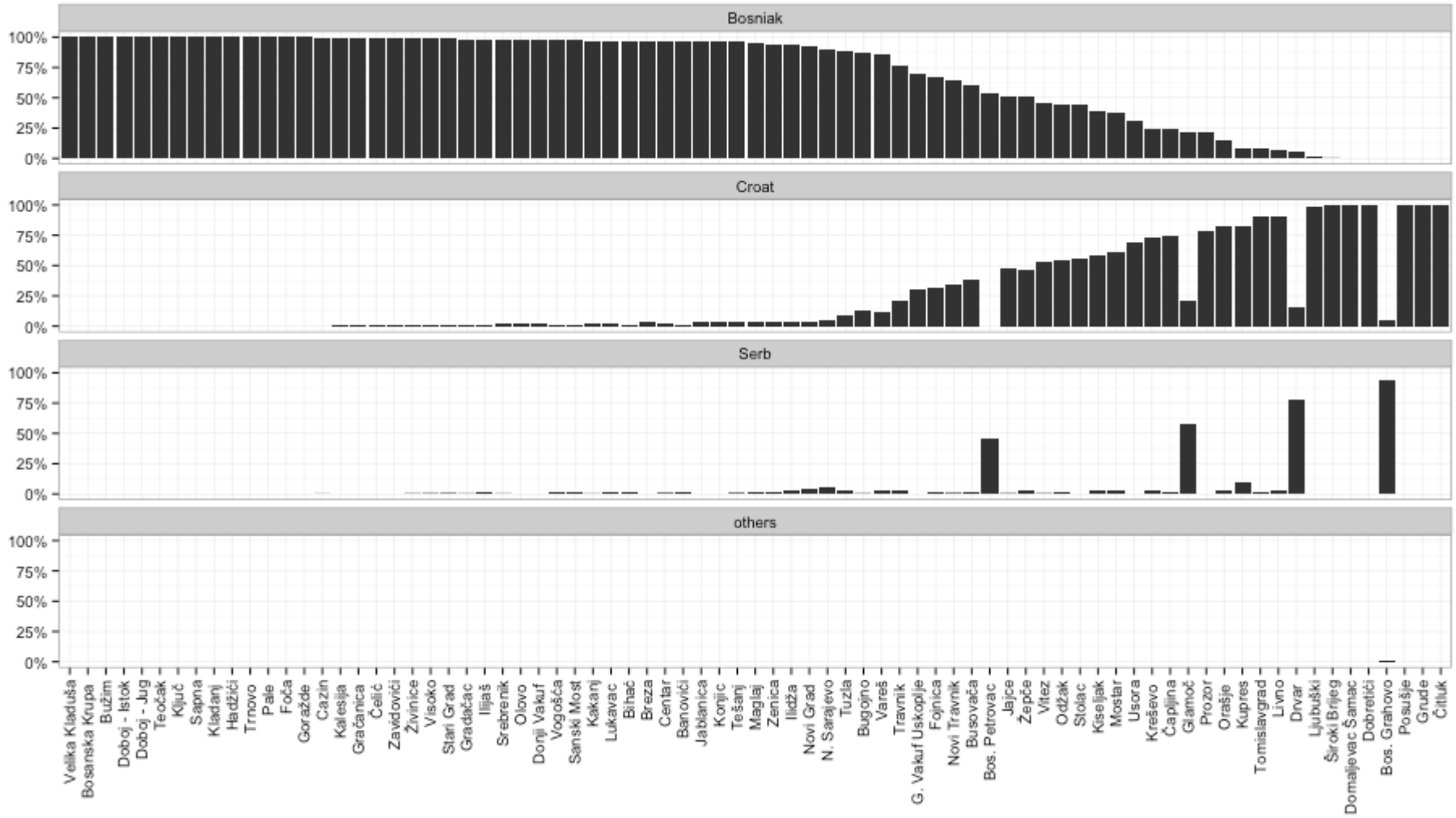


Figure 2: Ethnic map of the Federation of Bosnia and Herzegovina, 2008. Full model estimation.
Source map: GDi GISDATA d.o.o. Sarajevo (Gauss-Kruger - Zone 6). Data: own estimation.

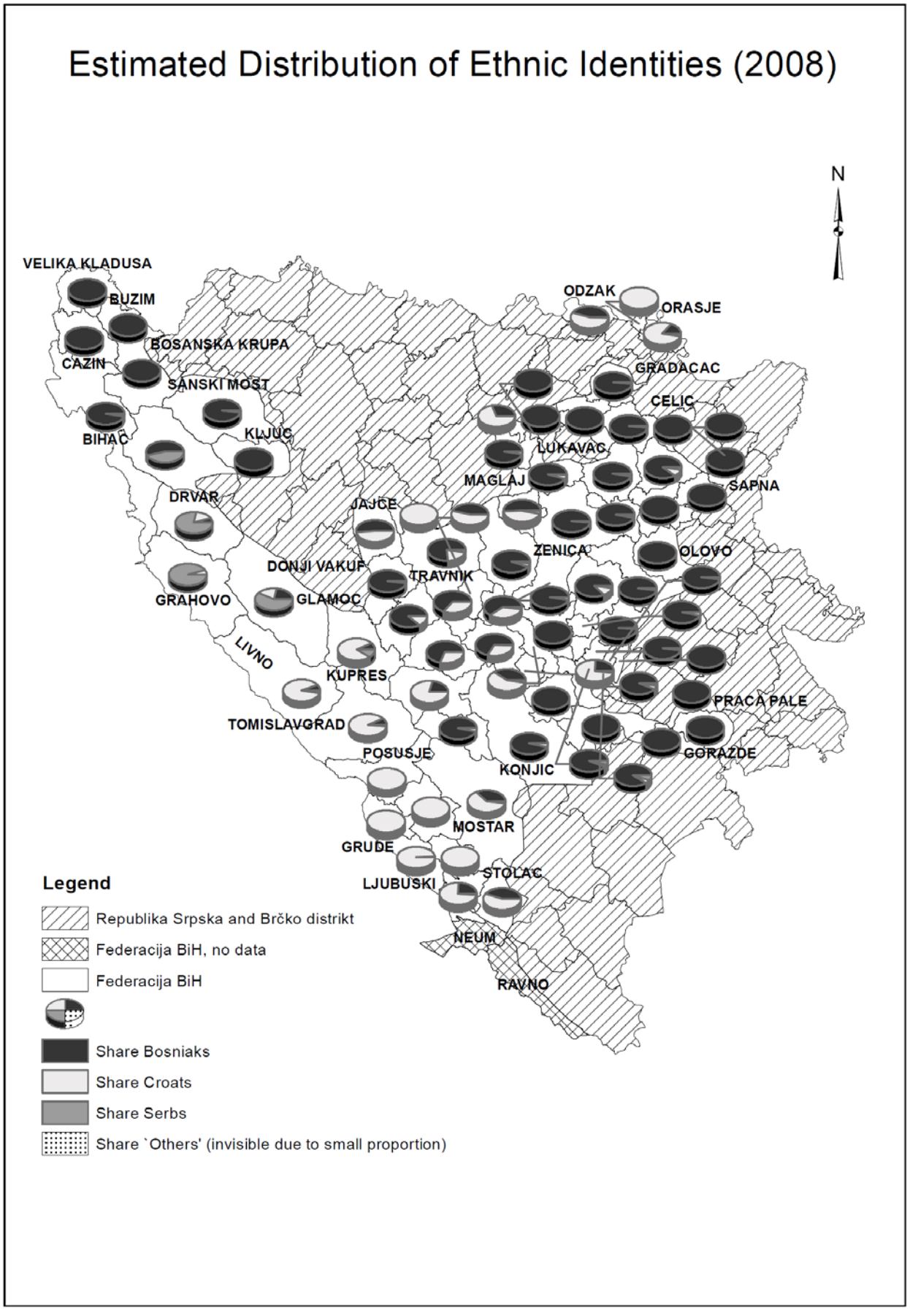


Figure 3: Estimated share of Croat population and vote share for Croat nationalist parties, 2008 municipal elections. (Full estimation model)

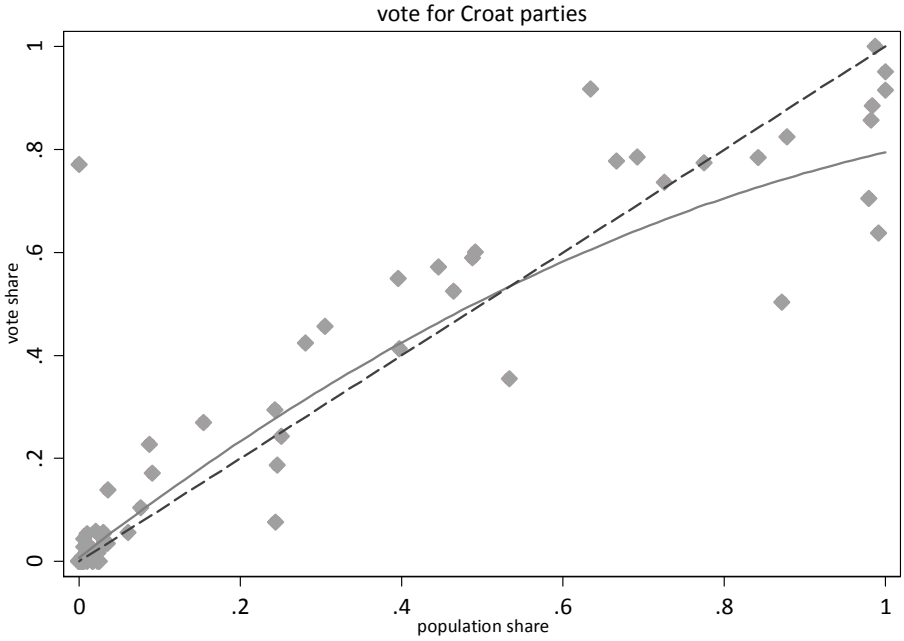


Figure 4: Results of construct validity tests, explanatory power (adjusted R^2)
Values close to 1 indicate high construct validity

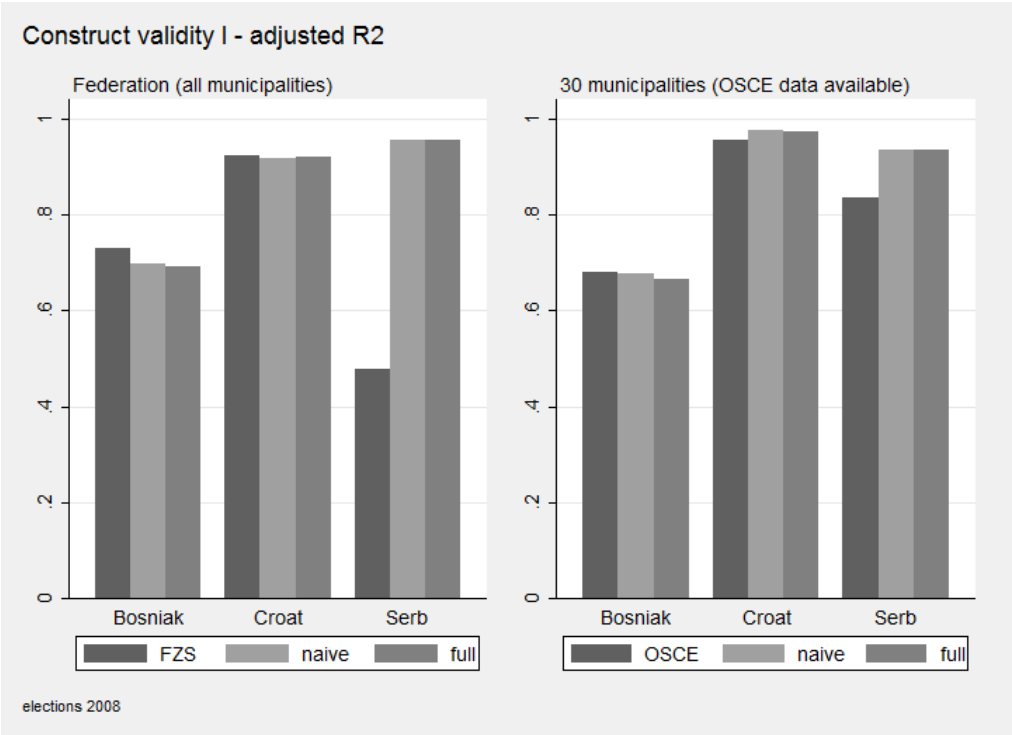
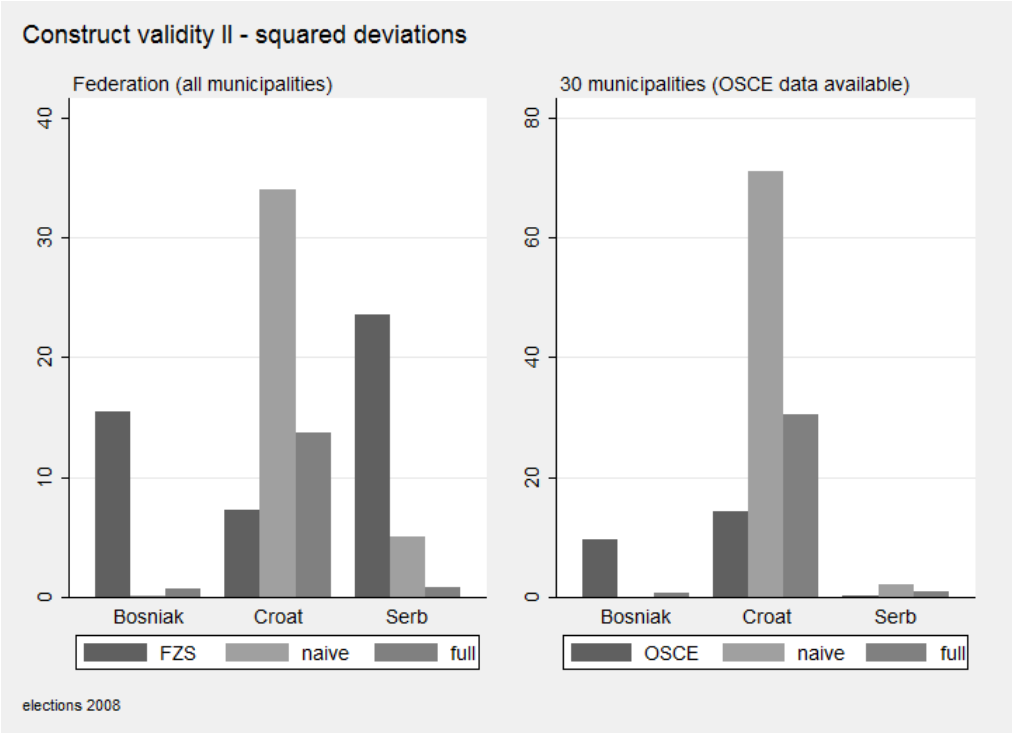


Figure 5: Results of construct validity tests, squared deviations (dev_e)
Values close to 0 indicate high construct validity



Appendix A: Categorisation of political parties and electoral lists

Bosniak	Bosanskohercegovačka Patriotska Stranka-Sefer Halilović
	Bošnjačka Seljačka Stranka
	BOSS - Bosanska Stranka-Mirnes Ajanović
	Demokratska Narodna Zajednica BiH
	Lista za Čapljinu SBiH-BPNS
	Narodna Bošnjačka Stranka
	Savez SDA-S BiH
	Socijaldemokratska Unija Bosne i Hercegovine
	Stranka Demokratske Akcije
	Stranka Demokratske Aktivnosti za Evropsku BiH A-SDA
	Stranka za Bosnu i Hercegovinu
Croats	HDZ 1990 HSS-NHI HSP
	Hrvatska Demokratska Zajednica 1990
	Hrvatska Demokratska Zajednica BiH
	Hrvatska Koalicija – HSS - NHI HDZ BiH HDZ 1990 HNZ
	Hrvatska Koalicija (HDZ BiH HDZ 1990 HKDU BiH)
	Hrvatska Koalicija HDZ 1990 HSS – NHI
	Hrvatska Koalicija HDZ 1990 HSS-NHI Za Žepče
	Hrvatska Koalicija HDZ 1990-HKDU
	Hrvatska Koalicija Kantona Sarajevo
	Hrvatska Koalicija za Drvar HDZ 1990-HDZ BiH-HSP Đapić-Jurišić
	Hrvatska Koalicija za Pougarje HSP-Đ-J HDZ1990 HSS-NHI
	Hrvatska Koalicija za Ravno HDZ1990 HDZBiH
	Hrvatska Koalicija za Travnik
	Hrvatska Koalicija za Zenicu HDZ BiH HDZ 1990 HSS-NHI HNZ
	Hrvatska Koalicija-HDZ BiH-HSS-NHI-HDZ 1990
	Hrvatska Kršćanska Demokratska Unija BiH
	Hrvatska Narodna Zajednica
	Hrvatska Stranka Prava Bosne I Hercegovine - Đapić – Dr. Jurišić
	Hrvatsko Zajedništvo Herceg-Bosne
	Hsp BiH Đapić-Dr.Jurišić-HNZ BiH
	Hss – NHI
	Koalicija Hrvatskih Stranaka HDZ BiH HSS-NHI HDZ 1990 HKDU BiH
	Koalicija: HNZ-HDU
	Koalicijska Lista HDZ BiH i HDZ 1990
Serbian	Demokratska Stranka Naroda Srpske
	Demokratski Narodni Savez
	Demokratski Pokret Srpske-Depos
	Koalicija Za Bolji Život-DNS i DS
	Narodna Demokratska Stranka
	Partija Demokratskog Progresa
	Savez Nezavisnih Socijaldemokrata - SNSD - Milorad Dodik
	Savez Srpskih Stranaka Srebrenice: SDS-PDP-SP i SRS RS
	SDS-DNS-PDP-SRS RS
	SNSD-PDP
	Socijalistička Partija
	Srpska Demokratska Stranka
	Srpska Demokratska Stranka 1990 - Izvorna - Pokret za Srpsku
	Srpska Napredna Stranka Republike Srpske
	Srpska Narodna Radikalna Stranka - Banja Luka
	Srpska Radikalna Stranka Dr Vojislav Šešelj
	Srpska Radikalna Stranka Republike Srpske
	Srpska Stranka Republike Srpske
	Srpski Pokret Obnove Republike Srpske
Mixed-ethnic	Građanska Demokratska Stranka BiH
	Narodna Stranka Radom za Boljitak
	Naša Stranka
	Socijaldemokratska Partija Bosne i Hercegovine

Appendix B: Results of construct validity tests

Construct validity tests, comparing the population estimates with the 2006 elections, are provided, in order to allow for a better time match with the 2005 estimates of the Statistical Office of the Federation (FZS). Therefore, the comparison to the 2006 elections is only included for the models for all 77 municipalities, and not for the subset of 30 municipalities.

Table B1: Results of construct validity tests, explanatory power (adjusted R^2)
Values close to 1 indicate high construct validity.

group	compared to elections of ...	FZS estimates (2005)	Naive model	Full model	OSCE estimates	Naive model	Full model
		All 77 municipalities			Only 30 municipalities		
Bosniak	2006	0.921	0.987	0.983			
Croat	2006	0.924	0.990	0.989			
Serb	2006	0.389	0.965	0.962			
Bosniak	2008	0.730	0.699	0.692	0.680	0.676	0.665
Croat	2008	0.922	0.916	0.921	0.955	0.975	0.973
Serb	2008	0.478	0.956	0.956	0.835	0.935	0.935

Table B2: Results of construct validity tests, squared deviation (dev_e)
Values close to 0 indicate high construct validity.

group	compared to elections of ...	FZS estimates (2005)	Naive model	Full model	OSCE estimates	Naive model	Full model
		All 77 municipalities			Only 30 municipalities		
Bosniak	2006	130.01	12.37	2.30			
Croat	2006	22.78	10.29	1.73			
Serb	2006	17.68	0.58	0.05			
Bosniak	2008	15.53	0.09	0.73	9.56	0.00	0.79
Croat	2008	7.32	33.94	13.69	14.40	70.98	30.52
Serb	2008	23.58	5.06	0.83	0.00	2.05	1.04