Milcher, Susanne; Fischer, Manfred M.

Conference Paper

On labour market discrimination against Roma in South East Europe

50th Congress of the European Regional Science Association: "Sustainable Regional Growth and Development in the Creative Knowledge Economy", 19-23 August 2010, Jönköping, Sweden

Provided in Cooperation with:

European Regional Science Association (ERSA)

On labour market discrimination against Roma in South East Europe

Susanne Milcher and Manfred M. Fischer
Vienna University of Economics and Business

Abstract. This paper lies in the tradition of statistical decomposition analysis popularized by Blinder (1973) and Oaxaca (1973), and aims to measure labour market discrimination against Roma in five South East European countries (Albania, Bulgaria, Croatia, Serbia and Kosovo). We use microdata from the 2004 UNDP household survey and apply a Bayesian approach, proposed by Keith and LeSage (2004), for the decomposition analysis of income differentials. Statistical inference for both discrimination and characteristics effects estimates are based on Markov Chain Monte Carlo (MCMC) estimation. Variance estimates derived from this method of estimation are known to reflect the true posterior variance when a sufficiently large sample of MCMC draws is carried out. The results obtained in this paper indicate the presence of statistically significant discrimination effects in Albania and Kosovo, but not so in Bulgaria, Croatia, and Serbia. In these countries differences in measured characteristics and not labour market discrimination are the overwhelming reason for the shortfall in incomes for Roma.

Keywords: Labour market discrimination, income differential decomposition, Bayesian regression model, Markov Chain Monte Carlo (MCMC) estimation, Roma, Europe

JEL classification: J71, C11, C50, O52

Paper to be presented at the 50th European Congress of the Regional Science Association, 19-23 August 2010, Jönköping, Sweden

1 Introduction

Roma are a unique minority in Europe. They have no historical homeland and are found in nearly all European countries. Current estimates suggest that seven to nine million Roma live throughout Europe, making them the largest minority in Europe. While some Roma groups are nomadic, the vast majority have settled in South East Europe, some during the Austrian-Hungarian and Ottoman empires, and others more recently under socialism (Revenga et al. 2002).

The collapse of the socialist regimes in South East Europe created new opportunities for all citizens, including Roma. For the first time in decades, minorities were able to express their ethnic identity, participate in civil society, and engage in previously forbidden economic activities. But these gains have been offset by a dramatic reduction in opportunities in many respects. For many Roma, the collapse of the socialist system has led to an erosion of security in jobs, housing and other services, and in the absence of viable economic opportunities to increasing poverty.

Indeed, Milcher (2009) identified Roma as one of the main poverty risk groups in South East Europe. Roma are both poorer than other population groups and more likely to fall into poverty and remain poor. The causes of Roma poverty are intertwined with many of the factors that are correlated with poverty throughout South East Europe, including low education levels, unemployment and large household sizes. The unfavourable starting point of Roma at the outset of the transition period – with low education levels and an overrepresentation among low skilled jobs – did also lead to disadvantages on the labour market.

Despite a general awareness of labour market discrimination against Rome in these countries, information on labour market discrimination is scarce, fragmented and often anecdotal. Measurement problems are daunting and include undersampling in censuses and household surveys, the reluctance of many Roma to identify as Roma, and the incredible diversity of Roma groups and subgroups. One notable exception is the cross-country study of Roma poverty by Revenga et al. (2002) that provides evidence for labour market discrimination in Bulgaria, Hungary and Romania.

With this study we share the ambition to measure labour market discrimination against Roma in South East Europe, based on statistical decomposition analysis, first employed in demography by Kitagawa (1955) and later popularized by Oaxaca (1973) and Blinder (1973) in the economics literature. But we depart from this previous work in several major respects. *First*, we note that Revenga et al. (2002) do not report standard errors or confidence intervals for the decomposition components. Indeed, it is hard to evaluate the significance of reported decomposition results without knowing anything about their sampling distribution. This motivates us to use a Bayesian approach to decomposition analysis, based upon Markov Chain Monte Carlo (MCMC) estimation. And this approach allows – without relying on asymptotic theory – to test the significance of characteristics and discrimination effects estimates. Variance estimates derived from MCMC estimation are known to reflect the true posterior variance when a sufficiently large sample of MCMC draws is carried out (Gelfand and Smith 1990).

Second, in order to account for the impact of non-constant variance in the semi-log regression relationships[1] on the resulting inferences, we apply the improved Bayesian approach suggested by Keith and LeSage (2004) to statistical inference for both characteristics and discrimination effects estimates based on MCMC estimation. *Third*, we use the most recent available data source, the 2004 **U**nited **N**ations **D**evelopment **P**rogramme (UNDP) dataset, a survey designed and implemented by a team of researchers of the UNDP Bratislava Regional Centre. This survey provides information on the living conditions of Roma across countries in South East Europe, including Albania, Bosnia and Herzegovina, Bulgaria, Croatia, Kosovo, Macedonia, Montenegro, Romania and Serbia. Roma households and individuals were identified using a multifaceted approach, including questions on self-identification, interviewer identification, language and parents' language, and family name. Finally, it is worth noting that the focus of our study is on labour market discrimination against Roma individuals. Thus, the observation units are individuals and not households as in the above mentioned previous work.

The remainder of the paper is organised as follows. The section that follows briefly describes the standard Blinder-Oaxaca approach to decomposition analysis. Section 3 outlines the

[1] A phenomenon that frequently occurs in the case of small samples.

Bayesian approach that is pertinent to the decomposition analysis in this study. Section 4 proceeds to describe the variables and data, and presents the empirical results. Finally, Section 5 closes the paper.

2 The standard approach to decomposition analysis

Since its popularization by Oaxaca (1973) and Blinder (1973), wage decomposition methodology has been the standard approach to estimating the extent of labour market discrimination on the basis of gender, race and/or ethnicity. Empirically, researchers attempt to apportion the gross wage [income] differentials among demographic groups into three components that represent the characteristics effect, the coefficients effects and the residual effects. The coefficients effect is interpreted as an estimate of the labour market discrimination coefficient.

Characteristically, the Blinder-Oaxaca decomposition of wage differentials between two demographic groups, which we label *j*=1 [Roma in the context of this study] and *j*=2 [non-Roma], is based on semi-log linear regressionships shown in Eq. (1)

$$Y_j = X_j \beta_j + \varepsilon_j \qquad j \in \{1,2\} \tag{1}$$

where Y_j is the *nj*-by-1 vector of log-wages for *nj* individuals in demographic group *j*. The matrix X_j contains a set of *k* column vectors representing characteristics (such as job experience and education) that purport to explain wage variation over the two samples of type 1 (Roma) and type 2 (non-Roma) individuals, as well as an intercept term. The *k*-by-1 parameter vector β_j provides a measure of the responsiveness of wages to the various characteristics for the two demographic groups. The disturbance vector ε_j is typically assumed to follow a constant variance normal distribution, $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2 I_{nj})$ where I_{nj} denotes the *nj*-by-*nj* identity matrix.

The wage decomposition differentials of interest[2] are obtained as

$$\overline{Y}_2 - \overline{Y}_1 = \left(\overline{X}_2 - \overline{X}_1\right)\hat{\beta}_2 + \overline{X}_1\left(\hat{\beta}_2 - \hat{\beta}_1\right) + \overline{\hat{e}}_2 - \overline{\hat{e}}_1 \qquad (2)$$

where $\overline{Y}_j$ denotes the average log-wages of group j=1 (Roma) and j=2 (non-Roma), $\overline{X}_j$ is a 1-by-k vector of consistent estimates (either by using maximum likelihood or Bayesian estimation) that we denote as $\hat{\beta}_j$ in Eq. (2), and $\overline{\hat{e}}_j$ reflects the mean of the residual vector from each of the relationships. The first $[(\overline{X}_2 - \overline{X}_1)\hat{\beta}_2]$, the second $[\overline{X}_1(\hat{\beta}_2 - \hat{\beta}_1)]$, and the third $(\overline{\hat{e}}_2 - \overline{\hat{e}}_1)$ components of the average log wage differential $(\overline{Y}_2 - \overline{Y}_1)$ represent the characteristics effect (denoted by *C*), the coefficients or discrimination effect[3] (denoted by *D*), and the residuals effect (denoted by *R*), respectively. Most applied work ignores the third component.

The computation of the decomposition components is straightforward, but it is less clear how one should estimate their standard errors. All three effects represent scalar quantities with extremely complicated statistical distributions sensitive to the assumption of homoscedastic disturbances, the lack of omitted variables or simultaneity bias (Keith and LeSage 2004). Oaxaca and Ransom (1998) suggest an asymptotic approximation to the variance of the effects based on a linear Taylor series expansion around the true, but unknown parameter vector. This approximation requires an assumption of an asymptotic multivariate normal distribution for the parameter vector and the use of the variance-covariance matrix for the parameter estimates. These assumptions, however, may not be valid in the face of small samples and outliers.

---

[2] Note that this standard approach is a special case of the more general decomposition suggested by Oaxaca and Ransom (1994): $\overline{Y}_2 - \overline{Y}_1 = (\overline{X}_2 - \overline{X}_1)\beta^* + [\overline{X}_2'(\hat{\beta}_2 - \beta^*) + \overline{X}_1'(\beta^* - \hat{\beta}_1)]$ where β^* is a set of benchmark coefficients (i.e. the coefficients from the non-discriminatory wage structure). Examples for β^* are $\beta^* = \hat{\beta}_1$ or $\beta^* = \hat{\beta}_2$ (Oaxaca 1973; Blinder 1973), and $\beta^* = 0.5\hat{\beta}_1 + 0.5\hat{\beta}_2$ (Reimers 1983).

[3] If in the absence of discrimination Roma and non-Roma would receive identical returns for the same characteristics, and differences in wages would thus be due only to differences in pay-related characteristics, then this coefficients effect can be interpreted as the part of the log wage differential due to discrimination. This is the essence of the Blinder-Oaxaca approach (Neumark 1988). However, unobserved factors, such as cultural differences, lifestyle, work ethics or prior discrimination in the education system are not accounted for in the wage equation but may have influence on wages and thereby overestimate the discrimination estimate. Therefore, it is suggested to consider this component of the wage gap as an 'upper bound' estimate of labour market discrimination.

The approximate variance of the discrimination effect suggested by Oaxaca and Ransom (1998) is given by

$$\text{var}\left(\hat{D}\right) = \left(\hat{D}+1\right)^2 \ \bar{X}_2' \left(\Sigma_1 + \Sigma_2\right) \bar{X}_2 \tag{3a}$$

$$\Sigma_1 = \hat{\sigma}_{\varepsilon 1}^2 \left(\bar{X}_1' \ \bar{X}_1'\right)^{-1} \tag{3b}$$

$$\Sigma_2 = \hat{\sigma}_{\varepsilon 2}^2 \left(\bar{X}_2' \ \bar{X}_2'\right)^{-1} \tag{3c}$$

where the noise variance estimates $\hat{\sigma}_{\varepsilon 1}^2$ and $\hat{\sigma}_{\varepsilon 2}^2$ are typically constructed using the least-squares residuals from the group 1 (Roma) and group 2 (non-Roma) regressions.

3 A Bayesian approach

In a Bayesian setting, we can replace the scalar quantities *C*, *D* and *R* representing the characteristics, discrimination and residual effects with samples from the posterior distribution constructed using Markov Chain Monte Carlo (MCMC) estimation, as proposed by Radchenko and Yun (2003). This approach produces a sample of MCMC draws for the parameter vectors β_1 and β_2 that reflect the entire posterior distribution for these parameters. These draws can be used to construct the complete posterior distribution for the characteristics, coefficient and residual effects that are of interest in the wage differential decomposition analysis. But this approach does not overcome sensitivity to outliers or non-constant variance that often arises in small samples of individuals' wages obtained from a household survey.

Following Keith and LeSage (2004) we rely on an extension of the Bayesian regression model by Geweke (1993), given by

$$Y_j = X_j \beta_j + \varepsilon_j \qquad j = \{1,2\} \tag{4a}$$

$$\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2 V_j) \qquad j = \{1,2\} \tag{4b}$$

$$V_j = \text{diag}(v_1, ..., v_{nj}) \qquad j = \{1, 2\}. \tag{4c}$$

This Bayesian variant of the regression model introduces a set of variance scalars $(v_1, ..., v_{nj})$ representing unknown parameters to be estimated. These variance scalars can accommodate heteroscedastic disturbances and/or outliers.

In accordance with Keith and LeSage (2004) we use the following prior distributions for the model

$$\pi(\beta_j) \sim \mathcal{N}(c_j, T_j) \tag{5}$$

$$\pi(r / v_j) \sim \text{IID } \chi^2(r) \tag{6}$$

$$\pi(1/\sigma_j^2) \sim \Gamma(d_j, v_j). \tag{7}$$

$$r \sim \Gamma(m, h) \tag{8}$$

where the prior distributions are indicated using $\pi(.)$. Given the small sample sizes typically encountered in wage decomposition studies, non-informative prior assignments seem reasonable for β_j and σ_j in our study. β_j is assigned a *normal* conjugate prior, which can be made almost diffuse by setting the vector of the prior means $c_j = 0$ and the prior variance-covariance $T_j = I_k \cdot 1e+10$. The variances, σ_j^2 together with v_j $(j = 1, ..., nj)$ are given (conjugate) *inverse gamma* priors. A diffuse prior for σ_j^2 would involve setting the parameters $(d_j = v_j = 0)$ in Eq. (7).

Prior information concerning the variance scalars v_j that arise in the two equations take the form of *nj* (*j*=1, 2) independent, identically distributed $\chi^2(r)/r$ distributions, where *r* represents the single parameter of the χ^2 distribution. This allows estimating the additional *nj* non-zero variance scaling parameters in the diagonal matrix V_j by adding only a single parameter (*r*) to the model. Note that we will use the same value for this hyperparameter for both regression relationships during estimation. The values assigned to *r* are controlled by assigning a $\Gamma(m, h)$ prior distribution with a mean of m/h and variance m/h^2. This prior is

consistent with a prior belief in heteroscedasticity, or non-constant variance as well as outliers. If the sample data does not contain these problems, the resulting posterior estimates for the variance scalar parameters v_j will take values near unity.

Conditional posterior distributions for the parameters β_j, σ_j and the variance scalar v_j $(j = 1, ..., nj)$ are required for MCMC estimation of the model. This method of estimation became popular when Gelfand and Smith (1990) have shown that MCMC sampling from the sequence of complete conditional distributions for all parameters in a model generates a set of estimates that converge in the limit of the true (joint) posterior distribution of the parameters. Hence, if we can decompose the posterior distribution into a set of conditional distributions for each parameter in the model, drawing samples from these will yield valid Bayesian parameter estimates (LeSage and Pace 2009).

The conditional posterior density for β_j takes the form of a multivariate normal with mean and variance-covariance given by

$$\beta_j \mid (\sigma_j, V_j) \sim \mathcal{N}\left\{H_j \left(X_j' V_j^{-1} Y_j + \sigma_j^2 T_j^{-1} c_j\right), \sigma_j^2 H_j\right\} \tag{9a}$$

$$H_j = \left(X_j' V_j^{-1} X_j + T_j^{-1}\right)^{-1}. \tag{9b}$$

Let $e_j = Y_j - X_j' \beta_j$, then the conditional posterior density for σ_j takes the form of a $\chi^2(nj)$ distribution

$$\left\{\sum_{i=1}^{nj} (e_{ji}^2 / v_{ji}) / \sigma_j^2\right\} \Big| (\beta_j, V_j) \sim \chi^2(n). \tag{10}$$

The posterior distribution of V_j conditional on (β_j, σ_j) is proportional to a $\chi^2(r+1)$ distribution

$$\left\{(\sigma_j^{-2} e_j^2 + r) / v_j\right\} \Big| (\beta_j, \sigma_j) \sim \chi^2(r+1). \tag{11}$$

Finally, it is worth noting that we draw a value for the hyperparameter *r* from the prior distribution $\Gamma(m, h)$. Given the conditional posterior densities by Eqs. (9) through (11), we can formulate our MCMC sampler for the model by the following steps:

(i) Begin with arbitrary values for the parameters which we denote $\beta_j^0, \sigma_j^0, v_j^0$ and r^0, where r^0 is a value for the hyperparameter drawn from the prior distribution $\Gamma(m, h)$.

(ii) Calculate the mean and variance of β_j using Eq. (9) conditional on the initial values σ_j^0, v_j^0 and r^0.

(iii) Use the computed mean and variance of β_j to draw a multivariate normal random vector, labelled β_j^1.

(iv) Compute expression (10) using β_j^1 determined in *Step* (iii) and take this value along with a random $\chi^2(n_j)$ draw to determine σ_j^1.

(v) Using β_j^1 and σ_j^1, compute expression (11) and use the value along with an *nj*-vector of random $\chi^2(r^0 + 1)$ draws to determine v_j^1.

(vi) Draw a $\Gamma(m, h)$ value to update r^0 to r^1.

One sequence of steps (i) to (vi) constitutes a single pass through the sampler. We carry out a large number of passes building up a sample $(\beta_j^q, \sigma_j^q, v_j^q, r^q)$ of *q* values from which we can approximate the posterior distribution. Note that Gelfand and Smith (1990) have shown that MCMC sampling from the sequence of complete conditional distributions for all parameters in a model such as given by Eq. (4) produces a set of estimates that converge in the limit to the true (joint) posterior distribution of the parameters.

In addition to model parameters, we are interested in the posterior distribution of the characteristics effect that can be constructed as $(\bar{X}_2 - \bar{X}_1)\,\beta_2^q$ and the discrimination effect that can be found as $\bar{X}_1\,(\beta_2^q - \beta_1^q)$. Statistical significance can be ascertained using Bayesian *p*-level calculations that are Bayesian equivalents to *t*-statistics. These are based on an enumeration of the draws larger or smaller than zero, depending on the sign of the coefficient (see Gelman et al. 2003).

4 Data, variables and results

We apply this Bayesian approach to decomposition analysis to measure labour market discrimination against Roma in South East Europe. The analysis is based on the 2004 UNDP dataset. The survey questionnaire that was used to generate the data is based on the philosophy of an integrated household survey. Thus, for each country, the survey contains individual and household level questionnaires. The household level questionnaire provides general information about each individual within the household as well as detailed information on household consumption and access to basic infrastructures. The individual level questionnaire gathered more specific information about the individuals within each household. Questions related to incomes and expenditures were addressed in both, making it possible to cross-check the results.

The samples are representative of the Roma population living in Roma settlements or areas of compact Roma population. Such settlements and areas were defined as settlements where the share of Roma population at least equals the national share of Roma population in the given country, as reflected in the census data. The major drawback of this sampling methodology is related to its application to municipalities where the share of Roma in the total population is below national averages. Nevertheless, the samples cover roughly 85 percent of Roma in each country. In order to derive data for meaningful comparisons, control groups' samples – representing non-Roma households living in close spatial proximity to Roma households – were constructed in each country using similar procedures as for the Roma samples.

The survey was executed by agencies-members of BBSS Gallup International, coordinated by the regional office of BBSS Gallup International in Sofia. Using BBSS Gallup International member agencies made it possible to apply similar standards and procedures in all countries covered by the project, making cross-country comparisons possible and reliable. In order to overcome the possible distrust to enumerators, Roma interviewers were used for the interviews (see UNDP 2006 for more information on the survey).

Our analysis is based on the individual observation units. We restricted the Roma and non-Roma control samples to individuals with an age between 16-65 years who reported wage income as the main source of income. Missing data on some independent variables did lead to

a further reduction of the country-specific sample sizes[4]. The final samples selected for our study comprise 289 Roma and 570 non-Roma individuals in Albania, 241 Roma and 370 non-Roma individuals in Bulgaria, 77 Roma and 219 non-Roma individuals in Croatia, 123 Roma and 280 non-Roma individuals in Kosovo as well as 111 Roma and 353 non-Roma individuals in Serbia. The differences in sample sizes between Roma and non-Roma populations are due to smaller proportions of Roma with wage income as major source of income in the respective countries.

The UNDP survey does not provide information on actual wages but on income. Income may include diverse sources of non-labour income. Even though we restrict the analysis to those individuals who declared wage income as the main source of income, we are aware that income is not the ideal dependent variable in a Blinder-Oaxaca decomposition framework, and the use of the income variable can seriously bias estimates of the rates of return to education (see Blinder 1973).

Position Table 1 about here

We use six independent variables to specify the matrix X_j (j=1, 2) in Eq. (4a). The full list of variables employed in the analysis is given in Table 1. Education measured in terms of years of schooling in primary, secondary and higher education is used to control for human capital differencing the Roma and non-Roma population groups. Age is a reasonable proxy for actual work experience. The square of this variable is included to capture decreasing marginal returns to experience. In addition, we use two dummy variables to characterize the occupational status of the individuals. Full time takes the value of one if the individual indicated to work full time, and zero otherwise. High skills is a dummy variable that takes the value of one if the individual is engaged in a skilled (blue or white collar) occupation, and zero otherwise. Finally, a gender dummy is taken to control for gender-specific effects. Table 2 shows the average differences in characteristics between Roma and non-Roma in the respective countries.

Position Table 2 about here

[4] Note that Bosnia and Herzegovina, Macedonia, Montenegro and Romania were excluded from this study due to too small sample sizes.

Table 3 summarizes the country-specific results of the decomposition analysis, using a sample of q=12,500 MCMC draws, with the first 2,500 excluded for start-up[5]. The first four columns present the parameter estimates of the Bayesian semi-log regression models for the two ethnic groups (j=1: Roma, j=2: non-Roma) along with Bayesian p-level calculations (in brackets) and standard deviations[6]. The coefficients have the predicted signs, and are highly significant with a few country-specific exceptions. While Roma in Albania, Croatia and Kosovo, for example, receive positive, yet diminishing returns to work experience, Roma in Bulgaria and Serbia are not rewarded for work experience. Education is associated with positive and significant impacts on Roma income in all countries, but is not significant in Serbia. Working full time and in a skilled occupation increases the incomes of Roma in all countries. But the full time variable is not significant in Croatia. The absence of gender effects among Roma in Bulgaria, Kosovo and Serbia may result from relatively low labour market participation rates among Roma compared to non-Roma women.

The final four columns of this table show the country-specific decompositions of log income differential into characteristics and coefficients (discrimination) effects, based on the Bayesian MCMC estimation methodology. The Bayesian estimates reported are based on the mean of 10,000 MCMC draws for the method set forth in the previous section. Given the standard deviations, significance levels can be constructed to test the null hypotheses of no characteristics effect, $\mathrm{H0}_C: C = 0$, and no discrimination effect, $\mathrm{H0}_D: D = 0$.

Position Table 3 about here

Table 4 presents the results of these MCMC tests. The reported probabilities indicate the existence of significant characteristics effects in all the countries considered. They also show that the null hypothesis $\mathrm{H0}_D: D = 0$ is rejected in Albania and Kosovo at the one percent level, but not rejected in the case of the other three countries. From these results we conclude that there may exist discrimination against Roma and in favour of non-Roma in Albania and

[5] We used MATLAB Version 7.0 and the public domain MATLAB function 'ols_g' from LeSage's Econometrics Toolbox to generate the draws. This public domain set of econometric algorithms can be found at www.spatial-economtrics.com.

[6] The standard deviations were calculated using the sample of 10,000 MCMC draws. Statistical significance is ascertained using Bayesian p-level calculations that are Bayesian equivalents to t-statistics.

Kosovo, but not in the other three countries. These results can also be seen from inspecting Figure 1, a graphical illustration of the posterior distribution of the Bayesian MCMC estimates for the country-specific characteristics and discrimination effects along with their highest posterior density (HPD) regions. These densities are based on a kernel density estimate constructed using the MCMC draws.

Position Table 4 about here

Next we look at characteristics and coefficients effects of each variable, that is, at detailed decompositions as given in the final four columns of Table 3. There is no consistent pattern of the two effects across the countries. Although there are not many significant individual discrimination effects based on the hypothesis test, it appears, nevertheless, worthwhile to point to some country-specific features.

- In *Albania*, the aggregate characteristics and coefficients effects explain 54.5 (=0.380/0.697) and 45.5 percent (=0.317/0.697) of the log income differential (0.697). All individual characteristics effects are statistically significantly different from zero. Work experience and decreasing marginal returns to experience contribute most to the income differential. Education and full time work are also important for the explanation. In contrast to characteristics effects, there is only one individual discrimination effect that is significantly different from zero: skilled jobs. This variable contributes to levelling the income gap in Albania.

- *Bulgaria*: About 90 percent of the log income differences (0.459) between non-Roma and Roma groups is explained through differences in characteristics (education, skilled occupation), and through differences in returns to those differences (education). This suggests that differences in endowments do indeed explain a large fraction of the observed differences in income between non-Roma and Roma groups in this country. Much of this reflects huge differences in educational endowments and access to education.

- *Croatia*: The aggregate discrimination effect identified for this country is not significantly different from zero, but the aggregate characteristics effect is. This effect largely contributes to the ethnic income differential. At the individual variable level, we have two

strongly significant individual characteristics effects (education and full time work) and two weakly significant individual discrimination effects: Work experience and work experience to the square in 100 that captures decreasing marginal returns to work experience. Note that these discrimination effects appear to matter most.

- *Kosovo*: The aggregate characteristics and coefficients effects explain 32.2 and 67.8 percent of the log income difference, respectively. This clearly indicates that discrimination effects are highest in this country where Roma poverty is highest among the five countries. Four individual characteristics effects (work experience, work experience to the square, high skills and gender) and one individual discrimination effect (full time) are statistically significantly different from zero. The full time variable largely contributes to widening the income differential.

- *Serbia*: As in Bulgaria and Croatia, we see here an aggregate discrimination effect estimate that is statistically not significantly different from zero. And again as in Bulgaria, the income gap between Roma and non-Roma is largely explained through differences in education and differences in return to these differences.

Finally, it should be noted that the contributions of the individual variables to the aggregate coefficients (discrimination) effects are not invariant with respect to the choice of reference groups for dummy variables (see Oaxaca and Ransom 1999 for this identification problem). With a different normalization, the coefficients effects showing the contributions of each of the variables (full time, high skills and gender) to discrimination could change. Fortunately, however, the overall decomposition and the individual characteristics effects are invariant with respect to the choice of left-out reference groups (see Oaxaca and Ransom 1999).

5 Closing remarks

In this study, we used the robust Bayesian approach suggested by Keith and LeSage (2004) to a Blinder-Oaxaca type of decomposition analysis. The approach has been applied to the decomposition of income differentials among Roma and non-Roma population groups in five South East European countries, using samples from the 2004 UNDP survey.

One merit of this Bayesian approach is that it can accommodate non-constant variance or heteroscedasticity in the cross-sectional semi-log regression relationships. The posterior distributions of the characteristics and discrimination effects are easily obtained by using Markov Chain Monte Carlo estimation. Another merit is that – without relying on asymptotic theory – a hypothesis test of whether the characteristics and discrimination effects are significantly different from zero can easily be derived from the posterior distribution of the MCMC estimates for the two effects.

The results obtained indicate the presence of statistically significant discrimination effects in Albania and Kosovo, but their absence in Bulgaria, Croatia and Serbia. The discrimination effect explains 67.8 and 42.5 percent of the income differential between Roma and non-Roma population groups in Kosovo and Albania, respectively.

Labour market discrimination is apparently an important factor in explaining income differences among Roma and non-Roma groups in these two countries. But differences in measured characteristics and not labour market discrimination against Roma, are the overwhelming reason for the shortfall in incomes for Roma in Bulgaria, Croatia and Serbia. Of course, discrimination outside the labour market may affect the acquisition of human capital (i.e. education) by Roma and lead to differences in observed characteristics. Moreover, discrimination in the labour market, as it affects the returns to education, may induce some differences in educational attainment. Hence, discrimination may have indirect effects on incomes, as well as the direct effects estimated in this paper.

References

Blinder AS (1973) Wage discrimination: reduced form and structural estimates. *Journal of Human Resources* 8(4): 436-455

Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85(410): 398-409

Gelman A, Carlin JB, Stern HS, Rubin DB (2003) Bayesian data analysis. Chapman and Hall, London

Geweke J (1993) Bayesian treatment of the independent Student-t linear model. *Journal of Applied Econometric*s 8, Supplement: S19-S40

Keith K, LeSage JP (2004) Robust decomposition analysis of wage differentials. *Journal of Economic and Social Measurement* 29(4): 487-505

Kitagawa EM (1955) Components of a difference between two rates. *Journal of American Statistical Association* 50: 1168-1194

LeSage JP, Pace RK (2009) Introduction to spatial econometrics. CRC Press, Boca Raton, London, New York

Milcher S (2009) Household vulnerability estimates of Roma in Southeast Europe. *Cambridge Journal of Economics*. Advance Access published October 25, 2009. doi: 10.1093/cje/bep060.

Neumark D (1988) Employers' discriminatory behavior and the estimation of wage discrimination. *The Journal of Human Resources* 23(3): 279-295

Oaxaca RL (1973) Male-female wage differentials in urban labour markets. *International Economic Review* 14(3): 693-709

Oaxaca RL, Ransom M, (1994) On discrimination and the decomposition of wage differentials. *Journal of Econometrics* 61(1): 5-21

Oaxaca RL, Ransom M (1998) Calculation of approximate variances for wage decomposition differentials. *Journal of Economic and Social Measurement* 24(1): 55–61

Oaxaca RL, Ransom M (1999) Identification in detailed wage decompositions. *The Review of Economics and Statistics* 81(1): 154-157

Radchenko SI, Yun M-S (2003) A Bayesian approach to decomposing wage differentials. *Economics Letters* 78(3): 431-436

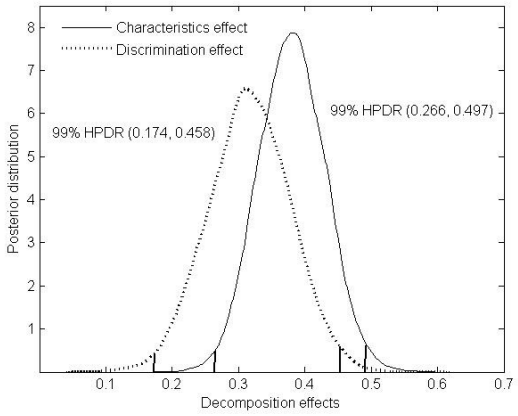
Reimers CW (1983) Labor market discrimination against Hispanic and Black Men. *The Review of Economics and Statistics* 65(4), 570-579

Revenga A, Ringold D, Tracy WM (2002) Poverty and ethnicity: a cross-country study of Roma poverty in Central Europe. World Bank Technical Paper no. 531. The World Bank, Washington DC

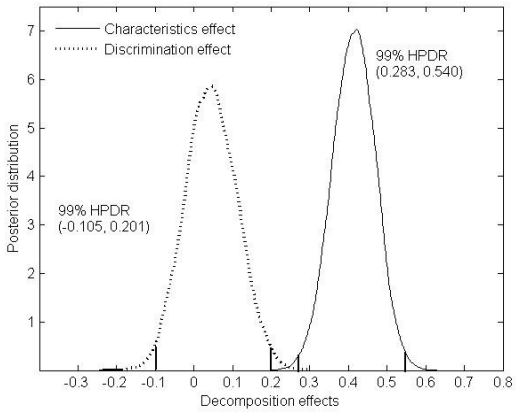
UNDP (2006) At risk: Roma and the displaced in Southeast Europe. United Nations Development Programme, Regional Bureau for Europe and the Commonwealth of Independent States, Bratislava

Figure 1 Posterior distributions of the Bayesian MCMC estimates for the characteristics and discrimination effects in (a) Albania, (b) Bulgaria, (c) Croatia, (d) Kosovo and (e) Serbia

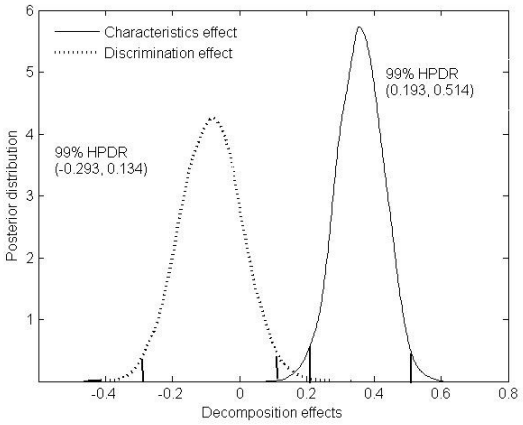
(a) Albania



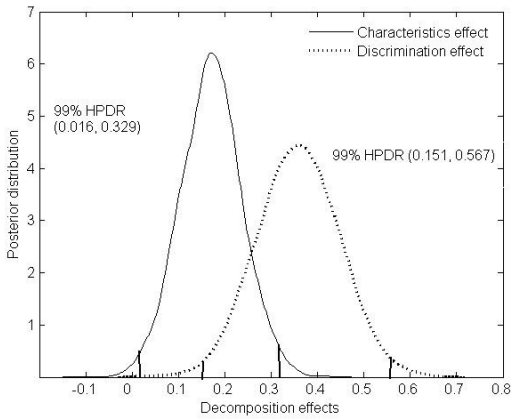
(b) Bulgaria



(c) Croatia



(d) Kosovo



(e) Serbia

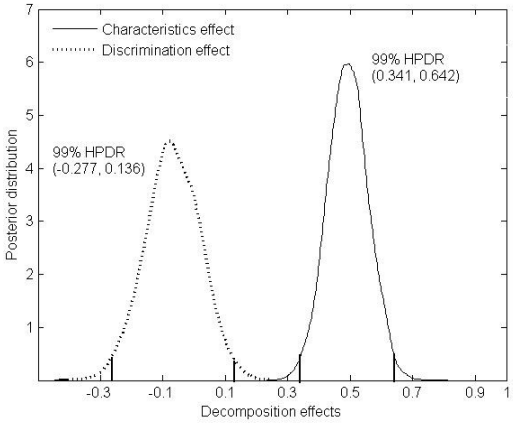


Table 1 Variables used in the analysis

Variable	Variable definition
Income	natural log of wage income [in Euro] per month
Education	number of years of schooling
Work experience	age of individual in years
Work experience squared	age (in years) squared in 100
Full time	a dummy variable taking the value of one if the individual works full time, and zero otherwise
High skills	a dummy variable taking the value of one if the individual is engaged in a skilled occupation, and zero otherwise
Gender	a dummy variable taking the value of one if male, and zero otherwise

Table 2 Description of the variables

	Albania		Bulgaria		Croatia		Kosovo		Serbia	
	Roma	Non-Roma	Roma	Non-Roma	Roma	Non-Roma	Roma	Non-Roma	Roma	Non-Roma
Variables (means and standard deviations in brackets)										
Log income	4.47	5.17	4.26	4.75	5.86	6.12	4.75	5.27	4.87	5.19
	(0.69)	(0.63)	(0.55)	(0.46)	(0.62)	(0.62)	(0.82)	(0.82)	(0.74)	(0.73)
Education	6	12	7	12	9	13	7	12	9	13
(no. of school years)	(3.65)	(2.83)	(3.09)	(2.60)	(3.05)	(2.69)	(3.16)	(2.54)	(3.09)	(2.55)
Work experience	36	41	38	40	32	38	35	38	39	41
(age in years)	(10.37)	(10.37)	(11.10)	(10.20)	(9.77)	(11.65)	(11.19)	(11.74)	(10.50)	(10.49)
Work experience	14	18	16	17	11	16	14	16	16	18
squared in 100	(7.99)	(8.22)	(8.69)	(8.27)	(6.45)	(9.36)	(8.59)	(9.50)	(8.12)	(8.47)
Dummy variables (percentage of sample, with each level of variable)										
Full time work										
yes	53	89	71	95	87	93	54	82	68	94
no	47	11	29	5	13	7	46	18	32	6
High skills										
yes	69	89	20	74	44	93	27	68	47	94
no	31	11	80	26	56	7	73	32	53	6
Gender										
yes	73	61	66	51	71	53	90	83	82	55
no	27	39	34	49	29	47	10	17	18	45

Table 3 Decomposition analysis: Bayesian approach

	Bayesian estimates				**Decomposition**			
	Roma (j=1)		Non-Roma (j=2)		Characteristics effect		Discrimination effect	
	Coefficient (p-level)	Standard deviation	Coefficient (p-level)	Standard deviation	Size (p-level)	Standard deviation	Size (p-level)	Standard deviation
(a) Albania (n1=289, n2=570)								
Constant	2.696	0.361	2.652	0.265			-0.044	0.447
	(0.000)		(0.000)				(0.922)	
Education	0.034	0.009	0.028	0.007	0.193	0.047	-0.031	0.062
	(0.000)		(0.000)		(0.000)		(0.619)	
Work exp.	0.045	0.020	0.068	0.013	0.342	0.066	0.808	0.863
	(0.013)		(0.000)		(0.000)		(0.350)	
Work exp.^{2}	-0.055	0.026	-0.079	0.164	-0.308	0.064	-0.331	0.436
	(0.019)		(0.000)		(0.000)		(0.448)	
Full time	0.386	0.063	0.443	0.066	0.160	0.024	0.030	0.048
	(0.000)		(0.000)		(0.000)		(0.535)	
High skills	0.445	0.066	0.185	0.067	0.038	0.014	-0.179	0.066
	(0.000)		(0.002)		(0.006)		(0.007)	
Gender	0.305	0.067	0.391	0.040	-0.045	0.005	0.063	0.057
	(0.000)		(0.000)		(0.000)		(0.267)	
Aggregate					0.380	0.050	0.317	0.061
					(0.000)		(0.000)	
(b) Bulgaria (n1=241, n2=370)								
Constant	3.755	0.293	3.256	0.353			-0.499	0.459
	(0.000)		(0.000)				(0.279)	
Education	0.020	0.009	0.045	0.009	0.249	0.047	0.177	0.085
	(0.009)		(0.000)		(0.000)		(0.041)	
Work exp.	0.009	0.016	0.021	0.015	0.040	0.029	0.466	0.842
	(0.292)		(0.087)		(0.180)		(0.581)	
Work exp.^{2}	-0.011	0.020	-0.025	0.019	-0.033	0.025	-0.218	0.440
	(0.289)		(0.093)		(0.195)		(0.621)	
Full time	0.219	0.063	0.198	0.154	0.046	0.036	-0.015	0.119
	(0.000)		(0.101)		(0.200)		(0.902)	
High skills	0.226	0.067	0.269	0.049	0.145	0.026	0.008	0.017
	(0.001)		(0.000)		(0.000)		(0.614)	
Gender	0.039	0.050	0.229	0.040	-0.033	0.006	0.125	0.042
	(0.224)		(0.000)		(0.000)		(0.004)	
Aggregate					0.414	0.055	0.045	0.067
					(0.000)		(0.502)	

Table 3 *ctd.*

	Bayesian estimates				**Decomposition**			
	Roma (*j*=1)		Non-Roma (*j*=2)		Characteristics effect		Discrimination effect	
	Coefficient (*p*-level)	Standard deviation	Coefficient (*p*-level)	Standard deviation	Size (*p*-level)	Standard deviation	Size (*p*-level)	Standard deviation
(c) Croatia (*n*1=77, *n*2=219)								
Constant	3.329	0.628	4.225	0.430			0.896	0.768
	(0.000)		(0.000)				(0.246)	
Education	0.052	0.020	0.074	0.012	0.318	0.051	0.188	0.201
	(0.005)		(0.000)		(0.000)		(0.352)	
Work exp.	0.100	0.035	0.023	0.020	0.121	0.104	-2.493	1.320
	(0.003)		(0.121)		(0.248)		(0.061)	
Work exp.^{2}	-0.130	0.052	-0.018	0.024	-0.074	0.100	1.283	0.666
	(0.007)		(0.231)		(0.459)		(0.056)	
Full time	0.089	0.200	0.380	0.166	0.023	0.010	0.253	0.224
	(0.338)		(0.014)		(0.024)		(0.260)	
High skills	0.243	0.120	-0.028	0.127	-0.013	0.062	-0.119	0.077
	(0.020)		(0.418)		(0.829)		(0.124)	
Gender	0.215	0.115	0.085	0.057	-0.015	0.010	-0.093	0.092
	(0.030)		(0.069)		(0.140)		(0.312)	
Aggregate					0.359	0.069	-0.085	0.092
					(0.000)		(0.360)	
(d) Kosovo (*n*1=123, *n*2=280)								
Constant	2.971	0.623	3.589	0.404			0.618	0.740
	(0.000)		(0.000)				(0.405)	
Education	0.022	0.017	0.012	0.014	0.059	0.066	-0.072	0.155
	(0.100)		(0.189)		(0.379)		(0.642)	
Work exp.	0.067	0.033	0.059	0.021	0.165	0.057	-0.284	1.375
	(0.021)		(0.002)		(0.005)		(0.837)	
Work exp.^{2}	-0.093	0.043	-0.073	0.025	-0.159	0.055	0.262	0.680
	(0.016)		(0.002)		(0.004)		(0.701)	
Full time	0.658	0.105	0.125	0.092	0.035	0.026	0.286	0.075
	(0.000)		(0.085)		(0.176)		(0.000)	
High skills	0.436	0.111	0.222	0.077	0.091	0.032	-0.057	0.036
	(0.000)		(0.002)		(0.004)		(0.117)	
Gender	0.088	0.173	0.284	0.091	-0.020	0.006	0.177	0.176
	(0.305)		(0.001)		(0.002)		(0.316)	
Aggregate					0.170	0.067	0.358	0.089
					(0.012)		(0.000)	

Table 3 *ctd.*

	Bayesian estimates				**Decomposition**			
	Roma (*j*=1)		Non-Roma (*j*=2)		Characteristics effect		Discrimination effect	
	Coefficient (*p*-level)	Standard deviation	Coefficient (*p*-level)	Standard deviation	Size (*p*-level)	Standard deviation	Size (*p*-level)	Standard deviation
(e) Serbia (*n*1=111, *n*2=353)								
Constant	3.344 (0.000)	0.662	3.288 (0.000)	0.404			-0.056 (0.943)	0.776
Education	0.016 (0.187)	0.018	0.080 (0.000)	0.010	0.314 (0.000)	0.040	0.593 (0.002)	0.192
Work exp.	0.029 (0.200)	0.035	0.001 (0.478)	0.018	0.001 (0.954)	0.025	-1.085 (0.479)	1.532
Work exp.^{2}	-0.016 (0.359)	0.045	0.005 (0.415)	0.022	0.005 (0.833)	0.025	0.345 (0.677)	0.827
Full time	0.389 (0.001)	0.119	0.561 (0.000)	0.123	0.147 (0.000)	0.032	0.116 (0.311)	0.114
High skills	0.323 (0.005)	0.120	0.161 (0.051)	0.098	0.075 (0.102)	0.046	-0.076 (0.294)	0.072
Gender	0.072 (0.297)	0.139	0.188 (0.000)	0.050	-0.051 (0.000)	0.014	0.095 (0.436)	0.122
Aggregate					0.492 (0.000)	0.066	-0.068 (0.441)	0.089

Note: The Bayesian estimates are based on the mean of 10,000 MCMC draws, with Bayesian *p*-level calculations that are Bayesian equivalents to *t*-statistics (in brackets)

Table 4 Country-specific MCMC discrimination effects estimates

Country	$\hat{C}$	Standard deviation	$H0_C$: C=0 Probability	$\hat{D}$	Standard deviation	$H0_D$: D=0 Probability
Albania $n1$=289 $n2$=570	0.380	0.050	0.000	0.317	0.061	0.000
Bulgaria $n1$=241 $n2$=370	0.414	0.055	0.000	0.045	0.067	0.502
Croatia $n1$=77 $n2$=219	0.359	0.069	0.000	-0.085	0.092	0.360
Kosovo $n1$=123 $n2$=280	0.170	0.067	0.012	0.358	0.089	0.000
Serbia $n1$=111 $n2$=353	0.492	0.066	0.000	-0.068	0.089	0.441

Note: $n1$=Roma, $n2$=non-Roma