Use and misuse of MAR *

Simon Hug†
Département de science politique et relations internationales,
Université de Genève
Centre for the Study of Civil War (CSCW),
International Peace Research Institute, Oslo (PRIO)

First version June 2011, this version, August 24, 2012

Abstract

The “minorities at risk” project and data have offered a first and innovative systematic view of communal and ethnic groups. Collecting detailed information on mobilized and/or discriminated groups it has allowed researchers to offer new insights in many areas dealing with relationships among groups. Nevertheless, scholars have used this data also in ways for which it was not designed or hardly appropriate. This paper offers an overview over the main contributions of the “minorities at risk” project and discusses for what research questions this data has to be used with care.

Keywords: Political violence, rebellion, civil conflict, discrimination, data

*An earlier version of this paper was presented at the CSCW Workshop “Theory and Methods in the Study of Civil War” at PRIO (Oslo June 9-10, 2011). Thanks are due to the participants, especially Nils Metternich, as well as to Johanna Birnir, Heidrun Bohnet, Nils-Christian Bormann, Dawn Brancati, Fabien Cottier, Nils Weidmann, and an anonymous reviewer for their helpful comments and to the Swiss National Science Foundation for financial support through the NCCR “Democracy in the 21st century” and Grant No. 105511-116795.
† Département de science politique et relations internationales, Faculté des sciences économiques et sociales; Université de Genève; 40 Bd du Pont d’Arve; 1211 Genève 4; Switzerland; phone ++41 22 379 83 78; email: simon.hug@unige.ch
1 Introduction

The “minorities at risk” (MAR) project has generated a widely used dataset and spurred research in a field of great policy relevance. Many findings have informed decision-makers, not least Gurr’s (2000a) influential Foreign Affairs article on “ethnic warfare on the wane,” and, given its innovative character, led many researchers to new findings in various important research areas. Thus, the creator of the “minorities at risk” data, Ted Robert Gurr (1993a, x), rightly claims in his first book-length discussion of the data that “. . . there is a systematic, empirical core to the Minority at Risk study that heretofore has been lacking from the literature on this topic.” This major contribution to the study of political violence and civil conflict has, however, not been fully appreciated. Tarrow (2007, 588) laments this “oblivious[ness],” especially among economists, and convincingly argues that much of the recent quantitative literature on civil wars is heavily indebted to Gurr’s (1993a) precursor work.

Given its innovative character the dataset has been used by many scholars to address a wide array of research questions. Not surprisingly, therefore, publications on the MAR-data by Ted Robert Gurr are widely cited in the literature. The Social Science Citation Index (SSCI, accessed August 18, 2012) reports several hundred citations to the two main books drawing on the MAR-data (Gurr, 1993a, 2000b). In addition between 1993 and 2012 at least 159 scientific publications (books, book chapters and journal articles) have used the MAR-data. As it often happens, the very wealth of the data has led scholars also to analyses exposing limitations of the “minorities at risk” data and thus questionable and problematic uses. In this paper I wish to offer a discussion of the findings that this important project and dataset have made possible and highlight the strengths and weaknesses of this endeavor. At the heart of this latter assessment is the question of how a “minority at risk” is defined, and how this definition is employed to constitute a list of cases.

Consequently in the next section I first present the way in which “minorities at risk” have been defined. I then discuss how this definition is operationalized and

---

1This information is based on two sources. For 1993-2004 the MAR-project lists 61 publications (http://www.cidcm.umd.edu/mar/publications.asp and accessed June 15, 2012). For the 2005-2012 period I searched http://scholar.google.com for all references to the MAR-data and checked each book and journal article whether the author(s) used this dataset empirically and found 98 additional publications.
how that dataset has been constructed. Already this operationalization should alert us to possible “misuses” of the data. In section three I present an assessment of the advantages and pitfalls of the MAR-data. These advantages and pitfalls relate closely to the definitional issues and the way in which the data are used. Section four discusses new developments both related to the “minorities at risk” project and other data collections with similar aims, while section five concludes.

2 Defining “minorities at risk”

The roots of the “minorities at risk” project can be traced back to the early work by Ted Robert Gurr (1968, 1970) assessing “why men rebel.” His path-breaking study showed, relying on a psychological argument dealing with relative deprivation, how country characteristics systematically relate to civil strife (Gurr, 1968). Even though the theoretical argument in Gurr’s (1968) study, as well as the related work by Nordlinger (1972) and Rabushka and Shepsle (2008, 1972), relied either on arguments related to ethnic and communal groups, empirical data used to evaluate these theoretical arguments usually relied on aggregate information (for examples, see Sigelman and Simpson, 1977; Hill and Rothchild, 1986) or focused on some select cases (see Rabushka and Shepsle’s (2008, 1972) discussion of unstable plural societies and Horowitz’s (1985) seminal study on “Ethnic Groups in Conflict”). While advancing significantly our knowledge about domestic conflict processes, a systematic empirical assessment of what leads particular groups to engage in civil strife was at that time still missing.

The initial data collection on “minorities at risk,” first presented in the early 1980s, was based in part on joint work published by Gurr and Gurr (1983), offering insights on groups politically discriminated against and engaged in separatist attempts. The data presented in Gurr and Scarritt (1989) and Gurr (1990, 1992) offered for all selected groups in five year intervals (changed to annual intervals in the 1990s) measures of particular group characteristics, like for instance the extent to which the group is economically and/or politically discriminated, whether it engaged in rebellious behavior etc.

As the goal was clearly to offer global coverage over a longer time span (i.e., 1945-1989 for the first phased presented in Gurr, 1993a) the data addressed shortcomings of previous studies and data collections on civil strife. First of all, by aiming at global coverage the data allowed for a systematic assessment of var-
ious hypotheses, as one had no longer to rely on a limited set of case studies
drawing on often uneven sources (see for instance Rabushka and Shepsle (2008,
1972), Horowitz (1985)). Second, as the data offers information at the group
level, analyses no longer had to rely on single measures at the country level, for
instance, to describe economic discrimination (and/or inequality, see Sigelman
and Simpson, 1977) or the ethnic makeup of a country (through an ethnono-linguis-
tic fractionalization index, like for instance in Hill and Rothchild, 1986). Third,
relying on precise coding instructions and drawing on a multiplicity of sources the
data offered comparable information, which has been updated in various phases
(see below) and corrected as well as improved previous codings. These codings
are carried out by “. . . well-trained students (who undergo a rigorous training
period) and all coding is reviewed by senior personnel” (see http://www.cidcm.
umd.edu/mar/margene/mar-codebook_040903.pdf), and an advisory board of-
fers scientific guidance (see http://www.cidcm.umd.edu/mar/about.asp).

Early insights into the wealth of information on “minorities at risk” appeared
in Gurr and Scarritt (1989) and Gurr (1990, 1992). While these early studies
revealed slightly different definitions and lists of “minorities at risk,” with the
publication of Gurr’s (1993a) seminal book (see also Gurr, 1993b) a clear-cut
definition established itself, which has largely remained unchanged through the
different phases of the project.\footnote{SIDEBAR NOTE: See Minorities at Risk Project. (2009) ”Minorities at Risk Dataset.” College Park, MD: Center for International Development and Conflict Management. Retrieved from http://www.cidcm.umd.edu/mar/ on June 15, 2012. In addition the groups also have to fulfill population criteria (see for instance Gurr and Gurr, 1983, 8). They have to reside in a country which had at least 1 million inhabitants in 1985 and their group size has to exceed 100000 in 1990 or form at least 1 percent of the total population in one country. As these population size criteria are of secondary importance I will not discuss them any further when assessing the strength and weaknesses of this dataset.}

Communal groups are politically salient . . . if they meet one or
both of two primary criteria: . . . 1. The group collectively suffers
or benefits from, systematic discriminatory treatment vis-à-vis other
groups in a state. . . . 2. The group was the focus of political
mobilization and action in defense or promotion of its self-defined
interests at some time between 1945 and 1989. (Gurr, 1993a, 6f)

This definition can be nicely formalized using Goertz’s (2006) recommenda-
tions for forming concepts and graphically illustrated. The definition relies on two
conditions, i.e. discrimination and mobilization, each of which are sufficient to
categorize a communal group as politically salient and thus as “minority at risk.”
Figure 1 offers a graphical illustration of this definition. The two axes reflect
the two dimensions of the definition, namely whether a group is discriminated
against, and whether it is mobilized. Both dimensions are probably continua (this
is implicitly assumed in the MAR-data, as the data comprises indicators which
reflect different levels of discrimination and mobilization), and for each dimension
implicitly a cutoff point is chosen (vertical line for the level of discrimination and
horizontal line for the level of mobilization) above which a group is considered to
satisfy one of these two criteria.

The initial data collection presented in Gurr (1993a) suggests that fulfilling
the two criteria often goes together. In the first book-length presentation of
the data Gurr (1993a, 6f) listed 233 “minorities at risk,” 166 thereof were both
discriminated against and mobilized (Gurr, 1993a). They thus fall in the top-right
rectangle of figure 1. Almost a quarter of all groups, namely 50 (Gurr, 1993a,
6), are mobilized but experience no discrimination (top-left rectangle in figure 1).
And finally roughly 10 percent, namely 27 groups, are discriminated against but
are not engaged in political mobilization (Gurr, 1993a, 7) (bottom-right rectangle
in figure 1). By virtue of the data collection strategy, for the remaining rectangle
(bottom-left) no indication can be given regarding the number of groups that
might fall in this category. Gurr (1993a, 5) refers to several authors (namely
Nietschmann, 1987; Nielson and Jones, 1988) estimating that there are “three
thousand to five thousand ‘nations’,” or “575 ethnic groups as actual or potential
joshuaproject.net/) with a list of 16598 ethnic groups.

Figure 2 reports how the number of minorities covered in the MAR-data has
changed over time (see for a similar depiction Sorens, 2010, 544). While the first
phase reported by the end of 1990 233 “minorities at risk” (however 227 such
groups are reported in Gurr, 1993a, 6, 9f) at the beginning of the post world
war II period, only slightly more than half of these groups fulfilled the criteria

---

3 The data used here and below come from Version 2 release (2005) of the “minorities at
risk” data covering 1940 to 2003. Thus, I do not cover the update bringing the coverage of the
data up to 2006. Also, I do not use the A-MAR (“All MAR”) data (Birnin, Wilkenfeld, Fearon,
Laitin, Gurr, Brancati, Saideman and Pate, 2012) as it uses a quite different definition and has
not yet led to any publications (see below).
discussed above. The second and third phases (1990 and 1996) expanded the list of groups by roughly 40 while the fourth phase (1998) increased the number of groups just by a handful.

Figure 2 about here

Figure 2 depicts the distribution of “minorities at risk” at four moments in time which correspond to the start dates of each of the four phases.\footnote{This figure was produced with Weidmann, Kuse and Gleditsch’s (2012) \texttt{cshapes} package (see also Weidmann, Kuse and Gleditsch, 2010).} I report these graphical illustrations as both Fearon and Laitin (1999) and Birnir, Wilkenfeld, Fearon, Laitin, Gurr, Brancati, Saideman and Pate (2012) (see also Birnir, Wilkenfeld, Brancati, Fearon, Gurr, Laitin, Pate and Saideman, 2011) suggest that particular world regions are not well covered by the “minorities at risk” data. Figure 3 shows that the “coverage” has changed most dramatically in Africa and Asia (India being the most notable example) where the number of “minorities at risk” increases considerably. Obviously this is in part due to the fact that many of the “countries” were colonies at the beginning of the period covered. At the same time, however, Birnir, Wilkenfeld, Fearon, Laitin, Gurr, Brancati, Saideman and Pate (2012) argue that while the original MAR-data had fewer observations in all world regions, particularly in Asia many more groups could be identified in the A-MAR project (see below).

This reduced coverage is likely due to the definitional structure and the data collection strategy. Relying on student coders to evaluate a group’s level of discrimination and mobilization suggests that it is at least in part a subjective assessment, as Fearon and Laitin (1997) rightly argue, and thus prone to measurement error. In a later paper Fearon and Laitin (1999, 4) even argue that “[t]he . . . criteria do not appear to have been operationalized and applied in any very systematic way.” The eventuality of measurement error is explicitly acknowledged in the most recent report on “Peace and Conflict” by Gurr, Marshall and Khosla (2005, 29): “Occasionally, new groups come to our attention as information improves and qualifying groups are added to the dataset.” More precisely, with respect to both criteria two types of error might occur. Type I error would consist of considering a group as discriminated against (or mobilized), even though it is not, while type II error would consist of considering a discriminated (or mobilized) group as not being discriminated against (or mobilized). Various formulations in the earlier studies (e.g., “[t]he roster of 233 groups is an
imperfect approximation of the universe of politically salient communal groups at the beginning of 1990.” (Gurr, 1993a, 10) suggest that the fear mostly focused on possible type II errors. In many writings the caution regarding the coverage was soon replaced by bolder statements suggesting that the MAR-dataset covers all politically salient communal groups. For instance Jenne, Saideman and Lowe (2007, 545) note that “[t]he dataset therefore comprises the universe of groups that could plausibly advance collective demands against the center.”

Whether this is problematic depends, however, on the research question at hand. Thus, for instance, Wimmer, Cederman and Min (2009, 317) argue that the “. . . Minorities at Risk data set . . . focuses exclusively on disadvantaged minorities and is thus unable to capture the dynamics of ethnic politics at the power center.” Consequently studies dealing with relationships among groups at the center of power might be ill advised to rely on MAR-data.

Fearon and Laitin (1999, 5) acknowledge the possibility of bias in their study on ethnic violence based on MAR-data: “Thus, the cases have been selected by criteria that may be associated with our dependent variable of interest, violence. This implies that we cannot generalize our empirical results beyond the set of ethnic groups already involved in some political activity or significantly deprived.” In an earlier study using MAR-data Fearon and Laitin (1997) mentioned selection bias in existing studies on ethnic violence, but did not consider it in the context of using MAR-data to study ethnic violence. Even more explicit is the assessment by Fearon (2003, 196) who states

The difficulty is the same as that of trying to learn the effect of SAT scores on academic performance by looking only at elite colleges. If we consider only oppressed or disadvantaged groups, we are truncating variation on the independent variable, and thus making it harder to detect a relationship between (say) discrimination and rebellion.

It is important to note that these three statements raise different concerns, all of which are, however, related to the research question at hand. The first concern suggests that for specific research questions information on non-discriminated groups (possibly majorities) is necessary. The second raises the issue that the set of “minorities at risk” might not be a complete list, and thus inferences generated on the basis of this subset might only describe this particular subset. Note,
however, that Fearon and Laitin (1999, 5) do not contemplate measurement error in the criteria used to select groups, which almost inevitably would lead to biased estimates in a study on violence. Finally, the last (and probably least important) is concerned about efficient estimation of effects. The latter is obviously increased if our independent variables have considerable variation. As the MAR-data includes also non-discriminated groups as long as they are mobilized, strictly speaking the concern raised by Fearon (2003, 196) is moot. For most research questions, however, lack of efficiency is the smaller of two problems compared to likely biases. Consequently in what follows I will focus on the problem raised in the quote by Fearon (2003, 196), but show that not efficiency is the main problem but biased estimates. Whether biases will result from using MAR-data, depends, however, crucially on the research question. Hence, it is not the MAR-data that inevitably produces selection biases in our estimates of causal effects, but MAR-data used for specific inferential tasks.

3 Advantages and pitfalls in studies using MAR

These issues have to be considered when assessing the contributions as well as the advantages and pitfalls of using the MAR-data in empirical studies. For many purposes the data provided offer all the information needed to draw correct and unbiased inferences, while for some research questions inferences based on MAR-data might very well result in biases. Needless to say that this applies probably to a broad majority of datasets widely used in the social sciences.

Generally scholars have used “minorities at risk” data to either describe and explain characteristics of various groups or study related phenomena at a country-level. Hence, studies can be distinguished on the one hand by their unit of analysis, i.e., the individual group or a country, and on the other by the descriptive or explanatory goal. Because of the development of the literature I will first discuss studies using groups as units of analysis before moving to work dealing with aggregated information at the country level. In both subsections, but mostly in the first, I will present first descriptive studies before moving to research with a more explanatory scope.
3.1 “Minorities at risk” as units of analysis

Not surprisingly, early studies presenting and using MAR-data had a largely descriptive focus. Scholars attempted to assess where “minorities at risk” were frequent, what characteristics they shared, and what means they employed to further their goals. Gurr and Scarritt (1989), one of the first articles in print giving a glimpse at “minorities at risk,” offered a detailed discussion of definitional issues and also motivated the label of “minorities” “because most are outnumbered by other groups” (Gurr and Scarritt, 1989, 375). As some discriminated groups are majorities, these are also included, as are some advantaged minorities (Gurr, 2000b, 7) which leads Fearon and Laitin (1997) to rightly note that this leads to confusion. In addition Gurr and Scarritt (1989, 387) state that “our initial roster . . . thus includes all types of advantaged minorities.” Nevertheless, Gurr and Scarritt (1989, 393) can offer precious insights, for instance that about a quarter of all “minorities at risk” have separatist goals, and that only roughly a tenth of all groups are advantaged minorities.

Both Gurr’s (1993b) article and his book-length treatment (Gurr, 1993a) offer various detailed insights on communal groups around the world (see also chapters 1 and 2 in Gurr, 1994). Most communal groups reside in African countries (Gurr, 1993a, 11ff), face considerable discrimination and are geographically concentrated, (Gurr, 1993a, 34ff) and in some instances they turn violent (Gurr, 1993a, 89ff). Later studies (e.g., Gurr, 2000b) have largely dwelt on this very politically relevant aspect. Similarly, several studies have attempted to assess how ethnic violence evolved over time (e.g., Gurr, 1993b; Gurr, 2000a; Olzak, 2006). Most notably, the article by Gurr (2000a) alerted scholars and policy-makers alike to a decrease in ethnic conflict in the last part of the 20th century. Similarly the reports on “Peace and Conflict” (for instance Gurr, Marshall and Khosla, 2005), provide detailed information on various aspects of the MAR-project.

The relevance of this data becomes apparent if these insights are compared with Rabushka and Shepsle’s (2008, 1972) and Horowitz’s (1985) influential studies on divided societies. While both books offer detailed case studies of particular countries and specific groups, the information presented does not offer the same systematicity in terms of comparable measures. Similarly, the MAR-project relies on a clear definition and operationalization to identify the relevant groups. While both Rabushka and Shepsle (2008, 1972, 20f) and Horowitz (1985, 51f) offer clear
definitions, given their largely qualitative perspective, their operationalization is much more lose-handed. Hence, the discussion of various groups in these studies and their characteristics does not attain the same level of systematicity as these initial descriptive studies of the MAR-data.

Such descriptive inferences based on MAR-data are certainly valid as long as the assumption holds that all groups are included in the data. As various revisions and extensions suggest, and as the basic criteria are subjective assessments, measurement error might lead to a slightly biased description of the population from which the “minorities at risk” are selected. In some sense these biases are akin to the consequences of sampling errors in survey research.

When it comes to causal inference, some formalism might help to understand under what circumstances the largely unavoidable measurement error of discrimination and mobilization might cause problems. Let us assume that we wish to explain a particular feature of a communal group, namely \( y \) with a set of explanatory variables \( x \). Thus, in a classical linear regression model we might have (see for related treatments Hug, 2003, 2006):

\[
y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} + \epsilon_i \tag{1}
\]

When using MAR-data to carry out our analyses only on the bases of the observed cases, namely those for which the measures indicate a sufficient level of either discrimination and/or mobilization implying that \( t \), the selection indicator, is equal to 1 (. indicating missing data):

\[
y_i = \begin{cases} 
y_{i,obs} & \text{if } t_i = 1 \\
\cdot & \text{else if } t_i = 0
\end{cases} \tag{2}
\]

From the definition of “minorities at risk” we know that \( t \) is related to some independent variables \( z \), most notably to the degree of discrimination and mobilization. Thus, we have

\[
t^*_i = \delta_0 + \sum_{j=1}^{l} \delta z_{ji} + \theta_i
\]

and

\[
t_i = \begin{cases} 
1 & \text{if } t^*_i \geq 0 \\
0 & \text{if } t^*_i < 0
\end{cases} \tag{3}
\]
It is easy to show that if the errors of equations 1 and 3 are uncorrelated (i.e., $\sigma_{\epsilon,\theta} = 0$) then assessing the effect of the $x$s on $y$ on the basis of the observed cases is unproblematic (e.g., Heckman, 1979). This may occur in two different ways. First, it might be the case that our independent variables in equation 3 perfectly explain the selection of cases, thus making $\theta$ equal to zero for all cases. As the revisions in the list of “minorities at risk” suggest, the selection of groups has been affected by some error. In the present case this comes about by the fact that the discrimination and mobilization variables are imperfect measures of the real values. This nicely transpires from figure 1. The reason why a group might not be selected (i.e., be in the lower-left rectangle) can only be a mismeasured level of discrimination or mobilization. Hence, why the independent variables in equation 3 might not perfectly explain the selection as a “minority at risk” is due to measurement error in these two key variables (and possibly of other ones that influenced coders and their supervisors to include a group).

Second, and more likely, if the error ($\theta$) in the selection equation is not at all linked to the variable ($y$, or more precisely $\epsilon$ in equation 1) we wish to explain, estimating with MAR-data the effect of the independent variables on $y$ would yield unbiased estimates, provided that no other assumptions of a classical linear regression model (in this case) are violated. If neither of these cases can be reasonably made, selection bias is likely and increases in the correlation between $\epsilon$ and $\theta$. In general the bias leads to an attenuation of the estimated effect, i.e., coefficients will be biased toward zero, either from below, if the estimated coefficient is negative, or from above if the coefficient is positive (for careful derivations and illustrations of this point, see Heckman, 1979; Achen, 1986; Geddes, 1991).

Two studies may help illustrate under what circumstances using MAR-data is unproblematic because the error in equation 1 is uncorrelated with the error in equation 3. Chai (2005), for instance, wishes to explain the characteristics of “minorities at risk,” like for instance whether they are religiously or linguistically distinct. As explanatory variables she uses, among other variables, the population size of the largest (religious, resp. linguistic) denomination and its square term. She argues if the largest religious (resp. linguistic) group approaches a size close to half the size of the population groups are more likely to distinguish themselves according to a religious (resp. linguistic) marker. In this particu-

---

5Obviously, the same would apply if our explanatory model would perfectly predict our outcome in equation 1 but this is even less likely.
lar case it is exceedingly unlikely that factors determining whether a group is religiously or linguistically distinct but not taken into account by Chai (2005) (i.e., omitted variables in equation [II] are somehow related to errors in the selection equation. Consequently, her finding that as a religious (resp. linguistic) denomination increases in size it becomes more likely that this particular marker defines the relevant groups, before decreasing this likelihood as the population share becomes dominant (above approximately 65 percent), is very likely to be unbiased.

One might expect the same thing to hold for Saideman’s (2002) study on what explains that some ethnic groups find support from external actors, while others do not. In this particular study external support, i.e., the dependent variable in equation [I] is most likely largely independent of any measurement error in the degree of discrimination and mobilization (as is implicitly assumed by Saideman, 2002). This largely insures that what one may call the selection equation [III] is independent of the outcome equation [I]. Hence, Saideman’s (2002) findings that ethnic kinship considerably affects the kind of support an ethnic group receives from external actors is most likely based on unbiased estimates.

More problematic are studies that focus directly on explaining either discrimination and mobilization or characteristics of groups intimately related to these two elements figuring prominently in the definition of “minorities at risk.” In these instances using MAR-data results in cases being selected on the dependent variable (see Achen, 1986; Geddes, 1991). For instance, in Cetinyan’s (2002) study on whether a group faces less political and economic discrimination if a kin-group lives in a neighboring country is likely to be affected by inferential problems. As discrimination is one of the criteria for selection (and probably measured with error), we have a case where observations are selected on the basis of the dependent variable. In this case the y in equation [I] is actually part and parcel of the z in equation [III]. In Cetinyan’s (2002, see especially 666) case this is likely to have attenuated the estimated effects of his various predictors, both at the country- and the group-level. Regarding his main finding, namely that the resources of a group, either its own or those provided by a kin, reduce its level of discrimination, is consequently in reality likely to be stronger than what he finds in his empirical analysis. The same should apply to Fox’s (2000) study on what affects the level of various forms of discrimination of ethno-religious and other
minority groups. As the dependent variable is again quite directly part of the selection mechanisms, his estimates are quite certainly biased.

Similarly, Gurr and Moore’s (1997) study assessing how rebellion, repression, mobilization and grievance interact is likely to be problematic as these various factors are, as the authors demonstrate, more or less directly related to the two dimensions used to select “minorities at risk” (see also Olzak, 2006). Consequently, the findings that mobilization positively influences rebellious behavior, but that grievances do not, has to be taken with a grain of salt. Hug (2003) offers an evaluation of this study and finds that grievances are quite strongly related to rebellion, while differences in mobilization cannot account for the level of rebellious behavior, contrary to Gurr and Moore’s (1997) results.

A host of studies have followed similar tracks. Cohen (1997) as well as Saide-man, Lanoue, Michael and Stanton (2002) propose to assess how political institutions explain the more or less violent behavior of “minorities at risk.” Both sets of authors find that groups in federal countries, for instance, are more inclined to protest behavior but engage much less frequently in rebellions. They also concur that democracies with a proportional representation electoral system are less conflict prone. Similarly, Toft (2002) (see also Toft, 2003) proposes to assess the relationship between a group’s geographical concentration and its rebellious behavior and finds a strongly positive effect. More precisely she finds that concentrated majorities are most likely to engage in conflict (see also Fearon and Laitin, 1997, 1999). As both protest and rebellious behavior almost by definition require some sort of mobilization, or at least is related to it, one must be worried about possible biased inferences in studies of this type. Birnir (2007, 185), on the other hand, uses MAR-data to assess whether the exclusion from the cabinet of a minority might be linked to violence. She finds that rebellions become more likely and violent the longer an ethnic group is excluded from the cabinet. As in the studies discussed above, this, and several other important findings, might be subject to questions, as the extent of violent behavior of a group is certainly not unrelated to its level of discrimination and mobilization, i.e., the factors used to select the observations in Birnir’s (2007) analyses.

Similarly, Saideman and Ayres’s (2000) findings that ethnic kins with separatist demands increase the likelihood of irredentism and secessionism, but does
not affect discrimination might be possibly problematic. As is their finding that
group concentration is important for secessionism but not irredentism. Again,
one might be worried that the selection of cases used might affect these results,
as those of Walter (2006). She assesses what explains the outbreak of secessionist
violence using “minorities at risk” as her unit of analysis. As such violence
has clearly to do with discrimination and mobilization, her findings, for instance
that the economic status of a group has no effect, might be questionable. Her
main finding, that previous accommodation by the state of demands for auton-
omy increases the likelihood of further conflict, is unlikely to be affected, as the
bias due to selection problems would be downwards. Related worries should also
apply to Fox’s (2003) study whether civilisational differences between minority
and majority group might explain rebellions. As the latter are likely to relate to
the selection of “minorities at risk,” the null-finding regarding these civilisational
differences should be taken with a large grain of salt.

Finally, Jenne, Saideman and Lowe (2007) offer a study of what explains
claims for autonomy and separation as well as rebellions of ethnic groups. They
find that concentrated groups and those with external military support make
more extreme demands, i.e., secessionism instead of autonomy. In parallel, groups
with some autonomy leads to more extreme demands, but does not affect the
likelihood and extent of rebellions. Contrary to their optimistic assessment (see
also above Jenne, Saideman and Lowe, 2007, 545) their analysis might well pose
problems, as making claims and rebellious behavior is quite likely related to the
two dimensions of the selection criteria. Thus again, the estimated effects on
which they rest their conclusions are probably biased.

3.2 Aggregated “minorities at risk” as units of analysis

In part due to the easy way in which information from the MAR-data can be
aggregated to the country level with the help of Bennett and Davenport’s (2003)
MARGene program a series of studies has proposed analyses at the country level.
Obviously, proceeding like this vitiates in part the main contribution of MAR,
namely the information at the level of groups instead of having of having to rely
on country-level characteristics. Nevertheless, already Gurr and Scarritt (1989)
offered important descriptive insights on the basis of aggregated MAR-data. For
instance, they highlighted that in more than forty countries at least 25 percent
of the population belonged to “minorities at risk,” reaching even 100 percent for South Africa and Chad (as well as the Israeli-occupied territories) (Gurr and Scarritt, 1989, 396).

More frequently, however, country level aggregations of MAR-data were used for explanatory purposes. Hence, the question arises again under what circumstances such analyses might lead to biased inferences. Using the same formalism as above, but using countries \((c)\) as units of analysis (possibly country-years) equation 1 becomes the following

\[
y_c = \begin{cases} y_{c,\text{obs}} & \text{if } t_i = 1 \text{ (for at least one group } i \text{ in country } c) \\ 0 & \text{else if } t_i = 0 \end{cases}
\]  

(4)

Here \(y_{c,\text{obs}}\) is an aggregate of \(y_{i,\text{obs}}\) from equation 1, for instance the mean level of rebellion among minorities at risk in a particular country, i.e., \(y_{c,\text{obs}} = \frac{1}{N_c} \sum_{i \in c} y_{i,\text{obs}}\) (where \(N_c\) is the number of groups in country \(c\)). In that case we implicitly assume that there are no rebellious activities in countries without “minorities at risk.” If, however, a “minority at risk” has been missed we have a dependent variable measured with error. The same problem may appear if the aggregation that is being used, e.g., maximum, median, etc., is affected by the omission of a group. Related problems appear if aggregated information from the MAR-data is used as independent variable(s). Omitted groups lead again to measurement error this time, however, on the right-hand side of our equations.

In classical linear regression models the consequences of such measurement error are reasonably well understood. If there is only measurement error in the dependent variable we will get less precise estimates of our coefficients if the error is random, and in case of systematic measurement errors, only our estimated constant will be biased. This might be considered as less problematic, but if models are used to make predictions, the latter will be biased as well. Should the measurement error occur in the independent variable(s), however, we have only detailed knowledge about the effects if a single variable is affected (see for instance Greene, 2003, 83ff). In this case, the estimated relationship between the affected independent variable and the dependent variable will be attenuated if the error is random, while additional bias may occur if this error is related with the error term of the regression. Most studies proceeding in this fashion, however, rely on nonlinear models, where measurement errors in the dependent and independent
variables have much more serious consequences (Hausman, 2001). As Hausman, Abrevaya and Scott-Morton (1998) show, measurement error in the dependent variable affects all estimated coefficients in nonlinear models (see also Hug, 2010).

This problem may affect several recent studies employing “minorities at risk” data. For instance, Regan and Norton (2005) employ this dataset aggregated to the country-level to assess whether the same set of factors explain the onset of protest, rebellions and civil wars in specific years. Controlling for lagged repression as measured by the Political Terror Scale (Poe, Tate and Keith, 1999) they find this factor to affect each level of activities, while discrimination only has a statistically significant effect on the onset of rebellions and civil wars. Rebellions appear also to be affected by the degree of democracy but the latter appears to be unrelated with the onset of civil wars. Similarly, Saideman and Lanoue’s (2005) study, focusing on elections and electoral institutions, provides apart analyses at the level of “minorities at risk,” also analyses at the country level. These authors find that having a fixed schedule for elections reduces the extent of protest and rebellions but come to this result by employing a classical linear regression model. Consequently, while the inferences from the former study are quite likely to be affected by biases due to the data and the nonlinear model employed, the latter study might be affected as well, but to a lesser degree given the linear model employed.

The previously discussed studies employ information from the MAR-project to create their dependent variable (and also some of their independent variables). Piazza (2011), however, relies mostly for his independent variables on this data project, more precisely, he measures the extent of economic discrimination of minorities by aggregating this information of all minorities residing in a particular country. He then finds that higher levels of economic discrimination of minorities increases the number of terrorist attacks. The presence of minorities that face no economic discrimination, however, reduces the number of such attacks. As possible problems of coverage in the MAR-data lead in this case to a mismeasured independent variable, and in addition the latter is used in a nonlinear model (an event count model), the biases introduced are difficult to characterize but likely to exist.
4 Developments

Several scholars attracted by the usefulness of the MAR-data also noted that for their research question it had to be broadened. This development took in part place in parallel to studies by economists arguing that heterogeneous societies had negative consequences, for instance for economic development. Hence, Alesina, Baqir and Easterly (1999) employ in their study on how ethnic divisions affect public good provisions, and thus economic growth, an ethno-linguistic fractionalization index based on the *Atlas Narodov Mira* (Bruk, 1964). This atlas provides a list of ethnic groups as determined by Soviet ethnographers in the 1960s and has been widely used in the literature. Obviously, compared to the MAR-data, it offers much sparser information on the groups covered.

Relying on a definition of ethnic groups (and data for that matter) of Soviet scholars raises some broader issues nicely discussed by Posner (2004) (for insightful discussions of how to define ethnic groups, see Laitin and Posner, 2001; Chandra, 2006). Criticizing the blind trust in the Soviet scholars’ work he argues that “. . . there is no single ‘correct’ accounting of the ethnic groups in a country . . .” (Posner, 2004, 850). As Alesina, Baqir and Easterly (1999) suggest that ethno-linguistic fractionalization affects public goods provisions, Posner (2004, 854) makes the point that only “politically relevant groups” should be retained when calculating a fractionalization index. In addition he convincingly argues that “an index of ‘politically relevant groups’ implies a subscript that specifies the mechanism for which the groups are relevant” (Posner, 2004, 854).

As a series of scholars has raised more generally the issue of selection biases in observational studies (e.g., Achen, 1986; Geddes, 1991; King, Keohane and Verba, 1994; Hug, 2003) some studies focusing on “minorities at risk” attempt to alleviate some of the problems discussed above. Brancati (2006) (see also Brancati, 2009), in her study on how decentralization through the creation of regionalist parties may increase political violence, complements the data with additional ethno-linguistic groups that are regionally based but not included in MAR. For example, the absence of “minorities at risk” in Belgium despite the

---

creation of regional parties mostly in the 1970s suggests that for Brancati’s (2006, 2009) theoretical argument a broader set of groups is relevant. Consequently she adds information on groups in Belgium, Czechoslovakia, Norway, Poland, Slovenia and Sweden to her dataset covering some thirty mostly western democracies. Based on her analyses she finds that decentralization increases the likelihood of conflict through regional parties, but that the emergence of the latter may be hindered by specific electoral rules.

Ishiyama (2009, 65) proceeds in a similar fashion by “. . . first identifying an appropriate universe of cases exogenously to MAR.” Unfortunately, identifying these “missed” cases is far from self-evident and believing that such a selection can be done “exogenously” is a bit of a profession of faith. Sorens (2010), in his study on what explains discriminatory policies, proceeds in a similar way, addresses the evident selection problem, however, more directly. More specifically, he collects information on some 500 additional groups and estimates a model that takes directly the selection process into account. His estimates suggest that the errors of the selection and outcome equations are not correlated, (i.e., the hypothesis that $\sigma_{\epsilon,\theta} = 0$ cannot be rejected). Both Christin and Hug (2003) and Hug (2003) proceed similarly, rely, however, on a model that does not require to collect information on “not selected” groups. In their replications of Cohen’s (1997), respectively Gurr and Moore’s (1997), studies, they find, contrary to Sorens (2010), considerable evidence for selection biases.

Based on the absence of a correlation between the error terms in the two equations Sorens (2010, 536) concludes that “. . . selection bias does not affect the results, indicating that groups excluded from the MAR project may be as likely to suffer discrimination as included groups.” This rather devastating conclusion for the MAR-project (there is as much discrimination among omitted groups as those included) is luckily enough not warranted. The absence of covariation in the error terms ($\sigma_{\epsilon,\theta} = 0$ ) simply suggests that taking into account the explanatory variables for the selection equation, the remaining variance in the selection variable is not related to omitted variables and measurement error in the outcome equation. Sorens’s (2010, 550) explanation that the selection equation might be “specified exceptionally well” is rather doubtful, as well. The percent of correctly predicted cases is 73 percent, but since the modal category of his “selection” variable makes up 59 percent (597 out of 847), the proportionate reduction
in error (PRE) is a modest 0.34. Nevertheless, Sorens’s (2010) study sketches one way to insure that possible selection biases do not affect the estimates of the substantively interesting effects. Thus, we can have more confidence in his results suggesting that executive constraints and economic development reduce somewhat discrimination, while ethnic control of a region reduces the maintenance of discrimination, but increases the likelihood of seeing the initiation of new discriminatory policies.

Following in Posner’s (2004) footsteps Wimmer, Cederman and Min (2009, 319) (see also Cederman, Wimmer and Min, 2010) created a dataset of “all politically relevant ethnic groups” (emphasis added) since World War II. These groups are defined based on the following conception of ethnicity and political relevance (Wimmer, Cederman and Min, 2009, 325):

[E]thnicity [is defined] as a subjectively experienced sense of commonality based on a belief in common ancestry and shared culture. . .

. An ethnic category is politically relevant if at least one significant political actor claims to represent the interests of that group in the national political arena, or if members of an ethnic category are systematically and intentionally discriminated against in the domain of public politics.

While this definitional structure is not very different from the one adopted by the MAR-project, its emphasis on “claims to represent the interests of [a] group” is much broader than MAR’s “political mobilization and action in defense or promotion of its self-defined interest” (Gurr, 1993b, 6f). Thus, it cannot surprise that the list of such “politically relevant ethnic groups” is larger than the one of “minorities at risk.” More precisely Min, Cederman and Wimmer (2011, 3) have information on “733 groups in 155 countries” and their access to executive power in their ethnic power relations (EPR) data.

An even more expanded list of ethnic groups is provided in the “A-MAR” project. Dropping from the definition both the discrimination and mobilization dimension Birnir, Wilkenfeld, Brancati, Fearon, Gurr, Laitin, Pate and Saideman (2011, 12f) they focus on a common descent of the members of the group, a

---

8I wish to thank Lars-Erik Cederman, Brian Min and Andreas Wimmer for having given me some background information on their paper comparing in part the EPR-data with the MAR-data.
recognized membership, a cultural distinction, a feature of which is practiced by majority, and a size of at least 100’000 or 1 percent of the population of a country. Using this definition these authors have found around 900 groups, half of which were not part of the MAR-data and did not appear in Fearon’s (2003) list of ethnic groups. As the A-MAR acronym, standing for ALL-MAR suggests, Birnir, Wilkenfeld, Fearon, Laitin, Gurr, Brancati, Saideman and Pate (2012) claim to be able to replace the question mark in the bottom-left rectangle of figure [1] with an exact number. Apart identifying this exact number the A-MAR project in a first steps also codes for a random sample of the additional groups identified (77 at the present time) most variables from the MAR project. Proceeding like this, provided that the “universe” identified by Birnir, Wilkenfeld, Fearon, Laitin, Gurr, Brancati, Saideman and Pate (2012) is accurate allows for an adequate correction of biases that might result from using the MAR-data. This, however, presumes that the list with approximately 900 groups has been based on information from which we can exclude measurement error.

Both Min, Cederman and Wimmer (2011) and Birnir, Wilkenfeld, Fearon, Laitin, Gurr, Brancati, Saideman and Pate (2012) compare the EPR-data with MAR, respectively A-MAR. While there are no glaring problems that would suggest an incorrect use of the definitions used, Birnir, Wilkenfeld, Fearon, Laitin, Gurr, Brancati, Saideman and Pate (2012) argue that when using A-MAR to explain the level of rebellion, the results differ from those obtained when using the EPR-data to explain civil war onsets. As the rebellion-variable from the MAR-project also covers low-intensity political violence, this has obviously less to do with the set of groups studied but the nature of the dependent variable. This transpires even more when the rebellion variable is limited to large scale onsets leading to reappearance of identical results as with the EPR-data.

These efforts to broaden the coverage of “minorities at risk” or “politically relevant ethnic groups” are certainly to be applauded, but Posner’s (2004, 854) incisive remarks should nevertheless not be forgotten. These data have been created for particular goals and might not be adequate for others. In addition, the pretension to cover all groups is rather ill-placed given the information and measurement problems we face in the social sciences. A bit of modesty might well be in order, as even Carl von Linné (1753) refrained from entitling his seminal work Omnes species plantarum.
5 Conclusion

Undeniably, the “minorities at risk” project and data have opened the way for much more systematic analyses of ethnic violence and mobilization that were not possible before. Especially, it allowed for the first time systematic empirical studies at the level of groups affected by discrimination or having mobilized politically. A series of studies has relied on this important data to uncover new facets of relationships between communal groups and the state. Some scholars have, however, also employed the data for studies well beyond the original domain for which it was designed. As this paper has tried to show, in some areas this may have led scholars to biased inferences. Why this might happen is closely related to the definitional structure chosen to identify “minorities at risk” and the unavoidable measurement error linked to the subjective criteria employed. As the definitional structure is quite explicit, it is straightforward to characterize the situations under which MAR-data can be used without any major inferential problems. Hence, contrary to some claims that can be found in the literature, the MAR-data itself is not affected by selection bias.

Nevertheless, the data has allowed researchers to document important trends based on the MAR-data and provide interesting insights into various characteristics of relationships among communal groups. The MAR-data has also been put to profitable use to assess what explains protest and rebellions of communal groups. Several scholars have highlighted that political institutions seem to play a considerable role, that ethnic kinship across borders affect the behavior of groups, etc. Many of these results need to be digested with caution, as inferences based on the MAR-data for such questions might be affected by selection biases. Again it is worth emphasizing that these biases depend on the research question and are not inherent in the data.

Finally, several studies have studied data stemming from the MAR-data at the country level. By aggregating data from the communal level to this higher level, the problems of measurement discussed above result in measurement error in either dependent or independent variable (or both). Again, some studies, given this measurement error, might face considerable inferential problems.

Based on this assessment it is quite clear that improvements undertaken in the data collection and the expansion of the list of cases to be included in the “All MAR” (Birnir, Wilkenfeld, Brancati, Fearon, Gurr, Laitin, Pate and Saideman,
2011; Birnir, Wilkenfeld, Fearon, Laitin, Gurr, Brancati, Saideman and Pate, 2012) is a laudable effort. Some notes of caution might all the same be advised. Referring back to figure 1 it seems illusionary to replace once and for all in the bottom-left rectangle the question mark with a precise number of communal groups that are not discriminated and not mobilized and thus do not belong to the MAR (or A-MAR) universe. Consequently, selection biases will not be eliminated by fiat through a longer list of cases. It is quite likely that a broader universe will alleviate or even eliminate possible inferential problems for studies focusing on particular research questions. Scholars dealing with other research questions will, however, still need to assess whether the new and expanded criteria to identify members of A-MAR might lead to inferential problems.
References


Nordlinger, E. A. (1972), *Conflict Regulation in Divided Societies*, Center for International Affairs, Harvard University, Cambridge.


Figure 1: Number of “minorities at risk” and the definitional structure
Figure 2: Number of “minorities at risk” over time
Figure 3: The geographical distribution of “minorities at risk” (1945, 1990, 1996, and 1998)